Effect of calculating Pointwise Mutual Information using a Fuzzy Sliding Window in Topic Modeling

Emil Rijcken*†, Kalliopi Zervanou‡,Marco Spruit ‡, Floortje Scheepers§, Uzay Kaymak*
*Jheronimus Academy of Data Science, Eindhoven University of Technology, Eindhoven, The Netherlands
Email: e.f.g.rijcken@tue.nl, u.kaymak@ieee.org
†Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands
‡Leiden Institute of Advanced Computer Science, Leiden University, Leiden, The Netherlands
Email: k.zervanou@liacs.leidenuniv.nl, m.r.spruit@liacs.leidenuniv.nl
§Psychiatry, University Medical Centre Utrecht, Utrecht, The Netherlands
Email: f.e.scheepers-2@umcutrecht.nl

2023 IEEE International Conference on Fuzzy Systems (FUZZ) | 979-8-3503-3228-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/FUZZ52849.2023.10309675

***Abstract*—Topic modeling is a popular method for analysing large amounts of unstructured text data and extracting meaningful insights. The coherence of the generated topics is a critical metric for determining the model quality and measuring the semantic relatedness of the words in a topic. The distributional hypothesis, a fundamental theory in linguistics, states that words occurring in the same contexts tend to have similar meanings. Based on this theory, word co-occurrence in a given context is often used to reflect word association in coherence scores. To this end, many coherence scores use Normalised Pointwise Mutual Information (NPMI), which uses a sliding window to describe the neighbourhood that defines the context. It is assumed that there is no other structure in the neighbourhood except for the presence of words. Inspired by the distributional hypothesis, we hypothesise the word distance to be relevant for determining the word association. Hence, we propose using a fuzzy sliding window to define a neighbourhood in which the association between words depends on the membership of the words in the fuzzy sliding window. To this end, we propose Fuzzy Normalized Pointwise Mutual Information (FNPMI) to calculate fuzzy coherence scores. We implement two different neighbourhood structures by the definition of the membership function of the sliding window. In the first implementation, the association between two words correlates positively with the distance, whereas the correlation is negative in the second. We compare the correlation of our proposed new coherence metrics with human judgment. We find that the use of a fuzzy sliding window correlates less with human judgment than a crisp sliding window. This finding indicates that word distance within a window is less important than defining the window size itself.**

***Index Terms*—Natural Language Processing, Distributional Hypothesis, Fuzzy Sliding Window, Topic Modeling**

I. Introduction

Topic modeling has emerged as a powerful technique for analysing unstructured text data and extracting meaningful insights. An essential aspect of topic modeling is evaluating the coherence of the topics generated by a model. Coherence refers to the semantic relatedness of the words in a topic and reflects how well these words support each other. A coherent topic has related words and conveys a clear and distinct theme. The coherence of a topic is a crucial metric to determine the overall model quality. Generally, the word co-occurrence is used to calculate coherence scores. The rationale for considering word co-occurrence comes from the distributional hypothesis, stating words occurring in the same contexts tend to have similar meanings [1], [2]. Normalised Pointwise Mutual Information (NPMI) [3] is often used to capture semantic relatedness when calculating the coherence in topic modeling. NPMI reflects the association between two words w_1 and w_2, which is the extent to which the joint probability of two words differs from what would be expected if w_1 and w_2 were independent. To calculate these probabilities, the metric uses sliding windows of size s to create different slices of a document. The probabilities are defined by the number of times a word occurs with another word, divided by the total number of windows. The window defines a neighbourhood that reflects a word's context; a larger window size can capture a broader context. The use of a sliding window to define a neighbourhood does not consider any structure in that neighbourhood. However, based on the distributional hypothesis, it seems intuitive to consider the distance between words to have an effect on word association. Consequently, we propose to use a fuzzy sliding window in which the membership to the fuzzy sliding window affects the computations for word association: the further apart two words are, the less their semantic relatedness is.

In our experiments, we focus on topic modeling to test whether word distances within a context matter, by using publicly available human judgment data. Our experiments compare three coherence scores; the c_v (crisp) score and two fuzzy alternatives that replace the NPMI calculation with FNPMI. One fuzzy approach, c_{fuzz}, assigns more weight to nearby words and the other, ${c_{fuzz}}^{\alpha}$, to distant words. Then, based on a corpus with topics and human judgment scores for each topic, we calculate the Spearman correlation between the human judgment- and the coherence scores.

