The Constructive Conundrum

Computational Approaches to Facilitate Constructive Commenting on Online News Platforms

Cedric Waterschoot

The Constructive Conundrum

Computational Approaches to Facilitate Constructive Commenting on Online News Platforms

Cedric Waterschoot

This thesis investigates whether AI-based moderation can assist in finding constructive discussion. The cover is created by prompting DALL-E3 to create an image consisting of colors and geometrical shapes representing constructive discussion. Backcover consists of shapes and colors representing non-constructive discussion. Just as with hybrid content moderation, a human touch might be needed to interpret the output.

DOI: https://doi.org/10.33540/2517

This thesis is part of the project Better-MODS, supported by the research programme ’Digital Society - The Informed Citizen’ (project number 410.19.006) which is financed by the Dutch Research Council (NWO).

SIKS Dissertation Series No. 2024 – 43
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

The Constructive Conundrum

Computational Approaches to Facilitate Constructive Commenting on Online News Platforms

Het Constructieve Raadsel

Constructieve Commentaren op Online Nieuwsplatforms Faciliteren door middel van Computationale Benaderingen

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het College voor Promoties
in het openbaar te verdedigen op

woensdag 20 november 2024 des middags te 2.15 uur

door

Cedric Danny K Waterschoot

geboren op 19 februari 1995
te Brasschaat, België

Promotoren:

Prof. dr. Antal van den Bosch

Copromotoren:

dr. Ernst van den Hemel

Beoordelingscommissie:

Prof. dr. W.M.P. Daelemans

Prof. dr. A. Fokkens

Prof. dr. E.J. Krahmer

Prof. dr. T. Poell

Prof. dr. J.F.T.M. van Dijck

Contents

1
Introduction

"I cannot tell you what a constructive comment is, I just know it when I see it."

NU.nl Moderator, NU.nl visit june 2022

Seated at the moderator's desk in the newsroom of The Netherlands' largest online newsroom *NU.nl*, an overwhelming stream of new user comments floods in faster than the content moderator can handle. Our project team is visiting *NU.nl* to conduct fieldwork[1]. We are interested in how online content moderation is operationalized and in the experiences of the moderators themselves. Our distracting presence contributes to the challenge, as we stare at the incoming comments within the moderation interface, soaking up the comments made by the moderators trying to explain to us their everyday world. Day in and day out, online platforms grapple with the influx of user-generated contributions, whether on social media or the comment section of their own webpages. *NU.nl*, the news outlet we are currently visiting, reports hosting over a million new user comments each month. To cope with this surge, moderators increasingly work alongside Artificial Intelligence (AI) models trained to filter out the most toxic or unwanted comments, alleviating some of the burden placed on the moderator's shoulders. Although such models reduce the pace at which comments appear on the monitor, the sheer volume of user contributions still poses a significant challenge for moderators. At *NU.nl*, the daily responsibilities of moderators at their comment platform *NUjij*, known as the 'interactieredactie' (interaction editors), involve not only removing toxic behavior but also actively promoting 'constructive' comments by pinning them to the top of the webpage – a manual and time-consuming task.

My research focused on developing computational tools to aid moderators in identifying the most constructive comments. However, before developing computational models to assist content moderators in efficiently filtering out such content, it was essential to understand the moderators' working definitions of constructive commenting. When a moderator scrolls through an online discussion, what factors guide their selection of comments to feature on the comment page? In more practical terms, based on which criteria can I annotate a collection of user comments? Such annotated datasets can subsequently serve to train models potentially capable of assisting moderators in more efficiently identifying the most constructive comments within a discussion. I was specifically interested in understanding how comments are evaluated in terms of constructive value in the context of fast-paced moderation that allows little to no room for in-depth analysis or discussion on comment quality. Moderators must constantly move on; on online platforms like *NUjij*, the discussion never ends.

Consequently, I asked the moderator how they would define a constructive comment. Without much hesitation, I was told that it was not possible to give me an exact definition. Moreover, it became apparent that moderators rely heavily on experience and intuition to identify constructive comments. "I cannot tell you what a constructive comment is, I just know it when I see it", the moderator said. Even though *NUjij* offers training and guidelines to their moderators, a lot of room for subjectivity remains in moderating online user comments. Through experience in overseeing online discussions and informal exchanges among moderators, wherein they share insights into how the user base engages in debates and aligns with the platform's objectives, a trained eye is developed to recognize constructive online comments. The issues and observations

[1] The fieldwork, together with Ernst van den Hemel and Liesje van der Linden, was conducted at the *NU.nl* offices in Hoofddorp, The Netherlands. We visited the moderation staff on June 8th 2022 and on March 15th 2023.

at *NUjij* are emblematic of a broader trend, where online news outlets aim to foster constructive engagement both between users and with them.

The ambiguity surrounding the concept of a constructive comment is also evident in the explanations provided by news outlets on their Frequently Asked Questions (FAQ) pages. For instance, *The New York Times* refers to constructive comments as “representing a range of views” (New York Times 2020), while *NU.nl* themselves seek out “substantiated” or “respectful” contributions (NUJij 2018). Furthermore, prior research on automatically classifying constructive user comments also adopts vague descriptors. Some speak of high-quality contributions (Yixue Wang and Diakopoulos 2022), others mention useful/helpful comments (Napoles et al. 2017) or refer to high-level ‘constructive characteristics’, which in turn may include solutions, evidence, or specific points (Kolhatkar, Thain et al. 2023).

The moderator’s remark, combined with the descriptors provided by news outlets and prior research, suggests that efforts to establish a universally applicable definition of a constructive comment may prove futile. Even a seasoned professional responsible for identifying such constructive comments cannot provide a clear-cut definition. The qualification of a comment may depend on various factors, including the context, such as the tone and topic of the discussion, the objectives of the hosting news outlet, the identity of the user, and the moderator’s own interpretation. Consequently, this thesis aims to explore content moderation and computational applications related to constructive commenting from diverse perspectives – an attempt to acknowledge and embrace the inherent ambiguity and contextuality involved in selecting constructive comments in online discussions.

The remainder of the introduction is structured as follows. First, I outline the motivation of this thesis, detailing the perspectives on constructive commenting employed in the included studies. Second, using these perspectives, I formulate the research questions underpinning the studied presented in thesis. Third, I discuss the thesis contributions to academic fields and stakeholders in regard to computational content moderation with the goal of promoting constructive discussion. The concluding paragraphs of the introduction provide an overview of the upcoming chapters.

1.1. Motivation

The Better-MODS project, of which this dissertation is a part of, aimed to contribute to content moderation and good online discussions on news articles by integrating computational models. My role in the project involved developing Natural Language Processing (NLP) models designed to identify and filter high-quality comments from online discussion data. I approached this task with the goal of preserving the inherent complexity and subjectivity involved in defining a constructive comment within the context of online content moderation. Over the past years, I have both sought and been asked to share my views on the blueprint of a constructive or high-quality comment. Predictably, the outcome is a mix of varied responses, sharing similarities in vague or descriptive terms like ‘good’, ‘informative’ or ‘interesting’, but diverging when translated into a practical setting. The response of a moderator employed at a news outlet differs from that of a linguist, data scientist, or the average reader on news platform. For a researcher focused on developing tools to assist moderators, this is a substantial challenge. There is no singular and universally agreed-upon conceptualization of constructive commenting. To address this, I have chosen a multifaceted approach. This involves formulating diverse perspectives on what qualifies as a constructive comment, integrating stakeholders' concerns and expertise. Following this, my goal was to develop computational models that utilize diverse definitions of constructive comments as a starting point.

To accomplish this goal, this thesis adopts a two-perspective approach to constructive commenting: an outside perspective from the viewpoint of researcher or external party, and secondly, an insider perspective, emulating the lens of content moderator themselves. To categorize and distinguish these two perspectives, I draw upon the terms *etic* and *emic* from the field of anthropology (Pike 1967). An etic perspective characterizes the outsider's view, studying behavior and practices as an external observer and noting what the researcher finds significant. Conversely, an emic perspective captures the insider's standpoint, recognizing the practices and values as conceptualized and understood by those engaging in them (Mostowlansky and Rota 2020). In the context of online content moderation, both perspectives differ in focus and measure distinct features. In the following chapters, I utilize both perspectives and reflect on their measurements and practical utility. Consequently, the incorporation of both perspectives enhances our understanding of the conceptualization of constructive commenting. Throughout this thesis, I develop, evaluate, and discuss computational models, using either an etic (outside) or emic (inside) perspective as starting point.

Scholars using an etic perspective on a constructive comment seek to formulate a practical working definition of such a contribution. In this thesis, I outline efforts to establish requirements for a constructive comment in the form of a checklist. Does a constructive comment present a clear line of argumentation? Must it introduce new information or express a different stance on the discussion topic to be deemed constructive? The objective I take here of using the outside perspective is to develop a clear annotation scheme, laying out the possible qualifying terms and enabling researchers to annotate a set of comments in terms of their constructive value. Subsequently, computational models are trained based on these features, capturing to some degree the specific constructive dimensions outlined in the annotation scheme. Much of prior research has followed this trajectory, devising operationalizable coding schemes tailored to a specific research question or aspect of an online discussion. In this thesis, I employ an etic perspective in Chapters 3 and 4 and discuss both advantages and shortcomings in the concluding chapter.

When utilizing an emic perspective on constructive commenting, these formal prerequisites are dropped. By adopting the decisions made by moderators in the past, scholars using this perspective encompass all contextual nuances of the discussion, along with the lived experience of the content moderator at the specific time the discussion in question was managed. I regard their decision-making as ground truth, attributing the concept of a constructive comment to those comments singled out by content moderators. Therefore, the goal of computational models employed in this context is to facilitate the decision-making by the content moderator themselves, as they understand constructive value as expected by their platform and editorial standards. An emic perspective is employed in Chapter 5, and I again evaluate its use in the concluding chapter.

In summary, through the incorporation of multiple perspectives on constructive commenting, I aim to embrace the ambiguity and contextuality of the concept, along with the experiences of the content moderators. The uncertainty and challenge in formulating clear-cut definitions of a constructive comment, and in turn the difficulty of developing targeted computational models, may be seen as a starting point of the included research in this thesis.

1.2. Research questions

Considering the current landscape of content moderation and the challenges associated with defining and discussing perspectives on constructive commenting, the main research question addressed in this thesis is as follows:

Main RQ: To what extent can computational models aid in the interpretation and identification of constructive comments by content moderators?

To tackle this question, the research presented in this thesis is divided into three themes: (1) the practice of promoting constructive comments, (2) computational models targeting constructive commenting and, (3) the impact of promoting constructive comments.

1.2.1. The practice of promoting constructive comments

The practice of moderating online discussions is continually evolving. Initially, moderators focused solely on filtering out toxic and undesirable content. However, as I further explore in Chapter 2, a new objective has emerged, which involves actively promoting constructive comments. Against the backdrop of the ‘post-truth era’ and scandals such as *Cambridge Analytica* on *Facebook*, the push towards positive and constructive engagement intensified. The operationalization of this concept on news platforms, however, remains unclear, particularly considering the fast-paced environment in which content moderators must process user comments. Furthermore, while news outlets commonly employ AI-based tools to identify toxic content, it is arguable whether similar AI-based tools could be utilized to identify and promote constructive comments. Determining what qualifies as constructive and desirable is subjective, leaving ample room for interpretation. Hence, examining the processes behind this moderation strategy is a valuable case study for understanding the concept of constructive online discussions. Consequently, there is a need for a cross-platform analysis to thoroughly examine and evaluate the practical processes implemented by news outlets to foster constructive user comments, comparing definitions of a constructive comment and the use of AI-based moderation tools. This research is discussed in Chapter 2.

RQ1: How is it decided what constructive comments are and how are they promoted on different news platforms?

1.2.2. Computational models targeting constructive commenting

Online content moderation operates within a hybrid framework, with human moderators collaborating with AI-based tools to execute their tasks. However, as I discuss in Chapter 2, these computational tools have a limited focus, primarily trained to assess comments for the presence of toxic content. The definition of toxicity in online content remains heavily debated, with a lot of scholarly scrutiny dedicated to developing and evaluating machine learning models focused on identifying undesirable content. In contrast, applications geared towards detecting high-quality or constructive comments are often discussed on a superficial level or are trained on restricted datasets, whether in size or additional metadata (e.g. Park et al. 2016; Kolhatkar and Taboada 2017). This thesis delves into efforts specifically directed at this subset of user comments, with each computational application grounded in one of the two perspectives on constructive comments previously discussed. Research using an etic perspective is presented in Chapters 3 and 4. Subsequently, I discuss research through the lens of an emic perspective in Chapter 5.

RQ2: Given either an etic or emic perspective on constructive commenting, to what extent can computational models detect constructive comments in an online discussion?

1.2.3. The impact of promoting constructive comments

I formulated research questions aimed at comparing how news outlets promote constructive discussion and exploring the potential contribution of AI-based tools may offer to this practice. Despite the widespread implementation of these moderation practices by news outlets, there is a lack of clarity regarding the impact of promoted constructive comments on the overall trajectory of an online discussion. The extent to which users utilize these comments as a template for future contributions has not been examined. Furthermore, it remains to be determined whether the promotion of constructive comments leads to increased engagement in the discussion. The data derived from Dutch comment platform *NUjij* allows for the comparison between online discussions where moderators promoted constructive comments and similar discussions where this moderation strategy was not employed. Such a comparative analysis is crucial for identifying potential differences in the evolution of online discussions following the promotion of constructive comments. This analysis is presented in Chapter 6.

In practical terms, this thesis describes the potential impact based on two distinct categories. The first area of focus centers on the quality of a discussion. Constructive comments are inherently deemed high-quality contributions according to the standards set by news outlets and content moderators. Therefore, if users perceive these promoted comments as a template for their further contributions to the discussion, the overall discussion quality may be impacted in comparison to discussions without promoted comments.

RQ3a: Comparing discussions where moderators promoted constructive comments to discussions lacking this moderation strategy, does the discussion quality increase after the comments were promoted?

The second aspect of potential impact on the discussion relates to user activity. Of particular interest is whether discussions featuring promoted constructive comments experience, on average, a rise in the number of new comments compared to discussions without promoted comments. Additionally, this study delves into whether the selection of constructive comments by moderators prompts a greater number of users to actively participate in the discussion.

RQ3b: Comparing discussions where moderators promoted constructive comments to discussions lacking this moderation strategy, does the discussion activity increase after the comments were promoted?

1.3. Thesis goals

Online content moderation operates as a hybrid process, wherein human moderators collaborate with AI-based tools to oversee discussions on news articles. Consequently, research in this domain spans multiple disciplines, merging insights from Natural Language Processing and fields like media and communication studies. In the following paragraphs, I outline the thesis goals related to academic fields as well as stakeholders involved in online commenting on news platforms.

The field of media studies has a keen interest in online commenting, and subsequently, in moderating these user contributions on their online platforms. The relationship to their active user-contributors plays a pivotal role in participatory journalism (Domingo et al. 2008) and constructive journalism (Løvlie 2018), fostering audience engagement and reader-journalist interaction. This thesis aims to contribute to the understanding of contemporary online content moderation, studying the dynamics between human moderator and AI-based applications. I investigate the various tasks assigned to the moderator within this hybrid setting and explore their understanding of constructive discussion. In turn, this understanding forms the basis of computational models capable of supporting the content moderator in their tasks. Furthermore, I will expand the discussion and analysis of content moderation performed by news outlets by explicitly focusing on constructive commenting and the effect of highlighting such content on the discussion.

Computational applications do not operate in a vacuum. The field of NLP and computational linguistics develops models and applications, and links theoretical advancements of natural language processing models to their practical use, in this case online content moderation and the promotion of constructive comments. This thesis attempts to establish a link between the input of content moderators and the development process of AI-based tools, valuing the insights and experience of those already working in the hybrid moderation setting. These insights are necessary to model the subjective processes involved in promoting constructive comments. Additionally, the output is assessed by these experts, and its practical contribution is discussed, bridging the gap between theory and application. Chapters 3 and 4 are particularly content driven by training NLP models. On the other hand, in Chapter 5, I discuss a content moderation application through the lens of an emic perspective. This approach, which does include textual representation, mainly makes use of a wide array of metadata.

The research presented in this thesis may be of interest to news outlets with online comment sections. Maintaining an advertiser-friendly comment space, capable of expanding editorial content and maximizing reader engagement, is vital for an online news outlet (Manosevitch and Tenenboim 2017; Paßmann, Helmond and Jansma 2023). The question arises: how can online content be curated, and discussion quality improved by encouraging constructive comments? To what extent can AI-based tools contribute to achieving this goal? Collaborating with news outlets allows researchers to access information that public datasets lack, including deleted comments, user flags, or specific timestamps. These features are essential building blocks for replicating real-life practices in academic studies. Consequently, news outlets may gain valuable insights into their moderation practices, such as a comprehensive impact analysis of featured comments or computational models specifically tailored to their context. Dutch online news outlet *NU.nl*, with its comment platform *NUjij*, is a suited partner for this studying online commenting and future possibilities regarding content moderation, as the company has invested in building up their comment section and interaction possibilities. The practical value of the work in this thesis became apparent due to working closely with *NUjij*, specifically my research on the featured comment group recommender system discussed in Chapter 5. Enthusiasm exists for this kind of research as well as resulting applications, combining scientific work with the interests of societal partners. In this thesis, Chapter 3 discusses research based on social media data, while all subsequent chapters utilize data obtained from comment platform *NUjij*, including non-publicly available variables.

And finally, this thesis is of interest to news readers and commenters themselves. Not all news readers participate in commenting, potentially due to toxicity in online debates. However, a recent survey conducted among the Dutch population, as outlined by L. v. d. Linden et al. (Forthcoming), reveals that a notable 11% of respondents participated in online news commenting within the past year. Moreover, individuals expressing interest in commenting emphasize the importance of transparency from both news outlets and moderators, viewing it as a crucial factor for establishing trust in moderation practices (Brunk, Mattern and Riehle 2019). This thesis provides a look into the world of online discussions and may inform commenters and non-commenters alike about the user interaction found within comment sections. Additionally, I discuss how their contributions are moderated and evaluated by news outlets. I raise questions regarding the impact of user participation on news outlets and their moderation processes, the criteria employed by moderators to assess contribution quality, and the potential for your constructive engagement to positively influence discussions. Despite being only partially answered, the inquiries discussed in this thesis have the potential to generate increased public interest in the online comment section and online discussions.

1.4. Outline

The remainder of the thesis is structured as follows. First, I describe and contextualize the promotion of constructive comments. Chapter 2 outlines how content moderators work alongside AI-based models. Additionally, I discuss the operationalization by five international news outlets of the concept 'constructive comments'. I refer to the outcome of this process as 'the third half of the internet' – a curated space of user comments aligning with editorial policies regarding constructive commenting.

I continue with computational models aimed at capturing what makes a comment constructive. First, I present research using an etic perspective. Chapter 3 discusses work addressing argument diversity and interactivity measurements in online discussion threads. Using data from *Reddit*, *Twitter* and *Gab*, we calculate potential skewedness in 'pro' and 'con' arguments regarding Black Pete (Zwarte Piet). In Chapter 4, the research makes use of data obtained from the Dutch comment platform *NUjij*. The editorial policy of *NUjij* explicitly prohibits the denial of human-caused climate change. Consequently, we explore an argument mining approach that seeks to classify the arguments presented in online comments posted on climate-themed news articles.

In Chapter 5, adopting an emic perspective, we present a recommender system trained to detect a selection of online comments deemed worthy of being featured by a content moderator. The models are trained on past moderation decisions. To evaluate the system's effectiveness, we make use of datasets containing *NUjij* comments from both 2020 and 2023. The testing involves examining the model's robustness against topic fluctuations caused by evolving news cycles as well as against platform growth. Additionally, content moderators at *NUjij* assess the performance of the best-performing model through a discussion-based evaluation, comparing its output against randomly selected online comments.

Then, I zoom out and examine the impact of promoting constructive comments on the discussion itself. Chapter 6 divides the *NUjij* dataset from 2023 into featured discussions, encompassing those with comments highlighted by a content moderator, and a control group comprising the discussions in where no comments were featured. We assess the impact of the moderation strategy by comparing the discussion activity in both groups in terms of unique users and comment count. Additionally, we explore the potential effects on discussion quality by examining the absence of negative content and the presence of good comments. This chapter takes into consideration quality evaluations from both the editorial standpoint and the view of the user base.

1.5. List of publications

The chapters in this thesis are based on the following publications:

2. **Waterschoot, Cedric**. Governing the 'Third Half of the Internet': The Dynamics of Human and AI-assisted Content Moderation (2024). Governing the Digital Society: Platforms, Artificial Intelligence, and Public Values, van Dijck, José; van Es, Karin; Helmond, Anne; van der Vlist, Fernando (eds.), Amsterdam University

Press (accepted for publication)

3. **Waterschoot, Cedric**; van den Bosch, Antal; van den Hemel, Ernst, Calculating Argument Diversity in Online Threads (2021). 3rd Conference on Language, Data and Knowledge (LDK2021). Vol. 93 Dagstuhl, Germany: OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl

4. **Waterschoot, Cedric**; Van den Bosch, Antal; Van den Hemel, Ernst. Detecting Minority Arguments for Mutual Understanding: A Moderation Tool for the Online Climate Change Debate (2022). In Proceedings of the 29th International Conference on Computational Linguistics, pages 6715–6725. International Committee on Computational Linguistics.

5. **Waterschoot, Cedric**; van den Bosch, Antal. A Time-robust Group Recommender for Featured Comments on Online News Platforms (2024). Frontiers in Big Data (accepted for publication)

6. **Waterschoot, Cedric**; van den Hemel, Ernst; van den Bosch, Antal. The Impact of Featuring Posts in Online Discussions. 16th International Conference on Advances in Social Networks Analysis and Mining – ASONAM-2024

2

Governing the ‘Third Half of the Internet’: The Dynamics of Human and AI-assisted Content Moderation

Content moderation almost feels like a computer game. Time and time again, new assignments pop up on the screen.

Fieldnotes, NU.nl visit june 2022

Based on: Waterschoot, Cedric. Governing the ‘Third Half of the Internet’: The Dynamics of Human and AI-assisted Content Moderation (2024). Governing the Digital Society: Platforms, Artificial Intelligence, and Public Values, van Dijck, José; van Es, Karin; Helmond, Anne; van der Vlist, Fernando (eds.), Amsterdam University Press (accepted for publication)

Abstract

In recent years, a major challenge for news outlets has been warding off toxic content from online spaces where they allow user contributions. The governance of these comments primarily focused on identifying and banning unwanted comments. This article highlights a more recent development: the promotion of constructive comments. It analyzes how banning toxicity and promoting constructive comments is performed across five international news outlets: The New York Times, The Guardian, Die Zeit, El País and Nu.nl. I conclude that keeping out toxicity is mainly assigned to AI-based tools. Such models are specifically trained to find and filter out unwanted contributions, but these tools are unfit to identify and promote constructive comments. This responsibility is assigned to the human moderators, who have to manually curate large numbers of user comments. The resulting collection of hand-picked contributions align with editorial guidelines, establishing a connection between editorial and user-generated content.

2.1. Introduction

User participation is essential for online news outlets, boosting revenue and community engagement (Ksiazek, Peer and Lessard 2016). Comment sections not only attract advertisers by increasing webpage activity but also build a loyal subscriber base. Additionally, these platforms utilize user contributions for content expansion and reader feedback (Manosevitch and Tenenboim 2017). However, open comments can lead to negative behaviors like trolling and harassment (Quandt, Klapproth and Frischlich 2022). Moderators face challenges in managing content, including combating fake news and misinformation (Meier, Kraus and Michaeler 2018; Tandoc, Lim and Ling 2018) and dealing with polarizing discussions that can escalate into toxicity (Strandberg, Himmelroos and Grönlund 2017). This negative aspect, termed 'dark participation' (Quandt 2018), has resulted in the comment section being pejoratively labeled as 'the bottom half of the internet' (Reagle 2015). Addressing these issues has become a priority for news outlets, leading to significant investment and scholarly scrutiny (Gollatz, Riedl and Pohlmann 2018; Wintterlin et al. 2020).

Besides deleting negative comments or eliminating comment sections altogether, another trend has emerged. Many news outlets and moderators are adopting methods to encourage constructive discussions (Yarnoz 2019; Diakopoulos 2015a). While much focus has been on countering hate speech and dark participation, strategies to foster positive engagement are less explored. However, these approaches could have substantial effects on online interactions.

In this chapter I aim to explore the evolution of news platforms' collective efforts to promote constructive discussions within their comment sections. I argue that this newfound emphasis has given rise to new configurations of hybrid moderation. While commonly used Artificial Intelligence (AI) tools in content moderation are adept at handling toxicity and incivility, they are unsuitable to promote constructive commenting. Consequently, news outlets task the human moderator with promoting quality comments. This involves manually sifting through growing discussions to identify user-generated contributions that align with the editorial vision of a constructive comment. This creates what I propose to call a 'third half of the internet': a space positioned between the outlet's journalistic content and user-generated comments, hand-picked by moderators, and guided by editorial preferences. It entails a big change in how the comment section is viewed. Traditionally the 'bottom half of the internet' was a disconnected space from the editorial work of journalists, where rowdy and wild but free exchange between non-professional commenters took place. It has, however, become more common for news outlets to see the comment section as integral to their journalistic responsibilities.

More specifically, I analyze how news outlets, aside from deleting unwanted content, promote constructive discussion, embedding it specifically within the context of hybrid content moderation. This work contributes to existing research through its focus on 'good' or constructive comments. I present five cases of major news outlets with large comment sections: *New York Times*, *El País*, *Die Zeit*, *The Guardian* and *NU.nl*. The emphasis is on how these news outlets have recently implemented content moderation to address toxicity as well as fostering increased constructive discussion. For this analysis, I compared platforms' public documents explaining their moderation policies in addition to analyzing how the promotion of constructive commenting is visually represented in the interface. The comparison highlights the diverse approaches these outlets adopt to cultivate a constructive comment section. It details the methods and strategies they use to mitigate toxicity and highlight constructive contributions. Finally, I discuss the interplay and division of tasks between human and non-human (AI) moderation, as this combination defines how comment spaces will be policed and shaped in the foreseeable future.

2.2. Challenges to the comment section

As mentioned, news outlets frequently encourage user participation on their online platforms for various reasons, such as boosting overall webpage traffic or generating new stories (Manosevitch and Tenenboim 2017). Moderating comment spaces aligns with the economic interests of news outlets, as dark participation tends to deter advertisers (Paßmann, Helmond and Jansma 2023). However, online journalism and comment sections on news platforms had to adapt to substantial challenges. One prominent obstacle is the growing presence and impact of online misinformation and disinformation (Lewandowsky, Ecker and Cook 2017). Misinformation, for instance, may overshadow valid information presented by journalists, prompting questions about the responsibility of those hosting comment spaces concerning the spread of potentially harmful content (S. v. d. Linden et al. 2017; McCright, Charters et al. 2016). In response to these challenges, content moderators and editors have advocated for more dialogue and increasing audience engagement (Meier, Kraus and Michaeler 2018).

Over time news outlets have shifted away from a strict top-down approach based on the lecturing of readers, which entailed, for example, the presentation of netiquette specifically telling users how to behave online and what not to do (Scheuermann and Taylor 1997). This was seen as a necessity for adapting to the changing online environment and, subsequently resulted in a community manager role for those in charge of the comment space of news outlets (Meier, Kraus and Michaeler 2018). Consequently, news outlets have explored various approaches for setting strategic and operative goals, including banning repeat offenders or, in some cases, completely abandoning the comment space (Meier, Kraus and Michaeler 2018).

Content moderation itself is frequently characterized as a gatekeeping role (Paasch-Colberg and Strippel 2022; Wolfgang 2018). This gatekeeping function is twofold. First, the moderator can delete toxic comments or block users. Second, constructive or beneficial content can be promoted (Wolfgang 2018). These two objectives are interconnected, as mitigating toxicity can create room for constructive discussion (Paßmann, Helmond and Jansma 2023). Such constructive dialogue and wider audience engagement are cornerstones of constructive journalism (Løvlie 2018). Online commenting facilitates valuable reader-journalist interaction and promotes connections among readers (Løvlie 2018). Enhancing these interactions, while simultaneously mitigating toxicity, creates a monetizable and constructive comment section. Additionally, interesting comments can also provide new story leads and enrich journalistic articles (Manosevitch and Tenenboim 2017). However, defining what constitutes a good discussion or constructive comment is challenging. In theory, constructive comments may be perceived as evidence-supported, well-written contributions that are relevant to the article (Kolhatkar and Taboada 2017). In practice, evaluating online comments in terms of constructiveness or quality proves to be much more complex. Furthermore, there has been relatively little research into what constitutes ‘constructive participation’ concerning online user comments, particularly in terms of how news platforms operationalize the promotion of such user content.

The introduction of AI systems has significantly reshaped the role of moderators. The sheer volume of comments and the possibility for storing data prompted platforms to integrate (semi-) automatic filtering tools, aiming to ease the moderators’ workload (Diakopoulos 2019; Paßmann, Helmond and Jansma 2023). However, moderators and publishers remain skeptical of these tools as they have not been designed with the practical human-computer interaction of hybrid content moderation in mind (Gollatz, Riedl and Pohlmann 2018). While AI nowadays has a firm presence in the practice of content moderation, many practitioners believe that AI must be limited to supporting human moderators, not replacing them altogether (Ruckenstein and Turunen 2020).

In what follows this chapter offers an analysis of five distinguished online news platforms. The chosen outlets, namely *The Guardian* (United Kingdom), *Die Zeit* (Germany), *El País* (Spain), *New York Times* (US), and *NU.nl* (The Netherlands), are characterized by their substantial online presence and commitment to upholding international journalism standards. These news organizations typically publish documents regarding their comment moderation policies. These documents shed light on the rationale behind their moderation guidelines and provide essential information for readers interested in contributing comments. I collected these documents during two periods: July-September 2021 and May-June 2023.

The analysis of these cases in the subsequent sections is structured around three main categories. The first examines technical aspects, such as login requirements, the comment interface, and user-interaction buttons. The second category investigates moderation features, focusing on how these outlets manage and filter out harmful or inappropriate comments. The final category addresses constructive commenting features, exploring the strategies these news outlets employ to encourage meaningful and constructive reader engagement.

2.3. The comment interface as a tool to stimulate user participation

This comparative analysis of five online news platforms addresses several technical aspects that may hamper or encourage user participation. The comment interface plays a pivotal role in shaping how users engage in online discussions (Stroud, Muddiman and Scacco 2017). Sorting comments by means of likes and popularity can reinforce partisanship (Shmargad and Klar 2020). I also consider if users can like or dislike others’ contributions, taking note of the specific semantics. Here the choice of terminology matters too; for example, a ‘respect’ button tends to foster fewer partisan comments compared to a ‘like’ option (Stroud, Muddiman and Scacco 2017).

2.3.1. Barriers to participation

All examined news outlets require a user account for individuals to comment on a news article, thereby imposing a restriction on participation. The *New York Times* has a paywall, requiring readers to subscribe not only to engage in commenting but also to access the article. Articles by *El País* become accessible when readers opt to allow advertisements on the webpage. However, commenting is restricted solely to users with a subscription. On the other hand, *Die Zeit*, *The Guardian* and *NU.nl* follow a less restrictive model, requiring a free user account for participation. During the sign-up process for such an account, the presentation of participation guidelines is a possibility. The Guardian does include them during the sign-up process. In contrast, *NU.nl* and *El País* display their ‘house rules’ above every comment section. *The New York Times* organizes its guidelines under the heading FAQs. Although *Die Zeit* maintains a netiquette page, it is not prominently linked on their comment interface, potentially affecting the visibility of these guidelines for users.

The majority of news outlets limit their comment space to pre-selected articles, such as *The Guardian*’s opinion and sports sections. *Die Zeit* and *NU.nl*, however, distinguish themselves by permitting commenting on all articles from their own editorial offices. This practice of pre-selecting articles serves the purpose of topic curation, enabling a conscious decision on which subjects are deemed suitable for online discussion. Additionally, it helps in managing the workload for moderators by constraining the number of open discussions that need simultaneous oversight.

2.3.2. Buttons & their semantics

In terms of buttons and their semantics, all platforms provide users with the opportunity to ‘like’ comments, but the terminology varies. *NU.nl* speaks of ‘respect’, *Die Zeit* has ’stars’, while *The Guardian* and *New York Times* opt for a ‘recommendation’. Notably, *El País* is the only included news outlet in this sample that has a dislike option, suggesting a deliberate choice. *NU.nl* explicitly states on its FAQ page that they aim to foster a positive environment where a dislike button has no place (NUJij 2018). As of April 2023, *Die Zeit* has expanded its options by introducing various emojis, in addition to the existing stars, for users to assign to a comment (Berresheim and Meyer 2023). For moderators, these ‘like’ features could also serve as markers for user reputation or signals of comment quality (Paßmann, Helmond and Jansma 2023).

Regarding sorting, all five news outlets provide users with a variety of sorting options to influence user behavior, with a common feature being the ability to sort comments by popularity. In addition to popularity-based sorting, platforms typically offer options to rank comments from oldest to newest and vice versa. *NU.nl* goes a step further by allowing sorting based on the number of replies. Upon opening the comment section, comments on *El País, New York Times*, and *The Guardian* are typically sorted from newest to oldest. However, *NU.nl* and *Die Zeit* adopt a unique standard approach by ranking user comments based on ‘respect’ points (likes). Consequently, readers initially encounter contributions with the highest number of ‘likes’ from other users when scrolling through comments.

The factors discussed above are intended to enhance the opportunities for positive user participation. The increasing number of commenting options, coupled with diverse ways of engaging with others’ comments, has resulted in a surge in activity and an ever-growing workload for moderators. Consequently, platforms found themselves compelled to expand and invest further in their moderation practices to effectively manage the sheer quantity of user contributions.

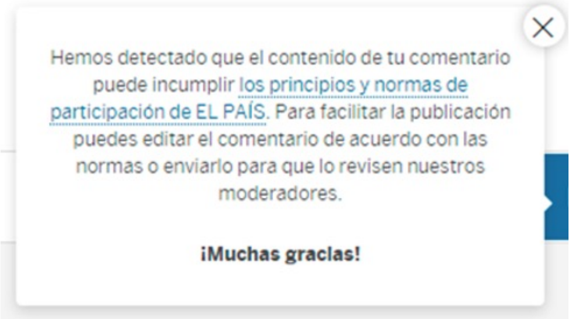
2.4. Combatting toxicity with AI-based moderation

(Partially) automating the moderation process provides the advantage of expanding comment and moderation possibilities, especially in terms of enabling more articles with open comment spaces. Prior to the integration of AI in comment sections, it was not uncommon for platforms to disable comment sections altogether (Goldberg 2018; Hoekman 2016). As an example, The *New York Times* only opened comment sections for approximately 10% of articles before implementing the Perspective API, primarily due to the manual workload associated with content moderation (New York Times 2017). By 2017, the implementation of AI tools had increased the comment space to 25%. Although AI-based tools alleviate some of the pressure on moderators, they still necessitate significant human judgment and expertise.

AI-assisted moderation has become a standard feature in the comment spaces of most major media outlets. They generally employ AI-assisted moderation in a limited and focused manner, primarily for detecting and preventing toxic content. The rapid increase in user comments necessitated the implementation of these systems, as human moderators were unable to manage the sheer volume. These AI tools are specifically trained to assess comments for toxicity, restricting their application to this area. In this hybrid setup, AI plays a specific role, allowing human moderators to concentrate on other aspects of moderation.

Additionally, we see that either they rely on pre-built solutions or develop their own solutions. As an example of a pre-built AI solution, the *New York Times* collaborated with *Jigsaw (Google)* in 2016 to develop the *Perspective API* (Salganik and Lee 2020; New York Times 2016). This API incorporates toxicity filtering in comments, empowering the *New York Times* to partly automate their moderation process within the ‘Moderator’ toolkit (Rieder and Skop 2021). Marked comments are evaluated by human moderators who determine whether they can be published (Salganik and Lee 2020). This approach has enabled the *New York Times* to open more comment sections (New York Times 2017). While *Perspective API* was originally based on English data, it has been subsequently expanded to encompass multiple languages. Notable, the Spanish newspaper *El País* has adopted the same system for filtering toxicity in their comment space since 2018 (Delgado 2019; El Pais 2018). *El País* utilizes a real-time evaluation to detect toxicity through a warning system (Figure 2.1). Users attempting to submit a comment flagged as toxic by the API receive a warning and are prompted to modify their comment appropriately.

It is, however, essential to acknowledge the limitations of such systems, as computational models may produce inaccurate or incorrect results. Analysts at the *New York Times* have raised concerns about identity bias in their use of the *Perspective API*, noting that identity statements such as "As a Jewish man" resulted in higher toxicity scores compared to comments without such identity markers (Salganik and Lee 2020). Dutch news outlet *NU.nl* utilizes a commercial toxicity filter for their comment sections as well, developed by *Utopia Analytics* and implemented since 2019 (Van Hoek 2020; Utopia 2021).



Hemos detectado que el contenido de tu comentario puede incumplir los principios y normas de participación de EL PAÍS. Para facilitar la publicación puedes editar el comentario de acuerdo con las normas o enviarlo para que lo revisen nuestros moderadores.

¡Muchas gracias!

