Future Generation Computer Systems 102 (2020) 437–452

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

On the use of distributed semantics of tweet metadata for user age prediction☆

Abhinay Pandya [a,*], Mourad Oussalah [a,*], Paola Monachesi [b], Panos Kostakos [a]

[a] Center for Ubiquitous Computing, Faculty of Information Technology and Electrical Engineering, University of Oulu, FI90014, Finland
[b] Department of Linguistics, University of Utrecht, 3512 JK Utrecht, The Netherlands

ARTICLE INFO

Article history:
Received 18 February 2019
Received in revised form 26 July 2019
Accepted 27 August 2019
Available online 2 September 2019

Keywords:
Social media mining
Twitter
Convolutional neural networks
Age prediction

ABSTRACT

Social media data represent an important resource for behavioral analysis of the aging population. This paper addresses the problem of age prediction from Twitter dataset, where the prediction issue is viewed as a classification task. For this purpose, an innovative model based on Convolutional Neural Network is devised. To this end, we rely on language-related features and social media specific metadata. More specifically, we introduce two features that have not been previously considered in the literature: the content of URLs and hashtags appearing in tweets. We also employ distributed representations of words and phrases present in tweets, hashtags and URLs, pre-trained on appropriate corpora in order to exploit their semantic information in age prediction. We show that our CNN-based classifier, when compared with baseline models, yields an improvement of up to 12.3% for Dutch dataset, 9.8% for English1 dataset, and 6.6% for English2 dataset in the micro-averaged F1 score.

1. Introduction

In the digital era, social media has become a ubiquitous part of our daily life where users constantly interact with Facebook, Twitter, Snapchat, among other social media platforms, sharing their experiences and opinions on various topics. The availability of many social media datasets (e.g., Twitter, public Facebook pages and blogs) offers golden opportunities to social scientists to study psychological and social questions at an unprecedented scale [1]. For instance, social media has been employed in stock market prediction [2], Oil price prediction [3], health monitoring [4], disaster management [5], forecast box-office revenues for movies [4], inferring national mood throughout the day [6], measuring behavioral risk factors [7], among others.

On the other hand, the open access of many of social media platforms has made it possible for people of every age to become author and reader without any formal restriction. This created an ideal environment for online predators to gain access to sensible user related information, which render internet activities of many vulnerable communities (e.g., kids, teenagers, females) at risk. Therefore, automatic identification of age groups from social media posts would offer an edge to crime prevention as well as many other activities – e.g., online tutoring, personalized advertisement, content personalization, and (intelligent) plagiarism detection to identify for instance whether the homework is performed by the student or another person. Besides, since the last decade, several large scale corporations, e.g., Amazon and Apple, have massively invested in human factors that comprehend consumer behaviors and predict consumer retention based on user's activity and registered profile [8,9], where age group plays key role in inferring the community membership of the user. Nevertheless in the absence of supporting biographical information, as it is the case in many social media platforms, inferring age attributes from textual posts only is rather challenging. Sociolinguistic theory advocates a strong connection between language use and age, or more generally between discourse and identity, where users employ language patterns to construct their identity [10]. In this context, age is considered as a social and fluid variable that is shaped depending on the societal context, the culture, individual experiences and social roles [11].

In the area of computational linguistics, there is an inherent interest in determining latent attributes of an author, which include his categorical age, where a variety of published work has been focused on linguistic analysis for author age in on- and offline texts, much of which corresponds to lexical and contextual clues, such as analyzing topic and genre or n-gram patterns. For instance, [12] analyzed online behavior associated with blogs (which are usually more comprehensive contents than tweets) and found that behavior (number of friends, posts, time of posts,

☆ This work is (partially) funded by the Marie Skłodowska-Curie, Finland grant (645706-GRAGE) and EU grant (770469-Cutler).

* Corresponding authors.
E-mail addresses: abhinay.pandya@oulu.fi (A. Pandya), mourad.oussalah@oulu.fi (M. Oussalah), p.monachesi@uu.nl (P. Monachesi), panos.kostakos@oulu.fi (P. Kostakos).

https://doi.org/10.1016/j.future.2019.08.018

etc.) could effectively be used in binary age classifiers. [13] identified a set of writing and speech patterns, referred to as age-grading, that changes over time as a person learns a language and develops socially. This suggests that the assumption of text-based age prediction is tenable, although no universally accepted solution emerged. Especially, it is widely debatable whether age patterns can be elicited from short social media related posts. [14] argued that at least 10000 words per author is needed (in [15], an estimate of 5000 words per author) in order to infer reasonable age patterns. This order of magnitude is widely unavailable in most of microblog posts, e.g., Twitter, Facebook messages where high proportion of noisy terms is expected in order to restrict the length of the message, which, in turn, challenges standard natural language processing tools [16,17]. In an attempt to tackle the above challenge, several works have investigated fine-grained specialized natural language processing tools that deal with microblog posts, construction of large scale corpora that includes a variety of noisy terms/abbreviations, and, therefore, devise a sound machine learning architectures for age categorization and prediction. [18] developed pre-processing techniques to normalize orthographic modifications as well as twitter-specific elements (@-usernames and #hashtags) that translate the noisy Twitter texts into standard English and found to work well in classification problems. [19] used [18] and [20] pre-processing techniques in order to identify types of creative lexical transformations resulting in OOV tokens in Twitter messages, which are then employed to differentiate between users' biographical attributes. A multi-corpus based approach was investigated by [21] that include transcribed telephone speech corpus and posts from breast-cancer online forum. Next, a domain adaptation approach was used to train a model on these corpora and separate the global features from corpus-specific features that are associated with age. [22] constructed a predictive lexica from a dataset of Facebook users who agreed to share their status updates and reported their age and gender. Age prediction has been widely considered as a classification problem where machine learning techniques were employed. For instance, [23] considered age as a latent attribute of Twitter user and SVM classification were employed to estimate the age class.

Moreover, computational linguistics competitions PAN 2013 and PAN 2014 of author profiling task [24,25] have seen a growing interest in the research community, where the participating teams have proposed various methodologies that vary widely in terms of pre-processing, choice of feature set and classification methods. Especially, in PAN 2014, the age prediction task was defined as a challenging multi-class prediction problem with five classes (18–24, 25–34, 35–49, 50–64, 65+). Notable observations of these works include the relative lack of predictive utility of n-gram based models, as well as the high level of accuracy achieved by a group using class similarity based features [26].

In summary, the preceding highlights two key findings. First, the issue of eliciting age from author's posts is tenable from both sociological and computational linguistics perspectives. Second, the issue of optimal configuration of generic estimation architecture remains widely open, which motivates further work on this matter. Previous work [27–34] has identified several features based on both language and social media specific meta-data that are relevant for the age prediction task including bag-of-words, linguistic features, stylometric features, profile features, (e.g., background color, profile image), social network, and preferences (e.g., liked tweets). Similarly, many approaches based on multi-class classification have been proposed, such as SVMs [23,35], logistic regression [12,27] and Naive Bayes [36].

As [28] pointed out, the problem of age estimation is very challenging due to the inherent variability of human language together with unstructured text in tweet messages, which raises the question of finding appropriate cues that elicit categorical age and the subsequent reasoning. The above studies revealed at least three key limitations and challenges. First, the variety of contexts and discourses poses serious challenges to inter-operability of linguistic features from one study to another. For instance, the application of the comprehensive WWBP linguistic database [29], which is developed using Facebook dataset, to Twitter has shown to be unsatisfactory. Second, although, many of these studies acknowledge the importance of metadata in social media posts, they do not explicitly make use of the content of this metadata. Third, the variety of estimation architecture ranging from the type of preprocessing, number and type of features employed, and machine learning or estimation algorithms testifies of the need for further research on the issue.

This motivates the current paper which focuses on Twitter data and builds on previous work on age prediction by relying on language related features and social media metadata to classify users in age groups, considering thus age as a categorical variable. Especially, the following contributions are highlighted.

1. Motivated by its sound theoretical convergence properties and good performance achievements in text-mining related applications [37–40] a CNN-based model for classification is adopted that integrates heterogeneous features for age-category classification.
2. New feature-set constituted of Hashtags and URLs content analysis is introduced. To the best of our knowledge, there is no previous work using the content of hashtags and URLs for age prediction. Although they have been considered as social media specific features (i.e. their number has been included in [27,30], their content has not been considered. We contend that hashtags and URLs in tweets are indicative of user's age since they reflect user's interests and activities [41]. We propose a novel method to derive relevant features from hashtags and URLs and incorporate these into our CNN-based classification model.
3. In order to tackle the lack of scalability and interoperability, an enhanced semantics through pre-trained word embeddings is introduced. Unsupervised learning of distributed representations (word embeddings) obviates the need for careful feature engineering and such representations are richer in semantic information than standard bag-of-words [42–44]. We propose to employ word embeddings for tweet texts, title text of the pages referred to by the URLs in the tweets, and for most frequently co-occurring words with each hashtag in the tweet. Furthermore, we pre-train word embeddings on different corpora to take the context into account. More specifically, word embedding vectors used for tweet texts and hashtags are pre-trained on large collection of tweets, and those used for URLs are pre-trained on blogs/news.
4. Comparison with some state-of-the-art estimation algorithms (SVM, logistic regression, random forest) is carried out in order to demonstrate the feasibility and good performance of our proposal.

We employ three existing datasets from two different languages (i.e. English and Dutch) to test the validity of our novel features and classification method against the versatile classification models SVM, Logistic Regression, and Random Forest as a baseline and show that our CNN-based model outperforms the baseline significantly. While our approach is easily extendable to include other demographic variables such as gender, ethnicity, etc., it is also adaptable to solve other related problems in social media mining ranging from online mental health surveillance, to social spammer detection, and to assist recommendation systems to find similar users on the social media. The rest of the

paper is organized as follows. In Section 2, we discuss related work. Section 3 depicts our approach to extract features for age prediction of Twitter users and describes our CNN-based model. Section 4 explains our experimental setup and results compared with baseline approaches. Finally, in Section 5, we present our conclusions.

2. Related work

In the context of Twitter based age estimation/prediction, one distinguishes at least two streams of research. The first one, sometimes, referred to as age-annotation, deals with the construction of dataset that genuinely relates author's posts to his categorical age. The second emphasizes the design and implementation of the software architecture that enables age estimation from pre-processed dataset.

Indeed, labeled demographic data, including age data in particular, are not systematically collected by Twitter when users set up new accounts. Even when very occasionally reported by the users, the latter often do not assign the correct age, which stresses on the need to use external resources or direct / indirect system query-based approaches. For instance, [31] employed the Twitter API to identify Twitter accounts that had tweets about birthdays mentioning the person's age (e.g., "Happy *XX* birthday *YY*") or individuals who sent birthday wishes (e.g.,"Wishing *@xxxxxx* a happy *XX* birthday"). This ultimately enabled linking the underlined Twitter user with the mentioning age. [27] inferred age estimate by adjoining LinkedIn profiles for youth who tweeted about a particular grade level in school. [45] inferred age attribute by looking at first names cross-referenced with baby name frequency data from Social Security Administration. [46] advocated the use of proxies and/or associated meta-data in order to derive demographic information of Twitter users. [47] put forward a distributed digital social research platform, referred to collaborative online social media observatory (COSMOS) that provides on-demand analytics including age attribute from Twitter stream by correlating it with other dataset and events. Other alternative work focused on the profile picture of the Twitter user, assuming the picture is genuine and include a clear face portrait, [48] used Face++, a free facial recognition service that can estimate a user's age within a 10-year span.

In the second stream of research related to software architecture for age prediction, the developed approaches vary according to type of pre-processing, input features, choice and number of age categories, supervision versus non-supervision scheme, type of supervision algorithm and validation strategy employed.