Our findings indicate that conventional sliding windows' coherence calculations correlate better with human judgment. Apparently, the size of a neighbourhood is more relevant than the distance structure within the neighbourhood. We believe this finding contributes to understanding the concept of context in Natural Language Processing (NLP).

The outline of the paper is as follows. Section II II provides background information on topic modeling. Section III

discusses how NPMI and the coherence score are calculated. Then, we describe how FNPMI is calculated from fuzzy sliding windows in Section IV. Subsequently, we discuss the experimental setup and data used for comparison in Section V. We discuss the results and its implications in Section VI and conclude our work in Section VII.

II. Topic Modeling

The distributional hypothesis is a fundamental principle in linguistics asserting that words occurring in similar contexts tend to have similar meanings. The underlying idea is that the context largely determines the meaning of a word it appears in. This principle is based on the observation that words sharing similar meanings tend to co-occur in similar contexts and that the distributional patterns of words can reveal their semantic properties [1], [2]. The distributional hypothesis has been widely used in NLP and computational linguistics, where it forms the basis for various techniques, such as word embeddings and topic modeling.

Topic modeling is a widespread task in (NLP) that entails discovering hidden semantic themes within a collection of text documents. Topic modeling can be used for a variety of purposes, including topic discovery [4], [5], text classification [6], [7], similarity analysis [8], and sentiment analysis [9], among others. The output generated by topic models can serve as input for several downstream applications and serve as the primary objective for latent topic exploration. In many topic modeling approaches, a user feeds the algorithm a corpus of documents and a number of topics to find. Then, the algorithm returns two matrices, $\mathbf{p}(\mathbf{W}|\mathbf{T})$ and $\mathbf{p}(\mathbf{T}|\mathbf{D})$. The former gives the propensity of word i given topic k, and the latter the propensity of topic k given document j. The highest propensity words are retrieved from $\mathbf{p}(\mathbf{W}|\mathbf{T})$ by picking the top-*n* words. There is a wide variety of topic modeling algorithms. Latent Dirichlet Allocation [10] is the most popular and has been the basis for other models such as ProdLDA and NeuralLDA [11]. More recently, fuzzy algorithms such as FLSA [12], FLSA-W [13], and FLSA-E [14] were proposed. Amongst all the former algorithms, FLSA-W has outperformed other topic models in various evaluation metrics [15]. More recently, BERTopic [16] has received much attention. This model produces intuitive topics. However, the user can not set the number of topics. Each run returns a different 'optimal' number of topics, and so a systematic comparison with other models remains challenging.

Topic models are typically evaluated based on their inter or intra-topic quality. A common metric for the former is the diversity score [17], which indicates how unique the words in different topics are. A common metric for the intra-topic score is the coherence score [18]. This can be calculated in various ways, all inspired by the distributional hypothesis. One way is using the cosine similarity between words in dense vector spaces [19], such as Glove [20] or Word2Vec [21], [22]. These approaches locate words nearby each other in a high-dimensional continuous space that also co-occur frequently in a corpus. Hence, co-occurrence is considered implicitly. Additionally, other approaches explicitly count words; many approaches calculate the coherence this way. In one approach [23], all coherence scores are considered to be a combination of configurations from a four-dimensional configuration space. After calculating a coherence score for each combination for various corpora and topics, c_v (Section III-B) is found to correlate the highest with human judgment. This score finds the association between two words by using normalized pointwise mutual information (Section III). This score uses a sliding window to represent the context of a word as a neighbourhood. It calculates the joint probability between two words by considering their co-occurrence in the sliding window. However, no further structure within the neighbourhood is considered. Using fuzzy sliding windows might be a more appropriate approach to represent the neighbourhood structure from the perspective of the distributional hypothesis. These windows treat words that are farther apart as being partially within the neighbourhood. Hence, we generalize the concept of a sliding window to a fuzzy sliding window, similar to fuzzy sets generalising classical sets [24]. In this case, different words within a sliding window have a different membership to the context, based on their distance from a target word.