Figure 2.1: Warning message while attempting to comment on an article *(El País)*

For news outlets and publishers, an alternative to outsourcing or purchasing pre-built AI solutions is to develop their own. Although this option demands expertise and investments, it offers a significant advantage. Platforms can maintain control and exert more agency over the processes that shape their comment space. *The Guardian* has been developing its own computational models for managing incoming comments since 2016. Their system, known as 'Robot Eirene', was described in a written statement to the Parliamentary Communications and Digital Committee in April 2021 (The Guardian News & Media 2021): "[...] Eirene does not replace human moderators, but rather it serves to reduce the volume of comments in our queues and to have high risk comments flagged to the moderation team." Interestingly, *The Guardian* suggests that the system could potentially be used to identify 'good' comments, a departure from the conventional focus on toxicity filtering (The Guardian News & Media 2021). However, any application to identify good behavior has yet to be developed and applied. Similarly, the German newspaper *Die Zeit* started developing their own AI-tool in 2016 under the name 'Robot Zoë' to handle the substantial increase in comments over time (Loos 2016). Nonetheless, they clearly state that detecting 'good' comments is not currently a technical option for such a system (Ogolla and Hard 2020).

An essential consideration when implementing AI-based moderation tools is system transparency, which is closely tied to user trust (Brunk, Mattern and Riehle 2019). Many existing systems function as black boxes, providing no insight into the algorithmic decisions they generate. News platforms must possess the expertise to maintain transparency in their hybrid moderation practices, clearly delineating the roles assigned to both 'humans' and 'machines.' Moreover, a strict distinction between the two actors in hybrid moderation can obscure how they converge and interact in practice (Rieder and Skop 2021). Demonstrating how certain moderation decisions are made and how AI systems evaluate incoming comments, is crucial for both moderators and readers. This transparency allows them to demand explainability as part of the hybrid decision-making processes (Molina and Sundar 2022).

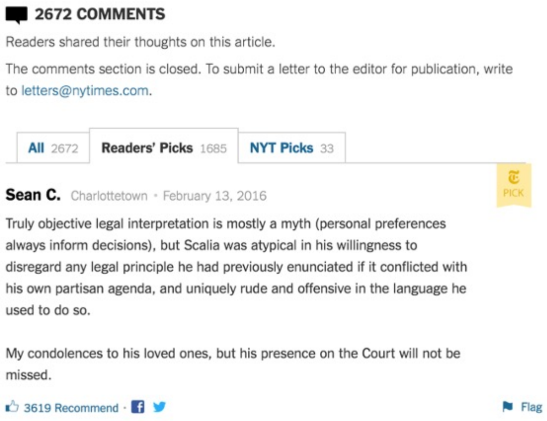
2.5. Promoting constructive commenting: the Third Half of the Internet

As discussed earlier, online news outlets hosting online comment spaces not only focus on filtering out unwanted comments but also increasingly strive to promote constructive discussion. This rather recent emphasis is distinct from toxicity filtering, as it specifically aims to encourage users to contribute what they perceive as constructive comments. In practical terms, this emphasis is operationalized by highlighting certain comments within a discussion. However, the AI-tools models discussed earlier are unsuitable for this task, as they are trained to assess comments in terms of toxicity. Consequently, the responsibility of sifting through discussions and identifying constructive comments often falls on the shoulders of human moderators. Moderators must make choices based on editorial standards and expectations. In the following paragraphs, I illustrate how each news outlet implements similar moderation strategies, mobilizing moderators to promote and feature desirable comments.

The New York Times employs the term 'NYT picks' to highlight selected comments. According to their FAQ page, these comments represent a range of views or are written by "readers with first-hand knowledge" (New York Times 2020). In addition to NYT picks, the news outlet features Readers' picks, defined as "a selection of comments with the highest amount of recommendations or upvotes" (New York Times 2020). These Readers' picks give users a sense of agency regarding elevating constructive comments. Both these categories are presented in separate tabs within the interface (Figure 2.2).

NU.nl designates their editorial selection of user comments as 'Highlighted comments'. According to their definition, these contributions are "well thought out and respectful" and "not selected based on political preferences" (NUJij 2018). Furthermore, the FAQ page specifies that they serve as an example to other users (NUJij 2018). Selected comments receive a star badge and are presented in a separate tab on the interface (Figure 2.3a). In addition to editor picks, *NU.nl* has implemented a user labelling system on their comment platform. The news outlet offers the possibility to add your job title as a so-called expert label (Figure 2.3b). To obtain such a label, visible on your comments, you will need to provide proof in the form of a contract, diploma, company website or a trustworthy *Linkedin* page (NU.nl 2020). This strategy aims to enhance the trustworthiness of comments and user-contributors. Furthermore, the *NU.nl* editors invite these experts to contribute to future stories (NU.nl 2020).

The Guardian calls their editor picks 'Guardian Picks' and prominently displays them



2672 COMMENTS

Readers shared their thoughts on this article.

The comments section is closed. To submit a letter to the editor for publication, write to letters@nytimes.com.

All 2672 | Readers' Picks 1685 | NYT Picks 33

PICK

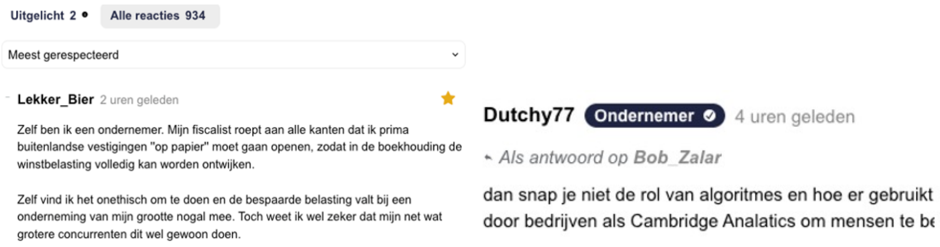
Sean C. Charlottetown · February 13, 2016

Truly objective legal interpretation is mostly a myth (personal preferences always inform decisions), but Scalia was atypical in his willingness to disregard any legal principle he had previously enunciated if it conflicted with his own partisan agenda, and uniquely rude and offensive in the language he used to do so.

My condolences to his loved ones, but his presence on the Court will not be missed.

3619 Recommend · Flag

Figure 2.2: Separate tabs with NYT picks and Readers' Picks



Uitgelicht 2 • Alle reacties 934

Meest gerespecteerd

Lekker_Bier 2 uren geleden

Zelf ben ik een ondernemer. Mijn fiscalist roept aan alle kanten dat ik prima buitenlandse vestigingen "op papier" moet gaan openen, zodat in de boekhouding de winstbelasting volledig kan worden ontwijken.

Zelf vind ik het onethisch om te doen en de bespaarde belasting valt bij een onderneming van mijn grootte nogal mee. Toch weet ik wel zeker dat mijn net wat grotere concurrenten dit wel gewoon doen.

Dutchy77 Ondernemer ✓ 4 uren geleden

Als antwoord op Bob_Zalar

dan snap je niet de rol van algoritmes en hoe er gebruikt door bedrijven als Cambridge Analatics om mensen te be

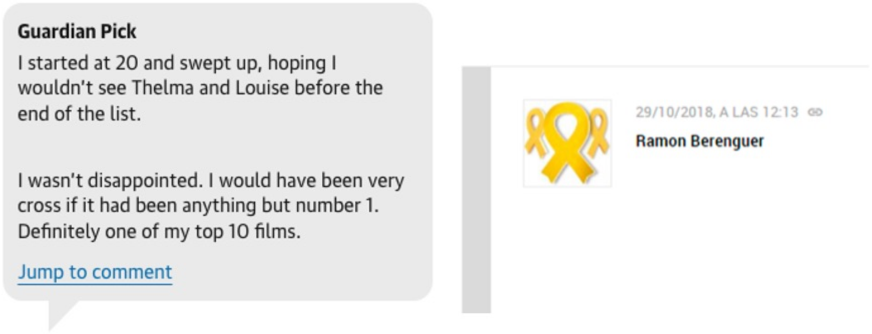
(a) Highlighted comment **(b)** Expert label

Figure 2.3: Promoting constructive commenting on NU.nl

at the top of the comment interface, presenting them in a speech bubble (Figure 2.4a). Interestingly, while the previous three platforms have a rather uniform implementation of promoting constructive comments, *Die Zeit* and *El País* differ. The former used to have editor picks ('Redaktionsempfehlung'), but this feature seems to be disabled without an editorial statement about its current status (D. Schmidt 2014)[1]. Browsing through *Die Zeit*'s sitemap[2], it seems that they may have partially or fully abandoned the approach in 2015 or 2016. Spanish newspaper *El País* has opted for a distinctive approach to highlighting content by awarding gold user badges to recognize outstanding, constructive users (El Pais 2015). To receive such a reward, users must have a history of 'beneficial participation' in the comment section (El Pais 2015). Distinguished users are granted extra visibility when commenting on news articles (Figure 2.4b). When these users make changes to their profile, the modifications must be pre-approved by moderators before becoming visible online (El Pais 2016).

[1] In their renewed comment interface announcement (4 April 2023), editor picks (Redaktionsempfehlung) are mentioned. However, there are no examples found within the comments on articles. `https://www.zeit.de/administratives/2023-04/kommentarbereich-design-struktur-emojis`

[2] `https://www.zeit.de/gsitemaps/index.xml`



Guardian Pick

I started at 20 and swept up, hoping I wouldn't see Thelma and Louise before the end of the list.

I wasn't disappointed. I would have been very cross if it had been anything but number 1. Definitely one of my top 10 films.

29/10/2018, A LAS 12:13

Ramon Berenguer

(a) Guardian Pick on The Guardian

(b) Highlighted user on El País

Figure 2.4: Promoting constructive commenting on The Guardian and El País

While the implementations for promoting constructive commenting have much overlap, their differences have important implications. First, awarding user badges instead of highlighting individual comments places a higher demand on user-contributors, as it considers their commenting history (El Pais 2015). Simply writing a qualifying comment is insufficient for recognition; users are encouraged to participate and contribute to constructive discussions consistently. Second, the direct visibility of highlighted content varies across news outlets. At *NU.nl* and *The Guardian*, the first comments encountered are those handpicked by moderators ensuring that readers initially interact with this filtered content. In contrast, at *The New York Times*, users need to navigate to the 'NYT picks' tab on the comment interface, giving them the option to avoid reading the specific content chosen by moderators. Finally, *Die Zeit* does not make use of 'Redaktionsempfehlungen' and did not make an editorial statement clarifying the abandonment of the moderation strategy. Other news outlets explicitly mention the moderation strategy and emphasize the importance of promoting constructive participation. At *Die Zeit*, this task may not hold similar significance, as discussions lack highlighted comments even though the term is still mentioned on the Netiquette webpage and throughout comment section updates by staff.

All in all, news outlets identify what they consider constructive comments and prominently feature them at the top of the comment section or on a dedicated page, creating a space between the editorial content (article) and the user-generated comments. Reagle (2015) describes the latter as the 'bottom half of the internet,' making this novel space the 'third half of the internet.' User-generated comments in this section build upon the news outlet's content, reinforcing or confirming the editorial view on constructive discussion.

Questions remain, however, regarding the effect of the 'third half' on the online discussion and the user base. In pursuit of the goal of editor's picks, has it succeeded in fostering a different kind of debate in comment sections compared to pages without highlighted comments? Evaluating specific interventions can assist news outlets in optimizing the human effort invested in the moderation process. Additionally, the rationale behind choosing what is deemed worthy of being featured remains unclear. News outlets often employ broad and ambiguous language to describe what constitutes a constructive comment. To achieve a clearer understanding of the universal characteristics of constructive commenting, it is essential to undertake a comparative analysis across various platforms. Such analysis should concentrate on pinpointing the types of user comments that are commonly highlighted or encouraged across different news organizations. By identifying these commonalities, we can better understand the general standards and expectations for constructive comments in online news forums.

2.6. Conclusion

In this chapter, I conducted a review of five different news outlets renowned for their prominent online comment section, aiming to grasp their recent strategies in managing user-generated content. My primary focus centered on their approaches to excluding toxic content and their emerging emphasis on fostering constructive discussion, all aimed at sustaining a monetizable and vibrant comment section. The conclusions are twofold.

First, the case studies reveal a clear trend of safeguarding the comment space from toxicity using (semi-)automated AI-based tools. These tools are specifically trained and implemented for this task, confining their scope to toxicity filtering. While some outlets have outsourced this practice to tech companies, others have opted to develop their own systems, affording them greater control and insight into the used models. The fast-paced evolution of these computational models has the potential to alter the current state of hybrid content moderation, possible reshaping the role AI models play in online content moderation once again. These moderation strategies will face challenges from new configurations of hybrid content moderation. The recent introduction of the newest generation of Large Language Models (LLM) including *ChatGPT* could further expand the use of automated content moderation, potentially using AI-based tools to detect constructive discussion as well. Given the highly subjective and context-dependent nature of promoting constructive discussion, along with the visibility and expressiveness that endorsed comments and their content receive, it is essential for news outlets to carefully consider the extent to which they integrate AI models into the hybrid moderation pipeline. At any rate, comment sections are still evolving at a fast pace, as seen in the recent revamp at *Die Zeit* (Berresheim and Meyer 2023).

Second, the emphasis on promoting constructive discussion takes the form of handpicking specific content, elevating it to greater visibility within the comment interface. This is commonly achieved through (human) editor's picks, while awarding user badges is an alternative strategy. Ethnographic fieldwork could provide insights into the operationalization of constructive commenting by human moderators and their interactions with users. Preliminary fieldwork with content moderators has indicated that they recognize constructive discussion even when it cannot be precisely defined, suggesting a high degree of subjectivity and contextual awareness. Elevating user-generated content that aligns with editorial standards establishes a distinct space between published journalistic articles and unfiltered user content — the 'third half of the internet'. Nevertheless, to maintain standards of quality journalism, moderation policies for the comment section need to articulate what the editorial staff defines as ‘constructive participation’ and discussion.

The shift towards promoting what is deemed constructive and the presentation of it in the ‘third half’ of news outlets raises unanswered questions and consequences. The task of filtering out the most constructive comments has so far been assigned to the human moderator, yet the definition of this concept is vague and often ill-defined. Evidently, the rather vague practice of manually picking out single comments may advance human bias in the comment section, as the moderators can act autonomously, evading discussion with colleagues due to time constraints or other factors. An open and transparent procedure of (human) moderation enhances checks and balances in the comment space. Constructive discussion, in this case, arises from the moderators’ perspective rather than reflecting the user base. In this thesis, I refer to this as the emic perspective on constructive commenting. There is clearly a point of friction when the users’ perspective does not align with the moderators’ definition of ‘constructive participation’. A more in-depth examination is necessary to understand precisely how online discussions are significantly influenced by (human) online moderation. A step towards this understanding is presented in Chapter 6.

3

Calculating Argument Diversity in Online Threads

The discussion is actually pretty simple.
On the one hand, a large group loathing a dubious 'convention'.
On the other, a group willing to adhere to their traditions.

Paraphrased NU.nl comment

Based on: Waterschoot, Cedric; van den Bosch, Antal; van den Hemel, Ernst, Calculating Argument Diversity in Online Threads (2021). *3rd Conference on Language, Data and Knowledge (LDK2021).* Vol. 93 Dagstuhl, Germany: OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl

Abstract

In this chapter, we propose a method for estimating argument diversity and interactivity in online discussion threads. Following the etic perspective on constructive commenting, comments contributing novel arguments and information to the discussion might be deemed constructive. Using a case study on the subject of Black Pete ('Zwarte Piet') in the Netherlands, the approach for automatic detection of echo chambers is presented. Dynamic thread scoring calculates the status of the discussion on the thread level, while individual messages receive a contribution score reflecting the extent to which the comment contributed to the overall interactivity in the thread. We obtain platform-specific results. Gab hosts only echo chambers, while the majority of Reddit threads are balanced in terms of perspectives. Twitter threads cover the whole spectrum of interactivity. While the results based on the case study mirror previous research, this calculation is only the first step towards better understanding and automatic detection of echo effects in online discussions.

3.1. Introduction

No shortage exists in regard to online discussions, whether raging on social media or on other websites including those of media outlets. A substantial amount of work has focused on particular aspects of such debates, such as filter bubbles, the purported consequence of personalization in search and recommendation algorithms (Pariser 2011), and echo chambers, clusters of like-minded individuals amplifying their unison reasoning (Flaxman, Goel and Rao 2016). What has been sparsely studied, however, is how individual messages contribute to the interactivity of an online discussion thread, either towards an echo chamber or balanced discussion.

This chapter presents a method for the automatic scoring of a discussion thread in terms of interactivity and argument diversity, as well as for grading each individual comment within the thread on the basis of interactive contribution at the time of posting. Taking an etic perspective of constructive online participation, contributing to a balanced debate containing differing arguments may be seen as constructive. On the other hand, an endless repetition of identical arguments cannot be deemed constructive. The starting point of the analysis is a dataset of messages where each sample has been labelled for the argument it presents. To illustrate the scoring of discussion threads, the case study in this chapter deals with the 'Zwarte Piet' (Black Pete) debate in the Netherlands, a topic with clear 'pro' sides, i.e. in favour of the figure, and 'con' side against the continued existence of 'Zwarte Piet'.

First, the literature on online discussions, echo chambers and argument diversity is discussed. Then, the scoring methodology is unpacked. The chapter ends by discussing the methodology, limitations and what to focus on in future research.

3.2. Background

Echo chambers and social media is a much discussed topic that has received ample attention from different perspectives, whether political, academic or from the media. An echo chamber is understood to be an enclosed, discursive space, online or based on other forms of media, which amplifies the uniform message encapsulated within. This process magnifies the shared opinion within the cluster while insulating it from rebuttal, creating an environment of positive feedback loops (Jamieson and Capella 2008).

Previous research tends to agree that echo effects exist on social media platforms, even though the concept remains contested (Flaxman, Goel and Rao 2016; Williams et al. 2015a; Du and Gregory 2017). A possible cause for such an echo effect is the fact that social media users have the tendency to discuss matters with like-minded individuals (Du and Gregory 2017). It has been concluded that this restricted debate increases polarization (Abramowitz and Saunders 2008; Sunstein and Vermeule 2009). However, others have criticised single media studies for echo chamber detection as it does not take into account the 'multiple media environment' that we find ourselves in today (Dubois and Blank 2018).

The notion of an echo chamber is seen as disadvantageous by dominant conceptions about democracy as well as by stakeholders in media and moderators. Discourse with those holding differing opinions increases understanding of the subject matter and tolerance for those who disagree (Mutz and Mondak 2006). This chapter aims to contribute to the development of information systems dealing with online discourse, by mapping interactivity of polarized debates.

The automated classification of echo chambers is not a much discussed topic, even though studies have focused on the subject, particularly in the field of politics. One study has outlined that homophily of social media feeds can be determined across groups by assigning users to either Democrats or Republicans (Colleoni, Rozza and Arvidsson 2014). Furthermore, network analysis has shown the online clustering of communities holding similar views regarding climate change (Williams et al. 2015a).

The current model aims to fill the gap and complement the research on echo chamber detection in pro/con-discussions by implementing domain unspecific calculations based on annotated data, meaning any labelled data can be used, regardless of the debate statement. The unit of analysis is the thread. Such discussions can either be balanced in terms of argumentation or skewed to one perspective. A second indicator is calculated at the message level, as every individual reply in a thread receives a contribution score.

From here on out, an *echo chamber* will refer to a thread in which the argumentative position presented in the parent message – the contribution starting the thread to which others have replied – is continued throughout the thread, per calculation. The opposite, in which the contrasting argumentative camp, whether pro or con, is the dominant presence in the thread, will be called an *opposition flood*. Equal presence of pro and con messaging results in a *balanced* discussion. A thread can be interpreted as a string of messages portraying an argument belonging to either the pro or con camp where all replies comment on the parent message. Simplified examples are as follows in the form $\{firstcomment \rightarrow replycomment \rightarrow replycomment \rightarrow ...\}$:

$$\begin{aligned} \text{Echo chamber} &:= X_{pro} \rightarrow Y_{pro} \rightarrow X_{pro} \rightarrow X_{pro} \\ \text{Opposition flood} &:= X_{pro} \rightarrow Z_{con} \rightarrow L_{con} \rightarrow M_{con} \\ \text{Balanced} &:= X_{pro} \rightarrow Z_{con} \rightarrow Y_{pro} \rightarrow M_{con} \end{aligned} \tag{3.1}$$

3.2.1. Case study

To illustrate the approach, an annotated dataset containing online threads discussing the controversial blackface figure of Black Pete in the Netherlands was created. This discussion has a clear pro/con divide. Those in favour of the figure, a component of the Dutch Sinterklaas festivities, argue that Black Pete ought to remain as it was celebrated throughout the last decades. The camp opposing the festivities assert that the character is a racist stereotype portraying people of colour and should not be celebrated. This debate ought to be seen more broadly in the discussion on racism in Dutch society (Balkenhol 2015). These threads were collected from social media platforms including *Twitter* (using the keyword ‘Zwarte Piet’), *Reddit*, by scraping the subreddit r/thenetherlands with ‘Zwarte Piet’, and finally *Gab*, also scraped using the hashtag ‘zwartepiet’ (Table 3.1).

Platform	Total Threads	Total Messages
Twitter	21	125
Reddit	7	39
Gab	7	22

Table 3.1: Threads and messages included, sorted by platform

Manual labelling with regard to the included arguments was performed, based on the outline presented in previous research (Schols 2020; Balkenhol 2015; Helsloot 2013; Helsloot 2012) and Table 3.2. Stance labelling of social media data is a challenging task and therefore, it is done at the level of argumentation presented in the literature (Kunneman et al. 2020; Küçük and C. A. Fazli 2020).

Level1 (l1)	Level2 (l2)
Pro	Dutch tradition, Christian tradition, Innocent, Intention, Pre-christian, Oriental
Con	Racial stereotype: historical, Racial stereotype: contemporary

Table 3.2: Arguments (Labels) in the Zwarte Piet discussion

Each comment in the data was labelled for the dominant argument (level2) that it presents in regard to the ‘Zwarte Piet’ discussion (Table 3.2). These labels have been derived from the extensive literature outlining this particular debate in The Netherlands. To test whether such argumentation can be clearly detected in online contributions, multiple annotators were employed to label all gathered comments. The annotation scheme is found in Appendix A. The annotators were familiarized with the discussion and arguments using the existing literature (Schols 2020; Balkenhol 2015; Helsloot 2013; Helsloot 2012). Furthermore, a sheet with all possible labels alongside a brief explanation was provided to guide the labelling process. A Krippendorff’s alpha of 0.745 was calculated, indicating that inter-rater agreement exists.

3.3. Methodology

We propose a calculation method for estimating indicators of interactivity in threads. A first indicator applies to the thread level; a second indicator relates to single messages.

The model created in this paper makes certain assumptions in order to compute interactivity. First, each comment contributes at least one argument in the discussion. Second, each argument can be assigned to a position in the discussion, whether it be ‘pro’ or ‘con’. Additionally, it is assumed that the more an argument is repeated, the smaller the contribution a new repetition will make in terms of diversity/interactivity on the individual message level. However, when calculating the state of the thread as a whole, a new repetition will weigh greater towards the extremes of echo chamber/opposition flood, i.e. constant repeating of identical reasoning will result in an echo chamber or opposition flooding faster.

3.3.1. Thread Interactivity Score

The thread as a whole receives a single score based on the interactivity and diversity detected in the comments. This real-valued indicator provides information on whether the presented collection of arguments constitutes an echo chamber, opposition flood or a balanced discussion. To compute the overall thread interactivity score, each message receives a cumulative log operator, which increases as an identical argument is repeated within the thread. Using this factor, repetition of a single reasoning weighs heavier towards the extremes, either echo chamber or opposition flood.

Calculating the log operator for both the echo and opposition scores requires the cumulative count of the argument (denoted as j) in each message at that point in time. Simply put, this variable equals the nth iteration of the particular argument represented in the sample at the order given in the data. To calculate the actual log operator, $log_{10}(j)$ is substracted. Dividing the log operator by the total number of messages in the thread (N) results in the message share. Per the assumptions, each argument can be assigned to either the 'pro' or 'con' side, which is notated as $l1$ of an argument, the deciding factor whether the share is negative or positive (denoted as multiplication by -1). The specific argument as presented in the case study is decoded as $l2$. The Thread Interactivity Score (TIS) is sum of all shares in thread T. An exception exists for replies where the specific argument is identical to the parent message. In this case, the share is multiplied by a weight and added to the parent message share that is not weighed down, with the result that a parent repetition impacts the echo score to a larger degree.

$$Share_i \begin{cases} \frac{j(x_i)-1-log_{10}(j(x_i)-1)}{N} * (-w) + \frac{1}{N} & \text{if } l2(x_i) = l2(x_0) \\ \frac{j(x_i)-log_{10}(j(x_i)}{N} & \text{if } l2(x_i) \neq l2(x_0) \wedge l1(x_i) \neq l1(x_0) \\ \frac{j(x_i)-log_{10}(j(x_i)}{N} * (-1) & \text{if } l2(x_i) \neq l2(x_0) \wedge l1(x_i) = l1(x_0) \\ 0 & \text{if } i = 1 \end{cases} \tag{3.2}$$

$$TIS_T = \sum_{i=1}^{N} share_i$$

A perfectly balanced discussion will have a TIS of 0, indicating that both the echo share and opposition are equal. An echo chamber is defined as a thread with a TIS below -0.5. Dipping below this threshold means that the share of echo comments is more than double that of the opposition comments. Threads with a TIS above 0.5 are overflooded with opposition messaging.

The opposition score is defined as the sum of shares of all messages from the opposite side of the parent argument on $level1$ (l1), while the echo score is the result of summing the shares in absolute value of all messages where $level1$ equals that of the parent.

To detect when a thread turns into an echo chamber or opposition flood, the TIS is calculated at each new posting in an iterative manner. Thus, it combines the log operator from the TIS with a time-dependent factor. This approach might enable future research to study trends in online discussions in regard to echo chamber prediction. The result is a matrix of message shares, calculated at each new posting in the thread at that point in time. Dynamic scoring follows equation 3.2 in which thread size N equals message index i at the point of calculation.

3.3.2. Message Interactivity Contribution

Alongside the indicators calculated at the thread level, individual comments receive a diversity score representing the extent to which this comment at the time of posting contributed to the thread in terms of interactivity. Simply put, if the new comment presents an argument that has not been part of the discussion, it contributes more to the thread compared to when perspectives are repeated. Subsequent repetition of identical arguments are downgraded by the individual log operator, which decreases the more an already presented argument is added. The message contribution of reply i is calculated as follows:

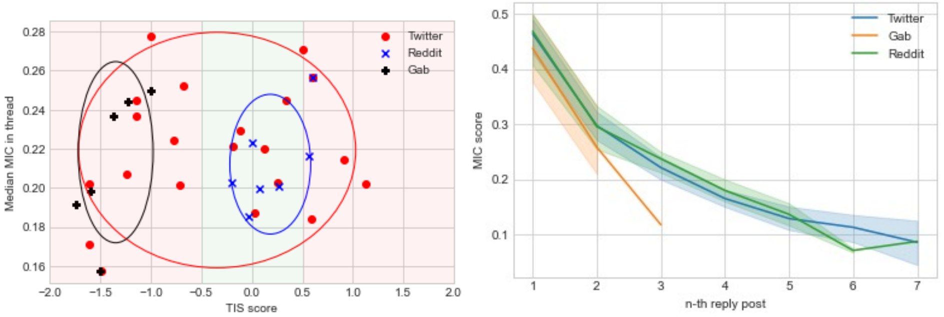
$$MIC_i \begin{cases} \frac{1-log_{10}(j(x_i))}{i} * w^{-1} & \text{if } l2(x_i) = l2(x_0) \\ \frac{(1-log_{10}(j(x_i))}{i} & \text{if } l2(x_i) \neq l2(x_0) \\ 0 & \text{if } i = 0 \end{cases} \tag{3.3}$$

To derive this MIC indicator, the message share at that point in time is calculated using the individual log operator, which decreases if an argument was already prevalent in the discussion. This share equals one minus the log of the cumulative count of the argument, i.e. j, divided by the number of arguments in the thread at the point in time of the message (i). The first comment of a thread always receives MIC equal to zero, as it is not a reply and due to the thread score remaining zero at that point in time. When the parent argument is repeated, the contribution is downgraded by the inverse of the weight. Large MIC values indicate greater contribution to the argument diversity within the thread. Following equation 3.3, the MIC in a thread converges to zero as the thread size grows.

To determine whether a message is an interactive contribution to the thread in terms of argument diversity, the current MIC value of comment i is compared to that one of the previous comment $i-1$. Replies with a greater MIC score than the previous comment are deemed interactive contributions. In case the first reply comment contains identical argumentation to the original comment, it cannot be seen as a contribution in terms of interactivity.

3.4. Results

The first obtained indicator is the Thread Interactivity Score (TIS), the overall score as a whole, plotted alongside the median MIC score in the thread (Figure 3.1a). TIS informs you whether the thread is an echo chamber, balanced debate or opposition flood. Balanced discussion is found when the TIS falls within the interval $[-0.5, 0.5]$, indicating a somewhat equal distribution of arguments. Threads with a score below -0.5 are deemed echo chambers, above 0.5 as opposition floods where the parent argument is overflooded by opposing messages. For this particular illustration, the weight for *punishing* repetition of the parent comment was kept at 1.1.



(a) Dynamic TIS

(b) MIC at the n-th reply, 95% ci

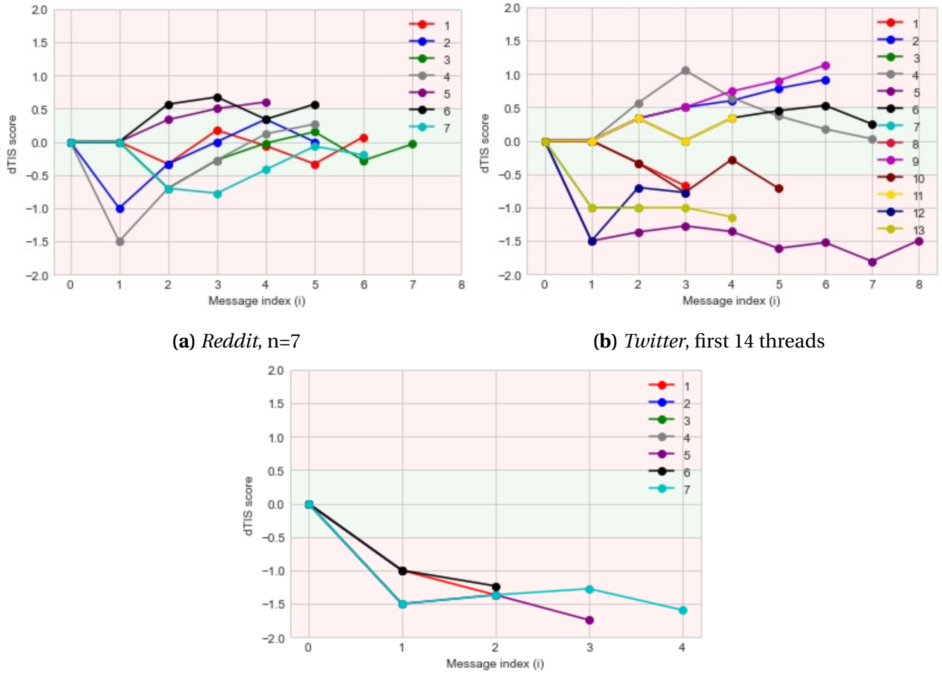
Figure 3.1: Dynamic TIS & MIC scores, Black Pete case study, by platform

The three online platforms showcase different characteristics in regard to overall thread status, at least in this dataset (Figure 3.1a). *Gab* appears to exclusively host echo chambers, confirming previous research (Lima et al. 2018). The ‘Zwarte Piet’ discussion on *Reddit*, however, results in balanced discussion with the exception of two threads. Finally, the TIS result indicates that one finds variability on *Twitter* regarding the thread status, with both echo chambers, balanced discussion and opposition flooding found in this dataset (Figure 3.1a). That being said, the 21 *Twitter* threads plotted here do collectively shift slightly towards echo chambers.

The dynamic TIS (dTIS) informs how a thread developed in terms of argument diversity and interactivity. Figure 3.2 visualizes threads from all included platforms. One can infer from the dTIS when a thread becomes an echo chamber (dipping below -0.5) or if it returns into the green zone, indicating a balanced discussion.

Figure 3.2 indicates that *Gab* lacks any argumentation from one side of the aisle, resulting in direct echo chambers. Secondly, threads on *Reddit* bounce back towards balanced discussion even when the first replies pull the thread towards an echo chamber. Furthermore, the variability in thread structure on *Twitter* are once again visible. Some discussions are echo chambers from the first reply onwards, never experiencing opposite messaging (e.g. thread 5, thread 13), others bounce back and forth between balanced and echo chamber (thread 10). On the other side of the spectrum, threads steadily grow towards opposition flood, meaning that every new reply to the thread argued against the parent message (thread 2, thread 9).

Moving on from the thread scoring, the MIC score reflects how much the comment in question contributed to the argument diversity at that point in time. Figure 3.1b



(a) *Reddit*, n=7

(b) *Twitter*, first 14 threads

(c) *Gab*, n=7

Figure 3.2: Dynamic TIS scores per platform, balanced discussion in $[-0.5, 0.5]$

summarizes this scoring by averaging the MIC score at each subsequent reply across platforms in the dataset.

In the case of *Gab*, where maximum thread size is four, it is clear that, due to the absence of diversity in arguments, replies quickly diminish in terms of contribution. Due to the linear MIC decline in the scraped threads, no reply comments can be deemed beneficial contributions in terms of argument diversity.

However, this cannot be said for the threads scraped from *Twitter* and *Reddit* (Figure 3.1b). The decline in message contribution is less steep compared to *Gab*. Furthermore, on *Reddit*, 14 replies were deemed interactive, meaning that the MIC was larger than the previous message. In the case of *Twitter*, 30 replies were found to be interactive, accounting for about a quarter of included comments.

In the case of the ‘Zwarte Piet’ dataset used for this calculation, one could infer that the most diverse debate in terms of argumentation is found on *Reddit*, due to the fact that a larger share of comments are deemed interactive, combined with the absence of a field dominated by echo chambers. However, this dataset is limited both in scope and size. While these indicators can be used to explore online discussions, in this instance it is a mere illustration of the calculation and variables.

3.5. Discussion & conclusion

This chapter presented a calculation procedure for two metrics for estimating echo chamber effects in online discussion threads. The case study, focusing on the ‘Zwarte Piet’ discussion in the Netherlands, illustrated how the debate exists on different online platforms. Threads belonging to the right-wing network *Gab* exclusively fall into the echo chamber category, in line with the literature (Lima et al. 2018; Zannettou, Bradlyn and Cristofaro 2018). In this specific dataset, the discussion around the ‘Zwarte Piet’ figure on subreddit r/thenetherlands falls mostly within the balanced category. Previous research put forward varied results in terms of echo chambers on *Reddit* depending on the subreddit in question (Mills 2018). Concerning the valuation of replies, the *Reddit* threads hold a larger share of interactive comments compared to *Twitter*. Furthermore, the discussion on *Twitter* experiences wide variability with a slight collective shift towards echo chambers. This divergence in thread status is reflected in previous research on the social media platform, as studies report a variety in results regarding bias and homophily on *Twitter* feeds (Bruns 2017; Williams et al. 2015a; Spohr 2017). Political studies as well as studies focusing on climate change tend to point towards echo effects on *Twitter* (Garimella et al. 2018; Williams et al. 2015b).

Comments deemed interactive by MIC calculation can be valuable for stakeholders. Journalists and moderators aim to have engaging forum discussions on their platform with a large number of participants. Academics might look at interactive comments to map out discussions, understand echo chambers and what effects they have on deliberative debate.

While the discussed indicators do confirm previous research, the approach has its limitations. First, for the approach to provide valid and qualitatively sound scoring, an annotated dataset is needed. This data ought to be labelled for the specific argument or debate stance put forward in the message. Without substantiated labelling, the scoring loses value and interpretability. However, as illustrated by the case study, when threads are well-annotated, the scoring yields understandable results.

The TIS and MIC scoring informs about the status of a thread and contribution of a message in the discussion in terms of argument diversity and interaction across argumentative camps. However, what it lacks is any indication on the quality of the interaction taking place. Understandably, a wide variety exists in terms of constructive communication among commenters on internet platforms and social media. This approach operates at the coarse pro/con and basic argumentative levels, ignoring further depth of the communicative discourse.

Further research is needed to address these limitations. The current chapter is small in scope and size. A larger case is needed to rigidly map out echo chambers on online platforms with the goal of being independent of topic, platform or language. Different weights for parent argument repetition ought to be included as well in order to pinpoint the effect. Additionally, the concept of interaction in online discussion needs to be unpacked in further detail by developing estimators for qualitative features of interaction. By introducing gradation in terms of discursive quality in the process of valuating reply contribution, the depth of such interaction can be included. Studies to come will could pinpoint just that aspect of online threads in order to fill this gap. Moreover, Chapter 4 will focus on the automatic labelling of online comments in regard to presented argumentation. While in this proof-of-concept study this was done manually, the automatic annotation of pro- and con-statements allows for a computational pipeline for echo chamber detection from the ground up. Upcoming research could focus on not only using the ‘Zwarte Piet’ case, but also other discussion cases to include broader topics that do not showcase such strong binary distinction between pro- and con-groups.

The concrete necessity to better outline and understand online discourse and echo chambers becomes more urgent as social media and other online platforms acquire dominance in societal conversation. As this trend progresses, so does the need for research to follow that path and develop automated methods that help detecting adverse and toxic discourse and communication. The presented calculation aims to contribute to this challenge by expanding the computational possibilities for forum and discussion moderation.