Pre-processing of Twitter messages allows to filter out the abundant noise present in social media, and to normalize the orthographic modifications as well as translate the various slang and SMS-like vocabulary into semantically meaningful text. Input features employed for age prediction vary from linguistic cues, network (e.g., number of friends, ratio of followers to friends), and user's profile related information (e.g., background picture, text color). Possibly due to difficulty in processing the network information in real time, unreliability of profile information together with advances in linguistics that distinguish language use of childhood, adolescence and adulthood, the quasi-majority of reported works employ linguistic features. In this respect, one of the most notable works was carried out as part of World Well-Being Project (WWBP) [29] where an open vocabulary analysis framework was advocated, whereby they link a series of individual words, phrases, and topics that emerge from open text context. Authors in [29] have shown clear distinctions across four age grouping categories (ages 13–18, 19–22, 23–29, 30–65) where they highlighted: (I) the greater use of emoticons and slang among younger groups and, (II) the developmental progression of individuals at different life stages (e.g., school, college, career, marriage, children, family).

Social media content, including Twitter, also exhibits medium-specific features (i.e. metadata) that show a different use wrt. age, as is the case for sharing links and/or images and tagging/hashtag use. [34] show that incoming communications from an individual's strong ties are more revealing of the individual's identity, but that this relationship only holds for publicly visible aspects of the identity. Authors in [49] have studied blog posts and showed that there is no trend in image sharing, but there is a gentle increase in usage of URLs in posts with respect to age: apart from an inexplicable peak at the age of 24, link sharing increases with age with users older than 35 posting the most. This result on URLs is supported also by [50], [51], who have continued research on blogging and found that the sharing of links increases with age. On the other hand, work in [27] demonstrated that a sharp rise in the use of links for Dutch Twitter users in their 20s, that stagnates in their 30s. They associated this finding with information sharing and impression management. [12] showed that the use of links and images in their blog data varies across all ages. In the case of hashtags, [27] found that hashtags are used more often by older Twitter users: low usage in teens, a steep climb in the 20s, the highest and continuous use through the years up until the oldest participants category (over 60 years of age). According to these authors, hashtags are, similarly to links, connected to the sharing of information and older tweeters apparently are more concerned with information sharing than younger users. Younger people seem to display a certain kind of online identity, something older people are less concerned with [52].

Computational work on age prediction has exploited these differences in the use of URLs and hashtags across age groups, but have not considered the content associated to them, that also reflects an age related use. For example, [49] have researched the behavior of two groups on Instagram: First, they found that the adult group (25–39) displays a wider range of interests in topics and are very diverse: arts/photos/design, locations, mood/emotion, nature, social/people. Second, the majority of the teens' (13–19) hashtags concern mood/emotion and follow/like. [41] concluded that hashtags are an important feature to discriminate age since older adults above 67 use mainly hashtags related to politics and leisure in Twitter while people below 55, use mainly hashtags in the context of work related activities and technology.

Table 1 summarizes the key results in terms of age prediction from online social media platforms (blogs, Facebook, and Twitter), highlighting the main features, approach employed and the level of accuracy obtained. It is evident from Table 1 that all approaches in the previous research considered bag-of-words (BoW) representation for features. Accuracy results on blogs are typically higher than Twitter because of more data available per user. The highest accuracy (94.13%) was reported by [56] on ICWSM 2009 blog dataset utilizing Naive Bayes model trained on content words, slang words and stylistic features. For Twitter users, [27] report micro-averaged F1 score for three age categories classification of 86.32% on their own dataset of 2,494 users. These results were obtained by Logistic Regression trained on unigram features. Table 2 presents a summary of previous work on the use of Twitter metadata for demographic attribute prediction. These approaches utilize Twitter metadata attributes such as friends/followers ratio, profile image, background color, etc. along with tweet content. However, only count values of hashtags and URLs were included ignoring the content thereof.

It is evident from the literature review tables (ref. Tables 1 and 2) that, while the previous research has exploited rich set of features such as linguistic, stylometric, and social network features, attention to the following predictors of age attribute is missing: (a) the content of hashtags and URLs embedded in posts (especially, on Twitter), and (b) the use of richer semantic models (such as word embeddings) instead of BoW representation.

Table 1
Summary of previous work on age prediction from social media.

Research work	Method used	Features employed	Size of dataset	Accuracy
[31]	SVM classification	k-top ngrams frequency statistics followers/friends ratio	400 Twitter users	0.805*
[27]	Logistic regression	unigrams	2,494 Twitter users	0.86+
[53]	SVM classification, f-divergence	unigrams	756 Twitter users	0.06%
[33]	SVM classification	unigrams, emoticons	324 Twitter users	0.77***
[46]	Pattern matching		1,471 Twitter users	1.0++
[54]	Linear Regression	Facebook likes	58,00 Facebook users	0.75
[48]	Face++	Profile images, User description	2,433 Twitter users	0.75++
[30]	Logistic regression	lexical features, Twitter metadata features	3,184 Twitter users	0.74**
[22]	Linear regression	unigrams	72,874 Facebookusers	0.831
[32]	Logistic Regression	unigrams	5000 Twitter users	unspecified
[50]	Multi-Class Real Winnow (MCRW)	unigrams, stylometric features	37,478 blogs	0.76+++
[55]	SVM classification	ngrams, POS ngrams, Wikipedia semantic	236,600 blogs	0.66+++
[56]	Naive Bayes	unigrams	75,558 blogs	0.95+++
[45]	Bayesian generative model	first names	1000 Twitter users	0.08$
[57]	Decision tree	unigrams, stylometric features	307 users	0.4++++
[58]	Bayesian Multinomial Regression	unigrams, stylometric features	19,320 blog authors	0.76++++++
[59]	semi-supervised learning Alternating Structure Optimization	unigrams	2000 blog users	0.64*****

* For two classes: 18–23, and 25+.
** For three classes: 13–17, 18–24, 25+.
*** For two classes: 0–20, 20 plus.
**** For five classes, KL-divergence.
***** For five classes, 10s 20s 30s 40s 50s.
\+ For three classes: 0–20, 20–40, 40+.
++ Manually verified.
+++ For three classes: 10s, 20s, 30s.
++++ For five classes 18–24, 25–34, 35–49, 50–64, 65+.
++++++ For three classes, 13–17, 23–27, 33–47.
% MSE for age class ratio prediction.
$ Pearson's correlation.

Table 2
Summary of previous work on use of Twitter metadata for demographic attribute prediction.

Research work	Meta-data Features Used	Method Used	No. of Users	Application
[60]	Count of hashtags, user mentions	Multinomial Naive Bayes	10,000	Geolocation Prediction in Twitter
[61]	Count of hashtags, user mentions	SVM	956	Predicting political alignment in Twitter
[62]	Count of hashtags, user mentions, URLs	Gaussian Processes	5,191	Income Prediction in Twitter
[30]	Count of Hashtags, user mentions, URLs	SVM, Logistic Regression, Random Forest	3,184	Age Prediction in Twitter
[63]	Background color, first name	Naive Bayes, Decision Trees	180,000	Gender classification in Twitter
[64]	User location, User name	SVM	7,977	Gender, Race/Ethnicity prediction in Twitter
[65]	User name, profile image, neighborhood info	SVM	1,495	Gender,Race/Ethnicity Prediction in Twitter
[66]	User name, profile image, account creation date	Gradient-boosted Decision Trees	3,000	Race/Ethnicity Prediction in Twitter

3. Materials and methods

3.1. Data

In order to classify a user into an age group we need a dataset with (a) Twitter user id's, and (b) corresponding ages. A major problem for the age prediction task in Twitter is the limited availability of validated data annotated with the age of users. We use three datasets: two in English language, and one in Dutch.

[27] sampled Dutch Twitter users in the fall of 2012. They employed external annotators to annotate the chronological age using information available through tweets, the Twitter profile and external social media profiles such as Facebook and LinkedIn. In total, over 3000 Twitter users were annotated. However, not all of these Twitter profiles are currently active, leaving us with the profile information of 2150 users that we have included in our Dutch dataset. For the English corpus, we used the datasets from [30] and [46]. The dataset from [46] was created by applying pattern matching rules to the profile descriptions of the Twitter users. Through a process of iterative testing and refinement, they derived three rules for age extraction using variations of the following phrases:

1. I am X years old
2. Born in X

Table 3
Dataset description.

Dutch Dataset		English1 Dataset		English2 Dataset	
Age group	No. of users	Age group	No. of users	Age group	No. of users
0–20	1134(52.7%)	13–17	259(24.11%)	13–17	472(26.3%)
20–40	689(32%)	18–40	761(70.8%)	18–24	1046(58.3%)
40 plus	327(15.2%)	40 plus	64(5.9%)	25 plus	276 (15.4%)
Total	**2150**	Total	1074	**Total**	**1794**

3. X years old

where X can be a (typically) two-digit number or a date of the form DD/MM/YY or DD.MM.YYYY. Applying these three age-extraction rules to each user description field in the order presented above, [46] had collected a dataset of 1470 users. However, we could find 1074 of them still active.

The dataset from [30] was created by capturing self-reported and congratulatory birthday announcements/wishings by using the search parameters "Happy *n*th Birthday". Birthday tweets for ages 13 to 50 were collected on August 22, 2014, September 29, 2014, April 2, 2015, and June 21, 2015. Each birthday tweet was manually reviewed to determine whether a user could be identified from the birthday message, to determine whether the declared age seemed reasonable (rather than a joke exaggerating the age of the user for comedic effect), and to exclude "celebrity" users whose content feed may be curated for promotional and endorsement reasons. Out of the 3184 labeled users, we could only find 1794 twitter profiles active at the time of writing this paper.

More information about the three datasets can be found in Table 3. Despite the availability of these three age-labeled datasets from [27,46], and [30], a direct comparison of our results with theirs on our re-created datasets is limited owing to: (a) the differences in size due to some users' data being not available because of account inactivation, privacy mode setting changes, etc., and (b) the unavailability of the actual tweets used by them since the datasets have only published Twitter user-id and age. In the rest of the paper, we refer to our re-created dataset from [27,46], and [30] as *Dutch*, *English1*, and *English2*, respectively.

For each user, we collected his/her recent 200 tweets. Although the Twitter API allows collection of up to 3200 most recent tweets, prior studies have shown that examining more than 100 to 200 posts per user provides minimal gain in model performance when predicting user demographics [22,67].

Table 4 shows the distribution of numbers of hashtags, URLs, and media in the tweets for each dataset in different age categories. It is evident from the table that people in the age categories of 20+ and 40+ frequently cite hashtags in their tweets. More than half of the total tweets from these age categories have at least one hashtag in them. Similarly, almost one third of these users cite at least one URL in their tweets. With an average length of 8.5 words per tweet in our dataset, it is evident from the table that ignoring hashtags and URLs in tweets for age prediction results in important information loss. Twitter allows up to 4 media files to be included in a tweet (photos, videos, or animated GIFs). However, while some users actively make use of this feature, other users do not use media in their tweets. We also notice that the age-group of a user is not correlated with the number of media included in the tweet.

3.2. Feature engineering

We identified two broad categories of features, namely, language features, and Twitter-specific features to detect age. Many of these features have individually been explored in the literature [27–34]. These features are mainly derived from tweet contents (tweet text) and meta-information, such as Twitter user profile, network, and activity. These features were used in conjunction with a supervised machine learning framework to create a model for age detection. Previous research has explored SVM, Random Forest, and Logistic Regression methods with these features. However, we show that our novel features engineered from hashtags and URLs, and feature transformation to distributed representations of words and phrases incorporated in our CNN-based model produce better results on age prediction than the state-of-the-art.

All proposed features are investigated (see Section 4.2) using the features analysis technique to determine the best combination of features with the highest discriminative power. The proposed features are discussed in detail in the following subsections.