III. Quantifying Topic Coherence

A. Normalized Pointwise Mutual Information

Pointwise mutual information measures the information that two words, w_n and w_m, share. It indicates how much knowing one of the words reveals information about the other. Hence, if both words are independent, knowing one does not give information about the other. Given two words w_n and w_m, marginal probabilities $p(w_n)$ and $p(w_m)$ and joint probability $p(w_n, w_m)$, the pointwise mutual information $PMI(w_n, w_m)$ is calculated as:

$$PMI(w_n, w_m) = log \frac{p(w_n, w_m) + \epsilon}{p(w_n) \times p(w_m)} \tag{1}$$

Where ϵ is a small number introduced to prevent dealing with 0 probability values. PMI can be considered to be an estimate of how much more the two words co-occur than we expect by chance. The ratio ranges between $-\infty$ and ∞. Its normalized variant *Normalized Pointwise Mutual Information* (NPMI) ranges between -1 and 1 [3] (Equation 2). In this case, a score of 1 indicates a perfectly positive association where the two words always occur together. A score of -1 indicates a perfectly negative association where the two words never occur together. The ε is added to avoid a logarithm of zero. NPMI is defined as:

$$NPMI(w_n, w_m) = \frac{log \frac{p(w_n, w_m)+\epsilon}{p(w_n) \times p(w_m)}}{-log(p(w_n, w_m)) + \epsilon} \tag{2}$$

These probabilities can be calculated at a document level or based on a sliding window on fractions of the text. In the latter case, a sliding window moves over the document, one-word token per step, where each step defines a new

virtual document. To ensure all words are weighted equally, the documents are padded so that the first sliding window only counts the first two words and the last sliding window only counts the last two words. For the calculations in this section, we formulate the following quantities:

D	the number of documents in the corpus,
$w_{d,i}$	corpus word at index i in document d
d	the document index in a corpus. $d \in \{1, 2, ..., D\}$,
θ_d	document d represented by a bag of words,
$\|\theta_d\|$	the number of words in document d
i	the word position in a document $i \in \{1, 2, ..., \|\theta_d\|\}$,
j	the word index in a padded document.
s	the size of the sliding window,

Then, the probability of word w_n and w_m cooccurring is:

$$p(w_n, w_m) = \frac{\sum_{d=1}^{D} \sum_{j=2-s}^{|\theta_d|+s-2} b_{d,j}(w_n, w_m)}{\sum_{d=1}^{D} |\theta_d| + s - 3} \tag{3}$$

where

$$b_{d,j}(w_n, w_m) = \begin{cases} 1, & w_n, w_m \in W_C \\ 0, & else. \end{cases} \tag{4}$$

with

$$W_C = \{w_{d,j}, w_{d,j+1}, ..., w_{d,j+s}\}. \tag{5}$$

B. Coherence (c_v)

Coherence is a measure to reflect how well within-topic words support each other. The c_v score was shown to correlate the highest with human judgment [23]. For this reason, we discuss the c_v metric in the remainder of this paper [23].
In addition to the variables introduced before, we define the following quantities:

k	topic index. $k \in \{1, 2, ..., K\}$,
K	the number of topics,
n	word index in a topic. $n \in \{1, 2, ..., N\}$,
N	the number of most probable words per topic that characterizes the topic,
$w^T_{n,k}$	topic word at index n in topic k,
$\overrightarrow{w}_{n,k}$	vector to represent topic word at index n in topic k. Where $\|\overrightarrow{w}_{n,k}\| = N$.

For each topic k, we create a $(N \times N)$ matrix in which each cell gives the NPMI (2) between the corresponding words. The probabilities in the NPMI are based on a boolean sliding window of size s[1] ((3)-(5)).

Each row[2] in this matrix is a word vector $\overrightarrow{w}_{n,k}$ corresponding to word n in topic k.

[1] $s = 110$, in case of c_v.

[2] Or column, because it is a symmetric matrix

$$\overrightarrow{w}_{n,k} = [NPMI(w^T_{m,k}, w^T_{n,k})], \forall m \in \{1, 2, ..., N\} \tag{6}$$

Then, we take the sum of all word vectors in topic k to calculate the *topic vector* $\overrightarrow{w}^*_k$.

$$\overrightarrow{w}^*_k = \sum_{n=1}^{N} \overrightarrow{w}_{n.k} \tag{7}$$

The topic coherence is measured by the dispersion of the topic-word vectors from the resultant topic vector, using cosine similarity. Then, the coherence metric c_v is calculated by:

$$c_v = \frac{\sum_{k=1}^{K} \sum_{n=1}^{N} s_{cos}(\overrightarrow{w}_{n,k}, \overrightarrow{w}^*_k)}{N \times K} \tag{8}$$

with

$$s_{cos}(\overrightarrow{v}, \overrightarrow{w}) = \frac{\overrightarrow{v} \cdot \overrightarrow{w}}{\| \overrightarrow{w} \| \times \| \overrightarrow{w} \|}. \tag{9}$$

IV. Topic Coherence with Fuzzy Sliding Windows

The probability estimation used in NPMI does not capture word distance. We propose *Fuzzy Normalized Pointwise Mutual Information* (FNPMI), a generalization of NPMI that takes into account the distance between words. More specifically, the weight assigned to two words is inversely related to their distance, indicated by the membership function of the fuzzy sliding window (Figure 1 for an example of the fuzzy sliding window).