4

Detecting Minority Arguments for Mutual Understanding: A Moderation Tool for the Online Climate Change Debate

Here we go again.
A discussion between scientists and
a group of people who simply say that it will all be okay.

Paraphrased NU.nl comment

Based on: Waterschoot, Cedric; Van den Bosch, Antal; Van den Hemel, Ernst. Detecting Minority Arguments for Mutual Understanding: A Moderation Tool for the Online Climate Change Debate (2022). In Proceedings of the 29th International Conference on Computational Linguistics (COLING2022), pages 6715–6725

Abstract

Moderating user comments and promoting healthy understanding is a challenging task, especially in the context of polarized topics such as climate change. In this chapter, we propose a moderation tool to assist moderators in promoting mutual understanding in regard to this topic. The approach is twofold. First, we train classifiers to label incoming comments for the arguments they entail, with a specific focus on minority arguments. We apply active learning to further supplement the training data with rare arguments. Second, we dive deeper into singular arguments and extract the lexical patterns that distinguish each argument from the others. Our findings indicate that climate change arguments form clearly separable clusters in the embedding space. These classes are characterized by their own unique lexical patterns that provide a quick insight in an argument's key concepts. Additionally, supplementing our training data was necessary for our classifiers to be able to adequately recognize rare arguments. We argue that this detailed rundown of each argument provides insight into where others are coming from. These computational approaches can be part of the toolkit for content moderators and researchers struggling with other polarized debates.

4.1. Introduction

Even though a consensus has existed within the scientific community on the topic of human-caused climate change for some time, the online debate remains very polarized. Online comment spaces are typically overwhelmed with a large quantity of contributions. This information flood hinders the promotion of mutual understanding and inclusivity in debate spaces. Additionally, climate change presents a splintered debate with niche opinions and many viewpoints. The recognition of these niche arguments are vital to support the moderator in adhering to the heterogeneous discussion that climate change presents. This setting presents opportunities for mutual understanding by improving issue awareness and the quality of deliberation.

In this chapter, we construct a twofold approach to support mutual understanding in the online climate change discussion, taking an etic perspective on what may be seen as constructive participation: the presence of clear argumentation leading to mutual understanding. First, we aim to classify comments for the argument they present. Second, we dive deeper into singular arguments to create an overview of the lexical patterns in each argument-specific sub-corpus. We conclude the chapter by discussing the limitations of modeling nuanced argumentation by a computational method and link our approach to fields struggling with content moderation and polarized debates.

4.2. Background

4.2.1. Argument Mining & Stance Detection

Our application falls under the umbrella of ‘argument mining’ and ‘stance detection’. Within Natural Language Processing, argument mining is defined as the automated identification and extraction of argumentation found in natural language (Lawrence and Reed 2019). Following the stark increase in the availability of textual data found on online fora and social media platforms, argument detection tasks have been receiving a lot of attention. The related task of stance detection is aimed at classifying the stance of the producer of a piece of text towards the target topic (Küçük and C. A. Fazli 2020). This result is often performed over three classes: *in favour (‘Pro’), against (‘Con’) or neutral.*

To define an argument, researchers often look towards the Toulmin model of argument (Toulmin 2003). Toulmin defined a formal argumentative model comprising of the following five elements: *claim, data, warrant, qualifier, and rebuttal* (Toulmin 2003). However, textual data from social media or comment platforms tend to fall short of fulfilling these formal requirements due to their briefness and elliptic nature. Researchers have therefore labeled tweets as argumentative when a portion of the formal argumentative structure was present (Bosc, Cabrio and Villata 2016). These portions can be a premise, a, conclusion, or the connecting relationship between these two argumentative parts. In this chapter, we follow the same operationalization of the definition of arguments.

One factor further complicating these tasks is the influence of context. Context may affect whether an utterance is interpreted as argumentative or not (Carstens and Toni 2015). Typically, the classification tasks are restricted to features intrinsic to the sentence, comment, or utterance, and are blind to context; therefore, resulting models may not be robust across different contexts (Lawrence and Reed 2019). What makes the contextual factor challenging is the fact that not all content and context is expressed explicitly (Moens 2018). A lot of this knowledge and expression remains "in the mind of communicator and audience" (Moens 2018). Some have even argued that, in particular cases, content can be less important than the context it resides in (Opitz and Frank 2019).

Related to the contextual factor is the importance of previous knowledge in stance detection and annotation. The complexity of stance-taking includes cultural and social aspects (AlDayel and Magdy 2021). Personal opinions and the aforementioned non-personal aspects make stance detection a non-trivial task (Du Bois 2007). In recent years, a range of work has focused on argument detection in online content. The first step of these approaches often relates to making the distinction between argumentative and non-argumentative samples. Previous research performed such a classification while subsequently classifying the evidence type presented within argumentative tweets with Support Vector Machines (SVM) and Decision trees (Addawood and Bashir 2016). Naderi and Hirst (2016) created a corpus of parliamentary discourse labelled as ‘Pro’ and ‘Con’ on the subject of gay marriage, alongside pre-defined argumentation specific to the topic. Cross-topic experiments pose even greater challenges than single topic argument classification. Stab, Miller and Gurevych (2018) annotated and classified web texts across eight different topics based on the three stance classes: pro, con and neutral. ‘Pro/Con’ classification on unseen topics has also been done using BERT models, which improved F_1-scores compared to attention-based neural networks (Reimers, Schiller et al. 2020). In this paper, we follow the methodology set out in the existing literature by creating a single-topic corpus (Naderi and Hirst 2016; Bosc, Cabrio and Villata 2016). The annotation scheme is based on pre-defined arguments in the discussion that are already explored in the wider literature on the selected topic of climate change.

4.2.2. Climate change argumentation

In the upcoming paragraphs, we outline the specific arguments that have been defined in the literature. Argumentation is divided between ‘Pro’, i.e. those acknowledging climate change is a human-caused threat, and ‘Con’, arguments that deny climate change as a problem caused by human action. The arguments are summarized in Table 4.1.

The latter seems to be the most diverse cluster. Previous research proposes a three-way distinction in climate change denial arguments: *(1) Impact scepticism, (2) Trend scepticism and (3) Attribution scepticism* (Rahmstorf 2004). Trend scepticism rejects the warming trend all together, while attribution sceptics question whether human activity is the cause (Rahmstorf 2004). The former seems to be an idea that is disappearing (Rahmstorf 2004; Dunlap and McCright 2012). On the other hand, impact scepticism states that the consequences from climate change might not be that bad (Rahmstorf

Stance	Argument (labels)	Explanation
Con	Impact scepticism	Denial of consequences
	Attribution scepticism	Denial of human influence
	Trend scepticism	Denial of warming trend
	No consensus	Denial of consensus among scientists
	Bad science	Accusation of bad models/ forecasts used in science
	Conspiracy theories	Umbrella category for all conspiracy-related content
Pro	Anthropogenic climate change (ACC)	Climate change is caused by human activity
None	No argument	No relevant argument is present/ comment is off-topic

Table 4.1: Climate change argumentation & annotation scheme

2004). Examples of this argument are statements detailing that a warmer climate is desirable or that we can simply mitigate the effects. Dunlap and McCright (2012) detail the same three movements against human-caused climate change: (1) no warming, (2) not caused by human activity and, (3) the 'non-problemacity' of climate change (Dunlap and McCright 2012). The latter focus of 'non-problemacity' seems to be based on a dominant social paradigm that our species is able to exert control over nature (McCright and Dunlap 2003). This control directly leads to the conclusion that climate change cannot pose a threat (Bord, O'Connor and Fisher 2000; Poortinga et al. 2011).

Aside from these three forms of scepticism, climate change denial also focuses on the scientific community. More specifically, the existence of a scientific consensus is often questioned (Leiserowitz et al. 2015). We label this argument *No consensus*. Interestingly, a consensus among scientists has long existed (Doran and Zimmerman 2009; Oreskes 2005). While it is uncertain as to why this consensus is questioned, a potential explanation lies in the fact that the scientists have long shied away from making dramatic warnings or conclusions in publications (Brysse et al. 2013). A second science-focused argument against climate change takes aim at the science and models themselves, which we label as *'Bad science'*. The claim posits that the complexity and uncertainty surrounding the climate system is a hurdle for scientists to make rigid forecasts (Poortinga et al. 2011). Pinpointing the exact cause for every reasoning disputing human-caused climate change is difficult if at all possible. However, a number of sources can be found, including organized anti-environmental movements like those found in the U.S. in the 1990s (McCright and Dunlap 2003), unreliable or incomplete interpretation of scientific evidence (Whitmarsh 2011) or online content like videos found on *Youtube* (Allgaier 2019). These sources are often presented as 'manufacturers of doubt' (van Linden et al. 2015).

A final category arguing against climate change is the *conspiracy-related* class. Content related to conspiracy theories often emerge in polarized debate in the online sphere, even in good-faith discussions (Samory and Mitra 2018). Similar to the definition of argument, we define ‘conspiracy’ loosely by not requiring all elements of a conspiracy theory, *agent, action and target*, to be explicitly present (Samory and Mitra 2018). References to conspiracies in user comments tend to be compact and make use of the most common denominator words for a conspiracy, and further rely on context to complete the conspiratory content.

Those arguing that the current climate crisis is caused by human activity find themselves in a more unified environment, which we label under the term anthropogenic climate change (ACC). By the late 1980s, and after the vast accumulation of evidence, the majority of academics had concluded that anthropogenic climate change was occurring (Leiserowitz 2007). The argument is in practice quite straight-forward and is reflected in the literature in the form of surveys of experts (Doran and Zimmerman 2009) or literature reviews of the field (Oreskes 2005). Additionally, references are often made to the reports from the Intergovernmental Panel for Climate Change (IPCC) (Masson-Delmotte et al. 2021).

4.2.3. Deliberation on online platforms

This chapter focuses on mutual understanding in the climate change debate in the setting of online comment platforms. In the previous paragraphs, we outlined the polarized argumentation that occurs in the discussion. Briefly, mutual understanding is established through comprehension of what others are trying to do or say as well as why (Margaret 1994). Exposure to other opinions can improve out-group tolerance, which in turn can facilitate mutual understanding (Mutz and Mondak 2006; Andersen and Hansen 2007). Evidence indeed shows that these heterogeneous environments are important for facilitating deliberative qualities (Suiter, Farrell and O’Malley 2016). A vital part of this process is the exposure to conflicting views, which promotes debate participation (Suiter, Farrell and O’Malley 2016). Online platforms can develop this deliberative atmosphere further. Hearing out marginalized argumentative camps through active facilitation may fundamentally improve the deliberative properties of a discussion (Strandberg, Himmelroos and Grönlund 2017). Experimental evidence indeed suggests that opinion polarization can be deconstructed through the implementation and facilitation of deliberative norms, as is the goal in moderated comment spaces (Grönlund, Herne and Setälä 2015). Thus, designing online fora with deliberative norms in mind, such as inclusion, justification, and equality of discussion, can result in a suitable comment space for mutual understanding in the climate change discussion (Wright and Street 2007).

4.3. Methodology

4.3.1. Data collection & annotation

We accessed a large dataset of comments from the platform *NUjij*, the discussion platform of online Dutch newspaper *NU.nl*. All contributions were posted in 2020, are in Dutch and include comments that were removed by moderators. The presence of these comments can be vital for our focus on minority classes, as we need training data for rare or unwanted arguments as well. First, we filtered out all comments that were not placed under articles with the tag *climate*. These tags originate from the journalists and editors themselves. This initial filtering step resulted in a comment pool of 43,106 comments.

From this climate dataset, we randomly sampled 3,000 comments for our initial annotation. Furthermore, we sampled 500 extra comments to create a separate validation dataset that will be used to validate each model in upcoming sections. Annotation was done following the scheme presented in Table 4.1. The full annotation guide is compiled in Appendix B. To derive inter-annotator agreement, subsets of the original data were labelled by two additional annotators. A subset of the original dataset ($n = 250$) was given to two independent annotators. To inform their choices, we created a document with the argumentation scheme. This sheet included clear explanations for each argument that we derived from the climate change literature, alongside examples of comments that contained the argument. These examples were not part of the annotated data. Following this procedure, we achieved a Krippendorf's alpha of 0.73.

4.3.2. Argument classification

Our particular task consists of a multiclass classification with eight different labels (see Table 4.1). We split the original dataset containing 3,000 comments into a training (80%) and test set (20%). This test set remained constant over all versions in this paper, similar to the validation data. As a classifier, we used a pre-trained Dutch transformer-based language model, RobBERT, and finetuned it on the training data (Delobelle, Winters and Berendt 2020). More specifically, we employed the version aimed at sequence classification, which adds a linear classification head on top of the pooled output (Wolf et al. 2020; Delobelle, Winters and Berendt 2020). The final models had a batch size of 32, a learning rate of $5e^{-5}$, optimized with AdamW (Loshchilov and Hutter 2019) and were trained for ten epochs. The best performing classifiers were achieved after two epochs.

4.3.3. Minority argument supplementation

During the annotation process, it became clear that certain argumentation classes were extremely rare in 'natural' discussion (Table 4.2). The bulk of comments were either 'no argument/ off-topic' or 'anthropogenic climate change'. The scarcity made classification of these nuanced cases difficult. With the specific goal of finding minority arguments to boost heterogeneous debate, it was vital to obtain and annotate more of these scarce comments. We opted for an active learning approach to get a better grip on minority classes and to counter possible frequency-related bias in our classification results.

In order to obtain more minority class comments for our training data, we employed a 'query-by-committee' active learning strategy (Zhao, Xu and Cao 2006). The goal is to filter out more minority arguments that will subsequently be added to the training data to finetune RobBERT further (Figure 4.1). First, we extract the BERT embeddings from

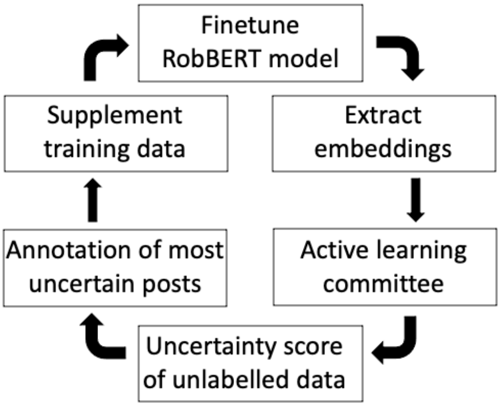

Finetune RobBERT model
Extract embeddings
Active learning committee
Uncertainty score of unlabelled data
Annotation of most uncertain posts
Supplement training data

Figure 4.1: Active learning approach and supplementation of training data

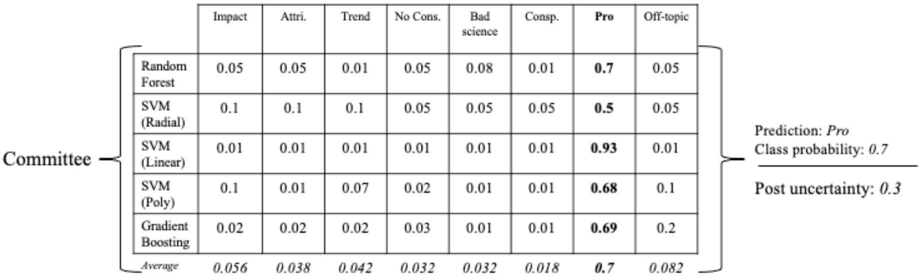
our primary RobBert model (finetuned on only the original data) as input for the first active learning committee. The committee is a collection of five classifiers implemented with Scikit-learn (Pedregosa et al. 2011; Danka and Horvath n.d.): *(1) Random Forest, (2) Support Vector Machine (SVM) (radial), (3) SVM (polynomial), (4) SVM (linear) and, (5) gradient boosting classifier.* Each learner within the committee starts with 10 labelled comments as initial training data. With every iteration, a new comment from the original data is queried based on the disagreement within the committee, calculated with Kullback-Leibler divergence (Zhao, Xu and Cao 2006). This sample is subsequently added to the training data. This process is repeated for 250 iterations.

Such a trained committee can be used for prediction, but more importantly for our application, we extracted the uncertainty measure for unseen comments. In this case, the uncertainty is computed as $1 - class\_probability$. This process is visualized in Figure 4.2. Each learner in the committee assigns probabilities to every comment for each of the eight classes. We obtain the $class\_probability$ by averaging these probabilities per class across the five learners, resulting in eight probability scores. We take the argument with the highest average class probability for the uncertainty calculation. For example, a comment that is difficult to classify may only have a class probability score of 0.3, which equals a high uncertainty score equal to 0.7 (Figure 4.2).

We randomly sampled $10,000$ unseen comments from the climate tagged dataset (containing in total $43,106$ comments) and extracted the uncertainty for each comment. Subsequently, we annotated the $1,000$ most uncertain comments from this collection. Table 4.2 indicates that this task achieved our goal, namely to relatively increase the number of arguments from minority classes compared to the non-argumentative/off-topic category. Uncertain comments were annotated by a human annotator. Predictions from the committee were disregarded. We repeated this circular procedure a second time *(wave 2)* to add more argumentative comments to the training data (Table 4.2).

Argument	**Original**	**Wave 1**	**Wave 2**
Impact scepticism	0.02	0.05	0.04
Attribution scepticism	0.03	0.09	0.11
Trend scepticism	0.01	0.01	0.015
No consensus	0.01	0.01	0.004
Bad science	0.01	0.04	0.057
Conspiracy theories	0.01	0.04	0.042
ACC	0.19	0.40	0.30
No argument/off-topic	0.72	0.36	0.42

Table 4.2: Original data versus uncertain comments. Numbers are fractions of 1 (e.g. 0.72 = 72%)



	Impact	Attri.	Trend	No Cons.	Bad science	Consp.	**Pro**	Off-topic
Random Forest	0.05	0.05	0.01	0.05	0.08	0.01	**0.7**	0.05
SVM (Radial)	0.1	0.1	0.1	0.05	0.05	0.05	**0.5**	0.05
SVM (Linear)	0.01	0.01	0.01	0.01	0.01	0.01	**0.93**	0.01
SVM (Poly)	0.1	0.01	0.07	0.02	0.01	0.01	**0.68**	0.1
Gradient Boosting	0.02	0.02	0.02	0.03	0.01	0.01	**0.69**	0.2
Average	*0.056*	*0.038*	*0.042*	*0.032*	*0.032*	*0.018*	***0.7***	*0.082*

Prediction: *Pro*
Class probability: *0.7*

Post uncertainty: *0.3*

Figure 4.2: Calculating uncertainty using the active learning committee (fictional comment)

After each wave of newly annotated data, we continued finetuning RobBERT using the previous version as the starting point (see Figure 4.1). Following this looping procedure, we obtained three versions:(1) *v1* based on the original data, (2) *v2* consisting of v1 supplemented with the first wave and, (3) a fine-tuned version of v2 using both waves of uncertain comments (*v3*). As stated in the previous section, these versions have a linear classification head. Additionally, we extracted the embeddings from all three RobBERT models as input for an active learning committee. Naturally, both the v1 and v2 embeddings are paired with the committees we had used to obtain the uncertain comments. To classify comments based on the v3 embeddings, we trained a third committee following the exact same procedure.

4.3.4. Patterns in argumentation

Previous sections outlined the automatic annotation of incoming comments for the argument it presents in order to aid moderators in balancing the discussion. Additionally, we aim to boost mutual understanding by diving deeper into what each argument brings to the table. It is important to comprehend the different viewpoints and arguments.

Unique patterns for each argument, i.e. those that have significant presence in one argument compared to all others, were analysed. First, we lowercased the entire corpus and removed stopwords. Subsequently, the corpus was split based on the eight argumentative classes. We used Colibri Core to collect recurring patterns in each subcorpus (Gompel and Bosch 2016). Following the outlined procedure by Gompel and Bosch (2016), the first step was to class encode the corpus. Subsequently, we created an unindexed pattern model entailing the word n-grams occurring at least twice and with a maximum length of eight tokens. We compared the collection of patterns belonging to a single argument with the seven other argumentative subcorpora taken together. To make this comparison, we utilized the log-likelihood (L) function outlined by previous research (Rayson and Garside 2000).

4.4. Results

4.4.1. Argument classification

The automatic labelling of comments for the argument it presents may assist moderators in maintaining the desired form of discussion. As outlined earlier, we finetuned a total of three RobBERT models alongside active learning committees that have been used to tag unseen comments for classification uncertainty. Additionally, these committees are used as a classifier on top of the embeddings from each RobBERT model. Each committee consists of five learners and predict arguments by averaging class probabilities within the committee.

Version	**Precision**	**Recall**	**F1**
RobBERT v1 (original data)	0.65	0.51	0.55
RobBERT v1 + committee	0.75	0.50	0.58
RobBERT v2 (original + wave 1)	0.65	0.62	0.62
RobBERT v2 + committee	0.81	0.60	0.64
RobBERT v3 (original + wave 1&2)	0.88	0.68	0.75
RobBERT v3 + committee	0.94	0.67	**0.78**
Random forest (Baseline)[1]			0.25

Table 4.3: Classification scores on validation data (macro scores)

Table 4.3 displays the classification metrics on the validation data. Classifying comments using the linear head on top of RobBERT underperforms the committee with each version. The latter improves the macro F1-score score by two to three percentage points by boosting the macro precision score slightly at the expense of the macro recall. RobBERT v3 paired with the committee of classifiers, which is trained on the original training data supplemented with two waves of uncertain comments, outperformed all other versions and achieves a macro F1-score of 0.78.

We constructed the active learning approach to improve the recognition of minority arguments. Table 4.4 shows that certain arguments like *'Consensus denial', 'Bad science' and 'Conspiracy theories'* posed severe problems for earlier versions. The third iteration of models, which included two waves of uncertain comments in the training data, produced improved F1-scores on the validation set (Table 4.4). The precision scores for each argument reaches very high levels. This is due to the fact that certain classes have

Version	Impact	Attribution	Trend	Consensus	Bad Science	Conspiracy
RobBERT v1	0.47	0.68	0.67	0.2	0.42	0.4
v1+committee	0.62	0.70	0.67	0.4	0.33	0.43
RobBERT v2	0.63	0.69	0.67	0.4	0.45	0.5
v2+committee	0.75	0.67	0.67	0.67	0.2	0.59
RobBERT v3	**0.8**	0.72	0.67	**0.8**	0.67	0.71
v3+committee	**0.8**	**0.79**	0.67	**0.8**	**0.71**	**0.75**

Table 4.4: F1-score per minority argument on validation data

a small number of comments in the data. Impact scepticism is found in 9 comments in the validation data, which is still more than trend scepticism ($n = 2$) and no consensus ($n = 3$). These minority arguments can lead to precision scores that are misleadingly high. For example, one comment labelled trend scepticism is the only comment that gets labelled as such by the classifier, leading to a perfect precision score, while recall (0.5 for each version) lacks due to the fact that the other comment belonging to the trend scepticism class is never correctly detected.

Figure 4.3 shows a two-dimensional representation of the embeddings extracted from RobBERT v3. The arguments, including the relatively rare ones, form noticeable clusters in the embedding space. In the next section, we look at the language and patterns within each argument. Patterns that are distinctively found in a single argument make these arguments distinguishable.

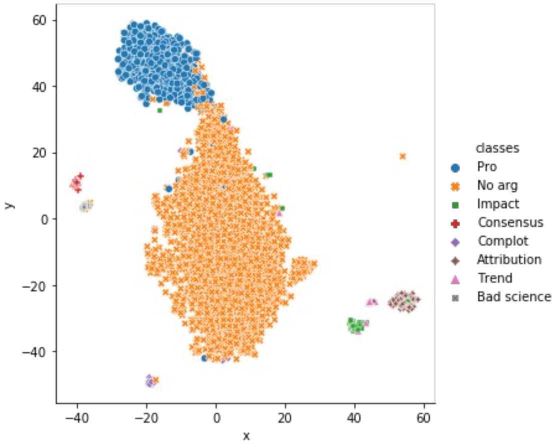


Figure 4.3: TSNE visualisation of RobBERT v3 embeddings

4.4.2. Argument vocabulary

We previously focused on the computational recognition of climate change argumentation presented in online comments. Additionally, Figure 4.3 shows that the arguments

Argument	Terms
No argument / off-topic	trash, solution, electricity, somewhere, powerplant, overpopulation, advantage, most people
Impact	worry, previous years, whine, be okay, good economy, say with certainty, measures, stop
Attribution	all times, billion years, speed up, earth sun, human influence, ice age, partly, million years ago
Trend	cold winter, volcanic, tree rings, religion, garbage, ice sheet, every year again
No consensus	consensus, prove, 40 years, 0 co2, assumption, prove hypothesis, phenomenon, expert
Bad science	prediction, assumption, study, grain of salt, case scenario, fearmongering, theories
Conspiracy	paris accord, pro, farmer, propaganda, acid rain, hoax, money, independent, manipulated
ACC (Pro)	use, less people, whole world, houses, voting, political, importance, inhabitants, 3 degrees

Table 4.5: Argument vocabulary: patterns with highest log-likelihood per argument

form visible clusters in the embedding space, hinting at unique vocabulary and patterns within the arguments. We particularly aimed to recognize minority standpoints in order to present the whole range of online opinions. Subsequently, it is important that users and moderators understand what is being said. The discussion of polarized discussion may be boosted by not only trying to comprehend others, but to invite them into an heterogeneous debate environment. To achieve this understanding of varying argumentative positions, we have derived the vocabulary and patterns within each argument to showcase what sets each of them apart. Table 4.5 presents the collected patterns with the highest log-likelihood per argument.

The first class, *no argument/off-topic*, contains a wide collection of patterns. comments within this category talk about a variety of topics, including the energy transition in The Netherlands and the potential effect of the growing population and consumption (Table 4.5). These comments are often adjacent to the discussion at hand, but do not present actual argumentation aimed at the cause of climate change.

The argument in favour of human-caused climate change (*ACC*) has a different focus. The political aspect comes to the forefront, expressed in terms like ‘voting’, ‘political’ and ‘importance’. These comments attempt to rally readers to take action. Another distinctive pattern in this argument details the global component of climate change. Commenters write about the need of unison action.

In our annotation scheme, climate change scepticism is broken down into three subcategories: *Impact, Attribution and Trend*. The latter has very clear patterns that sets this argument apart from the others. Trend sceptics on the comment platform often point towards cold winters, volcanic activity and the existence of ice sheets to reject the warming trend. Furthermore, these sceptics call human-caused climate change a religion and garbage. Generally, it seems that this scepticism is the most straight-forward rejection of human-caused climate change. On the other hand, attribution sceptics seem to be focused on the historical aspect of climate change. We recognize patterns like '[from] all times', 'billion years' and 'million years ago' (Table 4.5). Alongside this focal point, some attribution sceptics seem to concede that human influence might speed up natural processes ('human influence', 'speed up', 'partly'). These natural processes include the position of the planet relative to the sun and ice age cycles. These topics are not found in other arguments. The third and last argument within the scepticism umbrella, impact, mainly revolves around language claiming it is not necessary to worry about climate change ('worry', 'whine', 'be okay', 'measures' and 'say with certainty').

The two arguments rejecting climate science also rely on specific patterns. On the one hand, we see the dismissal of *scientific consensus* based on very distinctive patterns, e.g. 'prove hypothesis', 'expert' and 'consensus' (Table 4.5). The accusation of *bad science* revolves around the overarching notion of taking it with a 'grain of salt'. We detect among the patterns 'prediction', 'assumption', 'theories' and 'fearmongering'. These comments urge readers not to take these scientific models too seriously, as they are based on theories and assumptions which do not correspond to real-life circumstances.

The final argumentative class to break down into patterns is the conspiracy related content. We detected conspiracy terms like 'propaganda', 'hoax', 'money' and 'manipulated'. Other unique content references to the 'paris accord' of 2015 and 'acid rain', an environmental issue that received a lot of attention over the past decades.

4.5. Discussion

In this chapter, we presented an approach to automatically label online comments for the argument it entails, combined with a deeper dive into each argument in the discussion. In the upcoming paragraphs, we go through some methodological considerations and discuss our approach through the lens of content moderation. Furthermore, we reflect on the usefulness of our approach for other fields that struggle with mutual understanding and opinion polarization.

Translating detailed and nuanced concepts of argumentation into a computational labelling task requires generalization. Previous research makes the useful distinction between scepticism, uncertainty, and ambivalence (Poortinga et al. 2011). In our annotation scheme, we did not make this specific contrast. Whereas clear-cut scepticism can be rare, as is shown in our data, uncertainty about the anthropogenic causes of climate change might be much more widespread (Whitmarsh 2011). The dichotomy between uncertainty and scepticism may be an important aspect for mutual understanding and working towards the comprehension and acceptance of human-caused climate change. Unstable or uncertain beliefs can change through contact with scientific cues and information (Jenkins-Smith et al. 2020). In this chapter, uncertainty is included within the argument classes, even though the label refers to scepticism. Future research could include this distinction in the methodology to encompass the nuance of polarized debates into the computational approach.

Researchers in the field of content moderation and digital journalism struggle with the concept of mutual understanding, as well as with the implementation of computational technologies (Binns et al. 2017; Ruckenstein and Turunen 2020). The growing quantity of contributions threatens real-time curation efforts by human moderators. Automatic applications like the one we have presented in this paper are an avenue for assisting human moderators in curating the online comment space (Ruckenstein and Turunen 2020). While the moderators manage ongoing, interactive processes that are highly dependent on context, computational systems can assist this operation, for example in the form of argument classification and summaries.

Furthermore, research fields that specifically deal with polarized topics struggle with safeguarding civil discussion and mutual understanding. The climate change debate certainly falls within this category. Additionally, online debates on the topic of vaccination lack mutual understanding as well (X. Jiang et al. 2021). This discussion often lacks heterogeneous discussion due to so-called echo chamber effects (A. L. Schmidt et al. 2018). Computational moderation tools, like the one presented in this paper, are an asset for those invested in promoting mutual understanding in these polarized discussions. This approach can be expanded beyond the topic of climate change into other polarized topics. Clearly defined arguments are needed. An example of such a discussion is vaccination, in which clear pro and con sides can be detected (X. Jiang et al. 2021). Domain-specific research is a requirement to create annotation schemes that adequately entail all minority arguments.

4.6. Conclusion

In this chapter, we created a twofold approach to develop a moderation tool aimed at the climate change debate on online platforms, making use of the etic perspective on constructive commenting. First, we trained classifiers that label comments for the argument they present. Certain minority arguments, like trend scepticism and accusations of bad science, were very rare. An active learning approach was constructed with the goal of collecting more minority arguments to supplement into our training dataset. Our best model, after two waves of uncertain comments, achieved a macro F1-score of 0.78. Second, we dove deeper into singular arguments by extracting the lexical patterns that characterize each class. The arguments formed clusters in the embedding space, indicating that each reasoning may be characterized by specific vocabularies. These patterns serve as a swift and understandable view into each argument. Additionally, we formulated methodological considerations regarding the nuance in the annotation scheme and linked our approach to research fields that struggle with moderating online content while safeguarding understanding among participants. The computational approach presented in this chapter serves an assisting role to the human moderator, who in turn can deal with the contextual factors and decide what content is constructive for the discussion at hand.

5

A Time-robust Group Recommender for Featured Comments on Online News Platforms

AI has a prominent position in content moderation.
However, a lot still hinges on the interpretation by the moderator.

Fieldnotes, NU.nl visit june 2022

Based on: Waterschoot, Cedric; van den Bosch, Antal (2024). A Time-robust Group Recommender for Featured Comments on Online News Platforms. Frontiers in Big Data (accepted for publication)

Abstract

Recently, content moderators on news platforms face the challenging task to select high-quality comments to feature on the webpage, a manual and time-consuming task exacerbated by platform growth. This chapter introduces a group recommender system based on classifiers to aid moderators in this selection process. Utilizing data from a Dutch news platform, we demonstrate that integrating comment data with user history and contextual relevance yields high ranking scores. To evaluate our models, we created realistic evaluation scenarios based on unseen online discussions from both 2020 and 2023, replicating changing news cycles and platform growth. We demonstrate that our best-performing models maintain their ranking performance even when article topics change, achieving an optimum mean NDCG@5 of 0.89*. The expert evaluation by platform-employed moderators underscores the subjectivity inherent in moderation practices, emphasizing the value of recommending comments over classification. Our research contributes to the advancement of (semi-)automated content moderation and the understanding of deliberation quality assessment in online discourse.*

5.1. Introduction

Online news platforms allowing user-generated comments have been facing challenges in terms of content moderation due to the ever-increasing content stream (Meier, Kraus and Michaeler 2018; Wintterlin et al. 2020). Discussions are increasing in size and toxicity, described under the term ‘dark participation’, is omnipresent (Quandt 2018). The set of tasks assigned to the moderator has expanded, as well as the need to swiftly make difficult, interpretative moderation decisions (Paasch-Colberg and Strippel 2022). Platforms are increasingly interested in computational solutions to aid human moderators in tasks such as filtering out toxicity, countering misinformation, and promoting high-quality user comments (Gollatz, Riedl and Pohlmann 2018; Gillespie 2020). Broadly speaking, content moderation strategies revolve around two main approaches: maintaining a comment space free of toxic and unwanted content, and recently, highlighting what platforms consider as ‘good’ contributions, such as featuring them prominently on the webpage (Roberts 2017; Diakopoulos 2015a; Yixue Wang and Diakopoulos 2022). However, manually selecting comments to feature is labor-intensive and demands substantial attention and resources from editorial staff and content moderators. To address this issue, we propose a group recommender system capable of recommending a set of qualifying comments, potentially streamlining the decision-making process.

In this chapter, we introduce classifiers designed to rank comments based on class probability, aiding comment moderators in selecting featured comments. Using Dutch comment data with human labeling of featured comments, the operationalization of an emic perspective on constructive commenting, we train a series of models which present the human moderator with curated comments deemed qualified to be featured. These models, referred to as group recommenders, are not personalized for each moderator but instead represent the moderation strategy for content moderators as a collective entity.

Our contribution adds to the ongoing research on (semi-)automated content moderation and the evaluation of deliberation quality. We achieve this by training and examining a range of classifiers and by creating practical evaluation scenarios that mirror the real-world process of selecting individual comments based on online discussions and their context. In practical terms, we depart from evaluating on artificially split or balanced datasets and instead assess our models on unseen discussion articles spanning both 2020 and 2023. This approach mirrors the evolving news cycles and platform growth over time. Our findings indicate that our best-performing models maintain their ranking performance even on recent articles. The final step in our realistic evaluation scenario is performed by moderators currently employed at the platform in question. Their expert evaluation highlights the subjectivity inherent in the practice, thereby reinforcing the argument in favor of recommending comments rather than solely relying on classification.

5.2. Background

5.2.1. Online Content Moderation

The commenting environments on online news platforms and their user bases have been growing, driving content moderators to adapt and expand their moderation strategies. Studying user participation, and the moderation of this content, has grown into an important focus of digital journalism scholars (Quandt 2023; Gillespie et al. 2020). Initially, moderating online comments was focused on assessing the appropriateness of the comment in relation to the platform (Roberts 2017; Gillespie 2018). Dealing with such negative content has been a particular focus, e.g. (organized) misinformation campaigns (Meier, Kraus and Michaeler 2018; Zareie and Sakellariou 2021) or online harassment (Quandt, Klapproth and Frischlich 2022). Aside from such clear cases of toxicity, moderators were also tasked with dealing with grey cases, requiring a closer look at the perception of online incivility and hate comments (Paasch-Colberg and Strippel 2022). Online activity of this form has been described under the term 'dark participation' (Quandt 2018). Recently, however, a novel strategy emerged entailing the promotion of *high-quality* comments by content moderators (Yixue Wang and Diakopoulos 2022). In an attempt to counteract dark participation, moderators are selecting what they deem good, feature-worthy comments, and are flagging them to be moved (pinned) to the top of the comment space.

Outside the context of online news platforms and their moderation strategies, deliberation quality has been widely studied (Friess and Eilders 2015). However, it remains a struggle to define deliberation in diverse contexts (Jonsson and Åström 2014). The featuring of individual comments by content moderators may be seen as an operationalization of the concept in one specific context, purely based on the interpretation by the moderators and guidelines set by the platform.

Many platforms have been promoting what they see as high-quality contributions in recent years, for example in the form of *New York Times* (NYT) picks (Yixue Wang and Diakopoulos 2022), Guardian Picks at *the Guardian* or featured comments at Dutch news outlet *NU.nl* (NUJij 2018). Their FAQ pages describe such promotion-worthy comments as "substantiated", "representing a range of viewpoints" or "respectful" (New York Times 2020; NUJij 2018). Previous research has termed such efforts as *empowerment moderation*, an attempt to motivate the user base to discuss in a constructive manner (Heinbach, Wilms and Ziegele 2022). The authors concluded that these efforts did decrease perceived toxicity on online news platforms. Ziegele et al. (2020) link news value theory to deliberative factors found in the comments posted on news articles, studying how particular characteristics of news articles influence the deliberative quality of social media comments replying to the news article.

Diakopoulos (2015b) assigns a set of editorial criteria to such featured comments, in this case NYT picks. These range from argumentativeness to relevance to the discussion and entertainment value (Diakopoulos 2015b). Generally, this moderation practice can be seen as a "norm-setting strategy" (Yixue Wang and Diakopoulos 2022). Supplementary to the goal of promoting high-quality user-generated content and the positive normative effect that it may have on others, user engagement might increase as well. Previous research concluded that users who received their first featured comment subsequently increased their own comment frequency (Yixue Wang and Diakopoulos 2022).