3.2.1. Language features

- **Linguistic Features**: part-of-speech (POS) n-grams.
- **Stylometric Features**: average sentence length, average word length, ratio of the number of emoticons to the number of words, number of elongated words (and non-standard spellings) used, ratio of the number of hashtags to the number of words, number of slang words, number of acronyms.
- **Features from pre-trained lexica**: Researchers in sociolinguistics have derived lexicons of words and phrases that correlate with different age groups. We use two such resources: EMNLP2014 [22], and WWBP [29]. We use logit transformed values from these lexicons.
- **Sentiment scores as features**: average number of tweets with positive/negative/neutral sentiment.

3.2.2. Twitter-specific features

- **Tweet Features**: earliest, latest, and average timestamp from among all 200 tweets of a user, number of geo-locatable tweets of a user, number of tweets favorited, number of tweets which are in-reply-to, number of tweets which are re-tweets, number of user mentions, number of tweets with media (photo, video, or animated GIF), average number of media files per tweet.
- **Twitter user profile features**: account creation date, listed-count, verified or not, geo-enabled or not, status-count.
- **Twitter social network features**: number of friends, number of followers, ratio of the number of friends to the number of followers, number of friends or followers with directed tweet exchanges, number of friends that are also followers.
- **Twitter hashtag features (new)**: most frequently co-occurring words with the hashtags used in user's tweets (described in more details in Section 3.3).
- **Twitter URL features (new)**: words from the titles of the pages pointed to by the URLs in user's tweets (described in more details in Section 3.3)

3.3. Method to extract features from tweet text, hashtags, and URLs

Our approach to include the URL content and hashtag as novel features makes use of innovative deep learning approach that genuinely combines word embedding and convolution neural

Table 4
Distribution of hashtags, URLs, media counts, and sentiment scores in tweets from Dutch, English1, and English2 datasets.

	Dutch	English1	English2
Total no. of users	2150	1074	1794
No. of tweets that have at least one **hashtag**	100,206	47,760	35,595
Avg no. of hashtags per tweet (age: 0–20 (Dutch), 0–17 (English1), 0–17 (English2))	0.30	0.12	0.18
Avg no. of hashtags per tweet (age: 21–40 (Dutch), 18–40 (English1), 28–24 (English2))	0.64	0.50	0.22
Avg no. of hashtags per tweet (age: 40 plus(Dutch), 40+ (English1), 25+ (English2))	0.71	0.35	0.52
No. of tweets that have at least one **URL**	138,301	61,212	58,715
Avg no. of URLs per tweet (age: 0–20 (Dutch), 0–18 (English1), 0–18 (English2))	0.21	0.1	0.13
Avg no. of URLs per tweet (age: 21–40 (Dutch), 18–40 (English1), 28–24 (English2))	0.28	0.22	0.21
Avg no. of URLs per tweet (age: 40 plus(Dutch), 40+ (English1), 25+ (English2))	0.38	0.25	0.38
No. of tweets that have at least one **media (video, photo, or GIF)**	34,099	33,600	42,641
Avg no. of media per tweet (age: 0–20 (Dutch), 0–18 (English1), 0–18 (English2))	0.07	0.17	0.25
Avg no. of media per tweet (age: 21–40 (Dutch), 18–40 (English1), 28–24 (English2))	0.08	0.19	0.23
Avg no. of media per tweet (age: 40 plus(Dutch), 40+ (English1), 25+ (English2))	0.10	0.18	0.17
Tweets with positive sentiment			
Avg no. of tweets with positive sentiment (age: 0–20 (Dutch), 0–18 (English1), 0–18 (English2))	0.39	0.42	0.43
Avg no. of tweets with positive sentiment (age: 21–40 (Dutch), 18–40 (English1), 28–24 (English2))	0.42	0.40	0.41
Avg no. of tweets with positive sentiment (age: 40 plus(Dutch), 40+ (English1), 25+ (English2))	0.35	0.44	0.43
Tweets with negative sentiment			
Avg no. of tweets with negative sentiment (age: 0–20 (Dutch), 0–18 (English1), 0–18 (English2))	0.17	0.19	0.18
Avg no. of tweets with positive sentiment (age: 21–40 (Dutch), 18–40 (English1), 28–24 (English2))	0.20	0.20	0.19
Avg no. of tweets with positive sentiment (age: 40 plus(Dutch), 40+ (English1), 25+ (English2))	0.19	0.18	0.16

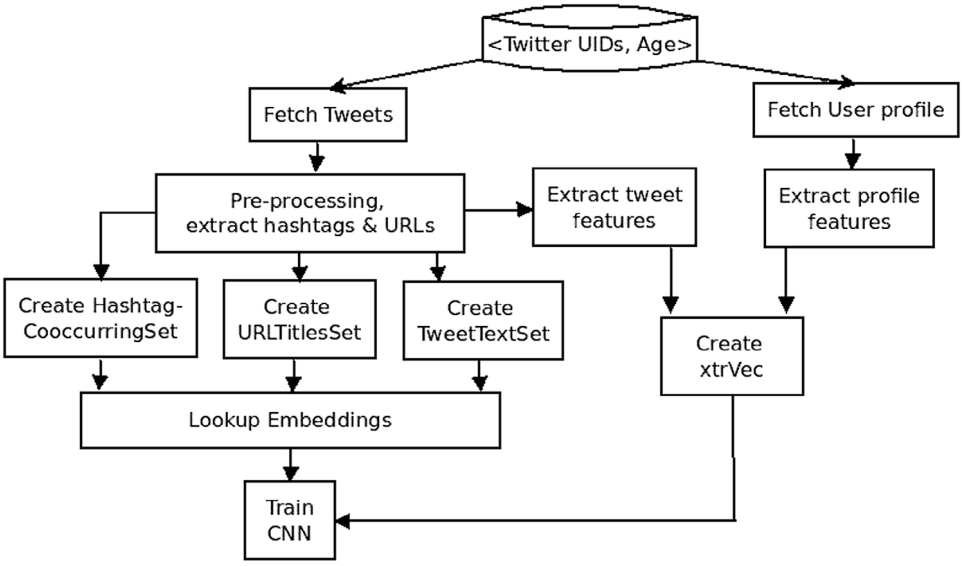

Fig. 1. Flowchart for feature extraction.

network architecture in order to extract useful patterns. A general skeleton of the approach is described in Fig. 1. We hypothesize that such features can play a role in the age detection task from Twitter data since it includes semantic information that reflects the interests of Twitter users that change with age. More specifically:

1. We extract the 200 most recent tweets after discarding re-tweets for each user to ensure a large enough dataset . We name this set the *TweetTextSet*.
2. We expand all URLs and hashtags in the *TweetTextSet* as follows:

Table 5
Example of feature extraction process.

Twitter user profile information	Tweet-specific features	Instances of the recent 200 Tweets of the user
UserID: 123xxxyy ScreenName: Location: Protected: Friends_count: Followers_count: Created_at:	created_at: 19.11.2017 in-reply-to-userid: geo-coordinates: retweet_status: retweet_count: reply_count: media_count_and_type: sentiment_score: __same_as_above__	*I am proud of you Google leadership! Thanks @sundarpichai and team for putting this in clear black and white.* https://www.blog.google/topics/ai/ai-principles/ .. *Especially the "AI applications we will not pursue".* *It is time for #BikeToWork guys. Let's save the environment.*
	__same_as_above__	*Summer has already started! Wooohooo ...*
	__same_as_above__	*Mother's day yet to come, but I cannot hold back.. Love you so much Mumma! Always an unwavering support #MothersDay*
Hashtags:	*#MothersDay, #BikeToWork*	
URLs:	https://www.blog.google/topics/ai/ai-principles/	

- We fetch the title of each linked web page. We name this set of titles the *URLTitlesSet*. We use a *metadata_parser*[1] for fetching the titles of the web-pages pointed to by the URLs in tweets.
- For each hashtag ht, we collect 1000 tweets containing this hashtag $HTTweets(ht)$. Since hashtags have different meanings at different times, we use a time window of +/- 10 days, counting from the date of the tweet mentioning the hashtag. The words/phrases in this set $HTTweets(ht)$ define the semantic context of that hashtag ht. We then find and select the most frequently co-occurring words/phrases with each hashtag and name these sets $CoOccurrences(ht)$. We name the union of all such sets $CoOccurrences(ht)$ as the set *HashtagCooccurringSet* which serves as a representative set of topics the user is interested in since it covers all hashtags cited in her recent 200 tweets.

3. We capture distributional semantics with word embeddings of words/phrases as follows:

   - We pre-train a word2vec model on a large collection of tweets and another word2vec model on GoogleNews/blogs/forums.
   - For each of the three sets *TweetTextSet*, *URLTitlesSet*, and *HashtagCooccurringSet*, we look up in pre-trained dictionaries to replace occurrences of words and phrases with their corresponding vectors. We ignore unknown words and phrases. For *TweetTextSet* and *HashtagCooccurringSet*, we use word2vec model trained on tweets, and for *URLTitlesSet*, we use the model trained on GoogleNews/blogs/forums.

   We use two different word2vec dictionaries to capture the semantics of the relevant contexts: for words and phrases in tweet texts, the model trained on tweets is found to be better fit than a model trained on generic texts such as blogs, news or forums. More specifically, in order to map Dutch *TweetTextSet* to the corresponding word embedding vectors, we pre-trained a word2vec model (200 dimensions) from 4.3 million Dutch tweets using Gensim library.[2] To map Dutch *URLTitlesSet* to the corresponding word embedding vectors, we utilize a pre-trained model from LREC2016 (320 dimensions).[3] These vectors were trained on Wikipedia, COW, Sonar500, and Roularta corpora. To map English *TweetTextSet* to the corresponding word embedding vectors, we use a Twitter word2vec model trained on 400 million tweets.[4] For mapping English *URLTitlesSet* to the corresponding word embedding vectors, we use pre-trained word2vec vectors (300 dimensions) on GoogleNews.[5]
4. We calculate the *min*, *max*, and avg vectors of the hashtag word embeddings (*HashtagCooccurringSet*) where *min* is the minimum value across all hashtag word embedding vectors in each dimension of the vector, and *max* and avg are maximum and average values, respectively. Avoiding computationally expensive option to find basis vectors for the span of word embedding vectors of *HashtagCooccurringSet*, we choose these three vectors as capturing the semantic space associated with user's hashtags which represents user's interests and activities.
5. We fetch Twitter metadata (e.g. tweet timestamps, see Section 3.2.2 for more details) for each user.
6. We extract additional features such as linguistic, stylometric, and Twitter-specific (see Section 3.2.2 for more details) from the metadata and *TweetTextSet*. These include the average number of media files in the users 200 most recent tweets, and the average sentiment score of these tweets. We create a normalized vector of these additional features. For English tweets, we use the Carnegie Mellon's TweetNLP suite[6] to tokenize the tweets and to assign Part-of-speech tags. For Dutch tweets, we use Frog[7] for tokenization and POS tagging. For sentiment score evaluation per tweet, we use NLTK's[8] implementation of VADER model [68].
7. We incorporate the word embeddings from *TweetTextSet* and *URLTitlesSet* and the additional features vector, as described above, into a Convolutional Neural Network based model for classification (see Section 3.4 for further information).

Our approach to include the URL content and hashtag as novel features in combination with word embeddings is summarized in Algorithms 13 and 2.

1 https://github.com/jvanasco/metadata_parser.
2 https://radimrehurek.com/gensim/index.html.
3 https://github.com/clips/dutchembeddings.
4 https://github.com/loretoparisi/word2vec-twitter.
5 https://code.google.com/archive/p/word2vec/.
6 http://www.cs.cmu.edu/~ark/TweetNLP.
7 http://languagemachines.github.io/frog/.
8 http://www.nltk.org.

Table 6
Some tweets for the hashtags #BikeToWork and #MothersDay.