In addition to the quantities formulated in Section III, we define:

$\delta_{m,n}$	the (minimal) distance between two words m and n in a sliding window (minimal when there are duplicates of m or n).

$$FNPMI(w_n, w_m) = \left(\frac{log \frac{a(w_n, w_m) + \epsilon}{p(w_n) \times p(w_m)}}{-log(a(w_n w_m) + \epsilon} \right). \tag{10}$$

where

$$a(w_n, w_m) = \frac{\sum_{d=1}^{D} \sum_{j=2-s}^{|\theta_d|+s-2} \mu_{d,j}(w_n, w_m)}{\sum_{d=1}^{D} |\theta_d| + s - 3}, \tag{11}$$

and

$$\mu_{d,j}(w_n w_m) = \begin{cases} \frac{s - \delta_{m,n} + 1}{s}, & w_n, w_m \in W^{'}, \\ 0, & else. \end{cases} \tag{12}$$

with

$$W^{'} = \{w_{d,j}, w_{d,j+1}, ..., w_{d,j+s}\}. \tag{13}$$

Hence, given the words in a sliding window, the association between two words w_n and w_m is weighted by the distance between them proportionally to the sliding window size s.

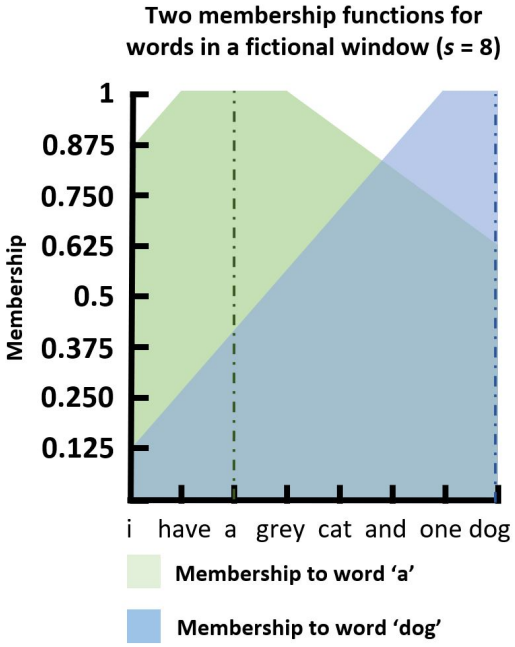

Fig. 1. A fictional window showing the membership functions for two words. For both words, it can be seen that the membership is inversely related to the distance between words.

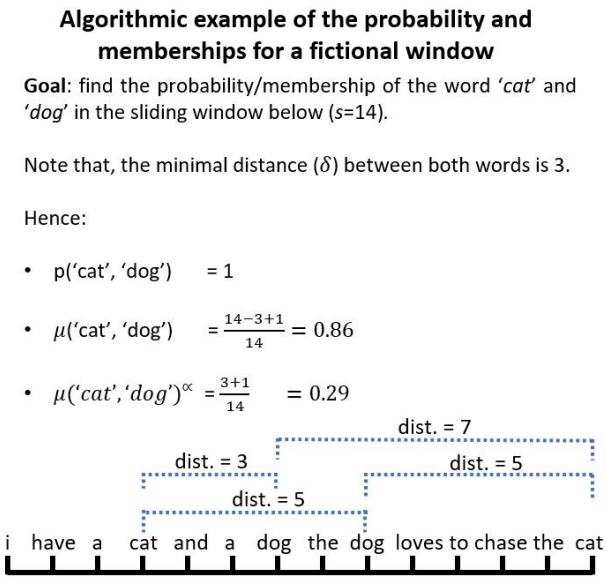

Fig. 2. An example showing how the probability and memberships are calculated for a fictional sliding window. The probability equals 1 if both words occur in a sliding window, regardless of the distance. Both memberships are based on the distance between both words and the length of the sliding window.