5.2.2. Hybrid Moderation

Our task of ranking featured comments within online discussions is rather novel, but it is adjacent to the line of research on news recommendation. However, the task of news recommendation often entails personalization aimed at readers on news platforms (Raza and Ding 2022). It differs from our application in that ours is aimed at improving the experience of the content moderators as opposed to that of the readers. In other words, our application supports the practice of content moderation, while news recommendation optimizes news consumption (Raza and Ding 2022). Another adjacent field of research combining the use of Natural Language Processing (NLP) and online discussion and deliberation is argument mining (Lawrence and Reed 2020). Aside from argumentative structure in online text samples, such applications have looked at possibilities to foster mutual understanding among discussion platform users and the evolution of quality deliberation among participants (Waterschoot, Hemel and Bosch 2022; Shin and Rask 2021). For example, previous research has used time series data to model the evolution of deliberation quality or adapters to model different quality dimensions (Shin and Rask 2021; Falk and Lapesa 2023). Building further on the usage of adapters for deliberation quality evaluation, previous research combines both expert and non-expert labelling, using the correlation between the two categories to derive a singular quality measurement (Behrendt et al. 2024). Our task differs from these applications as this study does not focus on argumentation as an indicator for deliberation quality or due to the fact that we do not aim to construct a metric for assessing comment quality as a general concept. In this study, we take the historical moderation choices as the standard of what constitutes, through the lens of a specific online platform, a quality comment. Additionally, as opposed to the mentioned applications, our framework includes comment and user information alongside text representation.

Hybrid moderation is the result of moderators at online news outlets increasingly working with computational systems to execute their tasks (Lai et al. 2022; Gorwa, Binns and Katzenbach 2020). The hybrid nature is causing the role of human moderator on the one hand, and the computational system on the other to be intertwined (Gillespie 2020). The goal of this approach is to make use of the strengths of both automated and manual content moderation (J. A. Jiang et al. 2023). Ideally, editors and moderators alike see the function of AI as offering decision support, instead of decision-making (Ruckenstein and Turunen 2020; J. A. Jiang et al. 2023). This AI assistance has also been referred to as a ‘machine-in-the-loop’ approach, elevating the human moderator to the central actor (Li and Chau 2023). Such support for the moderator in executing their tasks allows the moderators themselves to adapt to the nuances and rapid changes in online contexts (Park et al. 2016). In as much as AI could save time, moderators are able to invest the nuanced human interpretation and judgement that certain edge cases require (J. A. Jiang et al. 2023). The strength of the automated half of the hybrid pipeline is the quantity of comments that can be moderated, especially in terms of clear-cut decisions

(J. A. Jiang et al. 2023). Similar AI-assisted applications have been pursued on other types of online platforms, such as question answering platforms (Annamoradnejad, Habibi and M. Fazli 2022) and social media platforms (Morrow et al. 2022).

Automatically detecting toxicity in online comment sections has received substantial attention (Gorwa, Binns and Katzenbach 2020; S. Wang 2021). The classification of featured comments, however, has not been explored quite that often and has remained understudied. Diakopoulos (2015a) uses cosine similarity to calculate article and conversation relevance scores using *New York Times* editor picks. The study concludes that such relevance scores are associated with *New York Times* picks and computational assistance based on such scoring may speed up comment curation (Diakopoulos 2015a). Park et al. (2016) present their CommentIQ interface, which entails a Support Vector Machine (SVM) classifier on unbalanced, but limited, data (94 *NYT* picks, 1,574 non-*NYT* picks). The included classifier achieves a precision score of 0.13 and recall of 0.60. Their dataset includes both user features as well as comment-specific variables (Park et al. 2016).

Napoles et al. (2017) present their ERICs framework annotating *Yahoo News* comments in terms of "Engaging, Respectful, and/or Informative Conversations". Their work looks at constructive discussion at the thread level as opposed to singular comments. Additionally, their labeling is not based on editorial choices, as is the case for our featured comments or studies working with *New York Times* picks (Napoles et al. 2017). Kolhatkar and Taboada (2017) supplement those Yahoo comments with *NYT* picks. Using these picks as benchmark of constructive discussion, the authors achieve an F1-score of 0.81 using a BiLSTM on GloVe embeddings and a balanced training and testing set (Kolhatkar and Taboada 2017). Furthermore, the study combines a large set of variables, including comment length features and names entities, to train SVMs which reach an F1-score of 0.84 on balanced sets (Kolhatkar and Taboada 2017). In a follow-up study, the authors employed crowdsourced annotations and logistic regression to construct a similar tasks, yielding an F1-score of 0.87 (Kolhatkar, Thain et al. 2023).

In sum, classification of high-quality comments such as those featured by moderators is a task that has been explored relatively little. Aside from Park et al. (2016), classifiers proposed in earlier work lacked information outside comment content features, and focused on text representations or other comment features. Additionally, the validation of these models was performed on balanced test sets, which does not resemble the real-life practice of picking a few featured comments out of a discussion of a news article. The online content moderator chooses editor picks on the article level and, therefore, any model should be evaluated on this exercise. In this chapter, we aim to address this practice by putting together all information available to the moderator while they perform their tasks, including user information, comment statistics and text representation. Next, we replicate the task of picking a few featured comments out of many at the level of the discussion of a news article as an evaluation of our models.

5.2.3. Platform Specifics

The comment platform used in the current study is *NUjij*. This online reaction platform is part of the Dutch online newspaper *NU.nl*[1]. *NUjij*, which translates to 'now you', allows users to comment on a wide range of news articles published by the news outlet. Pre-moderation is set in place, consisting of automatic filtering of toxic content alongside the human moderators who check the uncertain comments (Van Hoek 2020). The platform has a moderation pipeline that includes multiple strategies, including awarding expert labels to select verified users and pinning featured comments at the top of the comment section (NU.nl 2020). As said, the latter is a moderation strategy also practiced at e.g. the *New York Times* or *the Guardian.*

Featured comments are chosen manually by the moderators at the platform. A comment is either featured or not. They define such comments as "substantiated and respectful" and "contributing to constrictive discussion" (NUJij 2018). The FAQ page informs users that moderators are aiming to present a balanced selection of featured comments in terms of perspectives and to not pick based on political affiliations. This chapter addresses the specific issue of picking featured comments using the information available to the human moderators while they perform their tasks. These variables include user information and their commenting history, for example whether their comments have been featured before. While highlighting quality content in the form of featured comments is a common moderation practice, other platforms might have different editorial guidelines in place. To best support the moderator in efficiently featuring comments they deem worthy of the status, it is vital that the computational approach is fully suited to their specific platform and context. This may include the intended human bias in choosing such comments. Therefore, we aim to train models that rank comments replicating the choices made by *NUjij* moderators in the past.

5.3. Methodology

5.3.1. Datasets and data splits

This chapter makes use of two datasets from the Dutch platform, the first one containing articles from 2020 and the second originating from 2023. Each dataset consists of a single file containing observations on the comment level. Each comment is timestamped and has a user and article ID number. Additionally, each comment has information on whether it was rejected by a moderator, whether it was featured, the number of replies, the number of likes and the actual comment text. On the article level, we discarded all comments within a discussion published after the timestamp of the final featured comment. This procedure mimics the time-related nature of picking featured comments. The moderator performs this task in the earlier phases of the discussion to present users with the featured content while they are still participating. Using the article ID, we scraped the topic of each article from the original web page. Each news article is given topical keywords by the editorial staff upon publication. A discussion

[1] `www.nu.nl`

refers to a collection of user comments published as response to a specific article. An article only has one discussion, which in turn can entail any number of comments. The goal of the study is to rank comments within a singular discussion to obtain the most 'featured-worthy' comments out of a specific article discussion.

The 2020 dataset contains a total of 528, 973 pseudonymized comments, spanning a total of 2, 015 articles from *NU.nl*. We limited the set by selecting only articles from three news topics, using topic labels manually assigned by the editorial staff: climate change, the 2020 US election and the Covid-19 pandemic. Other topics were relatively small in sample size. In total, the 2020 dataset contains 8, 354 featured comments. On average, a discussion consisted of 267 comments (median = 143), 4.14 of which were featured on average (median=3). This dataset was used for the training and testing of the models, as well as the initial evaluation on unseen discussions.

The second dataset contains discussions from 2023 spanning a wider range of topics: the nitrogen issue in the Netherlands, farmer protests, the local elections, climate change and the war in Ukraine. Similar to the 2020 data, the comments were pseudonymized and include a binary variable indicating whether these were featured. This dataset contains 538, 366 comments spanning 390 articles. On average, a discussion consisted of 1, 384 comments (median = 633), with the mean featured comment count at 3.73 (median = 4). Comparing to the means of 2020 it can be observed that the activity on the platform grew over the years, resulting in a much higher average comment count per discussion, while the average number of featured comments per discussion remained stable. These 390 articles from 2023 are used in the study as evaluation to test the time robustness of our models. Not only did the activity on the platform change, the content matter of the discussions from 2023 is substantially different. A featured comment recommender should be robust to topic changes over time; it should be context insensitive, obtaining similar ranking scores on the data from 2020 and 2023.

	Total Comments	**Featured**
Full training	295,678	4,903
Validation	36,946	627
Test	36,911	662
95/5 training set	97,660	4,903
Evaluation articles (2020)	159,438	2,162
Evaluation articles (2023)	538,366	1,453

Table 5.1: Data set distribution

The 2020 dataset was further split into a large set of articles for training and testing, alongside a smaller set of unseen articles for ranking and evaluation on similar content on which the models were trained. We grouped and chronologically sorted the comment data by article and split them 75%/25%. The first set (75%) contained 1, 511 articles up until October 23rd 2020. These comments were used for training and testing the classifiers. To achieve this, this dataset was further split into 80%/10%/10% generating a full training, validation and test set, respectively. The 25% set is referred to as evaluation articles (2020) in Table 5.1. Table 5.1 outlines the comment distributions in all the datasets used in this study. Thus, we work with three datasets. The first one consists of the training, testing and validation splits. The second dataset contains the unseen 2020 articles, while the third set consists of unseen 2023 articles.

Var category	**Var name**	**Description**
Comment info	Delta_minutes	Minutes between article and comment publication
	Reply_count	Absolute number of replies
	Respect_count	Absolute number of likes
	Wordcount	Number of words in the comment
User info	Total_posts_user	Total comments by user
	Featured_posts_user	Total featured comments by user
	Ratio_featured	Featured comments relative to total posts by user
	Ratio_rejected	Rejected comments relative to total comments by user
	Ratio_reply	Average reply count on comments by user
	Ratio_respect	Average number of likes on comments by user
	Avg_wordcount	Average wordcount of user
Context	Conversation similarity	Cosine similarity with mean discussion embedding
	Article similarity	Cosine similarity with article text
Content	Bag-of-Words	BoW representation of text
	RobBERT embedding	mean sentence embedding extracted from finetuned model

Table 5.2: Variable list: all variables used in the study

Table 5.2 summarizes the feature set, present in both 2020 and 2023 data, that we used to train and evaluate our models, including metadata. Several variables were calculated out of the original data. Each comment is accompanied by delta_minutes, which equals the difference between article and comment publishing timestamp. For each comment, we calculated the word count by simply counting the number of tokens in each comment text. We used the pseudonymized user IDs to aggregate user information by grouping all comments belonging to a single user. For each user in the data, we calculated their total comment count and total featured comment count. We calculated ratio_featured by dividing the latter by their total comment count. Such ratios were also calculated for the number of replies and respect points of a user. As a last user variable, we calculated the average word count across all their comments (Table 5.2).

To obtain the context variables, i.e. cosine similarities between the comments and their wider conversation and article it was commented on, we finetuned a pre-trained Dutch transformer-based language model, RobBERT (Delobelle, Winters and Berendt 2020). We finetuned the model on the default masked language task and trained it for 10 epochs with a batch size of 64, AdamW optimizer and a learning rate of $5e^{-5}$ (Loshchilov and Hutter 2019). Using the SentenceTransformers package, we obtained a vector representation of each comment and article by averaging the RobBERT embeddings across all 786 dimensions (Reimers and Gurevych 2020). The context variables were calculated following the procedure outlined by previous research (Diakopoulos 2015b). Similarity scores with the article were obtained by calculating cosine similarity between each comment and their article text. We obtained conversation similarity by calculating the mean embedding of each discussion and subsequently calculating cosine similarity between this embedding and each comment within the discussion (Diakopoulos 2015b). While not all models included text representation of the comment, we included certain iterations with either a Bag-of-Words representation or the vector representation of the comment embedding obtained from the RobBERT model (Table 5.2).

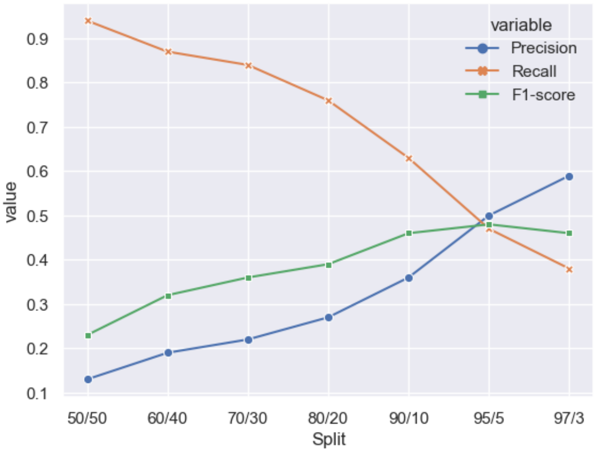


Figure 5.1: Training data splits: classification scores on validation set

The validation set was used to calculate the optimal downsampling of non-featured comments in the training set. Excluding the text representation variables outlined in Table 5.2, we trained a random forest to predict if a comment was featured on seven different downsampled training sets (Figure 5.1). These splits include all 4,903 featured comments found in the training set merged with a varying degree of non-featured comments. Using the scikit implementation[2], the downsampling was performed by randomly selecting the appropriate number of non-featured comments, relative to the total number of featured comments. For example, the 50/50 ratio includes all 4,903 featured comments along with a random selection of 4,903 non-featured comments. The ratios (Non-featured/Featured) that were tested are presented in Figure 5.1. To pick the best ratio, classification scores (Precision, Recall, F1-score) were calculated on the validation dataset. The 95/5 ratio, i.e. 95% non-featured comments and 5% featured comments, yielded the best result and is used as the training data henceforth. While the 95/5 ratio still remains unbalanced, the unsampled actual ratio approximates 98/2. Thus, the 95/5 training set constitutes a marked downsampling of non-featured comments.

[2] https://scikit-learn.org/stable/index.html, v1.2.0

5.3.2. Models

The upcoming paragraphs detail the models that were trained as part of this chapter. We have trained models without the text representation as well as a transformer-based model with purely textual input. Finally, we combined both by training two random forest models with both the non-content variables and text representation in the feature set. All models were trained on the 95/5 training set (Table 5.1).

Baseline

We created a threshold-based model as baseline. Specifically, to determine whether a comment is classified as featured, the comments are ranked in descending order by the featured comment ratio of the user. Users with a ratio above 3%, the 95th-percentile, received the featured label. The intuition behind this simple baseline model is that users with a history of writing featured-worthy comments might do so in new discussions as well.

Support Vector Machine

Using the variables described in Table 5.2 excluding the content category, we trained a Support Vector Machine (SVM) with the radial basis function (RBF) used in the scikit implementation.

Random Forest (RF)

We trained a random forest on the non-content variables outlined in Table 5.2. The standard sci-kit implementation of random forest was used and we perfomed a hyperparameter grid search. The final model has a max depth of 20, minimum samples to split a node of 5 and 1400 estimators.

Text representation baseline models

While previous models were trained on the set of variables excluding the content category, we also trained a set of model exclusively on the textual input. The training data consisted of only the tokenized comment text. We employed the pre-trained transformer-based RobBERT, a Dutch language model based on the robBERTa architecture (Delobelle, Winters and Berendt 2020). The sequence classification RobBERT model employs a linear classification head on top of the pooled output and was trained for 10 epochs (Wolf et al. 2020). The model had a batch size of 64, AdamW optimizer and a learning rate of $5e^{-5}$. The second model trained exclusively on textual data is a bidirectional LSTM. We trained this biLSTM for 10 epochs with Adam optimizer, a batch size of 32 and binary cross-entropy. The third and final model is a Convolutional Neural Network (CNN) trained on the tokenized training texts. We trained the CNN for 10 epochs with a batch size of 32. The latter two NLP models were implemented using the Keras python library[3]. These three models represent state-of-the-art text classification models, suited for comparing the performance of models trained on our non-textual datasets (Raza, Garg et al. 2024).

[3] `https://www.tensorflow.org/guide/keras`, v3.3.2

Rf_BoW & Rf_emb

The final two models combine text representation with the variables used in previously discussed classifiers. We extracted the embeddings from the RobBERT model by averaging them across all 768 dimensions using the SentenceTransformers package, resulting in a single vector per comment (Reimers, Schiller et al. 2020). This vector was combined with the feature set and used to train a random forest (Rf_emb). A hyperparameter grid search was performed resulting in a final random forest model with a max depth of 100, 600 estimators and a minimum of 5 samples to split a node.

For the final model (RF_BoW), we represented the content of each comment by a standard Bag-of-Words approach. This method simply counts the occurrences of the tokens in each comment. We lowercased the text and removed punctuation and stopwords. We used the sci-kit implementation of Bag-of-Words and included n-grams up to three words. To reduce the size of the word set, we kept only the tokens appearing in less than 5% of comments, thus removing common, and less informative, words and phrases. Once again, we performed a hyperparameter grid search to train the random forest which resulted in Rf_BoW with a max depth of 20, 1400 estimators and a minimum of 5 samples to split a node.

5.3.3. Ranking and evaluating discussions

While the initial testing of the models is done by calculating standard classification scores, the goal of the study is to rank comments within their discussion to provide the moderator with the comments most likely to be featured based on the predictions by the model. To achieve this, we ranked comments in a discussion based on the class probability of being featured in descending order. Each discussion ranking is evaluated by calculating Normalized Discounted Cumulative Gain (NDCG) at three sizes: at 3, 5 and at 10 (Jarvelin and Kekalainen 2000). An article has on average 3 featured comments, while 5 and 10 allows for the moderator to have a somewhat larger pool of options to choose from. NDCG is an often used metric to evaluate recommendation or ranking models and evaluates the top comments within the ranking, i.e. those that are shown to the moderator, in relation to the 'ideal' ranking (Yining Wang et al. 2013). In this case, the ideal ranking (Ideal Discounted Cumulative Gain, IDCG) is one that returns all correctly featured comments before showing non-featured comments within the ranking size.

NDCG is a useful metric since it takes into account the order within the ranking, meaning that comments high up in the ranking have a higher weight than those ranked lower. Therefore, models correctly ranking featured comments high in their output are rewarded, while incorrectly classified comments with a high class probability are penalized most. Scores range from 0 to 1 with a result of 1 indicating the best possible ranking. In practice, article discussions are handled one at a time. Subsequently, NDCG scores are averaged across all articles in the particular evaluation set, be it the unseen articles from the 2020 data or the recent 2023 evaluation articles. This procedure resulted in mean NDCG scores at the three ranking sizes for each trained model.

5.4. Classification results

Before ranking unseen discussion, we perform a standard evaluation using a test set. This initial evaluation of the previously discussed models is done on the set of comments that we obtained out of the original 80/10/10 split that produced the training, validation and test sets. The latter contained 36,911 non-featured alongside 662 featured comments published in 2020. The imbalance between featured and non-featured content illustrates the difficulty of the classification task, as merely 1.7% of the set belong to the featured class. Before moving on to ranking, the test set follows the standard procedure of a classification problem, not yet ranked by class probability. The classification scores are summarized in Table 5.3.

Model	Precision	Recall	F1
Baseline	0.15	0.34	0.20
SVM	0.61	0.42	0.50
RF	0.52	0.52	0.52
RobBERT	0.17	0.29	0.21
CNN	0.10	0.24	0.14
biLSTM	0.11	0.14	0.12
Rf_BoW	0.57	0.55	0.56
Rf_emb	0.59	0.48	0.53

Table 5.3: Classification results on initial, unbalanced test set (1.8% featured comments)

The baseline model achieved an F1-score of 0.20. The model lacks in precision (0.15), while achieving a slightly better recall-score of 0.34.

Our SVM model achieved the highest precision score at 0.61, although it lacked in recall (0.42) resulting in an F1-score of 0.50. RF, the random forest lacking text representation achieved a higher F1-score of 0.52 based on a balance in its precision and recall scores. Training a model purely on the textual content produced poor classification results. Compared to the baseline, our finetuned RobBERT model performed only slightly better (0.17). However, in terms of recall, the transformer-based model achieved a score of 0.29, even underperforming relative to the simple baseline model. Similarly, the CNN achieved an F1-score of 0.14, achieving the lowest precision (Table 5.3). The biLSTM achieved an even worse performance on the highly unbalanced test set, yielding an F1-score of 0.12. These results indicate that classifying featured comments based on text representation alone does not produce a working solution, potentially due to the fact that identical comments are not always featured.

Rf_emb was trained on the combination of the previous RF with the averaged embeddings of the comments derived from the RobBERT model. This newly obtained feature set did improve the precision-score of the original RF model by 7 percentage points (Table 5.3). However, this improvement was at a cost in terms of recall, achieving a recall-score of 0.48. This trade-off meant a F1-score boost of just a single percentage point. Finally, the model Rf_BoW, which combines the RF model with Bag-of-Words text representation, achieved the highest F1-score at 0.56. RF_BoW yielded the highest recall score (0.55) and a precision score of 0.57 (Table 5.3).

5.5. Evaluation of comment rankings

Besides the standard classification of rare featured comments, the models ought to be able to correctly rank those featured comments in the shown set of comments. As outlined in earlier sections, we evaluated our models at different ranking sizes: 3, 5 and 10. On average, an article had 3 featured comments, while the ranking sets consisting of 5 or 10 comments give the moderator the opportunity to pick and choose. The rankings were created by sorting all comments within a discussion based on probability of belonging to the featured class. The comments with the highest probability were ranked first.

5.5.1. Precision@

Before evaluating the ranking models by calculating NDCG scores giving higher ranked comments more weight, we calculated average precision scores at sizes 3 and 5. We omitted ranking size 10 in this intermediate step due to the fact that articles with more than 5 comments labelled as featured are very rare. This greatly affects the precision score due to the fact that it no longer has correct comments to present. It does not affect NDCG scores in similar fashion, due to the fact that earlier ranked comments receive much more weight. After calculating the precision scores for each article, they were averaged to obtain a mean precision@3 and mean precision@5 for each presented model (Table 5.4).

Model	Precision@3	Precision@5
Baseline	0.22	0.19
SVM	0.62	0.53
RF	0.64	0.53
RobBERT	0.18	0.17
CNN	0.15	0.14
biLSTM	0.14	0.13
Rf_BoW	0.67	0.57
Rf_emb	0.31	0.26

Table 5.4: Mean Precision@3 and mean Precision@5 calculated on the 2020 evaluation set

The data used for this evaluation step was the collection of unseen 2020 articles. This set contained 471 unseen articles (159,543 comments, 2,162 featured) with similar content matter compared to the data that were used in training and previous testing. In total, this set consisted of 351 articles on the Covid-19 pandemic, 25 on climate change and 95 on the US election in 2020.

Overall, precision decreased when the ranking size increased (Table 5.4). Reflecting the good performance on the initial test set, RF_BoW achieved the highest precision scores at both size 3 (0.67) and size 5 (0.57). The SVM and random forest (RF) achieved similar precision scores. The former yielded a precision of 0.62 at ranking size 3 and 0.53 at ranking size 5. RF achieved identical precision at ranking size 5, but achieved a precision of 0.64 when taking into account 3 comments with the highest class probability (Table 5.4).

The baseline model only taking into account the history of being often featured in the past outperformed the models exclusively trained on textual data. The baseline model achieved a precision score of 0.22 at size 3, a higher result than robBERT (0.18), CNN (0.15) and the biLSTM (0.14). Similar results were found at ranking size 5 (Table 5.4). However, these precision scores do not take into account the position of a comment within the ranking. To evaluate whether the discussed models achieve such correct positioning, in which the moderator first reads correctly recommended comments, we calculated NDCG scores.

5.5.2. Evaluation on unseen 2020 articles

Rankings were evaluated on an article basis by calculating NDCG scores at every ranking size. Subsequently, NDCG scores were averaged across all articles, producing a mean NDCG@3, 5 and 10 per model. The evaluation of the ranking capabilities of the models is threefold. First, we evaluated the models on the unseen 25% split of the 2020 dataset. This set deals with content similar to the training and testing data that we previously used. Second, we moved on from the content from 2020 and evaluated our models on unseen discussions originally published in 2023. On average, these discussions are much longer than those from 2020 and deal with a different range of topics. It is important that our models can deal with changing contexts, as the focus of news articles continuously changes. To probe the context-sensitivity, we present ranking scores per topic for both the 2020 and 2023 evaluation articles. Last, the current moderators employed at the NUjij platform evaluated the output of our best performing model in an offline, survey-style evaluation by choosing which comments to feature from a randomized list including highly ranked comments and random non-ranked comments within a random set of discussions.

Model	NDCG@3	NDCG@5	NDCG@10
Baseline	0.42	0.47	0.50
SVM	0.86	0.86	0.85
RF	0.89	0.88	0.86
RobBERT	0.43	0.46	0.51
CNN	0.25	0.30	0.37
biLSTM	0.26	0.30	0.34
Rf_BoW	0.90	0.89	0.88
Rf_emb	0.71	0.72	0.72

Table 5.5: Average ranking scores calculated on unseen 2020 articles

The simple baseline model achieved an optimum NDCG score at ranking size 10, reaching 0.50 (Table 5.5). At smaller ranking sizes, the baseline model achieved lower NDCG scores. The SVM model outperformed the baseline, achieving a better ranking @3 and

@5 (0.86) compared to @10 (Table 5.5). Subsequently, the random forest model without text representation (RF) performed better, achieving its optimum mean NDCG@3 equal to 0.89. Mimicking the poor performance on the initial test set, the NLP models yielded poor ranking results.The RobBERT model underperformed compared to the other trained classifiers, achieving an optimum NDCG score at ranking size 10 of 0.51, merely an 0.01 increase relative to the baseline model. Similarly, the CNN and biLSTM models yielded poor ranking results, even underperforming compared to the RobBERT model (Table 5.5). The embeddings of the RobBERT model did not increase performance of the random forest, even decreasing the ranking scores. Rf_emb achieved a NDCG@3 of 0.71 and 0.72 for both ranking size 5 and 10. The final model, which combines the random forest with Bag-Of-Words text representation, slightly outperformed the others, achieving an optimum NDCG@3 of 0.90 and NDCG@5 of 0.89 (Table 5.5). This model had already achieved the optimum F1-score on the initial test set.

As context-independent ranking is the goal, we unpack the three main topics in the 2020 dataset (Figure 5.2). Using the output of the best-performing model RF_BoW, we found similar ranking scores across topics. Using the Kruskal-Wallis H-test, we compared each NDCG@ score between the article topics.[4] We found no significant difference between the topical groups at ranking size 3 ($H = 0.79$, $p = 0.67$), size 5 ($H = 1.28$, $p = 0.53$) and the largest ranking size 10 ($H = 0.30$, $p = 0.86$). Therefore, we conclude that our best-performing model does not perform better on any topic over the others (Figure 5.2). However, the models ought to be validated on articles covering topics not found in the training and initial testing data to fully test whether the rankings work in a context-independent manner.

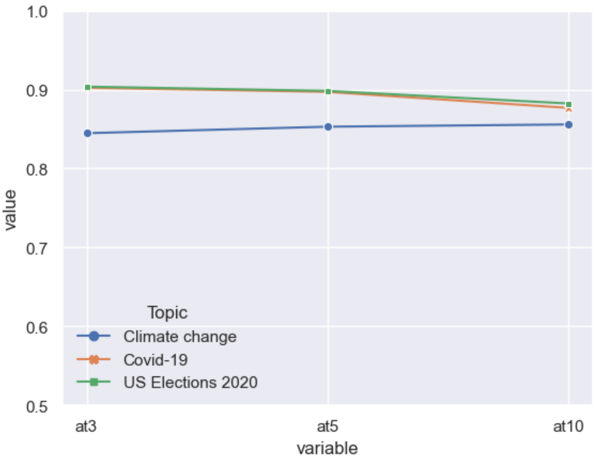


Figure 5.2: Performance on 2020 discussion topics

[4] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kruskal.html

5.5.3. Evaluation on recent data: different topics

As discussed earlier, the platform in question, Dutch *NUjij*, saw a stark increase in user activity recent years. Additionally, news cycles rapidly introduce novel topics to the online comment section. Ranking models supporting the content moderator ought to be able to cope with these changes in content matter and be generalizable across contexts. The goal of this second evaluation is to probe whether the models achieve similar mean NDCG scores compared to the 2020 data, as well as a topical breakdown of the results. The latter is used to analyze whether the ranking models are adequately resistant to unseen topics.

For this particular reason, we included a second dataset in the evaluation of our ranking models. This dataset contains a total of 390 unseen *NUjij* articles published throughout 2023. More specific, this second dataset contains 538,366 comments, of which 1,453 were featured by a moderator. In total, this evaluation ranked 47 articles on the topic of climate change, 112 on the farmer protests in the Netherlands, 33 on the nitrogen issue, 20 articles discussing the Dutch local elections and 35 discussions on the topic of the war in Ukraine.

Model	**NDCG@3**	**NDCG@5**	**NDCG@10**
Baseline	0.48	0.50	0.52
SVM	0.78	0.79	0.79
RF	0.86	0.86	0.86
RobBERT	0.15	0.17	0.21
CNN	0.09	0.11	0.14
biLSTM	0.05	0.07	0.10
Rf_BoW	0.88	0.88	0.87
Rf_emb	0.61	0.64	0.66

Table 5.6: Average ranking scores calculated on unseen 2023 articles

Overall, the ranking output of our models was not heavily impacted by the novel data (Figure 5.3). Interestingly, the simple baseline model yielded improved NDCG scores on the 2023 article set. While still lacking the ranking precision of other models, all ranking sizes did experience a slight increase in NDCG score (Table 5.6). The SVM model experienced a relatively large decline in performance, achieving an NDCG@5 and NDCG@10 of 0.79, a decline of respectively 0.07 and 0.06. The basic random forest model (RF) achieved NDCG scores of 0.86 across the board. This result constitutes a small decline at smaller ranking sizes, while the NDCG@10 score remained equal. Our best-performing model, RF_BoW did not experience a stark decrease in performance. At ranking size 3, the model lost 0.02. At the larger ranking sizes used in the evaluation, Rf_BoW lost only 0.01 in terms of NDCG score, which still yielded the highest ranking metrics across all trained models (Table 5.6).

The models that did break as a result of the topical changes were those trained on textual data: the transformer-based RobBERT model, CNN and biLSTM (Figure 5.3). Due to their sole input being text representation, the model did not make accurate rankings

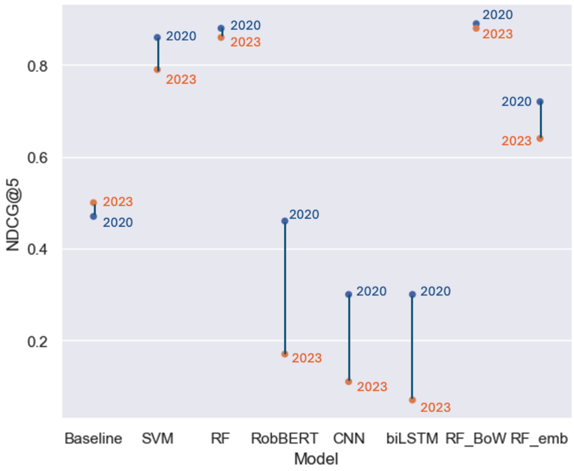


Figure 5.3: NDCG@5: Performance on evaluation sets

when the context changed drastically in the 2023 dataset. For example, the RobBERT model experienced a decrease of 0.28 at ranking size 3, 0.29 at size 5 and even 0.30 in terms of NDCG@10 (Table 5.6). This drop in performance also affected RF_emb. This model combining the set of variables with the mean text embeddings from RobBERT achieved an NDCG@3 score of 0.61 on the 2023 set, a drop of 0.10. At larger ranking sizes, the model achieved a score of 0.64 and 0.66, a drop in performance of 0.08 and 0.06 respectively (Figure 5.3).

Once more, we took a closer look at the specific distribution of topics found in the data (Figure 5.4). The scores were derived from the best-performing model, Rf_BoW. We conclude that, while all topics produced good NDCG scores, the ranking on certain topics yielded slightly better results. Using the Kruskal-Wallis test, we found significant differences between the set of topics ($H = 29.35$, $p < 0.01$). Using Dunn's test for post-hoc testing, we conclude that the ranking on the topics 'Farmer protests' produced less accurate rankings, as well as the topic on the war in Ukraine compared to the articles discussing the 'Nitrogen negotiations' in the Netherlands. However, even the NDCG@ scores on those particular topic are still high, hovering around those found in the 2020 data. It is the case that the other 2023 topics produced very accurate rankings, leading to the significant differences among unseen topics.

We conclude that our models, aside from those based on text embeddings, were robust against the changes within online comment sections, both originating from topical focus of the articles as well as the stark increase in activity. The best-performing model on the initial test set, RF_BoW, also achieved the highest ranking scores for both the 2020 and 2023 evaluation articles. While some topics in the more recent 2023 article set outperformed others present in the set, the ranking scores mimic those calculated on the unseen 2020 articles (Figure 5.2).

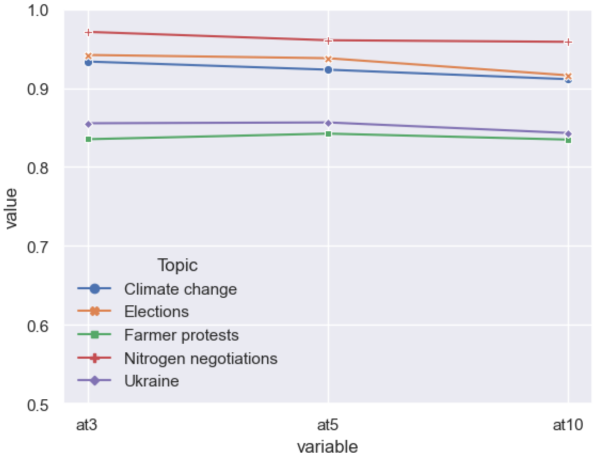


Figure 5.4: Performance on 2023 discussion topics

5.5.4. Expert evaluation by content moderators

Moderators currently employed at *NU.nl* were contacted through their team supervisor and agreed to evaluate the output of the study. In total, four content moderators separately participated in the expert evaluation and validated the output of a selection of ranked online discussions from the 2020 dataset. Each individual moderator first evaluated a shared set of articles, used to calculate inter-rater agreement. Afterwards, a unique set of news articles alongside the corresponding online discussion was evaluated by each moderator to maximize the evaluation size. In total, each moderator read and evaluated the discussion of 15 articles, 5 shared between the four moderators and 10 unique discussions.

Using the output of Rf_BoW, a random set of unseen articles from the 2020 dataset was collected alongside the rankings made by the model. An online survey was created consisting of 30 news articles combined with a set of comments. These user comments comprised of the top ranked comments by our model (comments with class probability above 0.50 and maximum 10 comments per article), alongside an equal number of randomly selected non-ranked comments from the same online discussion. These were shuffled randomly so the moderators did not know which comments belonged to the top ranking. To replicate the real-life practice in which content moderators at *NU.nl* pick featured comments out of a discussion, moderators were first presented with the actual news article. Underneath, the set of comments was shown. Each comment was supplemented by information that moderators have access to in the real-life practice: the total number of previously posted and featured comments by the user, the rejection rate of the user and the number of respect points the comment received. This procedure replicates the real-life process in which comments are judged individually, while the human moderator takes the discussion context into account.

The survey question presented to the four moderators was to decide for each individual comment, within the context of the article, whether they though it was a candidate to be featured on the comment platform. The expert evaluation showed the large variation and subjectivity that this practice entails. Using the shared set of articles, we calculated a Krippendorff's alpha inter-rater agreement of 0.62. Additionally, we compared the choices made by moderators during the expert evaluation with the featured picks in the original 2020 data. We found that 42.3% of included comments that were featured in 2020 were not chosen as featured-worthy comments during the survey. These numbers indicate that the moderation practice entails a notion of subjectivity that the moderator brings to the table themselves, strengthening the concept of ranking and recommending a set of comments to choose from.

While the variation in selected comments by human moderators may pose difficulty to computational models, the expert evaluation validated the ranking performance presented in this study. In all but one article did the moderators find comments to feature among those coming from the ranking, resulting in a NDCG score of 0.83. They did not always decide on the exact same comments. Even though a large portion of subjectivity and context-specificity is involved in the process of picking featured comments, the moderators did consistently feature comments from the set produced by the model. They were more likely to feature comments which were recommended by the model (64%) compared to comments belonging to the random non-ranked set (36%).

5.6. Discussion of the results

The presented approach differs from previous research in terms of time and topic changes as well as evaluation. Previous research mainly used artificially created test sets, while we opted to evaluate on an article basis. The latter mimics the practical setting of online content moderation and picking featured comments in particular. The lower classification scores derived from the initial evaluation on the highly unbalanced test set underscores the need to evaluate featured comment classification on data that follows the real-life distribution. Training and testing on balanced datasets, as done in most research on featured comments, produces better classification scores. However, this does not adequately portray the task of content moderation and can lead to models overestimating the number of featured-worthy comments. Moving forward, models aimed at online content moderation ought to be evaluated within the specific context that they would be used. This entails a large set of separate discussions with changing topics and participation rates. This factor was overlooked up until now in research on featured comments. It is important that models aimed at featured comments are not overfitted on the specific topics from which the training and testing data were derived.