Hashtag	A sample of tweets collected for the hashtag
#MothersDay	*So far my family's had our 1st #Easter, 1st #MothersDay and now 1st #MemorialDay without **Mom**. Also two **family** birthdays and a wedding anniversary. Just **another** day.* *Spring has sprung! Are you ready? #handmadejewelry #style #fashion #mothersday #jewelryfashion* *Piter @pawlicki777 on victory lap with his **mum** after yesterday's match. In Poland we had #mothersday this Sunday. Round of applause to Piter!* *An **event** doesn't have to be large to be a success. **Mother's Day** outing to Kinky Boots. We only do events, but we do them all!! Small, big, mainstream and outside the box. #events #eventplanner #mothersday* *It's **Mother's Day** today in France - this is me apparently #MothersDay #France*
#BikeToWork	*Good morning! Did you know today is #BikeTOWork day? Biking to **work** can save money, promote **health**, and is considered a good environmental option* *SWEEEEET! **kids** got served tickets for wearing helmets #biketowork for ice cream (he's tried to eat the one -does not taste like ice cream)* *The start of a new **week**..and not just any week, it's **bike** to **work** week!! Get those bikes shined up and hit the trails, road or whatever's between you and the **office**! #takeonPG #princegeorgenow #cityofpg #biketowork*

URL resolution for the link: https://www.blog.google/topics/ai/ai-principles/	
	"AI at Google: our principles"

Algorithm 1: Extract features

Input: Labeled training set of Twitter *User*s and corresponding ages.
Output: Trained CNN model.
Data: TweetW2V: word vector embedding dictionary trained on a large number of tweets,
GenericW2V: word vector embedding dictionary trained on generic text such as GoogleNews, Blogs,
N: number of tweets to fetch for each hashtag,
M: number of most-frequently co-occurring words to extract from N tweets for hashtags.

ProfileMetadata ← fetch metadata for each *User* from Twitter profile
Tweets ← fetch 200 recent tweets for each *User* after removing re-tweets
TweetVec ← *getEmbeddingVectors(Tweets, TweetW2V)*
URLs ← extract URLs in *Tweets*for each*User*
Titles ← fetch titles of *URLs* for each *User*
URLVec ← *getEmbeddingVectors(Titles, GenericW2V)*
hashtags ← *extracthashtagsinTweetsforeachUser*
HTTweets ← *fetchNt*w*eetsforeachhashtaginhashtagsforeachUser*
HTCoOcc ← *CoOccurrences(HTTweets, M)* for each *User*
HTVec ← *embed(HTCoOcc, TweetW2V)*
min, max, avg ← min, max and avg of *HTVec*
xtrVec ← extract additional features from Linguistic, Stylometric, Twitter profiles, ProfileMetadata, Average no. of media files, Average sentiment score
CNN ← *TRAIN_CNN*w*ithTweetVec, URLVec, HTVec, xtrVec*

Algorithm 2: Select the words that co-occur most with hashtags mentioned by a User.

Input: *HTTweets(ht)* for each hashtag *ht*.
Output: A set of most-frequently co-occurring words with all hashtags *ht*.
Data: M: Number of words/phrases to select.

coOccMap ← empty map of *User* to words
for *all hashtags ht* **do**
  Bag(ht) ← words and phrases from *HTTweets(ht)*
  Sort the *Bag(ht)* by number of co-occurrences.
  coOccMap ← *coOccMap* ∪ select top X words in *Bag(ht)*

Example

In order to exemplify the process of feature construction set, let us consider an example of Twitter Id 123*xxxyy* from one of the datasets.

- We present in Table 5, the profile information, some of the user's 200 most recent tweets, and their tweet-specific metadata we fetch using the Twitter API. In this example, we show only 4 of the user's 200 most recent tweets which have 2 hashtags and 1 URL altogether: #BikeToWork, #MothersDay, and URL: https://www.blog.google/topics/ai/ai-principles/.
- Next, for each of these 2 hashtags, we collect 1000 tweets from a time-window of [9.11.2017 – 29.11.2017]. Table 6 shows some of these tweets.
- Following the process in Algorithm 2, we find the most-frequently co-occurring terms with these hashtags in tweets. These are shown in bold-face letters in Table 6. For the hashtag #MothersDay, most-frequently occurring terms are: {*another, mother's day, event, mum, mom*}. And

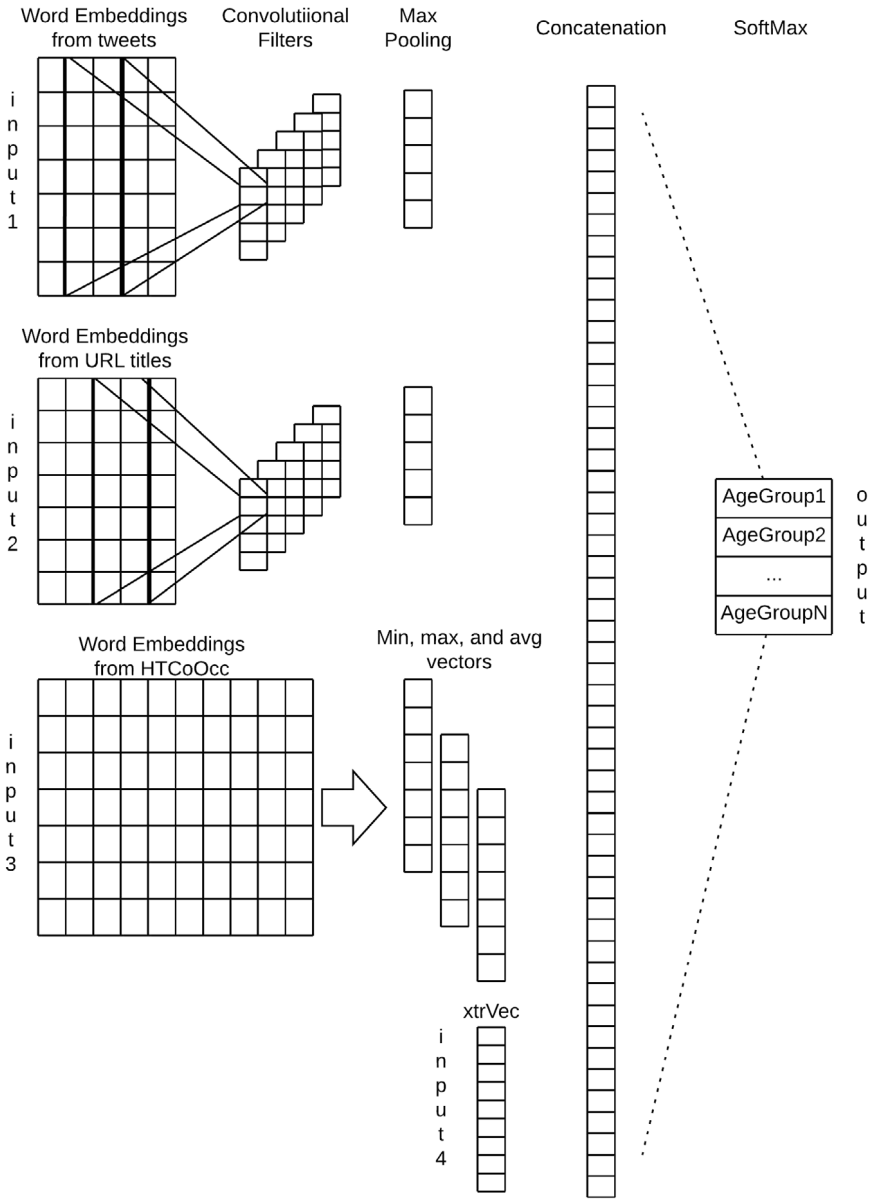

Fig. 2. Our CNN Model for age-category classification.

for the hashtag #BikeToWork, they are: {*health, work, kids, week, bike, office*}. Together, for the two hashtags, the set {*mom, event, work, office*} represents the most-frequently co-occurring terms.

- Next, we look-up these terms in a pre-trained word-embedding model trained on Twitter dataset.[9] and obtain a set of vectors $H_{1...M}$ and *min*, *max*, and *avg* vectors are computed as outlined in Section 3.3(4).
- We fetch the title of the page pointed to by the URL https://www.blog.google/topics/ai/ai-principles as "AI at Google: our principles". We look-up these terms in a pre-trained word-embedding model trained on GoogleNews[10] and obtain a set of vectors $U_{1...L}$.
- For each of the 200 recent tweets, after pre-processing and normalizing the length of a tweet to 30 terms, we look up the terms in a pre-trained word-embedding model trained on Twitter dataset. Together, word embedding vectors from 200 tweets form a set $T_{1...6000}$.

3.4. Convolutional neural network model

Inspired by [37,69], Fig. 2 illustrates our model based on convolutional neural network (CNN) for predicting the age category of a Twitter user. Our model is **innovative in two aspects**:

- Two separate input channels receive inputs from word embeddings of *TweetTextsSet* and *URLTitlesSet* and separate convolution filters of various sizes were used. This innovative design is proposed to prevent learning false associations between tweet words and words from the titles of webpages. The output of convolutional layer is passed through a non-linear activation function *ReLU*. Pooling layer aggregates vector elements by taking the maximum from each element of the convolutional feature map. Thus, these two output vectors after max-pooling represent features extracted from tweet texts and URLs for age-category prediction.

9 https://github.com/loretoparisi/word2vec-twitter.
10 https://code.google.com/archive/p/word2vec/.

- Since CNNs require fixed-sized homogeneous data sources, in order to utilize additional features, we propose another design innovation: the two vectors described above are concatenated with (a) a vector representing the additional features (Section 3.2.2), (b) the three vectors *min*, *max*, and *avg* calculated from word embeddings of *HashtagCooccurringSet*, and (c) a vector of features from pre-created lexicons after normalizing values to logits (-1 to +1). This concatenated vector is then fully connected with the output layer in soft-max setup. Since this vector is huge, we use drop-out method for regularization.

The details of the layers of the CNN architecture are as follows:

1. **Tweet sentences CNN (Model 1):**
   (a) Tweet Sentences Matrix (Input)
   All 200 recent tweets of a given user are represented by horizontal concatenation of d-dimensional word embeddings of its n constituent tokens. These word embeddings are pre-trained on a large twitter corpus from a given language (English, and Dutch). This generates a matrix $S \in R^{d \times n}$ which is inputted to the convolutional neural network model.
   (b) Convolutional Layer
   Convolution layer comprises of multiple filters of fixed length which are convolved with the input sentence matrix to extract discriminative word sequence patterns useful for classification. The convolution operation is defined as:
   $$c_i = \sum_{k,j} (S_{[i:i+h]})_{k,j} \cdot F^m_{k,j} \quad (1)$$
   where S is input sentence matrix, h is filter width, and $F^m_{k,j}$ are mth filter's coefficients. c_i is the value of the learned feature. The entire convolution of the mth filter with the input tweet produces $n - h + 1$ values which are concatenated together to produce a vector $\mathbf{c} \in R^{n-h+1}$. The vectors $\mathbf{c}$ are then aggregated over all m filters into a feature map matrix $C \in R^{m \times (n-h+1)}$.
   (c) Max Pooling
   The output of the convolutional layer is passed through a non-linear activation function such as *hardTanh* or *sigmoid* or *ReLU*. Pooling layer aggregates vector elements by taking the maximum from each element of the convolutional feature map. The resulting vector is $\mathbf{C_{pooled}} \in R^{m \times 1}$.
2. **URL titles sentences CNN (Model 2):**
   (a) URL Titles Sentences Matrix (Input)
   All title sentences of URLs in all of a given user's tweet are represented by horizontal concatenation of k-dimensional word embeddings of its n constituent tokens. These word embeddings are pre-trained on a large corpus of blogs, news-posts, and generic text in a given language (English, and Dutch). This generates a matrix $S \in R^{k \times n}$ which is input to the convolutional neural network model.
   (b) Convolutional Layer and Max Pooling layers identical to Tweet sentences.
3. **Hashtag vectors**
   From the *HashtagCoocurringSet* comprising of most frequently co-occurring words with all hashtags of from the recent 200 tweets of a given user, we choose three embedding vectors *min*, *max*, and *avg* from the pre-trained word embedding model on generic text (English, and Dutch) as explained in Section 3.3. This generates three vectors $V1, V2$, and $V3 \in R^k$ which is input to the concatenation layer (*merge* layer of keras[11]).
4. **Concatenation Layer and Dropout**
   Output of max pooling from (i) Tweet Sentence CNN and URL titles CNN, (ii) three hashtag vectors, and (iii) other features (stylometric, social network features, etc.) are concatenated. Since this vector is huge, in order to avoid overfitting, we use the dropout method proposed by [70]. Each dimension is randomly set to 0 using a Bernoulli distribution $B(p)$ where p is a hyper-parameter. In addition, we complement this method of regularization with $L2$-Regularization of softmax parameters. After dropout, the vector is passed onto the softmax layer.
5. **Softmax** Output from the concatenation layer $\mathbf{C_{concat}} \in R^m$ is used for softmax regression which returns the class $\hat{y} \in \{1, K\}$ with largest probability. i.e.,
   $$\hat{y} = \arg\max_j P(y = j \mid \mathbf{x}, \mathbf{w}, \mathbf{a}) \quad (2)$$
   $$q = \arg\max_j \frac{e^{(\mathbf{C_{concat}} \mathbf{w_j} + a_j)}}{\sum_{k=1}^{K} e^{(\mathbf{C_{concat}} \mathbf{w_j} + a_j)}} \quad (3)$$
   where $\mathbf{w_j}$ denotes the weights vector of class j and a_j the bias of class j.