To obtain a better picture of the effect of word distance, we also experiment with an implementation that assigns more weight to words further away in the sliding window[3]. We refer to this method as FNPMI$^{\alpha}$. Equations (10) and (11) also apply to FNPMI$^{\alpha}$. However, the membership is calculated as (14).

$$\mu_{d,j}(w_n w_m)^{\alpha} = \begin{cases} \frac{\delta_{m,n}+1}{s}, & w_n, w_m \in W^{'}, \\ 0, & else \end{cases} \tag{14}$$

where $W^{'}$ is calculated the same as in (13). Figure 1 shows a fictional example of the membership calculation.

We implement c_{fuzz}, the coherence score with fuzzy sliding window, the same as the c_v configuration, but replace NPMI (2) with FNPMI (10). Figure 2 shows an example that illustrates how the probability and memberships are calculated for a word pair in a given sliding window.

V. Experimental Setup & Data

Our experiments aim to determine whether the coherence scores using fuzzy sliding windows correlate higher with human judgment than their crisp variant. We follow a similar experimental setup to Röder et al. [23]. We start with a corpus, a list of topics, and human evaluation scores per topic. Then, we calculate coherence scores for each topic based on FNPMI, FNPMI$^{\alpha}$, and NPMI. Subsequently, we calculate the Spearman rank correlation between the coherence and human interpretation scores. If the coherence metric with the highest human judgement uses fuzzy sliding windows, this indicates it is more intuitive to account for word distances than using a crisp sliding window. We use the c_v metric (Section III-B) as the benchmark for the crisp coherence metric, as it was shown to correlate the highest with human judgment in [23]. For all metrics, we calculate scores based on sliding windows of size 10, 20, ..., 300, similar to the original work [23]. We use the movie dataset [25] for validation. This dataset comprises 125,409 articles with an average length of 283.8 words and has 100 topics with five words per topic[4]. 19 volunteers, all fluent in English, have created golden labels by rating the topics (κ = 0.29) [25].

VI. Results & Discussion

Figure 4 shows the Spearman correlations between human judgment data and the different coherence scores. We observe that all scores show a moderate positive correlation with human judgment scores. The conventional (boolean) coherence score, based on NPMI, correlates more with human judgment than the coherence scores using fuzzy windows (based on FNPMI), for all sliding windows. Both c_v and c_{fuzz} show a steep decrease in correlation for window size 100, whereas ${c_{fuzz}}^{\alpha}$ shows a steep increase for this sliding window. The variation in correlation scores implies that researchers and practitioners should carefully consider the window size when using coherence scores. Lastly, both coherence scores based on fuzzy sliding windows correlate similarly to human judgment. Although only based on two neighbourhood structures, this means that the shape of the structure seems to have little impact on human judgment.

[3]Note that more complex membership functions can be used in general.

[4]One topic contained the word 'comic', which does not appear in the corpus. For this reason, we removed it from the topic.

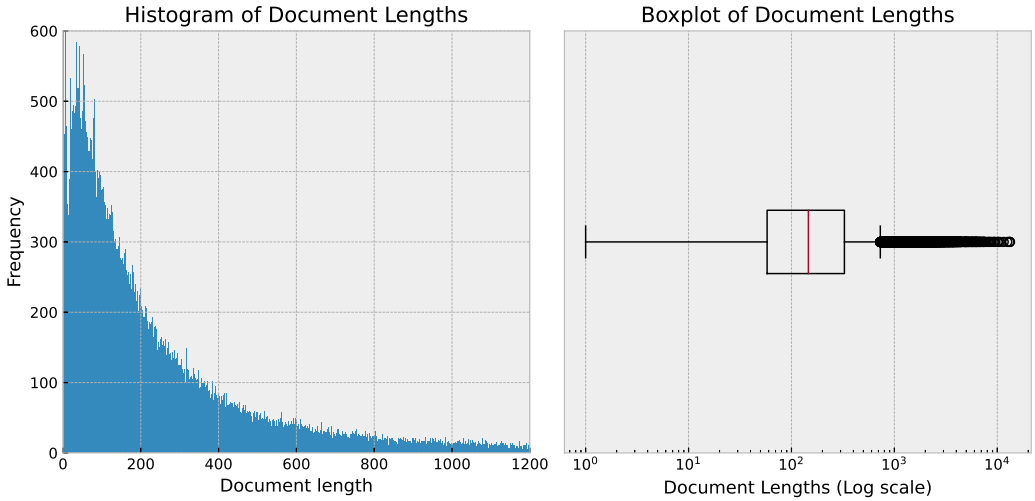

Fig. 3. The representations of the article length distribution in the corpus. Left is a histogram that shows the frequency of the article lengths up to a length of 1200. Right is a boxplot that uses a log scale to show the word length distribution in a boxplot.