Practically speaking, the expert evaluation underscores the importance of recommending a set of suitable comments to the human moderator as opposed to classification. While the content moderators did consistently feature comments presented within the top ranking presented to them, they did not agree with the entire set as produced by the model. Additionally, they did not always agree among themselves. Thus, presenting a selection of the most suitable comments within an online discussion allows the human moderator to apply necessary subjective and contextual judgement.

5.6.1. Robustness for time and context

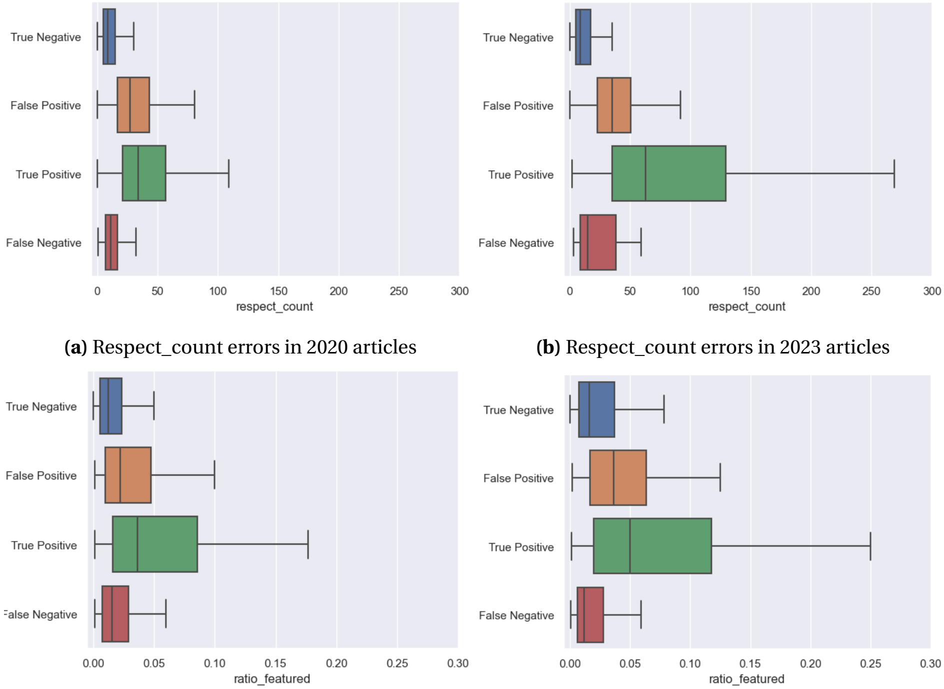
The context of hybrid content moderation requires models to function across changing content matters. Furthermore, platforms evolve over time, as seen in the difference in activity between the 2020 and 2023 datasets. Discussions grew larger, while featured comment counts remained stable. On top of that, moderators and users alike demand explainability of computational models used in the moderation pipeline (Ruckenstein and Turunen 2020; Molina and Sundar 2022). Such transparency is a prerequisite for user trust in online content moderation (Brunk, Mattern and Riehle 2019).

Online platforms can change a lot over relatively short periods of time, an aspect of content moderation that should be taken into account when developing computational models for use in this context. For instance, our datasets from 2020 and 2023 showcased stark differences in factors such as discussion size. The average discussion in 2020 consisted of 267 comments (median = 143). However, three years later the average discussion in our dataset comprised 1,384 comments (median = 633). While some slight variability can exist due to the topical differences and the public's interest in them, such a stark difference indicates growth in platform activity. The number of featured comments per discussion remained stable, pushing the discussions towards a larger class imbalance in regard to featured and non-featured comments. Other activity-related features that were influenced by platform growth were the average respect count of a comment, which was 3.66 in 2020 and which grew to 4.87 in 2023. Another interesting change in discussion dynamics existing in our feature set is the fact that on average, comments received fewer replies. In terms of wordcount, comments became on average shorter in 2023 compared to 2020. In 2020, the average comment counted 52 words, while we calculated an average wordcount of 40 in the 2023 dataset. While all of these discussion factors influenced our dataset and the feature set used by the classifiers, it did not strongly impact the ranking performance of our better performing models.

A closer look at the correctly and incorrectly ranked comments from both the 2020 and 2023 data provide insight into the behaviour of our best-performing model. More specifically, we explored whether certain features repetitively contributed to false positives (FP), and false negatives (FN). For this error analysis, we processed all unseen 2020 and 2023 articles and collected the ranked comments within each discussion at ranking size 5. The false negatives were collected from the entire discussion, since false negatives are by definition not part of the ranking. We used the python library 'treeinterpreter' to collect for each prediction the feature contribution[5]. The two most decisive variables for our model were respect_count and ratio_featured, the share of comments by a user that have been featured in the past. Interestingly, in light of the dynamics in discussion features between both datasets, the contributing factors to the incorrect predictions remained exactly the same. Figure 5.5 outlines the distribution of

[5] https://github.com/andosa/treeinterpreter

these variables across error categories. We conclude that the model is biased towards comments with a high number of respect points and users that have more often been featured in the past. And while these features were heavily impacted by the time-related differences between the 2020 and 2023 data, similar error patterns were found for both article sets (Figure 5.5). For example, featured comments with a low number of likes were missed, while non-featured comments with a relatively high respect count were ranked too high.



(a) Respect_count errors in 2020 articles

(b) Respect_count errors in 2023 articles

(c) Ratio_featured errors in 2020 articles

(d) Ratio_featured errors in 2023 articles

Figure 5.5: Error analysis including both 2020 and 2023 articles

The previously presented discussion dynamics are not the only factors that have rapidly changed over time. Changing topical focus in comment sections is a given due to it following news cycles. Robustness against such fluctuation in content matter is a necessity and, as shown in earlier paragraphs, our models are capable of dealing with this aspect of online content moderation. Interestingly, however, our results indicate that, even though our best-performing model did incorporate text representation, it is not a prerequisite for achieving accurate rankings of featured comments. First and foremost, the RF model achieved slightly worse, yet similar results to its variant with Bag-of-Words text representation included in its feature set. Using the Wilcoxon signed-rank test on the NDCG scores of both models, calculated on the unseen 2020 articles, we tested whether the performance was statistically different. We found no such significant difference for ranking size 3 ($W = 12,124$, $p = 0.54$), for size 5 ($W = 26,615$, $p = 0.19$)

and size 10 ($W = 41,761$, $p = 0.14$). Second, the models which only used the comment text as input achieved poor results, implying that text features only offered few clues as to whether a comment was featured. Comments with identical text were sometimes featured, sometimes not. Furthermore, the labelling is not exhaustive, meaning that not all quality comments received the featured label. Only a small selection of comments were chosen per article, creating a classification task in which textually similar comments were labelled with differing classes. Combined with the topical variety found across discussions, it posed difficulty to text-based classifiers, namely RobBERT, CNN and biLSTM (Figure 5.3). Thus, this result indicates the power of non-textual features. The models were robust to topical changes due to the fact that text representation only accounts for a minor share in performance.

5.6.2. The human bias of content moderation

The error analysis uncovered clear patterns in certain discussion aspects regarding featured comments. These patterns within the predictions and rankings of the classifiers arose from the already existing bias in the data. Table 5.7 summarizes some of the most important discussion variables, averaged for both the featured and non-featured comments in the entire 2020 dataset.

As already briefly mentioned in earlier paragraphs, a bias exists in which featured comments were written by users who have been featured in the past. While this finding strengthened the hypothesis on which our baseline model was based on, it uncovered the potential human bias in which moderators favour comments of users they know write featured-worthy comments. We found that featured comments tend to be longer than the average non-featured comment (Table 5.7). This can naturally result of the fact that to outline featured-worthy content, a larger word count is needed, aside from the fact that very short comments are not uncommon in responses to other users. The latter are never featured, as they are part of standard discussion thread and not standalone discourse.

	Respect count	**Ratio featured**	**Word-count**	**Ratio rejected**	**Non-replies**
Non-featured	3	0.5%	46	23%	26%
Featured	25	6%	100	14%	35%

Table 5.7: Mean discussion features (calculated on 2020 data)

Other predictive patterns are comments written by users that have a lower share of rejected comments and users that tend to comment directly on the article instead of replying to comments written by other users. On average, users that received featured comments in the 2020 data wrote 36% of their comments directly to the article, while the non-featured average was 26% (Table 3.2). These user features paint a picture of human bias towards certain users themselves. It may be the case that such bias is wanted as a consequence of this moderation strategy. Further ranking of the output can inform the content moderator of these tendencies. For example, by presenting the featured comment ratio of users within the ranking, moderators are able to opt for comments within the ranking which were written by users who have not received a featured comment before but tick all other boxes.

All in all, the human bias in picking featured comments in online discussions on NUjij can be seen as intentional. This chapter made clear that certain non-content aspects of online comments, be it a user who has often been featured before or a comment with a lot of respect points, can be used to reliably rank the comments. Content moderators at the platform in question use such variables to inform or speed up their manual comment curation, whereas others like wordcount can actually be a natural prerequisite. While theoretical definitions of a high-quality comment would probably focus on content matter and presentation, training classifiers purely on textual content did not withstand the topic fluctuations in our evaluation procedure. Additionally, the contextuality caused by other features, such as the content of other featured comments and the real-world position and tone of the discussion and article topic can only be accounted for by the human moderator. For example, an obituary or a scientific news report demand a different discussion character and will influence the featured comment selection. These contextual factors cannot be integrated easily into comment datasets. The classifiers therefore incorporated this intended bias in ranking user comments in order to present a selection of comments that the moderators at the platform deem featured candidates, mimicking their past decisions. This ranking and recommendation procedure attributes the final decision-making to the human moderator, who is able to take into account the contextuality not described in actual discussion datasets.

5.6.3. Limitations and future work

The previously discussed bias in picking featured comments might be platform specific. Other platforms, including the *New York Times* and *the Guardian* have employed the moderation strategy as well. The editorial interpretation of a featured-worthy comment, the emic perspective in practice, may differ from outlet to outlet. Future work should include a cross-platform analysis to more closely analyze the underlying comment and user distributions behind featured comments. However, the data requirements to adequately paint a full picture of online discussions are steep. Most comment datasets only entail public information. Thus, information on rejected comments would be missing. Furthermore, the aggregated user information may not be available to researchers, which forms an important component of understanding human bias in picking featured comments.

Another platform-related limitation is the language. All text used in this study was Dutch. Even though we did not test the outlined approach on data in another language, our approach, which assumes the presence of pre-labeled featured comment data and a transformer-based language model for said language, is entirely language-independent.

Future work should take a closer look at the practice of promoting good comments, as well as the human moderator making these decisions. Ethnographic fieldwork can inform researchers about the processes behind the actual featured comment choices, such as preferences for certain content or user profiles. Such insights would expand our understanding of the emic perspective on constructive commenting. Furthermore, such fieldwork can uncover at what times featured comment are chosen and whether it is a priority for content moderators. The context-specific nature of the moderation strategy also requires further research. Different article types or real-world setting of the story can influence the final decisions made by the human moderator, which is not described in comment datasets. A final point of focus of such future studies should be the detection of opportunities for computational models to support the human moderator within the hybrid moderation pipeline. Following the framework described in the current study, such computational approaches should focus on empowering the decision-making of the content moderator. This procedure supports the moderator and allows them to inject the contextual factors and interpretations that the computational models lack in a much needed and more efficient manner. Fieldwork is also needed to evaluate how content moderators perceive the use of computational systems within their hybrid context. Future work should ask the question whether moderators feel empowered by the use of such models and how the interaction between human moderator and computational model is perceived. While the current study did include an expert evaluation performed by a group of content moderators at the platform in question, we did not evaluate their perceptions of the hybrid moderation pipeline.

5.7. Conclusion

In this chapter, we presented a classifier-based ranking system aimed at supporting the online content moderator in picking featured comments, a widespread moderation strategy. Using comment and moderation data from a Dutch news platform, we showed that combining comment data with user history and contextual relevance achieves high ranking scores. More specifically, our random forest supplemented with Bag-of-Words text representation achieved the best ranking, achieving an optimum F1-score of 0.56 in the initial testing stage. While previous research focused on classifying constructive comments validated their models only on artificially balanced test sets, we validated our models on a large set of individual articles and their discussions. This evaluation setting replicated the real-life practice of content moderation.

To test the robustness of our ranking models against changing contexts and time-related platform growth, we performed ranking evaluations on two sets of unseen articles: (1) a set of articles published in 2020 with similar content compared to the training data and, (2) a more recent set of 2023 articles with a wide range of different topics. We showed that our rankings, aside from those solely based on text embeddings, are robust against these contextual and topic factors. Next, we unpacked the individual topics in both article sets and concluded that all topics achieved high ranking scores. Furthermore, content moderators currently employed at the platform in question evaluated the output of our best-performing model. This expert evaluation yielded an NDCG score of 0.83.

We unpacked our best performing model in terms of error analysis, showing that our model favoured comments from users with a history of being featured and might omit comments with a lower respect count. These findings opened up the discussion on the (intended) human bias in online content moderation, and the context-specificity that the human moderator brings to the table, a feature that cannot be extracted from comment datasets.

Based on the emic perspective on constructive commenting, we proposed a novel approach geared towards ranking feautured comments. In combination with model and decision-making transparency, we aim to support and empower the online comment moderator in their tasks. The human moderator plays and should play a vital role, bringing to the table contextual interpretation of an online discussion that any model lacks. With a clear and delineated role for the computational model in the hybrid moderation pipeline, we do not obscure the nuance and contextuality involved in choosing featured comments, while simultaneously improving both the experience and efficiency of online content moderation at a news platform.

6

The Impact of Featuring Comments in Online Discussions

"Rewarding good commenters might dissuade bad ones.
But I do not know for sure."

NU.nl moderator, NU.nl visit march 2023

Based on: Waterschoot, Cedric; van den Hemel, Ernst; van den Bosch, Antal. The Impact of Featuring Posts in Online Discussions. 16th International Conference on Advances in Social Networks Analysis and Mining – ASONAM-2024 (Shortened version accepted for publication)

Abstract

A widespread moderation strategy by online news platforms is to feature what the platform deems high quality comments, usually called editor picks or featured comments. In this chapter, we compare online discussions of news articles in which certain comments are featured, versus discussions in which no comments are featured. We measure the impact of featuring comments on the discussion, by estimating and comparing the quality of discussions from the perspective of the user base and the platform itself. We find that featured comments are relatively similar to each other and more similar to the article itself than to the average non-featured comment. Furthermore, our analysis shows that the impact on discussion quality is limited. However, we do observe an increase in discussion activity after the first comments are featured by moderators. Furthermore, we make the case for including ethnographic analysis into the study of online content moderation to understand the nuance with which the moderator is shaping the comment space.

6.1. Introduction

How to make online commenters behave? How can you prevent the comment section from becoming a toxic environment? Or, better yet, how can you foster constructive debate in the comment section? These questions become of increasing concern for online media outlets. For at least two decades, online news platforms have been struggling to curtail ‘dark participation’ and trolling in comment spaces (Quandt 2018). Moderators are tasked with moderating this comment space, which initially they did by keeping out all toxic or other unwanted content (Gillespie 2018; Quandt 2018).

As discussed in Chapter 2, in recent years the moderator received an increasingly wider range of tasks besides merely deleting undesired user content. Instead of just removing unwanted content, the moderator is now also tasked with recognizing and promoting ‘good’ comments. Featuring quality comments as a norm-setting strategy became widespread among large online outlets like, for example, the *New York Times* and *the Guardian* (Yixue Wang and Diakopoulos 2022; Diakopoulos 2015a). Presented as examples of constructive discussion among users, certain comments are pinned to a highly visible position within the comment interface. However, much remains unclear regarding the actual impact and implications of this moderation strategy on the discussion. What happens in the comment space when moderators start promoting certain comments and commenters? Is promoting quality comments an effective way to improve the quality of a discussion?

In this chapter, we analyze discussions in which moderators performed the moderation strategy of featuring high quality comments by comparing them to discussions in which no comments were featured. We call this latter set of discussions the control group. By further splitting online discussions, either with or without featured comments, in ‘before’ and ‘after’ subgroups based on the featuring time of comments in the data, we are able to pinpoint differences in discussion quality and activity between the control and the featured content discussions. Specifically, we aim to examine whether the discussion quality increased after comments were featured, as such comments may act as examples to other users. Quality is assessed from both the user and editorial perspective, widening the scope of the concept used in previous research. Additionally, we analyze whether the chosen content is (dis)similar to the wider discussion and the article itself by calculating textual similarities, estimated by computing cosine similarities on text embedding representations.

Overall, we found that featured comments themselves are textually more similar to the article text than to the average non-featured comment. Yet, discussion quality was mostly unaffected. Quality as estimated from the editorial perspective was not impacted by the presence of featured comments. Due to the fact that the control discussions' activity dwindled down faster over time, we end our analysis by hypothesizing whether featuring comments can be used to extend activity on the discussion platform. Indeed, we observe more engagement in terms of comments and involved users in discussions with featured comments.

Chapter 6 is structured as follows. We begin by introducing the moderation strategy of featuring quality comments on online news platforms and contextualize it within the practice of content moderation. Next, we discuss our 2023 news comments dataset from the Dutch online news platform NU.nl and present our discussion quality framework and similarity measurements. After presenting our results, we discuss the apparent lack of influence of featured comments on discussion quality as we measured it, combined with a discussion of the practice behind selecting high-quality user comments. We specifically discuss the need to study the human factor behind comment curation. We end this chapter by outlining several inquiries for future research, as well as the limiting factors of this study.

6.2. Background

On social spaces on the internet content moderation has always existed in some way, shape or form (Roberts 2017). In brief, the task of content moderation is defined as the screening of user-generated content to assess the appropriateness for the given platform, a practice adopted by all online platforms from news outlets to social media (Roberts 2017; Gillespie 2018). The practice has been evolving to address the needs of a growing, contemporary online environment and community. This development expanded the task set of the moderator, who needs to swiftly make interpretative moderation choices (Paasch-Colberg and Strippel 2022). Modern content moderation has grown to include a hybrid setting in which AI is employed alongside human moderators to cope with the sometimes unmanageable quantity of user-generated content (Ruckenstein and Turunen 2020; Gillespie 2020). The topic of content moderation touches many scholarly disciplines and remains full of unanswered questions (Gillespie et al. 2020).

Practically speaking, the moderation task has been described as a gatekeeping role (Wolfgang 2018). This function is twofold. First, moderators ought to keep the comment space clean of unwanted content (Paasch-Colberg and Strippel 2022). Described under the umbrella of ‘dark participation’, this content can take the form of trolling, cyberbullying or even organized misinformation campaigns (Quandt 2018; Lewandowsky, Ecker and Cook 2017; S. v. d. Linden et al. 2017). The moderator is charged with deleting or reducing the visibility of such content (Gillespie 2022). Additionally, coping with these negative influences required the platforms to expand the online content moderation practice (Wintterlin et al. 2020). This expansion, among other things, led to the promotion of *good* content (Wolfgang 2018; Diakopoulos 2015a). The bulk of the literature on content moderation focuses on the bad and unwanted comments and actors within the comment space. Relatively little is known about the active promotion of good commenting behaviour, even though the practice by now is widespread among online news platforms.

As a part of the modern comment section, platforms have been highlighting quality comments on their discussion page (Park et al. 2016). In a practical sense, it takes the form of ‘NYT Picks’ at the *New York Times* (Diakopoulos 2015a), ‘Guardian Picks’ at *The Guardian* (The Guardian 2009) and featured comments at Dutch news platform *NU.nl* (NUJij 2018), for instance. Roughly speaking, the platforms define such quality comments as "substantiated", "most interesting and thoughtful" or "presenting a range of perspectives" (NUJij 2018; Diakopoulos 2015b). In general, featuring what they deem high-quality content is an attempt by platforms at norm-setting (Yixue Wang and Diakopoulos 2022). Dutch online platform *NU.nl* specifically states that these comments serve as examples for other users (NUJij 2018).

Even though most research on content moderation focuses on unwanted comments, some have specifically looked at aspects of featured comments. NYT picks in particular have been used as examples of constructive comments in classification tasks (Kolhatkar and Taboada 2017). Yixue Wang and Diakopoulos (2022) use a classifier trained on NYT picks to assign quality scores to other comments, concluding that users who receive a NYT pick subsequently write higher quality comments, an effect that diminished over time. Yahoo News comment threads have been used for the annotation of good content, more specifically in terms of "ERICs: Engaging, Respectful, and/or Informative Conversations" (Napoles et al. 2017). The authors focuses on the thread level rather than on the comment and did not use an editorial standard as their labelling, as is the case with NYT picks (Napoles et al. 2017). Additionally, research has annotated constructive comments as containing specific evidence or solutions, as well as personal anecdotes or stimulating dialogue (Kolhatkar, Thain et al. 2023).

Focusing on the real-life practice, research has specifically been centred around helping moderators select these kind of comments. Diakopoulos (2015b) aims to reduce curatorial workload for the moderator by examining the relation between the relevance of a comment and their potential selection as a NYT pick. The author introduces article relevance as the cosine similarity between a comment and the article, as well as conversational relevance as the cosine similarity between a comment and the discussion as a whole (Diakopoulos 2015b).

As part of their visual CommentIQ interface, Park et al. (2016) classify and rank comments using comment and user history criteria. These criteria included readability scores and the number of likes a comment had received and relevance scores introduced in earlier research by Diakopoulos (Park et al. 2016; Diakopoulos 2015b). The visual interface allows for different plots and ranking possibilities and uses NYT picks (Park et al. 2016).

With the similar goal of supporting the moderators to pick featured comments, Chapter 5 discussed work in which we trained classifiers to rank comments based on the probability that moderators picked them as featured. Using data from the Dutch news platform *NU.nl*, we supplemented comment and user information with text representation. The models were tested on the discussion level on unseen articles from the platform and evaluated by the *NUjij* moderators themselves, yielding positive results in regard to the ranking of comments.

In sum, while literature on online content moderation is mostly aimed at toxic or other unwanted content in the comment space, moderation strategies aimed at promoting good user-generated content are widespread. While previous work did look at practical support for the moderator in picking content and the effect of highlighted comments on the user and replies to the comment, an analysis on the discussion level comparing discussions with featured content to those without has not yet been performed. This may help identify and quantify the impact that the presence of featured content has on the discussion that continued afterwards. Furthermore, almost all previous research uses the same data source, namely NYT picks. This work contributes results on a different data set and a different language, using language-independent methods.

This chapter aims to address the open questions regarding the impact and similarity of highlighted quality comments on the discussion by comparing them to a control set of discussions in which this strategy was not performed. Furthermore, we broaden the concept of discussion quality by including a user perspective as well, aside from the editorial definition used in previous work.

6.3. Methodology

In this chapter, we use a 2023 Dutch language dataset from the comment platform NUjij, part of the Dutch online newspaper NU.nl[1]. The platform allows users to comment on the news articles published by the outlet. A twofold process of pre-moderation is set in place, combining automatic toxicity filtering with human moderators tasked with checking uncertain outcomes (Van Hoek 2020).

Aside from the set of articles in which moderators picked featured comments ($n = 143$; 86,157 comments, on average 602 comments per article, 1,235 featured comments), we also obtained a control set of articles in which no featured content was chosen ($n = 66$; 32,862 comments, on average 498 comments per article). Articles in both groups have publication times spread throughout the day and follow a similar pattern in regard to comment activity (Figure 6.1). Included in the data are comments that were rejected by the moderators. The articles cover a range of topics such as climate change, the local elections, the nitrogen issue in the Netherlands and the war in Ukraine. Both the featured group of discussions as well as the control group includes articles covering these topics. In the case of featured comments, a timestamp indicates the exact time that moderators highlighted the comment.

In order to assess whether the presence of featured content had any impact, each discussion is split in two subgroups: (1) comments before featured content was chosen and (2) comments after this content was featured on the platform. For the discussions in which comments were featured by moderators (group ‘Featured’), the cut-off was made at the featuring time of the first featured comment per specific discussion. In the case of control discussions (group ‘Control’), which lack featured comments, the split was made at the median time of these first featured comments relative to the publication time of the article (123 minutes). We decided for the median as opposed to the mean (231 minutes) due to the impact of articles published late at night, for which the comment sections only opened up in the morning, leading to an outlier group in comment publication times relative to the article publication time.

This procedure resulted in five different subgroups of comments: (1) ‘Control before’, (2) ‘Control after’, (3) ‘Featured before’, (4) ‘Featured after’ and finally, (5) featured comments themselves.

[1] https://nu.nl

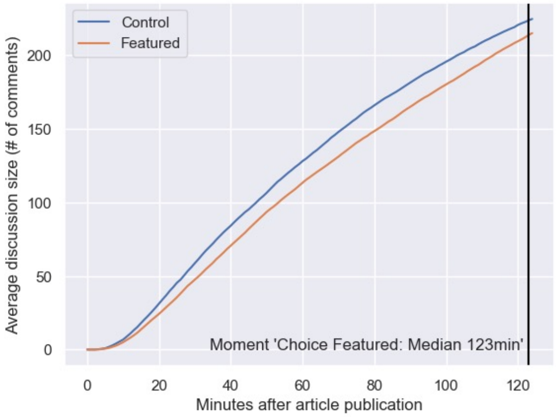


Figure 6.1: Activity in Control/Featured discussions up until cut-off of the control group

To test the validity of the comparison between the featured and control group, we constructed a logistic regression model based on the 'before' data. As dependent variable, we included the group identifier, either control or featured. Thus, the model tests whether it is capable to predict if a 'before' discussions belongs to the featured or control group. If this would be possible, the groups show different discussion characteristics. Independent variables are the studied features discussed in the upcoming sections. These variables were averaged across all comments before the cut-off, i.e. for the control group the 123 minute mark and for the featured group comments posted before the first featuring timestamp in the discussion. Significant effects of these discussion characteristics imply that we cannot conclude that the discussion groups showed similar discussion characteristics before the moderation strategy was performed. This result would suggest the invalidity of our between-group testing. All comparisons between the after subgroups were made using Mann Whitney U tests with Bonferonni correction to correct for Type I error.

6.3.1. Similarity between (featured) content

We focus first on the similarity of (featured) content within the discussion. To calculate the similarity, we finetuned a transformer-based Dutch language model RobBERT on all comments (Delobelle, Winters and Berendt 2020). More specifically, the final model had a batch size of 32, a learning rate of $5e^{-5}$, optimized with AdamW and was trained for ten epochs (Loshchilov and Hutter 2019). For each accepted comment, we derived a vector representation by averaging the BERT embeddings across all dimensions using the SentenceTransformers package (Reimers and Gurevych 2020). The similarity between two or more comments is derived by calculating cosine similarity based on these vectors (Reimers and Gurevych 2020; Diakopoulos 2015b). The higher the score (maximum of 1), the more similar comments are.

In total, we calculated three measurements of similarity. Article similarity describes the similarity of each comment compared to the article itself, while centroid similarity equals the cosine similarity compared to the mean embedding of all comments in the discussion. These two measures correspond with article relevance and conversational relevance, respectively, as operationalized in previous research (Diakopoulos 2015b). Finally, average cosine similarity was calculated comparing each comment with the set of featured comments within their discussion. To obtain each featured comment vector, we averaged the embeddings of the featured comments within the specific discussion. In the case of featured comments themselves, we omitted their own embedding when calculating average similarity to the featured subset, as they have a similarity score of 1 with themselves. This similarity measure indicates whether a comment has similar content compared to those featured by the moderator.

Category	Perspective	Variable	Explanation
Absence of bad content	User	Flagged comments	Number of flagged comments by users
	Editorial	Rejection rate	Share of comments deleted by moderators
Presence of quality comments	User	Respect count	Number of likes on a comment
	Editorial	Featured candidates	Number of featured-worthy comments

Table 6.1: Discussion Quality Framework

6.3.2. Measuring influence: quality & activity

Discussion Quality

We aim to analyze whether discussions in which featured comments are highlighted contain more quality comments compared to discussions without the moderation strategy. We operationalized the concept of *discussion quality* based on two categories, each further broken down into two perspectives (Table 6.1). The two categories of discussion quality are (1) the absence of bad content and (2) the presence of quality comments. Each category is analyzed from both the user and editorial perspective (Table 6.1). We contrasted these markers of discussion quality between the before and the after subgroups for the control and featured discussions to analyze whether discussions with featured content evolved differently.

The absence of bad quality from the user perspective is tested by averaging the percentage of comments that were flagged in both the before and after subgroups. Users are able to flag singular comments indicating that the comment is toxic or inappropriate for the discussion. A higher rate of flagged comments could be seen as an indication that the user base decided the discussion contained less quality comments. The editorial perspective in this category is operationalized through the rejection rate. This variable captures the percentage of comments in each discussion that moderators decided to delete. The need to reject incoming content indicates that it contained bad quality or off-topic content through the eyes of the moderators. Discussion quality would increase if the need to delete unwanted content decreases.

The second category related to discussion quality aims to capture the opposite, i.e. the presence of high quality comments (Table 6.1). The user perspective is defined through the average number of likes comments received, given other users. On the platform NUjij, likes are referred to as respect points. For both the before and after subgroups within the control and featured discussions, we calculated the average number of likes the comments received across the discussions.

Var name	**Description**
Reply_count	Number of replies to the comment
Respect_count	Number of likes comment received
wordcount	Number of words in the comment
wordspersentece	Mean sentence length within the comment
Total_posts_user	Total posts by user
Featured_posts_user	Total featured posts by user
Ratio_featured	Featured posts relative to total posts by user
Ratio_rejected	Rejected posts relative to total posts by user
Ratio_reply	Average reply count on posts by user
Ratio_respect	Average number of likes on posts by user
BoW	Bag of Words: text representation

Table 6.2: Scoring comments for quality: variables in the dataset

We assessed discussion quality from the editorial perspective by following the procedure outlined in Yixue Wang and Diakopoulos (2022). Using a different NUjij dataset containing comments from 2020 (previously used in Chapter 5), we replicated the model for scoring unseen comments in terms of comment quality, operationalized as the class probability for being featured (Yixue Wang and Diakopoulos 2022). A different dataset was necessary because all comments included in this study which are to be scored should not be included in the training and initial testing of the model. The variables used for training the model are summarized in Table 6.2 and are a subset of the variables used in Chapter 5. For training and testing, we split the 2020 dataset into an 80/20 split, resulting in $6,679$ featured comments in the training and $1,661$ featured comments in the test set. A random forest model was trained on a balanced set containing the $6,679$ featured comments alongside an equal number of random non-featured comments. The final model, calculated on unseen test set, achieved an F1-score of 0.86, a similar result as reported in previous work (Yixue Wang and Diakopoulos 2022). Each comment from the current dataset received a probability of being featured-worthy, a proxy for the editorial view of comment quality (Yixue Wang and Diakopoulos 2022). To assess the discussion quality, we counted the occurrences in which the classifier assigned a probability to be featured higher than 0.5 per discussion, the threshold for belonging to the featured class according to the model. Finally, we calculated the averaged percentage of featured-worthy candidates, contrasting the before and after subgroups. The assumption, again, is that higher discussion quality through the editorial lens is approximated by a higher number of comments that are classified as to be featured by moderators.

Discussion Activity

The final point of focus relates to the evolution of activity within discussions on the platform. We assessed how the discussions evolved in the first 10 hours by counting the average number of unique users and mean total comments, comparing the before and after subgroups for both the control and featured groups. For this analysis, we only included the accepted comments in the discussion, leaving out the comments that were deleted by a moderator. We are particularly interested in analyzing whether the featured content caused a different evolution in discussion growth.

6.4. Results

6.4.1. Validity of 'before' subgroups

	Coeff (std er.)	**p-value**
Respect count	-0.055 (0.039)	0.165
Featured candidates	0.038 (0.030)	0.209
Flagged comments	0.034 (0.035)	0.338
Rejection rate	-0.006 (0.015)	0.678
User count	0.007 (0.005)	0.158
Comment count	-0.003 (0.002)	0.172
Dependent variable: group (control before/featured before)		

Table 6.3: Logistic regression on before data: characteristics before cut-off

We examine whether the discussions in both groups before the cut-off have different characteristics to begin with, which would invalidate the comparison between the control and featured group. The variables we included in this analysis were those in the discussion quality framework as well as those representing discussion activity. Our comparison between the featured and control group was valid if no significant differences were found, as that would indicate that the discussions were already different before the cut-off. This analysis was done on the discussion level and included the average discussion characteristics calculated across comments posted before the cut-off, i.e. the 123 minute mark for the control group and comments posted before the first featuring timestamp in the featured group.

We fitted a logistic regression model with the group identifier as dependent variable. The results are summarized in Table 6.3. The included independent variables (averages on discussion level) are: *respect count, featured-worthy candidates, flagged comments, rejection rate, unique user count*, and *comment count*. The results show that before featured comments were chosen in the discussion, the included discussion characteristics were not statistically significant between the two groups. Therefore, we could safely assert that discussions before featured comments were selected are not intrinsically different.

6.4.2. How similar are featured comments to articles and other comments?

Table 6.4 summarizes the mean similarity scores across all subgroups in the data. The results show relatively high similarities across all categories, potentially due to the fact that off-topic or unwanted comments were already rejected by the moderators, combined with the topical focus within discussions on the platform, as those comments were all posted on a single article.

The overall outcome indicates that, on average, the set of featured comments within a discussion were distinct from the rest of the comments. With a cosine similarity of 0.903, they are significantly more similar to the article itself compared to the average non-featured comments ($U = 69272.0, p < 0.001$). Furthermore, these featured comments were also highly similar to the other comments within the featured set with a cosine similarity of 0.929. This similarity to featured content was significantly higher ($U = 64830.0, p < 0.001$) than that of the non-featured comments in the discussion group (cosine similarity of 0.828 before, and 0.827 after). Additionally, non-featured comments were more similar to the centroid of the non-featured comments than to those that were featured by a moderator (Table 6.4). Within the featured group, we did not find any indication that non-featured comments posted after the moderator highlighted quality comments are more similar to the featured content. This similarity before (0.828) was virtually the same as in the after group (0.827).

		Centroid	**Article**	**Featured**
Control	Before	0.939	0.883	N/A
	After	0.936	0.887	N/A
Featured	Before	0.938	0.868	0.828
	After	0.934	0.875	0.827
Featured comments		0.889	0.903	0.929

Table 6.4: Mean cosine similarity within online discussions

Additionally, we did not find any differences comparing non-featured comments in the featured group with the control discussions (Table 6.4). In the control group, the centroid similarity did not change after the time-cut off, which was also found in the featured discussion. Finally, the control and featured group followed a similar pattern in regard to article similarity. Comments are as similar to the article text in the control group as they are in the featured group

Summing up, we do not find different patterns between the non-featured comments in the featured group and the comments in the control group. Both groups of comments did not become more or less similar to the conversation itself or the article text after the time cut-offs. However, we do find that the featured content itself is significantly more similar to the editorial output. These sets of high quality comments were also significantly more similar to each other compared to the other, non-picked comments in the discussion.

6.4.3. Discussion Quality

Discussion quality is studied comparing the 'before' and 'after' groups and by contrasting the control group with the discussions in which featured content was chosen. Previously, we concluded that the before groups showed no difference in discussion characteristics. We divided quality into two categories: (1) absence of bad content and, (2) presence of high-quality comments. Each category was operationalized through the user and editorial perspective. This discussion quality framework resulted in four variables calculated for the before and after subgroups for both the control and feature sets (Table 6.5).

Category	Perspective	Variable		Before	After	Δ
Absence of bad content	User	Flagged comments	Control	9.08%	10.07%	+0.99pp
			Featured	9.74%	9.7%	-0.04pp
	Editorial	Rejection rate	Control	23.25%	22.20%	-1.05pp
			Featured	23.54%	20.91%	-2.63pp
Presence of quality comments	User	Respect count	Control	5.69	3.61	-2.08
			Featured	5.65	3.43	-2.22
	Editorial	Featured candidates	Control	14.04%	8.90%	-5.14pp
			Featured	14.39%	8.30%	-6.09pp

Table 6.5: Discussion Quality: differences between the before and after subgroups

The user perspective on the absence of bad quality content is expressed through the average percentage of flagged comments by the user base across all discussions. In the control group, we found that 9.08% before the 123 minute mark and 10.07% after the cut-off were flagged by at least one user (+0.99pp). For the featured group, we calculated a decrease from 9.74% before to 9.70% after moderators highlighted quality comments (-0.04pp). The results indicate that, while featuring content did not decrease the flagging of bad comments by users, it could prevent more flagged comments later on in the discussion, as is seen in the control group. However, the difference between the average number of flagged comments in the control (9.70%) and featured (10.07%) 'after' discussions is not significant ($U = 4830.50$, $p = 0.42$).

The editorial perspective on absence of bad content was calculated by the average percentage of comments deleted by moderators in a discussion at the time. Both the control group (−1.05pp) and the featured group (−2.63pp) showed a decline in the need to reject comments by the moderators as time went on (Table 6.5). Comparing the after groups of both the featured on control group, we did not find a significant difference (U=4021.5, p=0.38). This result suggest that the presence of featured comments did not reduce the need for moderators to reject incoming comments.