3.4.1. CNN training details

In order to train the above model, for each dataset, data was split into 85% training and 15% for validation sample. Batch size for training CNN was kept at 200 tweets. Since tweets differ in length, we limited each tweet to a maximum of 30 words (excluding emoticons, hashtags, and URLs). Tweets with shorter than 3 words or longer than 30 words were discarded from further processing. While the average length of a tweet in our datasets was small, some users have availed the newly-introduced (in November 2017) feature of Twitter supporting longer tweets (up to 280 characters). Since the limit of 30 words covers 100% of our collected tweets, we normalize the length of each tweet to be 30 words. Tweets shorter than 30 words were padded with a special PAD token. Since each tweet is different in content from others even by the same user, we carefully adjusted the sizes of kernel masks in order not to learn spurious features. In this way, we ensured that the convolution kernel masks did not move over different tweets. We limited URL title words to 25. If a URL title was shorter, we again padded the title with a special PAD token. If a title was longer than 25 words, we ignored the rest of the words. Since search engines typically display the first 50–60 characters of the titles, most web pages do not have titles exceeding the length of 25 words. Also, in our two datasets, we found that 93% titles are of shorter length than 25 words. Similarly, we found only 0.8%, 1.2%, and 2.5% words from tweets and URLs not available in pre-trained dictionaries for English1, English2 and Dutch datasets, respectively. We found that for Dutch dataset, the number of words not available in pre-trained word embedding resource was almost double than that for English datasets. This is indicative of lack of adequately sized resources for Dutch. The following choice of **hyper-parameters** was made: for both datasets, we use *ReLU* as non-linear activation function, filter windows of sizes 3, 4, 5, 6 with 128 feature maps each, and mini-batch size of 200. The loss is minimized using the Adam optimizer [71]. For **regularization**, we use the dropout method proposed by [70] after the max pooling layer with $p = 0.5$. In addition, we complement this method of regularization with $L2$-Regularization of softmax parameters.

[11] https://keras.io.

Table 7
Feature correlation analysis on Dutch, English1, and English2 datasets.

Features	Dutch dataset			English1 dataset			English2 dataset		
	0–20	21–40	40 plus	13–17	18–40	40 plus	13–17	18–24	25 plus
Linguistic features:									
Count of the term "family"			0.13(+)			0.23(+)			0.16(+)
Count of the term "college"		0.22(+)						0.17(+)	
Count of the term "lol"	0.18(+)		0.08(-)						
Stylometric features:									
Ratio of emoticons to words	0.09(+)					0.11(-)	0.18(+)		
Number of non-standard spellings			0.19(-)	0.14(+)					
Tweet features:									
No. of tweets favorited									
No. of user mentions		0.14(+)			0.21(+)				
Twitter user profile features:									
Age of Twitter account	0.31(-)		0.13(+)	0.27(-)		0.19(+)	0.24(-)		0.11(+)
Statuses count	0.10(+)								
Twitter social network features:									
Ratio of friends to followers		0.19(+)			0.24(+)				0.09(+)
No. of friends that are also followers	0.12(+)								
Hashtag words									
Count of the term "music"	0.23(+)				0.11(+)			0.13(+)	
Count of the term "love"			0.12(+)						0.08(+)
URL title words									
Count of the term "news"		0.19(+)							0.1(+)
Count of the term "gadget"					0.16(+)				
Media Count									
Value of average count of media files in tweets								0.01(+)	
Average sentiment score									
Value of average sentiment score									

4. Experiments and results

In order to show the effectiveness of our novel features, we conduct both regression and classification (into pre-defined age category bins) experiments. Below we define various combinations of features as baseline features to compare performance of regression and classification models on the datasets. The feature representation for tweet text, URL titles, and *HashtagCoocurringSet* in the baseline setting is bag-of-ngrams. While we show the improvement in results using these features, further improvement is yielded by using distributed representations of words incorporated in our 2-channel novel CNN model.

4.1. Baseline features

- B1: Lexical (Bag-of-words(BoW)) features: Unigrams, bigrams, and trigrams extracted from only tweet texts.
- B2: B1 and BoW representation of most frequently co-occurring words with hashtags in user's tweets (*HashtagCooccurringSet*)
- B3: B1 and BoW representation of words from titles of the URLs in user's tweets (*URLTitlesSet*)
- B4: B1 and Linguistic, Stylometric, Twitter-specific features
- B5: all features from B1, B2, B3, B4
- B6: B5 and features derived from external pre-trained age-specific lexica

4.2. Feature analysis

Following [30], we use Cohen's d measure to show the effectiveness of our features in classification of age-groups of Twitter users. We first convert Chi-square values into correlation coefficient r by using the formula $r = \sqrt{\frac{X^2}{N}}$. This value was then converted into a Cohen's d effect size per the formula $\frac{2r}{1-r^2}$. p-value was chosen as 0.001. Table 7 shows the top predictive features for each age category in all three datasets. The plus sign (+) shows the direction of association.

Table 8
Ridge regression results with Baseline features on Dutch, English1, English2 datasets.

	Dutch Dataset	English1	English2
B1	0.53	0.58	0.64
B2	0.56	0.61	0.67
B3	0.58	0.62	0.69
B4	0.55	0.67	0.70
B5	0.64	0.67	0.77
B6	0.65	0.69	0.81

4.3. Linear regression (ridge) with baseline features

We use ridge regression for predicting age as a continuous variable with different feature sets (see baseline features above). Feature sets $B1$ and $B2$ include the content of hashtags and URLs respectively in their bag-of-ngrams representation. Since the features as discussed above are very high dimensional, we used principal component analysis (PCA) to reduce the number of features. With using 10% of data as validating sample, we set the regularization parameter for ridge regression. As goodness-of-fit statistics, we have used R-squared (R^2) statistic. The results of regression are shown in Table 8.

4.4. Classification results

In order to evaluate the performance of novel 2-channel CNN model with the new features proposed on the above datasets, we compare the results with those obtained from (a)Support Vector Machines (SVM) classifier (with linear kernel), (b) Logistic regression, and (c) Random forest, using various combinations of features. We also use the additional language and social media specific features that have been utilized in the previous works. While the feature representation for SVM, Random Forest, and Logistic Regression is bag-of-ngrams, for CNN, we used distributed representations to capture semantic correlations between words in a dense semantic vector space induced by word embedding models. Feature sets used in our CNN model are:

- W1: word embeddings (pre-trained on tweets) of words/phrases from *TweetTextSet*
- W2: W1 to the input layer of CNN and *min*, *max*, and *avg* vectors derived from word embeddings of *HashtagCooccurringSet* concatenated with the max-pooled vector before soft-max
- W3: W1 to one branch of the input layer of CNN and word embeddings of *URLTitlesSet* to the other branch of input layer
- W4: W1 to the input layer of CNN and a vector constructed from Linguistic, Stylometric, and Twitter-specific features concatenated with the max-pooled vector before soft-max
- W5: W3 to the input layer of CNN, and all other features including *min*, *max*, and *avg* from *HashtagCooccurringSet* concatenated with the max-pooled vector before soft-max
- W6: same setup as W5 but with additional features derived from external pre-trained age-specific lexica concatenated with the max-pooled vectors

Tables 9–11 collect results on Dutch and English datasets, respectively. Included are the precision, recall, and micro-averaged F1 scores obtained with SVM, Logistic Regression, and Random Forest classification with various feature combinations. Because of the significant differences in the number of samples in each age-group, micro-averaged F1 score serves as a better indicator of classification accuracy. Micro-averaged precision and recall are defined as:

$$Precision_{\mu} = \frac{\sum_{ag \in AgeGroups} TP_{ag}}{\sum_{ag \in AgeGroups} TP_{ag} + \sum_{ag \in AgeGroups} FP_{ag}} \tag{4}$$

$$Recall_{\mu} = \frac{\sum_{ag \in AgeGroups} TP_{ag}}{\sum_{ag \in AgeGroups} TP_{ag} + \sum_{ag \in AgeGroups} FN_{ag}} \tag{5}$$

where TP_{ag}, FP_{ag}, and FN_{ag} are the true positives, false positives, and false negatives for age group class ag, respectively. Micro-averaged F1 score is defined as the harmonic mean of $Precision_{\mu}$ and $Recall_{\mu}$:

$$F1_{\mu} = \frac{2 \times Precision_{\mu} \times Recall_{\mu}}{Precision_{\mu} + Recall_{\mu}} \tag{6}$$

4.5. Discussion

It is evident from Table 7 that lexical features from tweet text, most frequently co-occurring words with hashtags, and words from URL titles indeed are discriminative for age-group classification. For example, number of occurrences of the word "family" is much higher among the older age-groups than among the younger. Similarly, slang words (such as, "lol") are more frequently observed in tweets of younger age-group. Number of words with non-standard spellings was found to be negatively correlated with age among the older age-group of Twitter users while it was positively correlated in the younger group. For the millennials (ages between 18 and 40), the number of user mentions in the tweets was observed much higher than among other groups suggesting that this generation practices engaging online social communication. This is further evidenced by noting their friends to followers ratio: higher values of the ratio indicating their proclivity to be a part of a social network. Use of hashtags and URLs is prevalent across all age categories (ref. Table 4), but the 'topics' as indicated by the hashtags or URLs differ among different age groups. For example, the frequency of the word 'music' was more pronounced in the *HashtagCooccurringSet* of younger users; whereas the word 'news' was found more prominent among the older groups.

Average number of media files per tweet of a user was not found to be discriminative for age-category classification (ref. Table 7. Very few users availed the feature of attaching media to their tweets and such users are from all age groups rather than being restricted to a particular age group (ref. Table 4. Similarly, we found that average sentiment score of all 200 tweets of a given user does not help in classification. Distribution of number of tweets with positive, negative, and neutral sentiment is almost similar across all age groups (ref. Table 4. Sentiment scores were evaluated for each tweet and while the average of sentiment score across all 200 recent tweets of a user may give insight into the personality of that user, does not characterize the behavior of his/her group.

Encouraged by the findings of feature importance evaluation, we carried out ridge regression experiments (ref. Table 8). While the results indicate performance improvement by including *HashtagCooccurringSet* and *URLTitlesSet*, the final results of including all features (after PCA) are far from desirable. This shows that the underlying assumption of multivariate Gaussian distribution of lexical features is not accurate for tweets and regression model trained on reduced feature set fails to achieve the desired fine-grained age prediction. In order to capture non-linear correlations between semantic dimensions of different words in a principled manner, we propose to use our CNN model trained with distributed representations (word embedding vectors) of words.