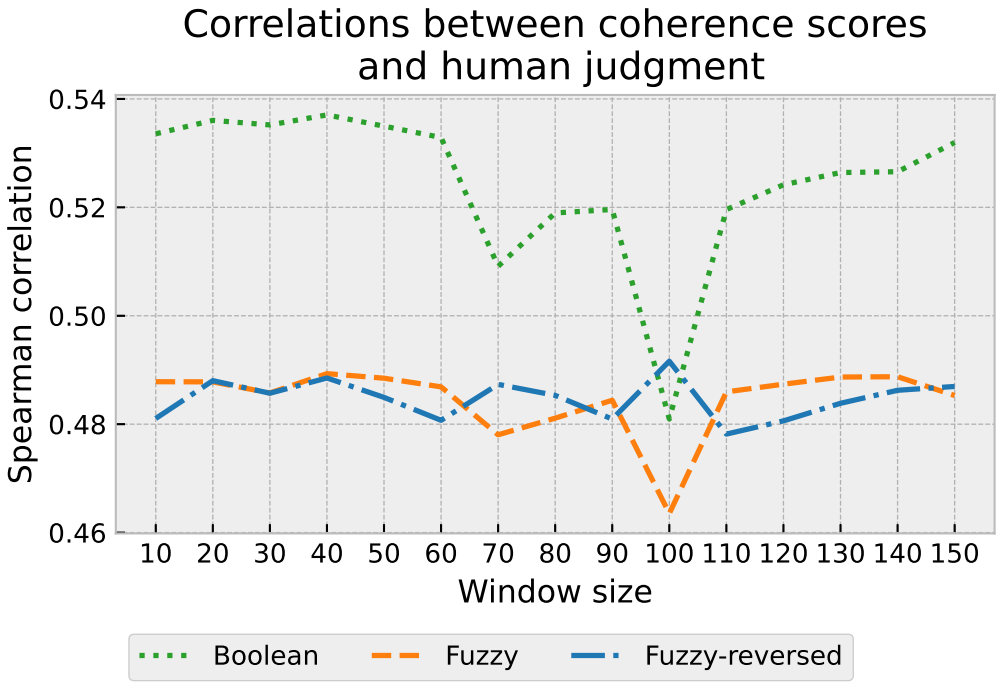

Fig. 4. A graph showing the Spearman correlations between the different coherence scores and human judgment for different sliding windows.

Based on the distributional hypothesis, we evaluate whether word distance helps to create better coherence scores in topic modeling. We hypothesized that word distance would impact word association and influence the coherence score. The conventional coherence method weights all within-window words equally. Hence, it does not consider the distance between words in a window, whereas the fuzzy alternative window does consider distance. However, the conventional method correlates higher with human judgment scores than the proposed fuzzy alternatives. Combined with the varying correlations with different window sizes, this finding suggests that the window size might be more critical than the distance between words within that window. This work gives a more detailed understanding of the meaning of context in NLP.

Note that we have only used one specific corpus of text data; this corpus may not generalize to other types of text data or domains. Moreover, the inter-annotator agreement (κ) for creating the human evaluation scores was 0.29. A score of 0 means no agreement, and 1 is perfect agreement. Hence, this score implies that humans are not consistent in rating topics. The inter-annotator agreement for the human judgment scores must be higher to draw conclusions about the correlation to human judgment. We only compare the fuzzy coherence scores to the c_v coherence score because, on average, this metric was found to correlate the highest with human judgment scores. However, there are many other approaches for calculating coherence scores. An evaluation comparing various metrics, such as C_P, C_{NPMI}, C_{UCI} [23], would provide a more comprehensive overview.

VII. Conclusion

This work evaluates whether the distance between words helps to create better coherence scores in topic modeling. We propose a fuzzy alternative to the NPMI metric, based on which we formulate a fuzzy coherence metric inspired by the distributional hypothesis. We calculate scores for various sliding window sizes based on these coherence metrics.