The second category of discussion quality aimed to capture the presence of quality content (Table 6.5). The user perspective of this category was operationalized by calculating the average number of respect points that comments received in a discussion before and after the cut-off. Over time, the average number of likes declined, potentially due to the fact that comments posted late in the discussion are read less often. We specifically looked at whether this decline would be less steep in the featured discussion. This was not the case (Table 6.5). In the featured group, the average number of likes declined by 2.22 respect points. While in the control group, the average declined by 2.08 respect points. This difference between the control and featured group after the cut-off is not significant ($U = 4420.0, p = 0.8$). This result implies that featuring content had no impact on the average number of likes other users gave to comments.

The final marker for discussion quality was the percentage of featured candidates (class probability > 0.5) from the classifier, representing the editorial perspective on quality comments. As shown in section 4.1, the average percentage of candidates before the cut-off showed no difference comparing the featured group (14.39%) and the control group (14.04%). As the discussion continued, the average share of featured candidates comments decreased in both featured ($-6.09pp$) and control ($-5.08pp$) groups (Table 6.5). The results imply that the editorial view on discussion quality was not affected by the presence of featured comments, as the difference in average number of featured candidates in the after subgroups were not significant ($U = 4093.5, p = 0.49$). The average percentage of featured candidates in the featured and control groups after the cut-off were 8.30 and 8.90, respectively (Table 6.5).

6.4.4. Discussion Activity

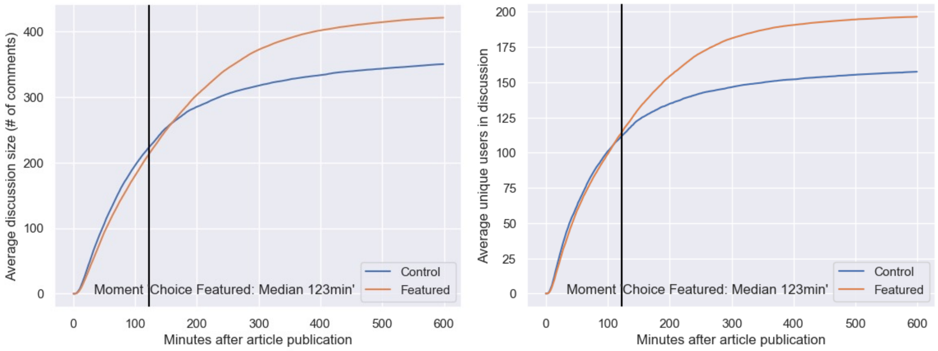
Finally, we analyzed the discussion activity before and after the cut-off based on two factors: (1) the set of users commenting and, (2) the average number of comments in the discussions (Table 6.6). This analysis only included the accepted comments in the discussion, omitting those that were rejected by the moderators. Previously we have shown that before the cut-off, the discussions showed similar activity characteristics.

		Unique Users	**Comment count**
Control	Before	112	223
	After	45	126
Featured	Before	115	213
	After	82	207

Table 6.6: Average platform activity in the before and after subgroups (first 600 minutes)

Unlike the quality markers discussed earlier, the discussion activity progressed differently within the featured group as opposed to the control discussions (Figure 6.2). The speed of discussion activity slowed down over time, which is expected due to the fact that platform users move on to more recent articles resulting in a dwindled down discussion. However, the average discussion activity in the control group dwindled down much quicker; the featured discussions continued on and slowed down at a point later on (Figure 6.2).

Before the cut-off, an average of 112 users commented before the 123 minute mark in the control discussions, while 115 users participated before the first comment was featured in the featured group. In the case of the after groups, however, a significant difference was found, indicating that more unique users commented in the featured discussions after content was featured ($U = 3205.5, p < 0.001$). The same result was found for the comment count (Table 6.6). The average comment count before the 123 minute mark in the control discussions was 223, while in the featured group this was 213 comments. After the cut-off, the average control discussion went on for 126 more comments, while the mean comment count in the featured group was 207. This difference in after groups is statistically significant ($U = 3161.5, p < 0.001$).



(a) Evolution of average comment count

(b) Evolution of average number of users

Figure 6.2: Growth of discussion activity in featured and control group (first 600min)

6.5. Discussion of the results

Overall, the results suggest that the presence of featured content did not affect the quality markers. Both the rejection rate and respect count followed the same downward trend in the control and feature groups as time went on. Even though the average flagged comments per 100 did increase in the control discussions after the cut-off, we did not find a statistically significant difference compared to the featured discussion. Similarly, the number of featured candidates decreased when comparing the before and after subgroups, indicating that the content itself in terms of quality is not responsive to moderators highlighting examples of what good content in the editorial sense. However, we did find a difference in regard to the evolution of discussion activity within the featured group. The results indicate that discussions in this group continued to grow for a while longer compared to the control discussions. Yet, in the case of discussion activity, it may be unclear which other factors might exert influence.

In the following paragraphs, we discuss and contextualize the apparent inability to influence discussion quality with featured comments, along with the similarities that we found within sets of featured content and between those comments and the article itself. We end the discussion by outlining several open questions for future work and the limitations of the current study.

On the Frequently Asked Questions page on NUjij,[2] it is stated that the featured com-

[2] https://www.nu.nl/nujij/5215910/nujij-veelgestelde-vragen.html

ments serve as an example to other users, implying a function as template in such a way that users subsequently write comments of higher quality. The standard of what a high-quality comment looks like is decided by the moderators and editorial staff.

However, our results have shown that discussion quality was unaffected by comparing discussions with featured content to the control group. We operationalized discussion quality into two categories, further split into user perspective on the one hand, and the editorial perspective on the other. First, featuring content did not lower the necessity to reject incoming comments, implying that the highlighted examples did not deter people from posting uncivil or off-topic content. While the rejection rate did decrease over time, control discussions without featured comments evolved in the same manner. From the user perspective, the control group experienced an increase in user flagging in the after group, while this increase was not found in the group with featured comments. However, this difference was not significant.

Second, we found no difference between the two groups in regard to the user perspective on quality, represented in the study by the number of likes comments received. Over time, the average number of likes a comment received decreased significantly with or without featured comments. The final marker of discussion quality, from the editorial perspective, was operationalized by training a classifier to predict whether comments could be featured. If such highlighted comments would successfully serve as examples to users of the editorial standard of quality, one would expect the number of comments qualifying as featured-worthy to be significantly higher in the after group of featured discussion compared to the after group comprising of control discussions. However, we did not find such a significant difference. In both groups, the number of featured-worthy comments as seen in the output of the classifier decreased in the after subgroup. Overall, these results indicate that discussion quality decreased over time, whether or not high quality comments were highlighted by the moderators on the platform.

Even though the practice of featuring user comments is widespread among online news platforms, the inability to influence discussion quality in all but one of the variables (and not significantly so) indicates that users do not use these comments as examples. In particular the lack of change in the editorial perspective showcased that the platform does not succeed in shaping the discussion to what they themselves deem good discussion.

However, previous research has suggested that after a user's comment was featured by a moderator, the subsequent comments made by that specific user were of a higher quality (Yixue Wang and Diakopoulos 2022). The discussion itself might be unaffected, but a spillover effect may improve other discussions on the news platform. Moreover, elevating certain content within comment sections might serve a more broad purpose aside from improving the editorial standard of quality. For example, a collection of comments can serve as a discussion summary for readers.

6.5.1. Future work and limitations

While the moderation strategy of featuring 'good' content is widespread among online news outlets, questions remain about its specific goal: there are big differences between media outlets about what constitutes a ‘good’ discussion or a ‘constructive’ comment. Furthermore, additional research is necessary to pinpoint further effects on the discussion. The following paragraphs formulate four avenues towards these open questions, as well as some limitations of the described research.

This chapter showcased that activity decreased quicker in discussions without featured content. With future research in mind, we hypothesize that featured content could be used to postpone the natural decrease in discussion activity over time. Further studies should focus on this particular point, eliminating other factors through, among other things, A/B testing. In particular, the timing of picking featured comments can be a factor to be manipulated.

Second, cross-platform analysis is necessary to assess the general impact of the moderation strategy. Platforms like, for example, the New York Times and the Guardian highlight certain comments as well. Such an analysis can compare multiple platforms to test whether moderators at different outlets feature similar content. Additionally, the impact on the discussion and on the user base of their own platform can be assessed to test whether the reaction of different audiences shows similarities. While most previous research has focused on datasets containing NYT picks, we provided an analysis of a different platform. Future research should follow this procedure and include different, international platforms performing similar moderation strategies.

One other avenue for future research regards the balance of viewpoints in the set of featured comments. Similarity is high within the chosen content, but that does not necessarily mean that there is no diversity of argumentation or perspectives. Future research can address this by analyzing the viewpoints present in sets of featured content in relation to the discussion topic of the article.

Research aimed at analyzing the effects of online moderation has the inherent constraint of the availability of data and metadata. To replicate the current study, researchers require not only the comments that were published at the time, but also information about those which were rejected or later deleted by moderators. The latter is not publicly available, typically, cannot be published, and requires cooperation with online news platforms to be obtained, or working from within the platform’s organization. Furthermore, individual comments require metadata indicating to which article they were posted, such that researchers can separate different discussions. Timestamps indicating when comments were featured are also not publicly available, typically, as well as information as to how many times a comment might have been flagged by other users. All in all, without this crucial information, the potential effect of these moderation strategies cannot be adequately analyzed. Cooperation with the platforms is needed to obtain such unpublished information.

Fourthly, we advocate for a mixed-methods approach in future work. The curation of user-generated comments remains a highly contextual affair, involving highly subjective processes, from the perspective of moderator and user. The practice of picking featured content is a very human affair, with only limited assistance from AI. Studying the impact of featured content on a discussion and participating users should therefore include an inquiry into how the human moderator operationalizes the concept of a quality comment and the practical selection process itself. How does the subjective preference of the moderator play out? How is the moderator trained, instructed to promote constructive content and what is the moderator's interpretation of the training and guidelines? Many such nuances cannot be extracted from datasets alone. Although in this article we focused on the demonstrable effects of the promotion of constructive content in datasets, to understand the practice more fully, a combination of methods will be necessary. Ethnographic fieldwork can aid in mapping out the human processes involved in the content moderation practice.

This became apparent when we visited the offices of *Nu.nl* and sat next to a moderator. The moderator was working on multiple screens and the task at hand was a discussion about whether a quota needed to be put in place to improve the number of women that are hired by European companies. The moderator quickly scanned the discussion, made a coffee, discussed with colleagues about a concert they saw past weekend whilst scrolling through the comments to find a comment that fit what he felt was missing in the discussion. We asked him what he was looking for, how does he recognize a constructive comment? The response was: "I don't know, but I know what a constructive comment is when I see it". This left an impression on us of the many contextual and subjective factors that are at play in the process of selecting a comment. In future studies of moderator practices, we recommend combining data-analysis with an attunement to the lived reality of content moderation.

6.6. Conclusion

User-generated content has sometimes been referred to as 'the bottom half of the internet' (Reagle 2015). This description had both positive and negative connotations. The bottom half of the internet is associated with a sense of freedom: a free exchange of uncurated and unmoderated thoughts. It also has a negative connotation: it is also the place where hateful, lewd, toxic behavior is sure to be found. The tendency of news outlets to promote comments that are deemed constructive creates what can be seen as a third layer, between journalistic content and comment section. Online news platforms have been highlighting comments by picking them out of the bottom they were originally published in and elevating them into a curated space between the editorial output (the top end) and the other user-generated content. The current study provided a deeper insight into what this third space, between journalistic content and comment section, consists of.

To achieve this, we analyzed discussions in which moderators picked featured content and compared them to control discussions in which this moderation strategy was not performed. These two groups were further divided into a before and after set, separated by the moment the moderator chose the the first comment to be featured, operationalized by a proxy of 123 minutes, the median publication time of the first featured comment in the featured discussions. The study was structured based on three focus points: (1) similarity between different classes of content, (2) impact on discussion quality and, (3) discussion activity.

Our results indicate that featured content within the same discussion is highly similar to each other, and is also relatively similar to the article text itself. This curated collection of content seems to follow an explicit or implicit editorial guideline and sets itself apart from the rest of the discussion, visually separated from the rest of the user-generated content. Furthermore, the similarity between the average comment in the after subgroup and the set of featured content did not increase compared to the before group, indicating that users do not start to write comments more similar to what is being highlighted.

Additionally, we did not find evidence indicating an impact on discussion quality as a result of the featured content, especially from the editorial perspective. The rejection percentage was unaffected by featuring content, as well as the number of featured-worthy comments. Both aspects of discussion quality decreased in both the control and featured groups in similar fashion. A similar decrease was found in the average respect points comments had received, the user perspective of presence of quality comments. What did change, however, was the average number of flagged comments by users. As opposed to the other quality variables in the framework, the average number of flagged comments increased in the after group of the control discussion, while this increase was not found in the featured group. However, this difference was not statistically significant between both 'after' groups.

Finally, we did find differences in discussion activity between the featured and control group. The results show that discussion activity, expressed in the number of unique users and comment count before and after the moment of choice, declined slower in the featured group. It would seem that this moderation strategy can be used to postpone the decline in user activity on the commenting platform. However, future research is necessary to eliminate other factors potentially influencing discussion growth.

To conclude, modern comment sections operated by online news outlets have been, and still are, growing and evolving. Promoting highlighting quality user-generated content has become an important strategy performed by the content moderator. While there potentially are multiple goals behind this focus, the current study showed that the discussion quality itself is not impacted, positively or negatively, questioning the example-setting objective. And while the comment space continues to adapt itself to changing online environments, the study of these practices ought to be expanded to encapsulate the entire context, combining computational analysis with ethnographic fieldwork with the moderators.

7

Discussion and Conclusion

"Comment sections can only be opened if platforms are prepared to invest in manual and creative content moderation"

NU.nl moderator, NU.nl visit june 2022

In the previous chapters, I explored the content moderation strategy of promoting constructive comments and how it is operationalized in practice. I examined two distinct perspectives for studying the concept of constructive commenting and various computational models designed for filtering such comments. Additionally, I evaluated the effect of certain strategies used to foster and promote constructive commenting. I did this by investigating how on *Nu.nl* featuring constructive user comments influences the discussion quality and activity. In this concluding chapter, I aim to discuss and contextualize the results from the preceding studies by addressing the research questions. Furthermore, I will reflect on the two perspectives on constructive commenting and formulate several unanswered questions for future research.

7.1. Research questions

7.1.1. The ‘Third half of the Internet’

Chapter 2 presented work aimed at analyzing how various news outlets promote constructive comments and, more specifically, how they formulate what featured-worthy comments look like:

RQ1: How is it decided what constructive comments are and how are they promoted on different news platforms?

I answered this question by studying five news outlets renowned for their online comment section: The *New York Times*, *El Pаìs*, *Die Zeit*, *The Guardian* and *NU.nl*. Specifically, I studied the implementation of two distinct moderation tasks: keeping out toxic content and promoting constructive comments. Whereas the former is done with AI and hybrid moderating systems, I concluded that the latter mostly takes the form of manually picking specific comments. These picks are visually promoted and highlighted in the comment interface, either pinned above the other comments or displayed in their own tab. Studying the definitions of such content provided by news outlets, I found diverse and vague descriptors, such as high-quality or informative. Usually singular contributions are highlighted. An alternative is awarding users a badge, rewarding readers who have a history of writing comments that are deemed constructive.

I coined the phrase ‘third half of the internet’, a curated collection of user comments placed between the editorial content (‘top half’) and other user contributions (‘bottom half of the internet’). These promoted comments build upon the content put out by the news outlet, therefore reinforcing the editorial view on constructive content. However, as mentioned earlier, a clear blueprint of constructive commenting was not available. By providing vague definitions and instructions to moderators, it may be expected that existing personal biases are reproduced.

Hybrid content moderation is widespread among online news outlets, creating a practice in which moderators work alongside AI-based tools. However, these tools are limited in scope, trained only to classify comments in terms of toxic content. Thus, moderators are tasked with manually sifting through online discussions to decide which comments are worthy of being featured. News outlets have expressed an interest in employing

AI-based tools within the third half of the internet, as it is currently a laborious task, requiring a lot of time and human attention. To achieve the goal of employing AI-based tools to promote constructive discussion, models need to be able to recognize specifically what makes a comment constructive, for example in terms of linguistic content. To achieve this, AI models need to be trained on large datasets containing not only the textual content of comments, but also a wide array of metadata available to the content moderator. Nevertheless, the most important requirement for the computational classification of constructive comments is deciding what exactly a constructive comment is and how this is identified in the data.

7.1.2. Computational filtering of constructive comments

The subjectivity and vagueness around the concept of what constitutes a constructive comment was the starting point of the research described in this thesis. In order to train and classify comments in terms of constructive value, a perspective and working definition of constructive commenting has to be defined. In this thesis I made use of two distinct perspectives: an etic perspective as the viewpoint of external researcher formulating a formal definition of a constructive comment and secondly, an emic perspective, adopting the viewpoint and decisions made by the content moderators who are tasked with filtering out constructive comments out of online discussions. I answered the following research question in Chapters 3, 4 and 5:

RQ2: Given either an etic or emic perspective on constructive commenting, to what extent can computational models detect constructive comments in an online discussion?

In Chapters 3 and 4, I studied constructive commenting through the lens of an etic perspective. This perspective is defined as the outsider's view, establishing requirements of what makes a comment constructive. To achieve a formal definition on constructive commenting, I studied potential building blocks of such a definition. As a result, a constructive user contribution may be seen as a *sum of different constructive parts*.

In Chapter 3, I investigated diversity of argumentation and interactivity as potential indicators of constructive commenting. To achieve this, I created a case study on a highly polarized and contentious topic that lead to a lot of public comments in the Netherlands. More specifically, I looked at discussions about Black Pete (Zwarte Piet) in The Netherlands using social media data from *Twitter*, *Reddit*, and *Gab*. We formulated a thread interactivity score which indicates whether diversity exists within the argumentation presented in an online discussion thread. This score is calculated iteratively each time a new comment was added to the discussion thread. A positive score stipulates a larger share of comments disagreeing with the original (parent) comment, while a negative score informs us that the thread is flooded with comments enforcing the arguments already presented in the thread, an echo chamber on the discussion thread level. Additionally, I calculated a Message Interactivity Contribution score, an indicator of how novel the argument was in the thread at the time the comment was posted. Comments with an argument currently not present in the discussion thread receive a higher score.

The constructive element in this chapter can be described as the presentation of novel information. A constructive comment brings a new argument to the table and contributes to a balanced discussion in terms of argument diversity. However, not all arguments in a discussion may be evaluated as equal by news outlets. Some may be seen by editorial staff as undesirable. This could be because they are seen as toxic or contributing to polarization, or because they contain misinformation. Consequently, an approach was needed which adopts this evaluative dimension by editorial standards of arguments in an online discussion.

In Chapter 4, I expanded on the used etic perspective on constructive commenting by focusing specifically on minority argumentation and mutual understanding, an application falling under the umbrella of 'argument mining' in NLP research. I created a case study using climate change news articles published by Dutch news outlet *NU.nl*. I trained classifiers to label user comments for the specific argument it presents, with a particular focus on improving the capability to recognize minority arguments more accurately. For this case study, these minority classes were those denying anthropogenic climate change. An interesting aspect of this discussion on *NU.nl* is the current editorial guidelines stating that denying anthropogenic climate change is prohibited on the platform. Thus, the automatic labelling of comments belonging to these minority classes aids in the search for constructive comments, as they would in principle qualify for removal, and may require specific moderator attention. In the case of the climate change discussion on *NU.nl*, a constructive comment ought to present the argument accepting anthropogenic climate change.

Initial classification of the minority arguments was poor. However, using an active learning approach to filter out more minority training samples out of unlabeled data, I improved the models' capability to correctly label comments belonging to the minority classes. Two waves consisting of 1,000 additional training samples, specifically filtered to include relatively more minority samples, improved the overall F1-score with 23 percentage points. My analysis ended with an investigation of the textual patterns found within each argumentative cluster. Showing an understanding of what other readers are saying within each argumentative class in a polarized debate may boost overall constructive value within a discussion, as mutual understanding increases. I conclude that clear patterns exist which shape each argument in the online climate change debate, providing a coherent understanding of the subject matter within each class.

Overall, using an etic perspective on constructive commenting offers several advantages. This approach makes it possible to formulate an annotation scheme, clearly indicating what content qualifies as constructive. Additionally, by unambiguously formulating what a constructive comment entails as well as labelling individual comments, it is possible to create a framework that can seamlessly be applied to other platforms. Analyzing constructive commenting from an etic standpoint relies solely on textual content in the form of user comments, without requiring additional metadata or platform-specific factors that may be unattainable or not found elsewhere. It is possible to replicate the presented studies on differing news outlets as well as on social media platforms.

However, employing an etic perspective also has drawbacks. Although it offers a precise annotation scheme for identifying constructive comments, it relies on manually labelled data. Unfortunately, such datasets are not readily available and are time-consuming to compile. Moreover, the research discussed above relied on case studies. The approaches were heavily dependent on the context and structure of the discussion, with clear pro and con sides. Thus, these chapters were modeling a specific discussion, not necessarily the concept of constructive commenting. With changing news cycles, these approaches will not remain up to date as discussion topics rapidly change with time. Manually compiling datasets on novel discussions is generally too time-consuming to keep up with the rapid pace of contemporary online news cycles and discussions.

Additionally, an identical comment may be constructive in one discussion and not constructive in another. The presence of good argumentation or newly available information may not be a strict requirement. A lot comes down to interpretation by those tasked with and trained on choosing the featured comments – the content moderators.

Through ethnographic fieldwork with *NU.nl* moderators, I found that context does indeed influence the selection of featured comments. Some discussions require personal anecdotes, others a clear representation of facts. Additionally, the presence of featured comments selected earlier in the same discussion may also influence moderation choices. Specific arguments that are not presented in earlier featured comments may become more constructive as the moderator aims to create a balanced selection of promoted comments. Consequently, an etic perspective may not hold across different discussions. A different approach to modelling constructive commenting which overcomes the hurdle of generalizability was necessary.

The alternative perspective used in this thesis is an emic perspective, which, in contrast to the previous studies, does entail the contextual decisions moderators made. This perspective makes use of historical moderation data. The goal of using an emic perspective is to computationally select constructive comments based on what the data puts forward as decisive criteria in past moderation decisions. Chapter 5 continued to explore this different view on what constitutes a constructive comment. I introduced a group recommender system tailored to integrate seamlessly into the hybrid framework of content moderation. Specifically, the approach refrained from a definitive labeling of comments as constructive or not. Instead, it entrusted the decision-making process to human moderators, while the model assigned a rank to each comment in an online discussion based on its likelihood of belonging to the featured comment class. As a result, moderators could focus solely on the top-ranked comments rather than sifting through the entire discussion. Consequently, resulting decisions regarding which comments to feature consider the context of the discussion and subjective perspectives of individual moderators.

In practical terms, I found that incorporating the textual input from comments did not improve the models. Moreover, attempting to strictly model the moderator's perspective on featuring comments based on textual input yielded poor results. This model also lacked robustness against shifts in news cycles, with its performance dropping notably when evaluated on article topics different from the training and validation data. A challenge for the computational modelling is that not every potentially suitable comment received the distinction of being featured. This is specifically challenging for content-based modeling of constructive commenting. Identical comments in terms of textual input were inconsistently featured by moderators, demonstrating the complexity of the task. The metadata, both on the comment and user level, proved to be the most informative. More specifically, the most contributing variables were the number of respect points (likes) a comment received, the wordcount and the ratio at which the user had been featured before in the past. The best-performing model was a random forest trained on both comment and user metadata and textual representation of the comments. However, its performance was not a statistically significant improvement over a random forest implementation lacking the textual representation.

To further assess an emic perspective and its potential for computational modeling of constructive commenting, I conducted a two-step evaluation. This involved replicating the effects of topic changes due to news cycles. The models were evaluated on unseen discussions from 2020 and from 2023, each characterized by a distinct set of news topics. I found that, aside from the strict text-based model, a change in article topics did not impede the ranking of user comments. This is a clear advantage over the approaches discussed in Chapters 3 and 4, due to the fact that these were based on case studies.

Furthermore, I performed an expert evaluation together with four content moderators currently employed at *NUjij*. This yielded two key insights. Firstly, the models did indeed facilitate a more efficient process for content moderators in featuring constructive comments. Moderators were more inclined to feature comments provided by the model (64%) compared to randomly selected, non-ranked comments in the evaluation (34%). Notably, in all but one of the assessed articles, moderators chose to feature at least one comment recommended by the model. Secondly, I observed that, even among experienced content moderators working at the same online platform, clear variation exists in the understanding and promotion of constructive commenting. Analyzing the decisions of the four moderators, I calculated a Krippendorff's alpha inter-rater agreement of 0.62. Additionally, I found that 42.3% of the evaluated comments previously featured in 2020 were not selected this time around in this evaluation. This variability underscores the decision to prioritize ranking comments over classification, as moderators did not universally agree on the labelling.

While these results were promising, some questions remained. Did using an emic perspective bring us closer to a computational understanding of constructive commenting? Did we model the behavior of *NUjij* moderators or is this approach transferable to other platforms? How much does the context, in the form of platform specifics and editorial guidelines, influence an approach based on the emic perspective? Approaches based on the etic perspective do not suffer from this potential drawback, as it is not influenced by platform specifics. The labels used in Chapters 3 and 4 are not derived specifically from *NUjij*, but from previous research on the topic pertaining to the case study.

Another disadvantage of studying constructive commenting using an emic perspective are the steep data requirements. The metadata used for training the models in Chapter 5, both on the user and comment level, are not readily available in the case of most platforms. While some variables are accessible, such as the number of likes a comment received or the wordcount, others are typically not part of publicly available comment datasets. Cooperation with the platforms in question may be needed to obtain crucial information. For example, platforms do not publish comments that have been deleted by moderators. Without these comments, it is impossible to calculate the rejection rate of every user. Additionally, one of the most informative variables in Chapter 5, the share of comments featured for each user (ratio_featured), could only be calculated by using a user dataset. A final variable only available through cooperation with platforms are the timestamps of moderation actions, the crucial factor in splitting the data for Chapter 6. Without these timestamps, it would have been impossible to obtain before and after subsets, a strict necessity for the impact analysis presented in Chapter 6. These steep requirements are a hurdle for the applicability of the approach and might make using the emic approach in future research less feasible.

In sum, I argue that my approach has shown that constructive commenting consists of multiple distinct aspects. I highlighted the necessity for including a multifaceted approach to investigate the complex concept in practice. The strength of adopting the etic perspective lies in its independence from specific online platforms. However, the need for labeled data poses a significant challenge, particularly in the context of fast-paced news cycles. Additionally, the topic dependency of the presented studies limited their use, leaning towards modelling individual debates rather than a universal concept of constructive commenting. On the other hand, employing an emic perspective offers a context-free approach, as evidenced by evaluations across various news topics, yet it also has severe shortcomings, notable in terms of data requirements. Collaborating with and among news platforms, sharing their data in an agreed-upon format, may provide a sustainable solution. Further (cross-platform) research is needed to determine whether the approach outlined in Chapter 5 merely modeled behavior of a moderators at *NUjij* or moves closer to a conceptualizing constructive commenting.

After all these efforts directed towards promoting constructive comments, one question remains: to what extent does the promotion of such comments influence the dynamics of an online discussion? And, going further, How is the measurement of impact influenced by the perspective on and definition of constructive commenting? It was unclear whether users adhere to the example set by the selected comments. Are discussions pushed in a more constructive direction as a result of the work put in by the moderators? This question is interesting both from a practical standpoint for news outlets, as well as from an academic standpoint looking at how online users react to content moderation strategies.

7.1.3. The impact of promoting constructive comments

Chapter 6 delved into an analysis of the moderation strategy, dividing the data into a control group, consisting of discussions without featured comments, and a featured group. Each group was further split into a before and after section using the timestamps indicating when the first comment in a discussion was featured by a moderator. This set-up was used to investigate two research questions regarding the impact on discussion quality and activity.

RQ3a: Comparing discussions where moderators promoted constructive comments to discussions lacking this moderation strategy, does the discussion quality increase after the comments were promoted?

In Chapter 6, I presented a framework on online discussion quality based on two factors of quality: (1) the absence of bad content and, (2) the presence of high-quality comments. Each factor was further split based on the individuals participating in a discussion: the user base and the editorial staff. All in all, the discussion quality did not increase after constructive comments were promoted. The user perspective on bad content, the share of comments which were flagged by other users, did increase in the control group. However, the resulting difference between the control and featured group was not statistically significant. Other discussion quality variables evolved similarly in both the control and featured groups. The rejection rate, which constituted the editorial perspective on bad content, decreased over time whether comments were featured or not. A similar result was found for the user perspective on the presence of high-quality content, operationalized through the average number of likes. The final variable of the discussion quality framework was the presence of quality comments from the editorial perspective, expressed through the number of featured-worthy comments in the discussion. Similar to the previous variables, I found a decrease over time both in the control and featured groups that was not statistically significant.

In short, the presence of featured, constructive comments did not impact the discussion quality as described in the framework. I concluded that users did not use such promoted comments as examples when writing their own comments. Additionally, it did not reduce the workload of moderators in terms of deleting toxic or other unwanted comments, as the rejection rate per discussion trended similarly in both the control and featured group. Aside from measuring the quality within online discussions, Chapter 6 investigated potential differences in activity between the control and featured groups:

RQ3b: Comparing discussions where moderators promoted constructive comments to discussions lacking this moderation strategy, does the discussion activity increase after the comments were promoted?

Discussion activity was analyzed based on two categories: the number of users commenting in a discussion and the absolute number of comments in a discussion over time. In contrast to discussion quality, I did find differences in activity between the control and featured groups. While activity levels remained consistent across the before subgroups, significant differences were found in the after subgroups. The natural decline in activity was less pronounced in the featured group. A greater number of users continued to participate in discussions where moderators intervened and featured constructive comments, leading to a larger overall comment count as well. Thus, I concluded that featuring constructive comments may be used to postpone the decline in discussion activity over time and to engage more users in active participation.

In sum, the effectiveness of featuring constructive comments depends on the desired outcome of the moderation strategy. My findings suggest that the quality of online discussions remained unchanged with the presence of featured comments, as perceived by both users and editorial staff. However, featuring constructive comments may delay the natural decline in user activity within online discussions, resulting in increased comment counts and greater engagement from unique users on the platform. Increased discussion activity may also be seen by news outlets as a marker of constructive interaction or by scholars wishing to understand constructive online interaction.

7.2. Answer to main research question

The research described in Chapters 2 through 6 resulted in a more complex understanding of promoting constructive commenting, as well as how computational tools may be used by news outlets to aid in the selection process. However, I did not find, and consciously refrained from attempting to create, a blueprint of constructive commenting. An indirect result of the investigations into constructive commenting is a nuanced appreciation of the subjective processing and context specificity behind evaluating a discussion in terms of constructive value. Given these novel insights, I attempt to formulate an answer to the main research question introduced in this thesis:

Main RQ: To what extent can computational models aid in the interpretation and identification of constructive comments by content moderators?

The term 'constructive' essentially serves as a broad descriptor for a situational assessment, characterized by various subjective perspectives. In one context, the moderator might decide that the story benefits from a personal anecdote. In another discussion, additional facts might improve the already published material. Moreover, time constraints might be a deciding factor in the selection of comments, limiting the number of comments evaluated by the moderator. The presence of previously selected constructive comments has an influence as well, as the moderator might look for specific argumentation to balance out the selection. It is within these interpretations that the notion of constructive commenting takes shape. Within the parameters of platform guidelines and discussion topics, moderators query an internal mental model of constructive comments by asking themselves such questions. After selecting a number of comments fitting the temporary description, the moderator proceeds to the next discussion, only to start the process anew. In all scenarios described above, a fine-tuned understanding of constructive commenting is needed – a multitude of contextual factors influence what a constructive comment looks like.

Promoting constructive comments remains a manual task performed by content moderators, while AI-based tools are employed to filter out toxicity. Given the fine-tuned understanding of constructive commenting presented earlier, how can AI-based tools contribute to the promotion of constructive interaction? If constructive value is configured in the interpretation of the moderator, computational tools ought to facilitate and streamline this process. Using either etic or emic perspectives on constructive commenting provides distinct tools to support the moderator. An etic perspective provides tools suited to a particular niche purpose. For example, argument mining models trained for specific discussion topics may efficiently provide the moderator with comments presenting a needed (counter)argument to balance out the selection of featured comments. The moderator decides a certain argument is needed and, instead of scrolling endlessly down the discussion looking for it, the model provides a set of comments belonging to that argument class. I have observed the difficulty and tediousness of finding specific minority arguments firsthand during fieldwork at *NU.nl*.

However, the moderator is not looking for a specific type of comment in every discussion. Approaches using an emic perspective provide a wider selection of recommended comments. Utilizing this perspective provides the moderator with comments that have been ranked based on patterns in past moderation decisions, for example by users with a history of writing featured contributions or comments which already received a great number of likes by other users. Evaluation of such an application has been performed by moderators, providing insights into how an implementation based on an emic perspective may contribute to online content moderation.

Ultimately, the success of any computational tool designed for promoting constructive commenting hinges on meeting the specific needs of content moderators. I argue that AI-based applications for constructive comments should not be evaluated on their ability to independently identify such comments. Instead, their effectiveness should be judged by their usefulness to the content moderators. To paraphrase the quote from a moderator of *NU.nl*, at the end of the day, they know a constructive comment when they see one.

Practically speaking, AI-based tools are needed to process the rapidly growing online discussions. Computational models, whether taking an etic of emic perspective, offer valuable insights into constructive commenting and have the potential to support moderators in streamlining their tasks. However, given the inherent subjectivity in assessing user comments for their constructive value, I conclude that these models fall short of providing a definitive classification.

7.3. Thesis contributions

The following section details the contributions made in this thesis.

1. I conducted a cross-platform analysis on the implementation of content moderation. I outlined the division of tasks between the human moderators and AI-based tools. Together, they operate in a hybrid setting.

2. I adopted the concept of the ‘Third Half of the Internet’, based on earlier work by Reagle (2015), to designate user-generated content curated by editorial staff to fit the mold of constructive commenting. Visually, this content is typically presented between the published editorial content (news articles) and other user-generated comments.

3. To expand the analysis of constructive participation on online news platforms, I formulated two distinct perspectives to study the concept. To achieve this, I borrowed the terms etic and emic from the field of anthropology. Both perspectives may be used for the study of online commenting and as dual starting points for the creation of computational models.

4. We introduced a novel approach specifically focused on argument diversity and

calculating interactivity found within online discussion threads.

5. Contributing to existing work on argument mining, we introduced a system suited to label online comments on the topic of climate change. Additionally, we presented an approach based on active learning to supplement training data with more examples belonging to the minority classes, which in turn improved the argument mining models.

6. We introduced the first work on ranking comments in online discussions in terms of constructive value to the discussion. The system is constructed as a supporting role to the content moderator. Based on ethnographic fieldwork conducted as part of the Better-MODS project, this system is tailored to the practical needs, data availability and platform specifics observed at Dutch news outlet *NU.nl*.

7. I presented a framework to investigate discussion quality from the perspective of both editorial staff and user base. I used this framework to assess the impact of featuring constructive comments on the discussion.

7.4. Future work

In this thesis I explored the dynamics of online comment sections and their moderation, portraying them as complex ecosystems where users engage with editorial staff, published content, and AI-based tools. Furthermore, practices are continuously evolving as a result of new technology. Consequently, I see multiple avenues for further research. In the following paragraphs, I discuss potential for further research in terms of cross-platform analysis, the impact of content moderation, realistic evaluation scenarios and the use of ethnographic fieldwork.

Chapter 2 introduced a cross-platform analysis. A framework comparing multiple online comment platforms allows for an expansion in terms of language. This thesis solely made us of Dutch language datasets derived from social media platforms and *NU.nl*. Such inquiries may shed light on whether the recommender described in Chapter 5 strictly modeled behavior of a subset of moderators employed at the same online platform or whether the implementation is universally applicable. One hurdle for platform-independent models is obtaining of a common set of input variables available for each platform.

Chapter 6 detailed the impact of featuring constructive comments on the discussion, providing insights into how the user base reacts to and is influenced by content moderation strategies. I see three potential extensions to this avenue of research. Firstly, a continuation of the impact on discussion activity is warranted. In Chapter 6, we concluded that discussion activity goes on for a longer time in the presence of highlighted constructive comments. However, it is unclear, on a single discussion basis, how much activity can be attributed to the featured comments. Such information cannot be derived from typical comment datasets due to the fact that they do not contain different scenarios in which the moderators did not intervene. For example, A/B testing could be used comprising a commenter group in which respondents interact with featured comments and a second group in which participants do not see discussions with featured comments. Additionally, research could model and predict discussion activity, comparing discussions containing featured comments to discussions without such promoted content.

Secondly, impact may be measured with the individual user as starting point. Does the commenting behavior of a user change after their comments have been deleted or featured? Are subsequent comments accepted or does the user refrain from commenting altogether? Some evidence indicates that users write comments of a higher quality after being featured (Yixue Wang and Diakopoulos 2022). However, the concept of comment quality and the selection of platforms need to be expanded.

A third and final research avenue in terms of impact evaluation is an extension of the data and methodologies to assess discussion quality. The investigation of discussion quality in Chapter 6 was based on past comment data, for instance observed behavior in the form of rejection rates and respect counts. Further research could make us of user ratings in terms of quality of discussions as a whole, comparing sets of comments posted in both control and featured discussions.