Tables 9–11, show experiments with SVM, Logistic Regression, Random Forest, and our model. Each table shows results of these four experiments on each of the three datasets: Dutch [27], English1 [46], and English2 [30]. We show the effectiveness of various features in each of these models. It is observed that all approaches yield higher performance on English1 dataset which can be attributed to its much smaller size (almost half) compared to the other two datasets. Further, there are differences in the method of data collection: English1 dataset relied on self-declared age value in profile descriptions which was manually verified, while English2 dataset was created by applying pattern-matching rules on congratulatory tweets, and Dutch dataset creation relied on external sources such as Facebook or LinkedIn. Also, the dataset creation of Dutch users may introduce a sampling bias since they select the users who are in the same social subnetwork.

From Tables 9–11 it is observed that older age groups across all three datasets yield lower accuracy. On the other hand, better performance results are noticed for the 0–20 age group (Dutch dataset), 18–40 age group (English 1 data), and 18–24 age group (English 2 data). Clearly, our supervised machine learning approach yields better results when the amount of training data is higher: Table 3 shows that these age groups represent the largest proportions in the respective datasets; whereas for older age groups, the number of data samples is much smaller across all datasets.

As can be seen from the experimental results utilizing baseline features (Section 4.1) from the tables (A,B, and C in Tables 9–11), including the most frequently co-occurring words with hashtags in the user's tweets into the bag-of-words model (column 2) actually degrades the performance of age-prediction. Hashtags are used to index keywords or topics so as to categorize tweets and to allow people to easily follow the topics they are interested in. In this experiment, we attempt to capture the topics a user is interested in by finding hashtag-relevant words from other tweets that include the same hashtag. In order to overcome the problem of *topic drift*, we only use tweets in a window of −10 to +10 days from the tweet containing the hashtag. Based on the hypothesis that a person's age is correlated with the topics he is interested in, we expect to see improvement in the accuracy. However, we notice that by bringing in hashtag-relevant words

Table 9
Results on Dutch dataset. B1–B6 and W1–W6 are explained in the text.

(A)

SVM_Dutch	B1	B2	B3	B4	B5	B6
Age 0-20	P:0.80 R:0.81 F:0.80	P:0.78 R:0.83 F:0.80	P:0.77 R:0.82 F:0.79	P:0.81 R0.84 F:0.82	P:0.82 R:0.84 F:0.83	P:0.83 R:0.85 F:0.84
Age 20-40	P:0.58 R:0.64 F:0.61	P:0.52 R:0.69 F:0.59	P:0.51 R:0.63 F:0.56	P:0.60 R:0.66 F:0.63	P:0.62 R:0.65 F:0.63	P:0.61 R:0.67 F:0.64
Age 40 plus	P:0.84 R:0.36 F:0.50	P:0.78 R:0.40 F:0.53	P:0.80 R:0.41 F:0.54	P:0.85 R:0.42 F:0.56	P:0.85 R:0.44 F:0.58	P:0.87 R:0.45 F:0.59
Micro-averaged	P:0.76 R:0.65 **F:0.70**	P:0.73 R:0.67 **F:0.70**	P:0.72 R:0.68 **F:0.70**	P:0.76 R:0.67 **F:0.71**	P:0.76 R:0.69 **F:0.72**	P:0.77 R:0.69 **F:0.73**

(B)

RF_Dutch	B1	B2	B3	B4	B5	B6
Age 0-20	P:0.64 R:0.66 F:0.65	P:0.66 R:0.70 F:0.68	P:0.69 R:0.72 F:0.70	P:0.71 R:0.71 F:0.71	P:0.73 R:0.72 F:0.72	P:0.74 R:0.74 F:0.74
Age 20-40	P:0.41 R:0.44 F:0.42	P:0.43 R:0.42 F:0.42	P:0.41 R:0.42 F:0.41	P:0.43 R:0.45 F:0.44	P:0.48 R:0.50 F:0.49	P:0.51 R:0.54 F:0.52
Age 40 plus	P:0.68 R:0.41 F:0.52	P:0.69 R:0.42 F:0.55	P:0.69 R:0.45 F:0.60	P:0.70 R:0.47 F:0.59	P:0.71 R:0.48 F:0.61	P:0.71 R:0.50 F:0.62
Micro-averaged	P:0.68 R:0.63 **F:0.65**	P:0.69 R:0.63 **F:0.65**	P:0.69 R:0.66 **F:0.67**	P:0.70 R:0.71 **F:0.70**	P:0.72 R:0.70 **F:0.71**	P:0.72 R:0.69 **F:0.71**

(C)

LR_Dutch	B1	B2	B3	B4	B5	B6
Age 0-20	P:0.78 R:0.80 F:0.79	P:0.76 R:0.79 F:0.77	P:0.78 R:0.81 F:0.79	P:0.83 R:0.84 F:0.83	P:0.82 R:0.85 F:0.83	P:0.83 R:0.83 F:0.83
Age 20-40	P:0.50 R:0.60 F:0.56	P:0.52 R:0.65 F:0.58	P:0.53 R:0.64 F:0.59	P:0.56 R:0.63 F:0.63	P:0.60 R:0.65 F:0.63	P:0.61 R:0.67 F:0.64
Age 40 plus	P:0.80 R:0.40 F:0.52	P:0.81 R:0.42 F:0.56	P:0.82 R:0.45 F:0.61	P:0.82 R:0.43 F:0.59	P:0.82 R:0.44 F:0.60	P:0.85 R:0.50 F:0.62
Micro-averaged	P:0.74 R:0.63 **F:0.68**	P:0.71 R:0.65 **F:0.68**	P:0.73 R:0.70 **F:0.71**	P:0.74 R:0.68 **F:0.72**	P:0.79 R:0.67 **F:0.73**	P:0.78 R:0.68 **F:0.72**

(D)

CNN_Dutch	W1	W2	W3	W4	W5	W6
Age 0-20	P:0.82 R:0.87 F:0.84	P:0.84 R:0.89 F:0.86	P:0.85 R:0.91 F:0.88	P:0.83 R:0.85 F:0.84	P:0.87 R:0.91 F:0.89	P:0.88 R:0.92 F:0.90
Age 20-40	P:0.60 R:0.68 F:0.64	P:0.62 R:0.72 F:0.68	P:0.63 R:0.73 F:0.69	P:0.61 R:0.71 F:0.67	P:0.66 R:0.74 F:0.71	P:0.68 R:0.75 F:0.71
Age 40 plus	P:0.83 R:0.45 F:0.58	P:0.85 R:0.49 F:0.62	P:0.87 R:0.52 F:0.63	P:0.83 R:0.48 F:0.61	P:0.89 R:0.57 F:0.69	P:0.88 R:0.60 F:0.71
Micro-averaged	P:0.77 R:0.72 **F:0.74**	P:0.78 R:0.75 **F:0.76**	P:0.80 R:0.76 **F:0.78**	P:0.79 R:0.73 **F:0.76**	P:0.82 R:0.77 **F:0.80**	P:0.83 R:0.80 **F:0.82**

Table 10
Results on English1 dataset. B1–B6 and W1–W6 are explained in the text.

(A)

SVM_English1	B1	B2	B3	B4	B5	B6
Age 13-17	P:0.66 R:0.69 F:0.67	P:0.68 R:0.70 F:0.69	P:0.69 R:0.72 F:0.71	P:0.72 R:0.73 F:0.72	P:0.75 R:0.76 F:0.75	P:0.76 R:0.78 F:0.77
Age 18-40	P:0.76 R:0.78 F:0.77	P:0.78 R:0.79 F:0.78	P:0.79 R:0.80 F:0.79	P:0.79 R:0.81 F:0.80	P:0.82 R:0.83 F:0.82	P:0.83 R:0.82 F:0.82
Age 40 plus	P:0.59 R:0.65 F:0.62	P:0.61 R:0.64 F:0.63	P:0.64 R:0.63 F:0.63	P:0.65 R:0.61 F:0.63	P:0.68 R:0.63 F:0.65	P:0.68 R:0.63 F:0.65
Micro-averaged	P:0.71 R:0.74 **F:0.72**	P:0.76 R:0.77 **F:0.76**	P:0.76 R:0.78 **F:0.77**	P:0.76 R:0.79 **F:0.78**	P:0.80 R:0.81 **F:0.80**	P:0.79 R:0.81 **F:0.80**

(B)

RF_English1	B1	B2	B3	B4	B5	B6
Age 13-17	P:0.61 R:0.65 F:0.63	P:0.59 R:0.67 F:0.63	P:0.58 R:0.67 F:0.62	P:0.62 R:0.70 F:0.66	P:0.63 R:0.73 F:0.68	P:0.63 R:0.74 F:0.69
Age 18-40	P:0.79 R:0.82 F:0.81	P:0.76 R:0.85 F:0.80	P:0.73 R:0.83 F:0.78	P:0.75 R:0.88 F:0.81	P:0.76 R:0.89 F:0.82	P:0.77 R:0.88 F:0.83
Age 40 plus	P:0.58 R:0.60 F:0.59	P:0.57 R:0.62 F:0.60	P:0.56 R:0.61 F:0.59	P:0.59 R:0.64 F:0.62	P:0.60 R:0.67 F:0.64	P:0.61 R:0.66 F:0.64
Micro-averaged	P:0.77 R:0.79 **F:0.78**	P:0.74 R:0.84 **F:0.79**	P:0.71 R:0.80 **F:0.76**	P:0.72 R:0.84 **F:0.78**	P:0.73 R:0.83 **F:0.79**	P:0.74 R:0.82 **F:0.78**

(C)

LR_English1	B1	B2	B3	B4	B5	B6
Age 13-17	P:0.70 R:0.71 F:0.70	P:0.71 R:0.74 F:0.73	P:0.70 R:0.73 F:0.71	P:0.72 R:0.74 F:0.73	P:0.74 R:0.76 F:0.75	P:0.73 R:0.77 F:0.75
Age 18-40	P:0.81 R:0.84 F:0.82	P:0.80 R:0.86 F:0.83	P:0.79 R:0.88 F:0.84	P:0.83 R:0.88 F:0.85	P:0.84 R:0.89 F:0.87	P:0.85 R:0.87 F:0.86
Age 40 plus	P:0.61 R:0.64 F:0.62	P:0.61 R:0.63 F:0.62	P:0.59 R:0.62 F:0.61	P:0.61 R:0.62 F:0.61	P:0.63 R:0.63 F:0.63	P:0.61 R:0.63 F:0.62
Micro-averaged	P:0.77 R:0.80 **F:0.78**	P:0.76 R:0.81 **F:0.79**	P:0.74 R:0.82 **F:0.78**	P:0.78 R:0.82 **F:0.80**	P:0.79 R:0.83 **F:0.81**	P:0.79 R:0.82 **F:0.81**

(D)

CNN_English1	W1	W2	W3	W4	W5	W6
Age 13-17	P:0.74 R:0.77 F:0.76	P:0.75 R:0.77 F:0.76	P:0.76 R:0.80 F:0.78	P:0.78 R:0.82 F:0.80	P:0.80 R:0.84 F:0.82	P:0.80 R:0.85 F:0.83
Age 18-40	P:0.85 R:0.85 F:0.85	P:0.87 R:0.86 F:0.86	P:0.87 R:0.89 F:0.88	P:0.86 R:0.90 F:0.88	P:0.87 R:0.91 F:0.89	P:0.86 R:0.90 F:0.88
Age 40 plus	P:0.70 R:0.64 F:0.67	P:0.72 R:0.63 F:0.68	P:0.72 R:0.64 F:0.68	P:0.74 R:0.65 F:0.70	P:0.76 R:0.68 F:0.72	P:0.76 R:0.70 F:0.73
Micro-averaged	P:0.83 R:0.84 **F:0.84**	P:0.85 R:0.83 **F:0.84**	P:0.85 R:0.86 **F:0.85**	P:0.84 R:0.85 **F:0.84**	P:0.85 R:0.89 **F:0.87**	P:0.83 R:0.88 **F:0.86**

for all hashtags from all 200 recent tweets introduce too much noise: while false negatives decrease resulting in improved recall, false positives increase resulting in poorer precision. Similar observation is made about including words from the URL titles (column 3). Utilizing linguistic, stylometric, and Twitter-specific features along with 1- to −3-grams from tweet texts improve the precision and recall over the basic BoW model. This confirms the sociolinguistic hypothesis that linguistic and stylometric features serve as indicators of person's age. Finally, exploiting predictive lexica which are pre-trained for age, such as EMNLP2014 [22] and WWBP [29] help in improving the accuracy slightly. Despite such pre-created lexica having potential to improve the accuracy for age prediction, we observe that because such lexica were trained on Facebook and our data is from Twitter, they fail to achieve higher accuracy as expected (since there is a difference in discourse styles of Facebook and Twitter).