We find that a crisp definition of a sliding window correlates better with human judgment than a fuzzy definition. Even though we have considered only a single data set in a limited number of configurations, we think our finding is important for understanding the concept of context and its relation to a neighbourhood structure in topic modeling applications. In the future, we intend to conduct more detailed experiments exploring the relationship between topic coherence and neighbourhood structures for calculating coherence.

References

[1] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[2] J. Firth, "A synopsis of linguistic theory, 1930-1955," *Studies in linguistic analysis*, pp. 10–32, 1957.

[3] G. Bouma, "Normalized (Pointwise) Mutual Information in collocation extraction," *Proceedings of GSCL*, vol. 30, pp. 31–40, 2009.

[4] S. Syed, M. Borit, and M. Spruit, "Narrow lenses for capturing the complexity of fisheries: A topic analysis of fisheries science from 1990 to 2016," *Fish and Fisheries*, vol. 19, no. 4, pp. 643–661, 2018.

[5] J. Arendsen, E. Rijcken, K. Zervanou, K. Rietjens, F. Vlems, and U. Kaymak, "Analyzing patient feedback data with topic modeling," in *International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU)*, 2022.

[6] P. Mosteiro, E. Rijcken, K. Zervanou, U. Kaymak, F. Scheepers, and M. Spruit, "Machine learning for violence risk assessment using Dutch clinical notes," *Journal of Artificial Intelligence for Medical Sciences*, vol. 2, no. 1-2, pp. 44–54, 2021.

[7] E. Rijcken, U. Kaymak, F. Scheepers, P. Mosteiro, K. Zervanou, and M. Spruit, "Topic modeling for interpretable text classification from EHRs," *Frontiers in Big Data*, vol. 5, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fdata.2022.846930

[8] D. Spina, J. Gonzalo, and E. Amigó, "Learning similarity functions for topic detection in online reputation monitoring," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 2014, pp. 527–536.

[9] T. A. Rana, Y.-N. Cheah, and S. Letchmunan, "Topic modeling in sentiment analysis: A systematic review." *Journal of ICT Research & Applications*, vol. 10, no. 1, 2016.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning research*, vol. 3, pp. 993–1022, 2003.

[11] A. Srivastava and C. Sutton, "Autoencoding variational inference for topic models," in *5th International Conference on Learning Representations,ICLR*, 2017.

[12] A. Karami, A. Gangopadhyay, B. Zhou, and H. Kharrazi, "Fuzzy approach topic discovery in health and medical corpora," *International Journal of Fuzzy Systems*, vol. 20, no. 4, pp. 1334–1345, 2018.

[13] E. Rijcken, F. Scheepers, P. Mosteiro, K. Zervanou, M. Spruit, and U. Kaymak, "A comparative study of fuzzy topic models and LDA in terms of interpretability," in *Proceedings of the 2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2021.

[14] E. Rijcken, K. Zervanou, M. Spruit, P. Mosteiro, F. Scheepers, and U. Kaymak, "Exploring embedding spaces for more coherent topic modeling in electronic health records," in *IEEE International Conference on Systems, Man, and Cybernetics*, 2022.

[15] E. Rijcken, K. Zervanou, P. Mosteiro, M. Spruit, F. Scheepers, and U. Kaymak, "A performance evaluation of topic models based on fuzzy latent semantic analysis," 2022. [Online]. Available: https://research.tue.nl/en/publications/a-performance-evaluation-of-topic-models-based-on-fuzzy-latent-se

[16] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[17] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic modeling in embedding spaces," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 439–453, 2020. [Online]. Available: https://aclanthology.org/2020.tacl-1.29

[18] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the 2011 conference on empirical methods in natural language processing*, 2011, pp. 262–272.

[19] R. Ding, R. Nallapati, and B. Xiang, "Coherence-aware neural topic modeling," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 830–836. [Online]. Available: https://aclanthology.org/D18-1096

[20] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[23] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the eighth ACM International Conference on Web Search and Data Mining*, 2015, pp. 399–408.

[24] L. A. Zadeh, "Probability measures of fuzzy events," *J. Math. Anal. Appl.*, vol. 23, pp. 421–427, 1968.

[25] F. Rosner, A. Hinneburg, M. Röder, M. Nettling, and A. Both, "Evaluating topic coherence measures," *CoRR abs/1403.6397*, 2014.