In Chapter 5, I attempted to create a realistic evaluation scenario of online content moderation in which discussions were assessed individually as opposed to one large, artificially created test set. Further research should include these forms of evaluations, where the real-life practice of content moderation is simulated to better understand how the AI-based tools fit in the hybrid moderation pipeline. Following this approach allows us to judge applications in terms of usefulness to real-life practices, as opposed to artificial testing. I see several building blocks for these types of evaluation. First, it is necessary to continue the evaluation on a discussion basis as presented in Chapter 5, maintaining the highly unbalanced nature of featured and non-featured comments. Featured comments happen to be caught in the moderator's eye; other comments that remain unselected could be equally good, yet they join the majority of unselected comments that would not be selected by a moderator Furthermore, the article topic may influence the distinct type of comment the moderator looks for, making the discussion the appropriate scope for evaluation. Second, using timestamps is vital to replicate the influence of timing when it comes to content moderation. Comments published at a time the discussion is fizzling out will not easily qualify to be featured. Taking into account the timing factor may inform further research why some comments, which may have qualified, were not featured. Third, news cycles need to be replicated. Discussion topics change rapidly over time, with topics flowing in and out the scope of news websites. AI-based tools need to be able to cope with topic changes, as well as other factors influenced by the article's subject, such as increased interaction on popular themes as politics. Consequently, test datasets may need to include different topics compared to the training data to probe the robustness against such changes. A final step towards realistic evaluation scenarios is the inclusion of expert evaluation. Chapter 5 introduces such realistic testing by content moderators currently employed at *NUjij*, with promising results. Given that there is no universal blueprint of a constructive comment, an expert evaluation provides insight into the subjective and contextual choices moderators make and in how far the AI-based tools support their decision-making processes.

The final research avenue I see is the inclusion of ethnographic fieldwork to achieve a deeper understanding of how content moderators operate. Additionally, it provides insights into how moderation strategies are influenced by moderators interacting with individual users and their personal experiences on online platforms. In this thesis, I made extensive use of insights I gained as a result of visiting the editorial offices of *NU.nl*. I was able to cater to the needs of the moderators by identifying which of their tasks are translatable to computational support. Furthermore, ethnographic fieldwork informs data requirements and availability for realistic evaluation scenarios and provides a clear picture of the user interface with which moderators interact on a daily basis, contributing to a nuanced understanding of everyday content moderation practices. Further work should make use of fieldwork to establish a greater link between AI-based moderation tools and the lived experience of content moderators employed at online platforms.

References

Abramowitz, Alan I. and Kyle L. Saunders (2008). ‘Is polarization a myth?’ In: *Journal of Politics* 70.2, pp. 542–555. DOI: 10.1017/S0022381608080493.

Addawood, Aseel and Masooda Bashir (2016). ‘"What Is Your Evidence?" A Study of Controversial Topics on Social Media’. In: *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pp. 1–11. DOI: 10.18653/v1/w16-2801.

AlDayel, Abeer and Walid Magdy (2021). ‘Stance detection on social media: State of the art and trends’. In: *Information Processing and Management* 58.4, p. 102597. DOI: 10.1016/j.ipm.2021.102597.

Allgaier, Joachim (2019). ‘Science and Environmental Communication on YouTube: Strategically Distorted Communications in Online Videos on Climate Change and Climate Engineering’. In: *Frontiers in Communication* 4.July, pp. 1–15. DOI: 10.3389/fcomm.2019.00036.

Andersen, Vibeke Normann and Kasper M Hansen (2007). ‘How deliberation makes better citizens: The Danish Deliberative pool’. In: *European Journal of Political Research* 46, pp. 531–556.

Annamoradnejad, Issa, Jafar Habibi and Mohammadamin Fazli (2022). ‘Multi-view approach to suggest moderation actions in community question answering sites’. In: *Information Sciences* 600, pp. 144–154. DOI: 10.1016/j.ins.2022.03.085.

Balkenhol, Markus (2015). ‘Zwarte Piet, racisme en emoties’. In: *Waardenwerk* 62/63, pp. 36–46.

Behrendt, Maike, Stefan Sylvius Wagner, Marc Ziegele, Lena Wilms, Anke Stoll, Dominique Heinbach and Stefan Harmeling (2024). ‘AQuA – Combining Experts’ and Non-Experts’ Views To Assess Deliberation Quality in Online Discussions Using LLMs’. In: arXiv: 2404.02761. URL: http://arxiv.org/abs/2404.02761.

Berresheim, Simon and Julia Meyer (2023). *Wir haben einen neuen Kommentarbereich.* https://www.zeit.de/administratives/2023-04/kommentarbereich-design-struktur-emojis. (Visited on 13/06/2023).

Binns, Reuben, Michael Veale, Max Van Kleek and Nigel Shadbolt (2017). ‘Like trainer, like bot? Inheritance of bias in algorithmic content moderation’. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 10540 LNCS, pp. 405–415. DOI: 10.1007/978-3-319-67256-4_32.

Bord, Richard J., Robert E. O’Connor and Ann Fisher (2000). ‘In what sense does the public need to understand global climate change?’ In: *Public Understanding of Science* 9.3, pp. 205–218. DOI: 10.1088/0963-6625/9/3/301.

Bosc, Tom, Elena Cabrio and Serena Villata (2016). ‘Tweeties Squabbling : Positive and Negative Results in Applying Argument Mining on Social Media’. In: *Computational Models of Argument* 0, pp. 21–32. DOI: 10.3233/978-1-61499-686-6-21.

Brunk, Jens, Jana Mattern and Dennis M. Riehle (2019). 'Effect of transparency and trust on acceptance of automatic online comment moderation systems'. In: *Proceedings - 21st IEEE Conference on Business Informatics, CBI 2019* 1, pp. 429–435. DOI: 10.1109/CBI.2019.00056.

Bruns, Axel (2017). 'Echo Chamber? What Echo Chamber? Reviewing the Evidence'. In: *Future of Journalism 2017*, pp. 1–11.

Brysse, Keynyn, Naomi Oreskes, Jessica O'Reilly and Michael Oppenheimer (2013). 'Climate change prediction: Erring on the side of least drama?' In: *Global Environmental Change* 23.1, pp. 327–337. DOI: 10.1016/j.gloenvcha.2012.10.008.

Carstens, Lucas and Francesca Toni (June 2015). 'Towards relation based Argumentation Mining'. In: *Proceedings of the 2nd Workshop on Argumentation Mining*. Ed. by Claire Cardie. Denver, CO: Association for Computational Linguistics, pp. 29–34. DOI: 10.3115/v1/W15-0504.

Colleoni, Elanor, Alessandro Rozza and Adam Arvidsson (2014). 'Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data'. In: *Journal of Communication* 64.2, pp. 317–332. ISSN: 14602466. DOI: 10.1111/jcom.12084.

Danka, Tivadar and Peter Horvath (n.d.). *modAL: A modular active learning framework for Python*. available on arXiv at https://arxiv.org/abs/1805.00979. URL: https://github.com/cosmic-cortex/modAL.

Delgado, Pablo (2019). 'How El País used AI to make their comments section less toxic'. In: *Google News Initiative*. URL: https://blog.google/outreach-initiatives/google-news-initiative/how-el-pais-used-ai-make-their-comments-section-less-toxic/.

Delobelle, Pieter, Thomas Winters and Bettina Berendt (2020). 'RobBERT: A Dutch RoBERTa based language model'. In: *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*, pp. 3255–3265. DOI: 10.18653/v1/2020.findings-emnlp.292.

Diakopoulos, Nicholas (2015a). 'Picking the NYT Picks : Editorial Criteria and Automation in the Curation of Online News Comments'. In: *#ISOJ, the official research journal of ISOJ* 5.1, pp. 147–166. ISSN: 2328-0662.

– (2015b). 'The editor's eye: Curation and comment relevance on the New York Times'. In: *CSCW 2015 - Proceedings of the 2015 ACM International Conference on Computer-Supported Cooperative Work and Social Computing*, pp. 1153–1157. DOI: 10.1145/2675133.2675160.

– (2019). *How Algorithms Are Rewriting the Media*. Cambridge, MA and London, England: Harvard University Press. ISBN: 9780674239302. DOI: 10.4159/9780674239302.

Domingo, David, Thorsten Quandt, Ari Heinonen, Steve Paulussen, Jane B. Singer and Marina Vujnovic (2008). 'Participatory journalism practices in the media and beyond: An international comparative study of initiatives in online newspapers'. In: *Journalism Practice* 2.3, pp. 326–342. DOI: 10.1080/17512780802281065.

Doran, Peter T. and Maggie Kendall Zimmerman (2009). 'Examining the scientific consensus on climate change'. In: *Eos* 90.3, pp. 22–23. DOI: 10.1029/2009EO030002.

Du, Siying and Steve Gregory (2017). 'The Echo Chamber Effect in Twitter: does community polarization increase?' In: *Studies in Computational Intelligence* 693, pp. 5–7. DOI: 10.1007/978-3-319-50901-3.

Du Bois, John W. (2007). 'The stance triangle'. In: *Stancetaking in Discourse*, pp. 139–182. DOI: 10.1075/pbns.164.07du.

Dubois, Elizabeth and Grant Blank (2018). 'The echo chamber is overstated: the moderating effect of political interest and diverse media'. In: *Information Communication and Society* 21.5, pp. 729–745. DOI: 10.1080/1369118X.2018.1428656.

Dunlap, Riley E. and Aaron M. McCright (2012). 'Organized Climate Change Denial'. In: *The Oxford Handbook of Climate Change and Society* January 2012. DOI: 10.1093/oxfordhb/9780199566600.003.0010.

El Pais (2015). *EL PAÍS mejora el sistema de comentarios en sus noticias.* URL: https://elpais.com/elpais/2015/03/24/actualidad/1427229587%7B%5C_%7D101365.html.

– (2016). *Principios y normas de participación.* URL: %5Curl%7Bhttps://elpais.com/estaticos/normas-de-participacion/%7D.

– (2018). *Inteligencia artificial para elevar la calidad del debate digital.* URL: https://elpais.com/sociedad/2018/12/17/actualidad/1545081231%7B%5C_%7D439667.html.

Falk, Neele and Gabriella Lapesa (May 2023). 'Bridging Argument Quality and Deliberative Quality Annotations with Adapters'. In: *Findings of the Association for Computational Linguistics: EACL 2023.* Ed. by Andreas Vlachos and Isabelle Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 2469–2488. DOI: 10.18653/v1/2023.findings-eacl.187.

Flaxman, Seth, Sharad Goel and Justin M. Rao (2016). 'Filter bubbles, echo chambers, and online news consumption'. In: *Public Opinion Quarterly* 80.Specialissue1, pp. 298–320. DOI: 10.1093/poq/nfw006.

Friess, Dennis and Christiane Eilders (2015). 'A systematic review of online deliberation research'. In: *Policy and Internet* 7.3, pp. 319–339. DOI: 10.1002/poi3.95.

Garimella, Kiran, Gianmarco De Francisci Morales, Aristides Gionis and Michael Mathioudakis (2018). 'Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship'. In: *WWW '18: Proceedings of the 2018 World Wide Web Conference*, pp. 913–922. DOI: 10.1145/3178876.3186139.

Gillespie, Tarleton (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press. DOI: 10.12987/9780300235029.

– (2020). 'Content moderation, AI, and the question of scale'. In: *Big Data and Society* 7.2, pp. 1–5. DOI: 10.1177/2053951720943234.

– (2022). 'Do Not Recommend? Reduction as a Form of Content Moderation'. In: *Social Media and Society* 8.3, pp. 1–13. DOI: 10.1177/20563051221117552.

Gillespie, Tarleton, Patricia Aufderheide, Elinor Carmi, Ysabel Gerrard, Robert Gorwa, Ariadna Matamoros-Fernández, Sarah T. Roberts, Aram Sinnreich and Sarah Myers West (2020). 'Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates'. In: *Internet Policy Review* 9.4, pp. 1–29. DOI: 10.14763/2020.4.1512.

Goldberg, Jeffrey (2018). 'We Want to Hear from You'. In: *The Atlantic.* URL: https://www.theatlantic.com/letters/archive/2018/02/we-want-to-hear-from-you/552170/.

8

Gollatz, Kirsten, Martin Johannes Riedl and Jens Pohlmann (2018). 'Removals of online hate speech in numbers'. In: *HIIG Science Blog* August 2018. DOI: 10.5281/zenodo.1342324. URL: https://www.hiig.de/en/removals-of-online-hate-speech-numbers/.

Gompel, Maarten van and Antal van den Bosch (2016). 'Efficient n-gram, Skipgram and Flexgram Modelling with Colibri Core'. In: *Journal of Open Research Software* 4. DOI: 10.5334/jors.105.

Gorwa, Robert, Reuben Binns and Christian Katzenbach (2020). 'Algorithmic content moderation: Technical and political challenges in the automation of platform governance'. In: *Big Data and Society* 7.1. DOI: 10.1177/2053951719897945.

Grönlund, Kimmo, Kaisa Herne and Maija Setälä (2015). 'Does Enclave Deliberation Polarize Opinions?' In: *Political Behavior* 37.4, pp. 995–1020. DOI: 10.1007/s11109-015-9304-x.

Heinbach, Dominique, Lena Wilms and Marc Ziegele (2022). 'Effects of empowerment moderation in online discussions : A field experiment with four news outlets'. In: *72nd Annual Conference of the International Communication Association (ICA), 26-30 May 2022, Paris, France*. May.

Helsloot, John (2012). 'Zwarte Piet and Cultural Aphasia in the Netherlands'. In: *Quotidian: journal for the study of everyday life* 3, pp. 1–20. ISSN: 1879-534X.

– (2013). 'Contesting Ambiguity : the black peter mask in Dutch'. In: *The Power of the Mask*, pp. 124–132.

Hoekman, Gert-Jaap (2016). 'NU.nl stopt met open reacties onder artikelen'. In: *NU.nl*. URL: https://www.nu.nl/blog/4305300/nunl-stopt-met-open-reacties-artikelen.html.

Jamieson, Kathleen and Joseph Capella (2008). *Echo Chamber: Rush Limbaugh and the Conservative Media Establishment*. Oxford University Press. ISBN: 0195398602.

Jarvelin, Kalervo and Jaana Kekalainen (2000). 'IR evaluation methods for retrieving highly relevant documents'. In: *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, pp. 41–48. DOI: 10.1145/3130348.3130374.

Jenkins-Smith, Hank C., Joseph T. Ripberger, Carol L. Silva, Deven E. Carlson, Kuhika Gupta, Nina Carlson, Ani Ter-Mkrtchyan and Riley E. Dunlap (2020). 'Partisan asymmetry in temporal stability of climate change beliefs'. In: *Nature Climate Change* 10.4, pp. 322–328. DOI: 10.1038/s41558-020-0719-y.

Jiang, Jialun Aaron, Peipei Nie, Jed R. Brubaker and Casey Fiesler (2023). 'A Trade-off-centered Framework of Content Moderation'. In: *ACM Transactions on Computer-Human Interaction* 30.1. DOI: 10.1145/3534929.

Jiang, Xiaoya, Min Hsin Su, Juwon Hwang, Ruixue Lian, Markus Brauer, Sunghak Kim and Dhavan Shah (2021). 'Polarization Over Vaccination: Ideological Differences in Twitter Expression About COVID-19 Vaccine Favorability and Specific Hesitancy Concerns'. In: *Social Media and Society* 7.3. DOI: 10.1177/20563051211048413.

Jonsson, Magnus E. and Joachim Åström (2014). 'The Challenges for Online Deliberation Research'. In: *International Journal of E-Politics* 5.1, pp. 1–15. DOI: 10.4018/ijep.2014010101.

Kolhatkar, Varada and Maite Taboada (2017). 'Constructive language in news comments'. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 11–17. DOI: 10.18653/v1/w17-3002.

Kolhatkar, Varada, Nithum Thain, Jeffrey Sorensen, Lucas Dixon and Maite Taboada (2023). 'Classifying constructive comments'. In: *First Monday* 28.4, pp. 1–16. DOI: https://doi.org/10.5210/fm.v28i4.13163.

Ksiazek, Thomas B., Limor Peer and Kevin Lessard (2016). 'User engagement with online news: Conceptualizing interactivity and exploring the relationship between online news videos and user comments'. In: *New Media and Society* 18.3, pp. 502–520. DOI: 10.1177/1461444814545073.

Küçük, Dilek and C. A.N. Fazli (2020). 'Stance detection: A survey'. In: *ACM Computing Surveys* 53.1. DOI: 10.1145/3369026.

Kunneman, Florian, Mattijs Lambooij, Albert Wong, Antal Van Den Bosch and Liesbeth Mollema (2020). 'Monitoring stance towards vaccination in twitter messages'. In: *BMC Medical Informatics and Decision Making* 20.1, pp. 1–14. DOI: 10.1186/s12911-020-1046-y.

Lai, Vivian, Samuel Carton, Rajat Bhatnagar, Q. Vera Liao, Yunfeng Zhang and Chenhao Tan (2022). 'Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation'. In: *Conference on Human Factors in Computing Systems - Proceedings*. DOI: 10.1145/3491102.3501999.

Lawrence, John and Chris Reed (2019). 'Argument mining: A survey'. In: *Computational Linguistics* 45.4, pp. 765–818. DOI: 10.1162/COLIa00364.

– (Jan. 2020). 'Argument Mining: A Survey'. In: *Computational Linguistics* 45.4, pp. 765–818. ISSN: 0891-2017. DOI: 10.1162/coli_a_00364. eprint: https://direct.mit.edu/coli/article-pdf/45/4/765/1847520/coli_a_00364.pdf. URL: https://doi.org/10.1162/coli%5C_a%5C_00364.

Leiserowitz, Anthony (2007). 'Fighting climate change: Human solidarity in a divided world International Public Opinion, Perception, and Understanding of Global Climate Change'. In: *Yale Program on Climate Change Communication*, pp. 1–40.

Leiserowitz, Anthony, Edward W. Maibach, Connie Roser-Renouf, Geoff Feinberg and Peter Howe (2015). 'Climate Change in the American Mind: Americans' Global Warming Beliefs and Attitudes in April 2013'. In: *SSRN Electronic Journal*. June 2015, pp. 1–9. DOI: 10.2139/ssrn.2298705.

Lewandowsky, Stephan, Ullrich K.H. Ecker and John Cook (2017). 'Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era'. In: *Journal of Applied Research in Memory and Cognition* 6.4, pp. 353–369. ISSN: 22113681. DOI: 10.1016/j.jarmac.2017.07.008.

Li, Haoyan and Michael Chau (2023). 'Human-AI Collaboration in Content Moderation: The Effects of Information Cues and Time Constraints'. In: *ECIS 2023 Research-in-Progress Papers*. URL: https://aisel.aisnet.org/ecis2023%5C_rip/2.

Lima, Lucas, Julio C.S. Reis, Philipe Melo, Fabricio Murai, Leandro Araujo, Pantelis Vikatos and Fabricio Benevenuto (2018). 'Inside the right-leaning echo chambers: Characterizing gab, an unmoderated social system'. In: *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018*, pp. 515–522. DOI: 10.1109/ASONAM.2018.8508809.

Linden, Liesje van der, Cedric Waterschoot, Florian Kunneman, Ernst van den Hemel, Antal van den Bosch and Emiel Krahmer (Forthcoming). 'Who Are the Online Commenters: A large-scale representative survey to explore the identity of online commenters '. In.

Linden, Sander van der, Anthony Leiserowitz, Seth Rosenthal and Edward Maibach (2017). 'Inoculating the Public against Misinformation about Climate Change'. In: *Global Challenges* 1.2. DOI: 10.1002/gch2.201600008.

Loos, Andreas (2016). 'Mein Bot und Ich'. In: *Die Zeit Online*. URL: https://www.zeit.de/digital/2016-09/kuenstliche-intelligenz-kommentar-bot-zeit.

Loshchilov, Ilya and Frank Hutter (2019). 'Decoupled Weight Decay Regularization'. In: *ICLR 2019*. arXiv: arXiv:1711.05101v3.

Løvlie, Anders Sundnes (2018). 'Constructive Comments?' In: *Journalism Practice* 12.6, pp. 781–798. DOI: 10.1080/17512786.2018.1473042.

Manosevitch, Idit and Ori Tenenboim (2017). 'The Multifaceted Role of User-Generated Content in News Websites'. In: *Digital Journalism* 5.6, pp. 731–752. DOI: 10.1080/21670811.2016.1189840.

Margaret, Tan (1994). 'Establishing Mutual Understanding in Systems Design: An Empirical Study'. In: *Journal of Management Information Systems* 10.4, pp. 159–182. DOI: 10.1080/07421222.1994.11518024.

Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Pean, Berger S., N. Caud, Y. Chen, L. Goldfarb, M.i. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekci, R. Yu and B. Zhou (2021). *IPCC, 2021: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Tech. rep.

McCright, Aaron M., Meghan Charters, Katherine Dentzman and Thomas Dietz (2016). 'Examining the Effectiveness of Climate Change Frames in the Face of a Climate Change Denial Counter-Frame'. In: *Topics in Cognitive Science* 8.1, pp. 76–97. ISSN: 17568765. DOI: 10.1111/tops.12171.

McCright, Aaron M. and Riley E. Dunlap (2003). 'Defeating Kyoto: The Conservative Movement's Impact on U.S. Climate Change Policy'. In: *Social Problems* 50.3, pp. 348–373. DOI: 10.1525/sp.2003.50.3.348.

Meier, Klaus, Daniela Kraus and Edith Michaeler (2018). 'Audience Engagement in a Post-Truth Age: What it means and how to learn the activities connected with it'. In: *Digital Journalism* 6.8, pp. 1052–1063. DOI: 10.1080/21670811.2018.1498295.

Mills, Richard A. (2018). 'Pop-up political advocacy communities on reddit.com: SandersForPresident and The Donald'. In: *AI and Society* 33.1, pp. 39–54. DOI: 10.1007/s00146-017-0712-9.

Moens, Marie Francine (2018). 'Argumentation mining: How can a machine acquire common sense and world knowledge?' In: *Argument and Computation* 9.1, pp. 1–4. DOI: 10.3233/AAC-170025.

Molina, Maria D. and S. Shyam Sundar (2022). 'When AI moderates online content: Effects of human collaboration and interactive transparency on user trust'. In: *Journal of Computer-Mediated Communication* 27.4. DOI: 10.1093/jcmc/zmac010.

Morrow, Garrett, Briony Swire-Thompson, Jessica Montgomery Polny, Matthew Kopec and John P. Wihbey (2022). 'The emerging science of content labeling: Contextualizing social media content moderation'. In: *Journal of the Association for Information Science and Technology* 73.10, pp. 1365–1386. DOI: 10.1002/asi.24637.

Mostowlansky, Till and Andrea Rota (2020). 'Emic and Etic'. In: *The Cambridge Encyclopedia of Anthropology*, pp. 1–16. DOI: http://doi.org/10.29164/20emicetic.

Mutz, Diana and Jeffrey Mondak (2006). 'The workplace as a context for cross-cutting political discourse'. In: *The Handbook of Discourse Analysis* 68.1, pp. 398–415. DOI: 10.1002/9780470753460.ch21.

Naderi, Nona and Graeme Hirst (2016). 'Argumentation mining in parliamentary discourse'. In: *Lecture Notes in Computer Science* 9935 LNAI, pp. 16–25. DOI: 10.1007/978-3-319-46218-9_2.

Napoles, Courtney, Joel Tetreault, Enrica Rosato, Brian Provenzale and Aasish Pappu (2017). 'Finding good conversations online: The yahoo news annotated comments corpus'. In: *LAW 2017 - 11th Linguistic Annotation Workshop, Proceedings of the Workshop*, pp. 13–23. DOI: 10.18653/v1/w17-0802.

New York Times (2016). *The Times is Partnering with Jigsaw to Expand Comment Capablities*. URL: https://www.nytco.com/press/the-times-is-partnering-withjigsaw-to-expand-comment-capabilities/.

– (2017). *The Times Expands Comments to All Top Stories*. URL: https://www.nytco.com/press/the-times-expands-comments-to-all-top-stories/.

– (2020). *Comment FAQ*. URL: https://help.nytimes.com/hc/en-us/articles/115014792387-The-Comments-Section.

NU.nl (2020). 'NUjij laat met expertlabels de kennis en expertise van gebruikers zien'. In: *NU.nl*. URL: https://www.nu.nl/nulab/6093189/nujij-laat-met-expertlabels-de-kennis-en-expertise-van-gebruikers-zien.html.

NUJij (2018). 'NUjij - Veelgestelde vragen'. In: *NU.nl*. URL: https://www.nu.nl/nujij/5215910/nujij-veelgestelde-vragen.html.

Ogolla, Shirley and Vivien Hard (2020). *Examples of Artificial Intelligence in Business Practice*. URL: https://eu2020-reader.bmas.de/en/new-work-human-centric-work/examples-of-artificial-intelligence-in-business-practice/ (visited on 07/07/2021).

Opitz, Juri and Anette Frank (June 2019). 'An Argument-Marker Model for Syntax-Agnostic Proto-Role Labeling'. In: *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 224–234. DOI: 10.18653/v1/S19-1025.

Oreskes, Naomi (2005). 'Scientific consensus on Climate Change'. In: *Science* 306, pp. 2004–2005. DOI: 10.1126/science.1103618. URL: http://linkinghub.elsevier.com/retrieve/pii/000632079400057W.

Paasch-Colberg, Sünje and Christian Strippel (2022). '"The Boundaries are Blurry…": How Comment Moderators in Germany See and Respond to Hate Comments'. In: *Journalism Studies* 23.2, pp. 224–244. DOI: 10.1080/1461670X.2021.2017793.

Pariser, Eli (2011). *The Filter Bubble: What the Internet is hiding from you*. Penguin Press. ISBN: 1594203008.

Park, Deokgun, Simranjit Sachar, Nicholas Diakopoulos and Niklas Elmqvist (2016). 'Supporting comment moderators in identifying high quality online news comments'. In: *Conference on Human Factors in Computing Systems - Proceedings*, pp. 1114–1125. DOI: 10.1145/2858036.2858389.

Paßmann, Johannes, Helmond Helmond and Robert Jansma (2023). 'From healthy communities to toxic debates: Disqus' changing ideas about comment moderation'. In: *Internet Histories* 7.1, pp. 6–26. DOI: 10.1080/24701475.2022.2105123.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay (2011). 'Scikit-learn: Machine Learning in Python'. In: *Journal of Machine Learning Research* 12, pp. 2825–2830. DOI: 10.1289/EHP4713.

Pike, Kenneth L. (1967). 'Etic and Emic Standpoints for the Description of Behavior'. In: *Language and Thought: An Enduring Problem in Psychology*. Ed. by Donald C. Hildum. : Van Nostrand, pp. 32–39.

Poortinga, Wouter, Alexa Spence, Lorraine Whitmarsh, Stuart Capstick and Nick F. Pidgeon (2011). 'Uncertain climate: An investigation into public scepticism about anthropogenic climate change'. In: *Global Environmental Change* 21.3, pp. 1015–1024. DOI: 10.1016/j.gloenvcha.2011.03.001.

Quandt, Thorsten (2018). 'Dark participation'. In: *Media and Communication* 6.4, pp. 36–48. DOI: 10.17645/mac.v6i4.1519.

– (2023). 'Euphoria, disillusionment and fear: Twenty-five years of digital journalism (research)'. In: *Journalism* 0.0, pp. 1–18. DOI: 10.1177/14648849231192789.

Quandt, Thorsten, Johanna Klapproth and Lena Frischlich (2022). 'Dark social media participation and well-being'. In: *Current Opinion in Psychology* 45, p. 101284. DOI: 10.1016/j.copsyc.2021.11.004.

Rahmstorf, Stephan (2004). 'The climate sceptics'. In: *Weather catastrophes and climate change - Is there still hope for us?*, pp. 76–83.

Rayson, Paul and Roger Garside (2000). 'Comparing Corpora using Frequency Profiling'. In: *The Workshop on Comparing Corpora*, pp. 1–6. DOI: 10.3115/1117729.1117730.

Raza, Shaina and Chen Ding (2022). *News recommender system: a review of recent progress, challenges, and opportunities*. Vol. 55. 1. Springer Netherlands, pp. 749–800. DOI: 10.1007/s10462-021-10043-x.

Raza, Shaina, Muskan Garg, Deepak John Reji, Syed Raza Bashir and Chen Ding (2024). 'NBIAS: A natural language processing framework for BIAS identification in text'. In: *Expert Systems with Applications* 237.PB, p. 121542. ISSN: 09574174. DOI: 10.1016/j.eswa.2023.121542.

Reagle, J M (2015). *Reading the Comments: Likers, Haters, and Manipulators at the Bottom of the Web*. Business book summary. MIT Press. ISBN: 9780262028936.

Reimers, Nils and Iryna Gurevych (2020). 'Sentence-BERT: Sentence embeddings using siamese BERT-networks'. In: *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 3982–3992. DOI: 10.18653/v1/d19-1410. arXiv: 1908.10084.

Reimers, Nils, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab and Iryna Gurevych (2020). 'Classification and clustering of arguments with contextualized word embeddings'. In: *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 567–578. DOI: 10.18653/v1/p19-1054.

Rieder, Bernhard and Yarden Skop (2021). 'The fabrics of machine moderation: Studying the technical, normative, and organizational structure of Perspective API'. In: *Big Data & Society* 8.2. DOI: 10.1177/20539517211046181.

Roberts, Sarah T. (2017). ‘Content Moderation’. In: *Encyclopedia of Big Data*, pp. 1–4. DOI: 10.1201/9781003293125-7.

Ruckenstein, Minna and Linda Lisa Maria Turunen (2020). ‘Re-humanizing the platform: Content moderators and the logic of care’. In: *New Media and Society* 22.6, pp. 1026–1042. DOI: 10.1177/1461444819875990.

Salganik, Matthew and Robin Lee (2020). ‘To Apply Machine Learning Responsibly, We Use It in Moderation’. In: *New York Times*. URL: https://open.nytimes.com/to-apply-machine-learning-responsibly-we-use-it-in-moderation-d001f49e0644.

Samory, Mattia and Tanushree Mitra (2018). ‘’The Government Spies Using Our Webcams’’. In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW, pp. 1–24. DOI: 10.1145/3274421.

Scheuermann, Larry and Gary Taylor (1997). ‘Netiquette’. In: *Internet Research* 7.4.

Schmidt, Ana Lucía, Fabiana Zollo, Antonio Scala, Cornelia Betsch and Walter Quattrociocchi (2018). ‘Polarization of the vaccination debate on Facebook’. In: *Vaccine* 36.25, pp. 3606–3612. ISSN: 18732518. DOI: 10.1016/j.vaccine.2018.05.040.

Schmidt, David (2014). ‘Bitte beklatschen Sie mich’. In: *Die Zeit*. URL: https://www.zeit.de/community/2014-04/leserempfehlungen-funktion.

Schols, Heleen (2020). *Keeping things gezellig: Negotiating Dutchness and racism in the struggle over ‘Black Pete’*. ISBN: 9789463754828.

Shin, Bokyong and Mikko Rask (2021). ‘Assessment of online deliberative quality: New indicators using network analysis and time-series analysis’. In: *Sustainability* 13.3, pp. 1–21. ISSN: 20711050. DOI: 10.3390/su13031187.

Shmargad, Yotam and Samara Klar (2020). ‘Sorting the News: How Ranking by Popularity Polarizes Our Politics’. In: *Political Communication* 37.3, pp. 423–446. DOI: 10.1080/10584609.2020.1713267.

Spohr, Dominic (2017). ‘Fake news and ideological polarization: Filter bubbles and selective exposure on social media’. In: *Business Information Review* 34.3, pp. 150–160. DOI: 10.1177/0266382117722446.

Stab, Christian, Tristan Miller and Iryna Gurevych (2018). ‘Cross-topic Argument Mining from Heterogeneous Sources Using Attention-based Neural Networks’. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3664–3674. DOI: 10.18653/v1/D18-1402.

Strandberg, Kim, Staffan Himmelroos and Kimmo Grönlund (2017). ‘Do discussions in like-minded groups necessarily lead to more extreme opinions? Deliberative democracy and group polarization’. In: *International Political Science Review* 40.1, pp. 41–57. DOI: 10.1177/0192512117692136.

Stroud, Natalie Jomini, Ashley Muddiman and Joshua M. Scacco (2017). ‘Like, recommend, or respect? Altering political behavior in news comment sections’. In: *New Media and Society* 19.11, pp. 1727–1743. DOI: 10.1177/1461444816642420.

Suiter, Jane, David M. Farrell and Eoin O’Malley (2016). ‘When do deliberative citizens change their opinions? Evidence from the Irish Citizens’ Assembly’. In: *International Political Science Review* 37.2, pp. 198–212. DOI: 10.1177/0192512114544068.

Sunstein, Cass R. and Adrian Vermeule (2009). ‘Symposium on conspiracy theories: Conspiracy theories: Causes and cures’. In: *Journal of Political Philosophy* 17.2, pp. 202–227. DOI: 10.1111/j.1467-9760.2008.00325.x.

Tandoc, Edson C., Zheng Wei Lim and Richard Ling (2018). 'Defining "Fake News": A typology of scholarly definitions'. In: *Digital Journalism* 6.2, pp. 137–153. DOI: 10.1080/21670811.2017.1360143.

The Guardian (2009). *Frequently asked questions about community on the Guardian website.* URL: https://www.theguardian.com/community-faqs.

The Guardian News & Media (2021). *How we moderate comments on our site.* URL: https://committees.parliament.uk/writtenevidence/25757/pdf/.

Toulmin, Stephen (2003). *The Uses of Argument.* Cambridge University Press. ISBN: 0521534836.

Utopia (2021). *NU.nl first considered developing the AI tech by themselves.* URL: https://utopiaanalytics.com/case/case-nu-nl/ (visited on 15/09/2021).

Van Hoek, Colin (2020). 'Hoe NU.nl beter wordt van een robot'. In: *NU.nl.* URL: %7Bhttps://www.nu.nl/blog/6045082/hoe-nunl-beter-wordt-van-een-robot.html%7D.

van Linden, Sander L., Anthony A. Leiserowitz, Geoffrey D. Feinberg and Edward W. Maibach (2015). 'The scientific consensus on climate change as a gateway belief: Experimental evidence'. In: *PLoS ONE* 10.2, pp. 2–9. DOI: 10.1371/journal.pone.0118489.

Wang, Sai (2021). 'Moderating Uncivil User Comments by Humans or Machines? The Effects of Moderation Agent on Perceptions of Bias and Credibility in News Content'. In: *Digital Journalism* 9.1, pp. 64–83. DOI: 10.1080/21670811.2020.1851279.

Wang, Yining, Liwei Wang, Yuanzhi Li, Di He and Tie-Yan Liu (2013). 'A Theoretical Analysis of NDCG Type Ranking Measures'. In: *Proceedings of the 26th Annual Conference on Learning Theory.* Ed. by Shai Shalev-Shwartz and Ingo Steinwart. Vol. 30. Proceedings of Machine Learning Research. Princeton, NJ, USA: PMLR, pp. 25–54. URL: https://proceedings.mlr.press/v30/Wang13.html.

Wang, Yixue and Nicholas Diakopoulos (2022). 'Highlighting High-quality Content as a Moderation Strategy : The Role of New York Times Picks in Comment Quality'. In: *ACM Transactions on Social Computing* 4.4, pp. 1–24. DOI: 10.1145/3484245.

Waterschoot, Cedric, Ernst van den Hemel and Antal van den Bosch (Oct. 2022). 'Detecting Minority Arguments for Mutual Understanding: A Moderation Tool for the Online Climate Change Debate'. In: *Proceedings of the 29th International Conference on Computational Linguistics.* Gyeongju, Republic of Korea: International Committee on Computational Linguistics, pp. 6715–6725. URL: https://aclanthology.org/2022.coling-1.583.

Whitmarsh, Lorraine (2011). 'Scepticism and uncertainty about climate change: Dimensions, determinants and change over time'. In: *Global Environmental Change* 21.2, pp. 690–700. DOI: 10.1016/j.gloenvcha.2011.01.016.

Williams, Hywel T.P., James R. McMurray, Tim Kurz and F. Hugo Lambert (2015a). 'Network analysis reveals open forums and echo chambers in social media discussions of climate change'. In: *Global Environmental Change* 32, pp. 126–138. DOI: 10.1016/j.gloenvcha.2015.03.006.

– (2015b). 'Network analysis reveals open forums and echo chambers in social media discussions of climate change'. In: *Global Environmental Change* 32, pp. 126–138. DOI: 10.1016/j.gloenvcha.2015.03.006.

Wintterlin, Florian, Tim Schatto-Eckrodt, Lena Frischlich, Svenja Boberg and Thorsten Quandt (2020). 'How to Cope with Dark Participation: Moderation Practices in German Newsrooms'. In: *Digital Journalism* 8.7, pp. 904–924. DOI: 10.1080/21670811.2020.1797519.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest and Alexander Rush (2020). 'Transformers: State-of-the-Art Natural Language Processing'. In: pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6.

Wolfgang, J. David (2018). 'Cleaning up the "Fetid Swamp": Examining how journalists construct policies and practices for moderating comments'. In: *Digital Journalism* 6.1, pp. 21–40. DOI: 10.1080/21670811.2017.1343090.

Wright, Scott and John Street (2007). 'Democracy, Deliberation and Design: the case of online discussion forums'. In: *New Media & Society* 9.5, pp. 849–869. DOI: 10.1177/1461444807081230.