(D) in Tables 9–11 show results of our experiments based on the use of CNN as our classification model in combination with word embeddings for tweet words/phrases. We find that utilizing word embeddings into our CNN model improves on the baseline of using BoW (compare column 1 of A, B, and C with that of D in each of the tables) since CNN learns complex features associating different dimensions of word embedding vectors of a tweet word sequence. In the case of URL titles, replacing them with a

Table 11
Results on English2 dataset. B1–B6 and W1–W6 are explained in the text.

(A)

SVM_English2	B1	B2	B3	B4	B5	B6
Age 13-17	P:0.68 R:0.70 F:0.69	P:0.67 R:0.72 F:0.69	P:0.66 R:0.72 F:0.68	P:0.70 R:0.74 F:0.72	P:0.71 R:0.76 F:0.73	P:0.72 R:0.76 F:0.74
Age 18-24	P:0.77 R:0.73 F:0.75	P:0.75 R:0.76 F:0.75	P:0.74 R:0.75 F:0.74	P:0.79 R:0.77 F:0.78	P:0.80 R:0.80 F:0.80	P:0.81 R:0.80 F:0.81
Age 25 plus	P:0.61 R:0.64 F:0.62	P:0.59 R:0.67 F:0.63	P:0.59 R:0.64 F:0.62	P:0.63 R:0.67 F:0.65	P:0.66 R:0.70 F:0.68	P:0.65 R:0.72 F:0.69
Micro-averaged	P:0.71 R:0.71 **F:0.71**	P:0.70 R:0.73 **F:0.72**	P:0.70 R:0.72 **F:0.71**	P:0.73 R:0.74 **F:0.74**	P:0.74 R:0.76 **F:0.75**	P:0.75 R:0.76 **F:0.76**

(B)

RF_English2	B1	B2	B3	B4	B5	B6
Age 13-17	P:0.60 R:0.62 F:0.61	P:0.61 R:0.60 F:0.60	P:0.62 R:0.63 F:0.62	P:0.63 R:0.62 F:0.62	P:0.64 R:0.64 F:0.64	P:0.66 R:0.67 F:0.66
Age 18-24	P:0.63 R:0.64 F:0.63	P:0.65 R:0.66 F:0.65	P:0.64 R:0.65 F:0.65	P:0.67 R:0.69 F:0.68	P:0.69 R:0.71 F:0.70	P:0.69 R:0.71 F:0.70
Age 25 plus	P:0.52 R:0.57 F:0.54	P:0.53 R:0.58 F:0.55	P:0.53 R:0.59 F:0.55	P:0.54 R:0.59 F:0.57	P:0.57 R:0.60 F:0.58	P:0.59 R:0.61 F:0.60
Micro-averaged	P:0.61 R:0.61 **F:0.61**	P:0.61 R:0.63 **F:0.62**	P:0.62 R:0.62 **F:0.62**	P:0.63 R:0.64 **F:0.64**	P:0.64 R:0.66 **F:0.65**	P:0.65 R:0.66 **F:0.66**

(C)

LR_English2	B1	B2	B3	B4	B5	B6
Age 13-17	P:0.70 R:0.71 F:0.70	P:0.68 R:0.71 F:0.69	P:0.68 R:0.70 F:0.69	P:0.70 R:0.72 F:0.71	P:0.72 R:0.73 F:0.72	P:0.72 R:0.74 F:0.73
Age 18-24	P:0.73 R:0.73 F:0.73	P:0.73 R:0.74 F:0.72	P:0.74 R:0.75 F:0.74	P:0.75 R:0.74 F:0.74	P:0.77 R:0.76 F:0.76	P:0.79 R:0.80 F:0.79
Age 25 plus	P:0.55 R:0.60 F:0.57	P:0.56 R:0.62 F:0.59	P:0.56 R:0.61 F:0.58	P:0.57 R:0.61 F:0.59	P:0.57 R:0.62 F:0.60	P:0.59 R:0.63 F:0.61
Micro-averaged	P:0.70 R:0.69 **F:0.70**	P:0.71 R:0.70 **F:0.71**	P:0.71 R:0.72 **F:0.71**	P:0.72 R:0.71 **F:0.72**	P:0.73 R:0.73 **F:0.73**	P:0.75 R:0.77 **F:0.76**

(D)

CNN_English2	W1	W2	W3	W4	W5	W6
Age 13-17	P:0.70 R:0.73 F:0.72	P:0.73 R:0.75 F:0.74	P:0.72 R:0.74 F:0.73	P:0.71 R:0.74 F:0.73	P:0.75 R:0.77 F:0.76	P:0.76 R:0.79 F:0.78
Age 18-24	P:0.79 R:0.75 F:0.77	P:0.83 R:0.79 F:0.81	P:0.81 R:0.77 F:0.79	P:0.81 R:0.79 F:0.80	P:0.85 R:0.80 F:0.83	P:0.86 R:0.82 F:0.84
Age 25 plus	P:0.66 R:0.69 F:0.68	P:0.67 R:0.73 F:0.70	P:0.68 R:0.74 F:0.71	P:0.69 R:0.72 F:0.70	P:0.70 R:0.73 F:0.71	P:0.71 R:0.75 F:0.73
Micro-averaged	P:0.77 R:0.73 **F:0.75**	P:0.78 R:0.75 **F:0.77**	P:0.78 R:0.76 **F:0.77**	P:0.77 R:0.75 **F:0.76**	P:0.80 R:0.79 **F:0.79**	P:0.81 R:0.82 **F:0.81**

sequence of corresponding word embedding vectors and selecting features using convolutional filters improve the performance of the system (compare column 3 of A, B, and C with that of D in each of the tables). We also notice that instead of directly using *HashtagCooccurringSet* words for classification, utilizing three vectors *min*, *max*, and *avg* derived from word embeddings of *HashtagCooccurringSet* yields improvement in precision and recall both (compare column 2 of A,B, and C with that of D in each of the tables). Since much less noise is included as opposed to the method of including all words, both false positives and false negatives decrease resulting in improvement in precision and recall. Finally, similar to SVM baseline model, including linguistic, stylometric, Twitter-specific features and using external lexica help to improve the accuracy further.

Overall, using our CNN-based architecture along with novel features improves the micro-F1 score by 12.3%, 9.8% and 6.6% for Dutch, English1 and English2 datasets, respectively when compared against the best results of SVM, Random Forest, and Logistic Regression models employing bag-of-ngrams representation of baseline features.

5. Conclusion

In this paper, we proposed a novel way to include features derived from hashtags and URLs from tweets for age prediction of Twitter users. We show that using distributed representations incorporated into convolutional neural network improve the accuracy over the baseline bag-of-words model. Augmenting these features with features derived from URLs and hashtags further improves the precision and recall. We examined the effect of adding novel features incrementally and conclude that our model outperforms the baseline by 12.3%, 9.8% and 6.6% for Dutch [27], English1 [46], and English2 [30] datasets, respectively.

Present-day social media platforms facilitate effective social communication by offering several meta-data features that users may avail to make their messages more meaningful. The proposed method presents a way to include information from URLs and Hashtags for analytics of social media messages. While the evaluation of accuracy of our approach is limited by the amount of the labeled data available for age demographics, as a future work, we plan to utilize this approach for prediction of other demographical information such as gender, ethnicity, etc. Another limitation of the proposed work is in its partial reliance on the use of language to identify the age. Research has shown that since public messaging in social media may reveal significant information about a person, some users modulate their communication strategies to preserve privacy [72]. However, language can still reveal the identify of the users when they engage in one-to-one communication on a public forum; for example, when talking with a close friend [73], or parents' messages to/about their children violating their privacy [74]. Our approach to discard retweets but to preserve the reply-to messages helps capturing the linguistic signature of an individual. Further, we capture interest profile of an individual as manifested by hashtags and URLs in his/her tweets.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] D. Lazer, A.S. Pentland, L. Adamic, S. Aral, A.L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al., Life in the network: the coming age of computational social science, Science (New York, NY) 323 (5915) (2009) 721.

[2] E. Bothos, D. Apostolou, G. Mentzas, Using social media to predict future events with agent-based markets, IEEE Intell. Syst. (1) (2010).

[3] M. Oussalah, A. Zaidi, Forecasting weekly crude oil using twitter sentiment of us foreign policy and oil companies data, in: 2018 IEEE International Conference on Information Reuse and Integration, IRI, IEEE, 2018, pp. 201–208.

[4] S. Asur, B.A. Huberman, Predicting the future with social media, in: Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 01, IEEE Computer Society, 2010, pp. 492–499.

[5] A. Dittrich, C. Lucas, A step towards real-time detection and localization of disaster events based on tweets, in: Proceedings of the 10th International ISCRAM Conference, 2013.

[6] A. Mislove, Pulse of the nation: Us mood throughout the day inferred from twitter, 2010, http://www.ccs.neu.edu/home/amislove/twittermood/.

[7] M.J. Paul, M. Dredze, You are what you tweet: Analyzing twitter for public health, ICWSM 20 (2011) 265–272.

[8] J. Li, L. Zhan, Online persuasion: How the written word drives wom: Evidence from consumer-generated product reviews, J. Advert. Res. 51 (1) (2011) 239–257.

[9] S.M. Mudambi, D. Schuff, Research note: What makes a helpful online review? a study of customer reviews on amazon.com, MIS Quarterly 18 (2010) 5–200.

[10] M. Bucholtz, K. Hall, Identity and interaction: A sociocultural linguistic approach, Discourse studies 7 (4–5) (2005) 585–614.

[11] P. Eckert, Age as a sociolinguistic variable, Handbook Sociolinguist. (1997) 151–167.

[12] S. Rosenthal, K. McKeown, Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, Association for Computational Linguistics, 2011, pp. 763–772.

[13] S.E. Wagner, Age grading in sociolinguistic theory, Language Linguist. Compass 6 (6) (2012) 371–382.

[14] J. Burrows, All the way through: testing for authorship in different frequency strata, Literary Linguist. Comput. 22 (1) (2006) 27–47.

[15] C. Sanderson, S. Guenter, Short text authorship attribution via sequence kernels, markov chains and author unmasking: An investigation, in: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2006, pp. 482–491.

[16] M. Oussalah, F. Bhat, K. Challis, T. Schnier, A software architecture for twitter collection, search and geolocation services, Knowl.-Based Syst. 37 (2013) 105–120.

[17] M. Oussalah, B. Escallier, D. Daher, An automated system for grammatical analysis of twitter messages. a learning task application, Knowl.-Based Syst. 101 (2016) 31–47.

[18] M. Kaufmann, J. Kalita, Syntactic normalization of twitter messages, in: International Conference on Natural Language Processing, Kharagpur, India, 2010.

[19] S. Gouws, D. Metzler, C. Cai, E. Hovy, Contextual bearing on linguistic variation in social media, in: Proceedings of the Workshop on Languages in Social Media. Association for Computational Linguistics, 2011, pp. 20–29.

[20] D. Contractor, T.A. Faruquie, L.V. Subramaniam, Unsupervised cleansing of noisy text, in: Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010, pp. 189–196.

[21] D. Nguyen, N.A. Smith, C.P. Rosé, Author age prediction from text using linear regression, in: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Association for Computational Linguistics, 2011, pp. 115–123.

[22] M. Sap, G. Park, J. Eichstaedt, M. Kern, D. Stillwell, M. Kosinski, L. Ungar, H.A. Schwartz, Developing age and gender predictive lexica over social media, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2014, pp. 1146–1151.

[23] D. Rao, D. Yarowsky, A. Shreevats, M. Gupta, Classifying latent user attributes in twitter, in: Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents, ACM, 2010, pp. 37–44.

[24] F. Rangel, P. Rosso, M. Koppel, E. Stamatatos, G. Inches, Overview of the author profiling task at pan 2013, in: CLEF Conference on Multilingual and Multimodal Information Access Evaluation, CELCT, 2013, pp. 352–365.

[25] E. Stamatatos, W. Daelemans, B. Verhoeven, M. Potthast, B. Stein, P. Juola, M.A. Sanchez-Perez, A. Barrón-Cedeño, Overview of the author identification task at pan 2014, in: CLEF 2014 Evaluation Labs and Workshop Working Notes Papers, Sheffield, UK, 2014, pp. 1–21.

[26] A.P. Lopez-Monroy, M. Montes-Y.-G.omez, H.J. Escalante, L. Villasenor-Pineda, E. Villatoro-Tello, Inaoe's participation at pan'13: Author profiling task, in: CLEF 2013 Evaluation Labs and Workshop, 2013.

[27] D. Nguyen, R. Gravel, D. Trieschnigg, T. Meder, How old do you think i am? a study of language and age in twitter, in: ICWSM, 2013.

[28] D. Nguyen, D. Trieschnigg, A.S. Doğruöz, R. Gravel, M. Theune, T. Meder, F. De Jong, Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment, in: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014, pp. 1950–1961.

[29] H.A. Schwartz, J.C. Eichstaedt, M.L. Kern, L. Dziurzynski, S.M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M.E. Seligman, et al., Personality, gender, and age in the language of social media: The open-vocabulary approach, PLoS One 8 (9) (2013) e73791.

[30] A.A. Morgan-Lopez, A.E. Kim, R.F. Chew, P. Ruddle, Predicting age groups of twitter users based on language and metadata features, PLoS One 12 (8) (2017) e018353.

[31] F. Al Zamal, W. Liu, D. Ruths, Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors, ICWSM 270 (2012) (2012).

[32] S. Volkova, Y. Bachrach, M. Armstrong, V. Sharma, Inferring latent user properties from texts published in social media, in: AAAI, 2015, pp. 4296–4297.

[33] E. Siswanto, M.L. Khodra, Predicting latent attributes of twitter user by employing lexical features, in: Information Technology and Electrical Engineering (ICITEE), 2013 International Conference on, IEEE, 2013, pp. 176–180.

[34] D. Jurgens, Y. Tsvetkov, D. Jurafsky, Writer profiling without the writer's text, in: International Conference on Social Informatics, Springer, 2017, pp. 537–558.

[35] C. Peersman, W. Daelemans, L. Van Vaerenbergh, Predicting age and gender in online social networks, in: Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents, ACM, 2011, pp. 37–44.

[36] J. Tam, C.H. Martell, Age detection in chat, in: Semantic Computing, 2009. ICSC'09. IEEE International Conference on, IEEE, 2009, pp. 33–39.

[37] Y. Kim, Convolutional neural networks for sentence classification, 2014, ArXiv preprint arXiv:1408.5882.

[38] B. Hu, Z. Lu, H. Li, Q. Chen, Convolutional neural network architectures for matching natural language sentences, in: Advances in Neural Information Processing Systems, 2014, pp. 2042–2050.

[39] Y. Goldberg, A primer on neural network models for natural language processing, J. Artificial Intelligence Res. 57 (2016) 345–420.

[40] Y. Zhang, B. Wallace, A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification, 2015.

[41] P. Monachesi, T. de Leeuw, Analyzing elderly behavior in social media through language use, in: Proceedings of HCI International 2018. Communications in Computer and Information Science, 2018.

[42] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, A neural probabilistic language model, J. Mach. Learn. Res. 3 (Feb) (2003) 1137–1155.

[43] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, J. Mach. Learn. Res. 12 (Aug) (2011) 2493–2537.

[44] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.

[45] H. Oktay, A. Firat, Z. Ertem, Demographic Breakdown of Twitter Users: An Analysis Based on Names, Academy of Science and Engineering (ASE), 2014.

[46] L. Sloan, J. Morgan, P. Burnap, M. Williams, Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data, PLoS One 10 (3) (2015) e0115545.

[47] P. Burnap, O. Rana, M. Williams, W. Housley, A. Edwards, J. Morgan, L. Sloan, J. Conejero, Cosmos: Towards an integrated and scalable service for analysing social media on demand, Int. J. Parallel Emergent Distrib. Syst. 30 (2) (2015) 80–100.

[48] J. An, I. Weber, # Greysanatomy vs.# yankees: Demographics and hashtag use on twitter, 2016, ArXiv preprint arXiv:1603.01973.

[49] J.D. Burger, J.C. Henderson, An exploration of observable features related to blogger age, in: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, Menlo Park, CA, 2006, pp. 15–20.

[50] J. Schler, M. Koppel, S. Argamon, J.W. Pennebaker, Effects of age and gender on blogging, in: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, vol. 6, 2006, pp. 199–205.

[51] J.W. Pennebaker, L.D. Stone, Words of wisdom: Language use over the life span, J. Personal. Soc. Psychol. 85 (2) (2003) 291.

[52] U. Pfeil, R. Arjan, P. Zaphiris, Age differences in online social networking–a study of user profiles and the social capital divide among teenagers and older users in myspace, Comput. Hum. Behav. 25 (3) (2009) 643–654.

[53] H. Asoh, K. Ikeda, C. Ono, A fast and simple method for profiling a population of twitter users, in: The Third International Workshop on Mining Ubiquitous and Social Environments, Citeseer, 2012, p. 19.

[54] M. Kosinski, D. Stillwell, T. Graepel, Private traits and attributes are predictable from digital records of human behavior, Proc. Natl. Acad. Sci. (2013) 201218772.

[55] K. Santosh, A. Joshi, M. Gupta, V. Varma, Exploiting wikipedia categorization for predicting age and gender of blog authors, in: UMAP Workshops, 2014.

[56] M. Rustagi, R.R. Prasath, S. Goswami, S. Sarkar, Learning age and gender of blogger from stylistic variation, in: International Conference on Pattern Recognition and Machine Intelligence, Springer, 2009, pp. 205–212.

[57] S. Mechti, M. Jaoua, L.H. Belguith, R. Faiz, Machine learning for classifying authors of anonymous tweets, blogs, reviews and social media, in: Proceedings of the PAN@ CLEF, Sheffield, England, 2014.
[58] S. Argamon, M. Koppel, J.W. Pennebaker, J. Schler, Automatically profiling the author of an anonymous text, Commun. ACM 52 (2) (2009) 119–123.
[59] D. Ikeda, H. Takamura, M. Okumura, Semi-supervised learning for blog classification, in: AAAI, 2008, pp. 1156–1161.
[60] L. Chi, K.H. Lim, N. Alam, C.J. Butler, Geolocation prediction in twitter using location indicative words and textual features, in: Proceedings of the 2nd Workshop on Noisy User-generated Text, WNUT, 2016, pp. 227–234.
[61] M.D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, F. Menczer, Predicting the political alignment of twitter users, in: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, IEEE, 2011, pp. 192–199.
[62] D. Preoţiuc-Pietro, S. Volkova, V. Lampos, Y. Bachrach, N. Aletras, Studying user income through language, behaviour and affect in social media, PLoS One 10 (9) (2015) e0138717.
[63] J.S. Alowibdi, U.A. Buy, P. Yu, Empirical evaluation of profile characteristics for gender classification on twitter, in: 2013 12th International Conference on Machine Learning and Applications, vol. 1, IEEE, 2013, pp. 365–369.
[64] S. Bergsma, M. Dredze, B. Van Durme, T. Wilson, D. Yarowsky, Broadly improving user classification via communication-based name and location clustering on twitter, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013, pp. 1010–1019.
[65] X. Chen, Y. Wang, E. Agichtein, F. Wang, A comparative study of demographic attribute inference in twitter, in: Ninth International AAAI Conference on Web and Social Media, 2015.
[66] M. Pennacchiotti, A.-M. Popescu, A machine learning approach to twitter user classification, in: Fifth International AAAI Conference on Weblogs and Social Media, 2011.
[67] S. Volkova, B. Van Durme, D. Yarowsky, Y. Bachrach, Social media predictive analytics, in: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts, 2015, p. 9.
[68] C.J. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: Eighth International AAAI Conference on Weblogs and Social Media, 2014.
[69] Y. Sun, L. Zhu, G. Wang, F. Zhao, Multi-input convolutional neural network for flower grading, J. Electr. Comput. Eng. 2017 (2017).
[70] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.
[71] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, ArXiv preprint arXiv:1412.6980.
[72] K. Strater, H.R. Lipford, Strategies and struggles with privacy in an online social networking community, in: Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction, vol. 1, British Computer Society, 2008, pp. 111–119.
[73] F. Stutzman, J. Vitak, N.B. Ellison, R. Gray, C. Lampe, Privacy in interaction: Exploring disclosure and social capital in facebook, in: Sixth International AAAI Conference on Weblogs and Social Media, 2012.
[74] T. Minkus, K. Liu, K.W. Ross, Children seen but not heard: When parents compromise children's online privacy, in: Proceedings of the 24th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2015, pp. 776–786.

Abhinay Pandya is a doctoral researcher in the Centre for Ubiquitous Computing, Faculty of Information Technology and Electrical Engineering (ITEE) at the University of Oulu, Finland. He holds a masters degree in Computer Engineering from the Indian Institute of Technology (IIT) Bombay where he graduated securing top 1% position. His main interests include Natural Language Processing, Social network analysis, machine learning, and deep learning. He holds 2 US patents, and more than 10 international publications.

Mourad Oussalah is a newly appointed Senior Research Fellow and research Professor by the Centre for Ubiquitous Computing where he leads the Social Mining Research Group. Prior that he was a Senior Lecturer at University of Birmingham, Research Fellow at City University of London and Postdoctoral at KU Leuven. He conducted research in Data analytics, Information Fusion, Data Mining and mobile computing where he published more than two hundred international publications, presented more than thirty invited talks and supervised about a dozen PhD students in the past 10 years, awarded several best paper awards, and secured several national and EU funding. He is a Senior member of IEEE, Fellow of Royal Statistical Society and EU Cognition Group, and acted as an executive member of IEEE SMC UK and Ireland Chapter.

Paola Monachesi is an Assistant professor at University of Utrecht, Netherlands with Utrecht Institute of Linguistics where she conduct research in Language Technology research and analysis of language in communication contexts. She has been PI of several Dutch and EU projects for the creation of a Language Technology infrastructure. She has initiated a new research line aiming at applying Language Technology techniques in the area of eLearning. She has a consolidated project management experience matured as Project leader and PI of several STREP projects.

Panos Kostakos is a Postdoctoral researcher in the Center for Ubiquitous Computing, University of Oulu, Finland. He received the Ph.D. from the Department of Politics, Languages & International Studies at the University of Bath, UK. He has over 40 publications in the areas of organized crime, terrorism and radicalization using novel methodologies like, natural language processing, social media text mining, computer vision, and machine learning. He also pursues research in Big Data, data analytics, and behavioral modeling applications for smart and safe city solutions, and he is the chair of the first UNODC Education for Justice (E4J) Winter School on Transnational Organised Crime.