Yarnoz, Carlos (2019). *El lector gana protagonismo*. URL: https://elpais.com/elpais/2019/10/26/opinion/1572078022%7B%5C_%7D285897.html.

Zannettou, Savvas, Barry Bradlyn and Emiliano De Cristofaro (2018). 'What is Gab ? A Bastion of Free Speech or an Alt-Right Echo Chamber ?' In: *WWW '18: Companion Proceedings of the The Web Conference 2018*, pp. 1007–1014. DOI: 10.1145/3184558.3191531.

Zareie, Ahmad and Rizos Sakellariou (2021). 'Minimizing the spread of misinformation in online social networks: A survey'. In: *Journal of Network and Computer Applications* 186.May, p. 103094. DOI: 10.1016/j.jnca.2021.103094.

Zhao, Yue, Ciwen Xu and Yongcun Cao (2006). 'Research on Query-by-Committee Method of Active'. In: *Lecture Notes in Computer Science*, pp. 985–991. DOI: doi.org/10.1007/11811305_107.

Ziegele, Marc, Oliver Quiring, Katharina Esau and Dennis Friess (2020). 'Linking News Value Theory With Online Deliberation: How News Factors and Illustration Factors in News Articles Affect the Deliberative Quality of User Discussions in SNS' Comment Sections'. In: *Communication Research* 47.6, pp. 860–890. DOI: 10.1177/0093650218797884.

A

Black Pete Annotation

This appendix contains the annotation scheme used in Chapter 3. The original document was in Dutch. The overall annotation scheme consists of three classes: 0 - not relevant, 1x - Pro argument, 2x - Con argument.

Label	Explanation
0	Not relevant (No argument present)
10	Innocent children party (gezellig)
11	Children do not see racism (Not innate)
12	Black Pete is is simply not racist (no intention)
13	People of color also celebrate Black Pete/Sinterklaas
14	Christian tradition: Bishop of Myra descendant from devilish figure
15	Racial nativism: No racism in the Netherlands / you are racist for being against Black Pete
16	Defensive nationalism: Dutch culture ought to be protected
17	Germanic culture: Wodan, Krampus (Pre-Christian)
18	Black Pete is like a fairy tale figure (made-up)
19	Orientalism: Black Pete is a Moorish servant
21	Racist stereotype: historical perspective Slavery, Blackface, Colonisation
22	Racist stereotype: contemporary Effect on children, emblematic for institutional racism

Table A.1: Annotation cheat sheet Black Pete

B

Climate Change Annotation

Examples of climate change argumentation in online comments compiled for the annotation exercise used in Chapter 4. The overall annotation scheme consists of three classes: 0 - not relevant, 1x - Pro argument, 2x - Con argument.

Label 0 - No Argument/Off-topic: Strict category for all comments that do not contain an argument as outlined in the annotation scheme

---

Label 11 - Impact Scepticism: It is not that bad, there are measures against climate change. It will not impact our lives.

Er wordt nog steeds vanuit gegaan dat klimaat maakbaar is. Je moet je gewoon aanpassen aan de klimaatverandering. Nu wordt het wat warmer, straks weer wat kouder. Dat fatalistische is nergens voor nodig.

Sowieso gaat moeder Aarde zich herstellen van de te snelle co2 uitstoot. Als sommige volkeren moeten verhuizen naar elders is dat niet haar probleem. Na 2200 hebben we ook wel weer het ergste gehad ook.

Label 12 - Attribution Scepticism: Not caused by human activity, but by natural processes such as ice ages, position of the sun

Daarentegen is er wel overweldigend wetenschappelijk onderzoek dat aantoont dat de mens nauwelijks invloed heeft en dat er meer moet worden gekeken naar zaken als zon-

nevlekken. Er is nu eigenlijk een gebrek aan co2.

Hij weet dat ook wel, maar dat levert geen geld op. NASA geeft toe dat de klimaatverandering natuurlijk is en veroorzaakt wordt door de zon

---

Label 13 - Trend Scepticism: No warming trend, simply no difference compared to the past

Heel verhaal over dat het steeds erger wordt en dan 2020 erger was dan voorgaande jaren. En dan toch op een gedeelde eerste plek met 2014 als warmste jaar ooit in Nederland. In 6 jaar tijd dus niets veranderd.

Er worden al honderden en duizenden jaren records gebroken, al lang voordat we het weten enbewijzen zijn er wel. Opdrogen van de zeeën die nu woestijnen heten.

Label 14 - No consensus: No consensus exists among scientists whether human activity is the cause of climate change

Klimaatactivisten beweren dat "97% van de wetenschappers" het eens is over de oorzaken en urgentie van "klimaatverandering". Onzin. In dit filmpje laten we zien hoe het zit.

Dus omdat hij, een prominent wetenschapper, een andere mening had dan veel (maar zeker niet alle) wetenschappers van nu, meent u dat zijn titel tussen aanhalingstekens moet. Ik lees in het artikel niet dat zijn standpunt werd gevormd door de sponsoring.

Aha, dus jij weet als enige hoe het zit? Wie een andere mining heeft is niet goed geïnformeerd. Triest. Het gaat om theoretische wetenschap en die kan tot verschillende conclusies leiden, zeker met het versrijken van de tijd. Zo was men er vroeger heilig van overtuigt dat de aarde plat is.

Label 15 - Bad science: The science behind the issue is flawed. Incorrect predictions, wrong assumptions, too complex or wrong models used

Degene die jij aanhaalt zijn allemaal ecologen. Zelfs 1 die gepromoveerd is op klimaatverandering. Deze kijken dus ook met een gekleurde bril en zijn gebaat bij slechte cijfers wat uitstoot betreft van de landbouw. Deze zijn ook bevooroordeeld

Ze voorspellen niks, het zijn allemaal slechts theorieën

Label 16 - Conspiracy theories: Broad category for all content related to conspiracy theories. For example, NU.nl is propaganda, politicians paid by researchers, made up to be a way to earn extra money, etc.

Je mag hier ook helemaal niet zeggen dat de klimaatverandering misschien wel door andere factoren dan de mens kan komen. Het linkse bolwerk wil angst zaaien. Belastingen verhogen. "want jij bent de schuldige" Een andere mening wordt direct verwijderd.

Label 21 - Pro: Broad category for the acceptance of anthropogenic climate change

Ik ben het hiermee eens, maar ik zet hier een grote kanttekening bij: De huidige crisis en inactie om hier wat aan te doen komt door het neoliberalistische beleid van de overheid van de afgelopen 20 jaar. Die ideologie draait om "persoonlijke verantwoordelijkheid".

Slogans als "Een beter milieu begint bij jezelf" zijn enorme dooddoeners en leggen de verantwoordelijkheid bij de burgers, in plaats van de overheid. Ik bezit geen auto, eet geen vlees etc., net als veel vrienden van me, maar dat komt omdat wij in een situatie zitten waarin dat kan. Het is aan de overheid om ervoor te zorgen dat iedereen dat kan doen, om de verantwoordelijkheid te nemen en een wereld te creëren waarin duurzaamheid betaalbaar is voor iedereen, niet voor een select groepje welgestelden. Die verantwoordelijkheid nemen ze simpelweg niet. Zij kunnen belastingen/subsidies creëren om van NL een duurzaam land te maken.

Nederlandse samenvatting

Met het openen van discussieplatformen onder hun artikelen hoopten online nieuwsplatformen interactie te stimuleren tussen lezers en een constructieve discussie te stimuleren. Maar wat maakt een discussie net constructief? Hoe kan je zo een discussie promoten bij de gebruikers op een online platform? En in hoeverre kunnen computationele toepassingen hierbij helpen? Dit laatste wordt alsmaar belangrijker in de context van de sterk gestegen activiteit op deze platformen. Deze thesis neemt de onduidelijkheid rond het concept van constructieve discussie als startpunt. Door eerst te kijken naar hoe online nieuwsplatformen nu constructieve discussie hopen te stimuleren, kan een basis gelegd worden voor computationale toepassingen voor het modereren van gebruikerscommentaren.

Uit de antropologie leent deze thesis twee perspectieven om constructieve discussie te definiëren. Een *etic* perspectief werd gebruikt om constructieve discussie te beschrijven vanuit het oogpunt van externe onderzoeker. Via een etic perspectief hoopten we een formele definitie van het concept te formuleren die vervolgens gebruikt kan worden in een computationele toepassing of data annotatie. Vervolgens werd in deze thesis een *emic* perspectief op constructieve discussie geformuleerd. Deze benadering gebruikte de inzichten van de moderatoren, zij die dagelijks de schifting moeten maken tussen constructieve en niet-constructieve discussie.

Hoofdstuk 2 beschrijft onderzoek naar de uitvoering van online content moderatie door nieuwsplatformen. De focus lag op vijf nieuwsplatformen (*The Guardian, Die Zeit, New York Times, El País* en *NU.nl*). Het onderzoek keek naar hoe zij commentaren verwijderen en proberen om constructieve discussie te stimuleren. De resultaten tonen dat nieuwsplatformen individuele commentaren die binnen hun definitie van constructief passen vastpinnen aan de webpagina. Deze commentaren vormen de 'derde helft van het internet', gepositioneerd tussen, aan de ene kant, de inhoud geschreven door journalisten en redacteuren en, aan de andere kant, de overige commentaren die niet in het plaatje van constructieve discussie passen. Deze uitgelichte commentaren zijn een voorbeeld voor andere gebruikers en worden vaak '*featured comments*' of uitgelichte commentaren genoemd. Moderatoren krijgen de taak om deze zelf manueel uit de discussie naar voren te halen. Het verwijderen van ongewenste commentaren wordt steeds vaker overgelaten aan Artificiële Intelligentie (AI). Door deze ontwikkeling kreeg de moderator de tijd en ruimte om constructieve discussie te promoten.

Het onderzoek in hoofdstukken 3 en 4 maakten gebruik van een etic perspectief op constructieve discussie. In hoofdstuk 3 staat interactiviteit en diversiteit rond argumentatie centraal – een constructieve discussie bevat zowel voor- als tegenargumenten. In het hoofdstuk worden formules beschreven om de balans tussen '*pro*' en '*con*' argumenten te berekenen op drie sociale media platformen (*Reddit, Twitter* en *Gab*). De resultaten tonen dat de balans omtrent argumentatie verschilde tussen de platformen. Op een rechts platform als *Gab* vonden er uitsluitend 'echo chambers' plaats, discussies waarin tegenargumenten niet aanwezig zijn. Op de andere platformen was er volgens deze aanpak meer diversiteit te vinden.

Hoofdstuk 4 bouwt verder op de notie dat een constructieve discussie verschillende argumenten bevat. Rond het thema klimaatverandering werden commentaren van *NUjij*, het commentplatform van *NU.nl*, geannoteerd voor zowel voor- als tegenargumenten. Als gevolg van de imbalans tussen de verschillende argumenten – specifieke *fringe* argumenten kwamen niet vaak voor – sloegen de initiële taalmodellen er niet in om alle argumenten te herkennen. Om de resultaten te verbeteren bevat hoofdstuk 4 een *active learning* aanpak die de slechte classificatie van de initiële modellen gebruikte om de training data aan te vullen met voorbeelden van zeldzame argumenten. Evaluatie op basis van modellen getraind op deze aangevulde training data resulteerde in een betere classificatie waarin ook infrequente argumenten herkend werden. Deze *argument mining* methoden kunnen moderatoren ondersteunen in het filteren van genuanceerde discussies.

Vervolgens beschrijft deze thesis onderzoek gebaseerd op een emic perspectief voor constructieve discussie. Het onderzoek in hoofdstuk 5 gebruikte de keuzes van de moderatoren zelf om te modelleren wat een constructieve commentaar is. Op deze data werd een *group recommender system* getraind, een computationeel model dat de moderatoren kan ondersteunen in het kiezen van constructieve commentaren. De voorgestelde modellen rangschikken de commentaren op basis van de waarschijnlijkheid dat deze gekozen zouden worden als uitgelichte commentaren. In dit scenario krijgt de moderator enkel de top 5 of top 10 commentaren te zien, zodat zij niet de gehele discussie moeten doornemen.

Om te testen of de modellen robust waren tegen verandering in nieuwsonderwerpen werden de *group recommender systems* geëvalueerd op data van 2020 en 2023 – twee datasets met uiteenlopende nieuwsonderwerpen. Deze evaluatie toonde aan dat de modellen niet afhankelijk waren van de specifieke thema's in de training data. Ten slotte werd het best presterende model getoetst door de moderatoren zelf. Deze expert evaluatie toonde twee zaken aan. Eerst werd er geconcludeerd dat er nog veel variatie bestaan onder de moderatoren zelf. Vaak werden verschillende commentaren uitgekozen door de moderatoren in eenzelfde discussie. Ten tweede toonde de expert evaluatie aan dat moderatoren vaker commentaren kozen die voorgesteld werden door het model dan willekeurige commentaren. Uit dit laatste wordt in Hoofdstuk 5 de conclusie getrokken dat deze modellen de moderatoren bij platformen zoals *NU.nl* kunnen ondersteunen in het kiezen van constructieve commentaren door een voorselectie te maken.

Ten slotte bleef het een open vraag of het uitlichten van individuele, constructieve commentaren ook een impact had op een online discussie. Steeg de kwaliteit nadat moderatoren ingrijpen? Namen er meer gebruikers deel aan de discussie wanneer er uitgelichte commentaren op de webpagina te vinden waren? Hoofdstuk 6 beschrijft een analyse die deze vragen onderzocht. De data voor deze studie omvatte twee groepen: discussies met uitgelichte commentaren en discussies zonder (controlegroep). Om de impact te meten van de moderatie interventie werd elke discussie opgedeeld in een voor- en na subgroep. De discussies met uitgelichte commentaren werden gesplitst op het tijdstip toen de eerste commentaar gekozen werd door een moderator. De controlegroep werd opgedeeld op het mediaan tijdstip van alle eerst uitgekozen commentaren in de uitgelichte discussiegroep.

Om de impact te meten werden twee factoren getoetst: kwaliteit en activiteit. De resultaten tonen aan dat de kwaliteit van discussies niet beïnvloed werd door de aanwezigheid van uitgelichte (en constructieve) commentaren. Moderatoren moesten nog steeds een gelijkaardige hoeveelheid aan ongewenste commentaren verwijderen. Verder was er geen grotere aanwezigheid van goede commentaren. Dit laatste werd gemeten door het gemiddelde aantal *likes* per discussie en de aanwezigheid van potentiële commentaren om uit te lichten, berekend door het *group recommender system* van hoofdstuk 5. Activiteit werd gemeten door het gemiddeld aantal commentaren en gebruikers te vergelijken tussen de uitgelichte en controlegroep. De resultaten laten een potentieel effect van de moderatiestrategie zien. De verwachte vermindering van activiteit zette zich later in bij discussies waarin moderatoren uitgelichte commentaren kozen. Verder onderzoek is nodig om het verschil in discussie activiteit te linken aan de moderatie strategie.

Het onderzoek in deze thesis toont het belang van meerdere perspectieven aan. De termen etic en emic werden gebruikt om de distincte perspectieven te beschrijven. Wat constructief is in een discussie wordt beïnvloed door een reeks factoren, inclusief onderwerp, argumentatie, de visie van de moderatoren en dergelijke. Computationele toepassingen kunnen moderatoren ondersteunen in het uitkiezen van constructieve commentaren. Deze ondersteunende rol laat ruimte voor de moderator zelf om aan te voelen welke van de bestudeerde factoren van belang zijn in de specifieke context van een online discussie. Veldwerk werd ingezet om beter te begrijpen hoe computationele toepassingen ingezet kunnen worden in een complexe context als online moderatie op nieuwsplatformen. De combinatie met veldwerk was cruciaal om computationele toepassingen af te stellen op de context-specifieke factoren van praktische scenario's en de subjectieve visie en beleving van moderatoren. Uiteindelijk zijn zij diegenen die beslissen wat constructieve commentaren zijn op online nieuwsplatformen.

Curriculum Vitæ

Cedric Waterschoot

1995	Born in Brasschaat, Belgium.

Education

2019–2020	MSc. Public Policy and Human Development (Double degree) Maastricht Graduate School of Governance, Maastricht University UNU-MERIT, Maastricht *Thesis: : Climate Change, ideology and (social) media:* *a recipe for disaster of policy-relevant discussion?*
2015–2019	Bsc. Cognitive Science University of Osnabrück *Thesis: Political framing and sensationalism on American cable news:* *a corpus-based approach*
2017–2018	University of Amsterdam
2013–2015	Applied Linguistics KU Leuven

Work

2024–present	Postdoctoral Researcher Department of Advanced Computing Sciences Maastricht University
2020–2024	PhD Candidate KNAW Meertens Instituut Institute for Language Sciences, Utrecht University

Acknowledgements

Begin maar eens aan een PhD tijdens een lockdown. Wat doet een promovendus nu eigenlijk? Wordt er verwacht dat ik op het einde van de week een eerste studie klaar heb liggen? Zal ik dan maar direct aan dat boek beginnen schrijven? Met dit soort vragen begon ik vier jaar geleden aan deze onderneming. Gelukkig ondervond ik dat ik vanaf dag één kon rekenen op een club ondersteunende collega's, vrienden en familie.

Zenuwachtig wandelde ik op 1 september 2020 het Meertens Instituut binnen. Antal, dankzij jouw rondleiding door het gebouw vergat ik plotseling die zenuwen en kreeg ik direct motivatie om aan dit gigantisch werkstuk te beginnen. Ondanks jouw drukke agenda kon er altijd de tijd genomen worden voor een praatje of productieve vergadering. Je leerde mij de kunst achter het verzinnen van titels en de meer geavanceerde kneepjes om teksten in te korten. Door jouw ervaring en creatieve inzichten groeide ik in mijn onderzoek en voelde ik het vertrouwen om mijn eigen ideeën na te jagen. Deze ervaringen zorgen ervoor dat ik als onderzoeker steviger in mijn schoenen sta.

Ernst, je bent een motivational speaker pur sang. Zelfs na vijf uur op de trein en slechts één enkele meeting steeg mijn motivatie om verder te werken aan dit verhaal. Je liet zien hoe je enthousiasme voor jouw vak kan vertalen in productieve vormen van onderzoek. Van realistisch tot ambitieus, van praktisch tot zo conceptueel dat ik eerst een kop koffie nodig had om het allemaal te laten bezinken. Bedankt voor de brainstorm sessies, de kennismaking met veldwerk, de peptalks en de koffiepauzes. Al dit maakte niet alleen mijn werk beter, het zorgt er ook voor dat ik met vreugde terugkijk op de afgelopen vier jaar.

Dank aan mijn manuscript commissie, Walter Daelemans, Antske Fokkens, Thomas Poell, José van Dijck en Emiel Krahmer, voor het lezen en beoordelen van mijn dissertatie.

Veel heb ik te danken de fijne samenwerking binnen het Better-MODS team. Emiel en Florian, bedankt voor al de waardevolle inzichten. Het bespreken van mijn onderzoek met jullie leidde naar interessante onderzoeksvragen, creatieve studies en het uitbreiden van mijn interesses binnen de wereld van NLP en online media. Liesje, bedankt voor de vele gezellige meetings. Samen begonnen we bij het Better-MODS project en leerden we te navigeren binnen de onderzoekswereld. Jouw inzichten hielpen niet enkel bij het verbeteren van mijn werk; samen konden we de ups en downs van het PhD traject bespreken en relativeren. Ik kijk met veel vreugde terug aan al de Better-MODS meetings en hoop dat deze samenwerkingen nog lang blijven duren.

Thanks to all the friends I made during my time in Antwerp, Osnabrück and Maastricht. Despite the fact that you are a very diverse group, all of you motivated me to continue studying and chasing my interests even when I was thinking about calling it a day.

Speciale vermelding voor Jos, half beste vriend, half cheerleader, die tijdens een enkele trailrun al de stress en zorgen kan relativeren.

Dank aan mijn familie voor het vertrouwen over de jaren heen. Specifiek wil ik Luk en Chantal, mijn ouders, bedanken om mij de tijd en ruimte te geven om zelf te ontdekken waar mijn interesses en ambities liggen. Terwijl anderen dachten dat ik besluiteloos van de ene naar de andere plek dwaalde, zagen jullie het nut in van mijn zoektocht. Zonder die steun was ik nooit toevallig het vak *corpus linguistics* binnengewandeld. De energie die je terugkrijgt van zelfontdekking en het verdiepen van mijn eigen interesses was het wachten waard.

Ten slotte wil ik de persoon bedanken die mij dag in dag uit onvoorwaardelijk heeft gesteund en mij het gevoel gegeven heeft dat, zelfs tijdens lockdowns, ik er niet alleen voor stond. Vera, bedankt om steeds mee te willen denken en mij aan te moedigen tijdens de lastige tijden. Je hielp mij met zelfpromotie tijdens het schrijven van motivatiebrieven, het voorbereiden op interviews of bij het nalezen van teksten. Bij een publicatie herinnerde jij mij eraan om de positieve dingen te vieren, bij een afwijzing drukte jij mijn neus op de feiten. Soms vergat ik zelf dat ik op het juiste pad zat. Wat een fijn gevoel om te weten dat er altijd iemand voor je klaar staat.

Alphabetical Index

List of SIKS-dissertations

2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines

02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow

03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support

04 Laurens Rietveld (VUA), Publishing and Consuming Linked Data

05 Evgeny Sherkhonov (UvA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers

06 Michel Wilson (TUD), Robust scheduling in an uncertain environment

07 Jeroen de Man (VUA), Measuring and modeling negative emotions for virtual training

08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data

09 Archana Nottamkandath (VUA), Trusting Crowdsourced Information on Cultural Artefacts

10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms

11 Anne Schuth (UvA), Search Engines that Learn from Their Users

12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems

13 Nana Baah Gyan (VUA), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach

14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization

15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments

16 Guangliang Li (UvA), Socially Intelligent Autonomous Agents that Learn from Human Reward

17 Berend Weel (VUA), Towards Embodied Evolution of Robot Organisms

18 Albert Meroño Peñuela (VUA), Refining Statistical Data on the Web

19 Julia Efremova (TU/e), Mining Social Structures from Genealogical Data

20 Daan Odijk (UvA), Context & Semantics in News & Web Search

21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground

22 Grace Lewis (VUA), Software Architecture Strategies for Cyber-Foraging Systems

23 Fei Cai (UvA), Query Auto Completion in Information Retrieval

24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach

25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior

26 Dilhan Thilakarathne (VUA), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains

27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media

28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control

29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning

30 Ruud Mattheij (TiU), The Eyes Have It

31 Mohammad Khelghati (UT), Deep web content monitoring

32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations

33 Peter Bloem (UvA), Single Sample Statistics, exercises in learning from just one example

34 Dennis Schunselaar (TU/e), Configurable Process Trees: Elicitation, Analysis, and Enactment

35 Zhaochun Ren (UvA), Monitoring Social Media: Summarization, Classification and Recommendation

36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies

37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry

38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design

39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect

40 Christian Detweiler (TUD), Accounting for Values in Design

41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance

42 Spyros Martzoukos (UvA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora

43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice

44 Thibault Sellam (UvA), Automatic Assistants for Database Exploration

45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control

46 Jorge Gallego Perez (UT), Robots to Make you Happy

47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks

48 Tanja Buttler (TUD), Collecting Lessons Learned

49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis

50 Yan Wang (TiU), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime

02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation

03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines

04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store

05 Mahdieh Shadi (UvA), Collaboration Behavior

06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search

07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly

08 Rob Konijn (VUA), Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery

09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text

10 Robby van Delden (UT), (Steering) Interactive Play Behavior

11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment

12 Sander Leemans (TU/e), Robust Process Mining with Guarantees

13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology

14 Shoshannah Tekofsky (TiU), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior

15 Peter Berck (RUN), Memory-Based Text Correction

16 Aleksandr Chuklin (UvA), Understanding and Modeling Users of Modern Search Engines

17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution

18 Ridho Reinanda (UvA), Entity Associations for Search

19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval

20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility

21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)

22 Sara Magliacane (VUA), Logics for causal inference under uncertainty

23 David Graus (UvA), Entities of Interest — Discovery in Digital Traces

24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning

25 Veruska Zamborlini (VUA), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search

26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch

27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People’s Preferences, Perceptions and Behaviors

28 John Klein (VUA), Architecture Practices for Complex Contexts

29 Adel Alhuraibi (TiU), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"

30 Wilma Latuny (TiU), The Power of Facial Expressions

31 Ben Ruijl (UL), Advances in computational methods for QFT calculations

32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives

33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity

34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics

35 Martine de Vos (VUA), Interpreting natural science spreadsheets

36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging

37 Alejandro Montes Garcia (TU/e), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy

38 Alex Kayal (TUD), Normative Social Applications

39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR

40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems

41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle

42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets

43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval

44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering

45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement

46 Jan Schneider (OU), Sensor-based Learning Support

47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration

48 Angel Suarez (OU), Collaborative inquiry-based learning

2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations

02 Felix Mannhardt (TU/e), Multi-perspective Process Mining

03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction

04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks

05 Hugo Huurdeman (UvA), Supporting the Complex Dynamics of the Information Seeking Process

06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems

07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems

08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems

09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations

10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology

11 Mahdi Sargolzaei (UvA), Enabling Framework for Service-oriented Collaborative Networks

12 Xixi Lu (TU/e), Using behavioral context in process mining

13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future

14 Bart Joosten (TiU), Detecting Social Signals with Spatiotemporal Gabor Filters

15 Naser Davarzani (UM), Biomarker discovery in heart failure

16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children

17 Jianpeng Zhang (TU/e), On Graph Sample Clustering

18 Henriette Nakad (UL), De Notaris en Private Rechtspraak

19 Minh Duc Pham (VUA), Emergent relational schemas for RDF

20 Manxia Liu (RUN), Time and Bayesian Networks

21 Aad Slootmaker (OU), EMERGO: a generic platform for authoring and playing scenario-based serious games

22 Eric Fernandes de Mello Araújo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks

23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis

24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots

25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections

26 Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology

27 Maikel Leemans (TU/e), Hierarchical Process Mining for Scalable Software Analysis

28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel

29 Yu Gu (TiU), Emotion Recognition from Mandarin Speech

30 Wouter Beek (VUA), The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

---

2019 01 Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification

02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty

03 Eduardo Gonzalez Lopez de Murillas (TU/e), Process Mining on Databases: Extracting Event Data from Real Life Data Sources

04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data

05 Sebastiaan van Zelst (TU/e), Process Mining with Streaming Data

06 Chris Dijkshoorn (VUA), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets

07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms

08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes

09 Fahimeh Alizadeh Moghaddam (UvA), Self-adaptation for energy efficiency in software systems

10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction

11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs

12 Jacqueline Heinerman (VUA), Better Together

13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation

14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses

15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments

16 Guangming Li (TU/e), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models

17 Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts

18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication

19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents

20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence

21 Cong Liu (TU/e), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection

22 Martin van den Berg (VUA),Improving IT Decisions with Enterprise Architecture

23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification

24 Anca Dumitrache (VUA), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing

25 Emiel van Miltenburg (VUA), Pragmatic factors in (automatic) image description

26 Prince Singh (UT), An Integration Platform for Synchromodal Transport

27 Alessandra Antonaci (OU), The Gamification Design Process applied to (Massive) Open Online Courses

28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations

29 Daniel Formolo (VUA), Using virtual agents for simulation and training of social skills in safety-critical circumstances

30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems

31 Milan Jelisavcic (VUA), Alive and Kicking: Baby Steps in Robotics

32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games

33 Anil Yaman (TU/e), Evolution of Biologically Inspired Learning in Artificial Neural Networks

34 Negar Ahmadi (TU/e), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES

35 Lisa Facey-Shaw (OU), Gamification with digital badges in learning programming

36 Kevin Ackermans (OU), Designing Video-Enhanced Rubrics to Master Complex Skills

37 Jian Fang (TUD), Database Acceleration on FPGAs

38 Akos Kadar (OU), Learning visually grounded and multilingual representations

---

2020 01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour

02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models

03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding

04 Maarten van Gompel (RUN), Context as Linguistic Bridges

05 Yulong Pei (TU/e), On local and global structure mining

06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support

07 Wim van der Vegt (OU), Towards a software architecture for reusable game components

08 Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search

09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research

10 Alifah Syamsiyah (TU/e), In-database Preprocessing for Process Mining

11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data AugmentationMethods for Long-Tail Entity Recognition Models

12 Ward van Breda (VUA), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment

13 Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming

14 Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases

15 Konstantinos Georgiadis (OU), Smart CAT: Machine Learning for Configurable Assessments in Serious Games

16 Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling

17 Daniele Di Mitri (OU), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences

18 Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems

19 Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems

20 Albert Hankel (VUA), Embedding Green ICT Maturity in Organisations

21 Karine da Silva Miras de Araujo (VUA), Where is the robot?: Life as it could be

22 Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar

23 Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging

24 Lenin da Nóbrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots

25 Xin Du (TU/e), The Uncertainty in Exceptional Model Mining

26 Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization

27 Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context

28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality

29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference

30 Bob Zadok Blok (UL), Creatief, Creatiever, Creatiefst

31 Gongjin Lan (VUA), Learning better – From Baby to Better

32 Jason Rhuggenaath (TU/e), Revenue management in online markets: pricing and online advertising

33 Rick Gilsing (TU/e), Supporting service-dominant business model evaluation in the context of business model innovation

34 Anna Bon (UM), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development

35 Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production

---

2021 01 Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space

02 Rijk Mercuur (TUD), Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models

03 Seyyed Hadi Hashemi (UvA), Modeling Users Interacting with Smart Devices

04 Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning

05 Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems

06 Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot

07 Armel Lefebvre (UU), Research data management for open science

08 Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking

09 Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play

10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning

11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision

12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs

13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning

14 Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support

15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm

16 Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues

17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks

18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks

19 Roberto Verdecchia (VUA), Architectural Technical Debt: Identification and Management

20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems

21 Pedro Thiago Timbó Holanda (CWI), Progressive Indexes

22 Sihang Qiu (TUD), Conversational Crowdsourcing

23 Hugo Manuel Proença (UL), Robust rules for prediction and description

24 Kaijie Zhu (TU/e), On Efficient Temporal Subgraph Query Processing

25 Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications

26 Benno Kruit (CWI/VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables

27 Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You

28 Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs

---

2022 01 Judith van Stegeren (UT), Flavor text generation for role-playing video games

02 Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey

03 Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare

04 Ünal Aksu (UU), A Cross-Organizational Process Mining Framework

05 Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization

06 Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding

07 Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics

08 Maikel L. van Eck (TU/e), Process Mining for Smart Product Design

09 Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach

10 Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines

11 Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring

12 Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases

13 Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge

14 Michiel Overeem (UU), Evolution of Low-Code Platforms

15 Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining

16 Pieter Gijsbers (TU/e), Systems for AutoML Research

17 Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification

18 Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation

19 Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation

20 Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media

21 Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments

22 Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations

23 Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents

24 Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values

25 Anna L.D. Latour (UL), Optimal decision-making under constraints and uncertainty

26 Anne Dirkson (UL), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences

27 Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences

28 Onuralp Ulusoy (UU), Privacy in Collaborative Systems

29 Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality

30 Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays

31 Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management

32 Cezara Pastrav (UU), Social simulation for socio-ecological systems

33 Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management

34 Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change

35 Mike E.U. Ligthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction

---

2023 01 Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions

02 Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts

03 Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations

04 Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval

05 Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications

06 António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment

07 Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning

08 Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning

09 Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques

10 Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing

11 Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications

12 Tessel Bogaard (VUA), Using Metadata to Understand Search Behavior in Digital Libraries

13 Injy Sarhan (UU), Open Information Extraction for Knowledge Representation

14 Selma Čaušević (TUD), Energy resilience through self-organization

15 Alvaro Henrique Chaim Correia (TU/e), Insights on Learning Tractable Probabilistic Graphical Models

16 Peter Blomsma (TiU), Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters

17 Meike Nauta (UT), Explainable AI and Interpretable Computer Vision – From Oversight to Insight

18 Gustavo Penha (TUD), Designing and Diagnosing Models for Conversational Search and Recommendation

19 George Aalbers (TiU), Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals

20 Arkadiy Dushatskiy (TUD), Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning

21 Gerrit Jan de Bruin (UL), Network Analysis Methods for Smart Inspection in the Transport Domain

22 Alireza Shojaifar (UU), Volitional Cybersecurity

23 Theo Theunissen (UU), Documentation in Continuous Software Development

24 Agathe Balayn (TUD), Practices Towards Hazardous Failure Diagnosis in Machine Learning

25 Jurian Baas (UU), Entity Resolution on Historical Knowledge Graphs

26 Loek Tonnaer (TU/e), Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour

27 Ghada Sokar (TU/e), Learning Continually Under Changing Data Distributions

28 Floris den Hengst (VUA), Learning to Behave: Reinforcement Learning in Human Contexts

29 Tim Draws (TUD), Understanding Viewpoint Biases in Web Search Results

2024 01 Daphne Miedema (TU/e), On Learning SQL: Disentangling concepts in data systems education

02 Emile van Krieken (VUA), Optimisation in Neurosymbolic Learning Systems

03 Feri Wijayanto (RUN), Automated Model Selection for Rasch and Mediation Analysis

04 Mike Huisman (UL), Understanding Deep Meta-Learning

05 Yiyong Gou (UM), Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair

06 Azqa Nadeem (TUD), Understanding Adversary Behavior via XAI: Leveraging Sequence Clustering to Extract Threat Intelligence

07 Parisa Shayan (TiU), Modeling User Behavior in Learning Management Systems

08 Xin Zhou (UvA), From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation

09 Giso Dal (UT), Probabilistic Inference Using Partitioned Bayesian Networks

10 Cristina-Iulia Bucur (VUA), Linkflows: Towards Genuine Semantic Publishing in Science

11 withdrawn

12 Peide Zhu (TUD), Towards Robust Automatic Question Generation For Learning

13 Enrico Liscio (TUD), Context-Specific Value Inference via Hybrid Intelligence

14 Larissa Capobianco Shimomura (TU/e), On Graph Generating Dependencies and their Applications in Data Profiling

15 Ting Liu (VUA), A Gut Feeling: Biomedical Knowledge Graphs for Interrelating the Gut Microbiome and Mental Health

16 Arthur Barbosa Câmara (TUD), Designing Search-as-Learning Systems

17 Razieh Alidoosti (VUA), Ethics-aware Software Architecture Design

18 Laurens Stoop (UU), Data Driven Understanding of Energy-Meteorological Variability and its Impact on Energy System Operations

19 Azadeh Mozafari Mehr (TU/e), Multi-perspective Conformance Checking: Identifying and Understanding Patterns of Anomalous Behavior

20 Ritsart Anne Plantenga (UL), Omgang met Regels

21 Federica Vinella (UU), Crowdsourcing User-Centered Teams

22 Zeynep Ozturk Yurt (TU/e), Beyond Routine: Extending BPM for Knowledge-Intensive Processes with Controllable Dynamic Contexts

23 Jie Luo (VUA), Lamarck's Revenge: Inheritance of Learned Traits Improves Robot Evolution

24 Nirmal Roy (TUD), Exploring the effects of interactive interfaces on user search behaviour

25 Alisa Rieger (TUD), Striving for Responsible Opinion Formation in Web Search on Debated Topics

26 Tim Gubner (CWI), Adaptively Generating Heterogeneous Execution Strategies using the VOILA Framework

27 Lincen Yang (UL), Information-theoretic Partition-based Models for Interpretable Machine Learning

28 Leon Helwerda (UL), Grip on Software: Understanding development progress of Scrum sprints and backlogs

29 David Wilson Romero Guzman (VUA), The Good, the Efficient and the Inductive Biases: Exploring Efficiency in Deep Learning Through the Use of Inductive Biases

30 Vijanti Ramautar (UU), Model-Driven Sustainability Accounting

31 Ziyu Li (TUD), On the Utility of Metadata to Optimize Machine Learning Workflows

32 Vinicius Stein Dani (UU), The Alpha and Omega of Process Mining

33 Siddharth Mehrotra (TUD), Designing for Appropriate Trust in Human-AI interaction

34 Robert Deckers (VUA), From Smallest Software Particle to System Specification - MuDForM: Multi-Domain Formalization Method

35 Sicui Zhang (TU/e), Methods of Detecting Clinical Deviations with Process Mining: a fuzzy set approach

36 Thomas Mulder (TU/e), Optimization of Recursive Queries on Graphs

37 James Graham Nevin (UvA), The Ramifications of Data Handling for Computational Models

38 Christos Koutras (TUD), Tabular Schema Matching for Modern Settings

39 Paola Lara Machado (TU/e), The Nexus between Business Models and Operating Models: From Conceptual Understanding to Actionable Guidance

40 Montijn van de Ven (TU/e), Guiding the Definition of Key Performance Indicators for Business Models

41 Georgios Siachamis (TUD), Adaptivity for Streaming Dataflow Engines

42 Emmeke Veltmeijer (VUA), Small Groups, Big Insights: Understanding the Crowd through Expressive Subgroup Analysis

43 Cedric Waterschoot (KNAW Meertens Instituut), The Constructive Conundrum: Computational Approaches to Facilitate Constructive Commenting on Online News Platforms

44 Marcel Schmitz (OU), Towards learning analytics-supported learning design

45 Sara Salimzadeh (TUDelft), Living in the Age of AI: Understanding Contextual Factors that Shape Human-AI Decision-Making

Cedric Waterschoot