Demographic Word Embeddings for Racism Detection on Twitter

Mohammed Hasanuzzaman
ADAPT Centre
Dublin City University
Ireland

Gaël Dias
GREYC UMR 6072
Université de Caen Normandie
France

Andy Way
ADAPT Centre
Dublin City University
Ireland

Abstract

Most social media platforms grant users freedom of speech by allowing them to freely express their thoughts, beliefs, and opinions. Although this represents incredible and unique communication opportunities, it also presents important challenges. Online racism is such an example. In this study, we present a supervised learning strategy to detect racist language on Twitter based on word embeddings that incorporate demographic (Age, Gender, and Location) information. Our methodology achieves reasonable classification accuracy over a gold standard dataset (F_1=76.3%) and significantly improves over classification performance of demographic-agnostic models.

1 Introduction

The advent of microblogging services has impacted the way people think, communicate, behave, learn, and conduct their daily activities. In particular, the lack of regulation has made social media an attractive tool for people to express online their thoughts, beliefs, emotions and opinions. However, this transformative potential goes with the challenge of maintaining a complex balance between freedom of expression and the defense of human dignity (Silva et al., 2016). Indeed, some users misuse the medium to promote offensive and hateful language, which mars the experience of regular users, affects the business of online companies, and may even have severe real-life consequences (Djuric et al., 2015). In the latter case, (Priest et al., 2013; Tynes et al., 2008; Paradies, 2006b; Darity Jr., 2003) evidenced strong correlations between experiences of racial discrimination and negative mental health outcomes such as depression, anxiety, and emotional stress as well as negative physical health outcomes such as high blood pressure and low infant birth weight.

As the information contained in social media often reflects the real-world experiences of their users, there is an increased expectation that the norms of society will also apply in social media settings. As such, there is an increasing demand for social media platforms to empower users with tools to report offensive and hateful content (Oboler and Connelly, 2014).

Hateful content can be defined as "speech or expression which is capable of instilling or inciting hatred of, or prejudice towards, a person or group of people on a specified ground including race, nationality, ethnicity, country of origin, ethno-religious identity, religion, sexuality, gender identity or gender" (Gelber and Stone, 2007). While there are many forms of hate speech, racism is the most general and prevalent form of hate speech in Twitter (Silva et al., 2016). Racist speech relates to a socially constructed idea about differences between social groups based on phenotype, ancestry, culture or religion (Paradies, 2006a) and covers the categories of race (e.g. black people), ethnicity (e.g. chinese people), and religion (e.g. jewish people) introduced in Silva et al. (2016).

Racism is often expressed through negative and inaccurate stereotypes with one-word epithets (e.g. tiny), phrases (e.g. big nose), concepts (e.g. head bangers), metaphors (e.g. coin slot), and juxtapositions (e.g. yellow cab) that convey hateful intents (Warner and Hirschberg, 2012). As such, its automatic identification is a challenging task. Moreover, the racist language is not uniform. First, it highly depends on contextual features of the targeted community. For example, anti-african-american messages often refer to unemployment or single parent upbringing whereas anti-semitic language predominantly makes refer-

ence to money, banking, and media (Warner and Hirschberg, 2012). Second, the demographic context of the speaker may greatly influence word choices, syntax, and even semantics (Hovy, 2015). For instance, a young female rapper from Australia may not use the exact same language as an elder male economist from South Africa to express her racists thoughts (Figure 1).

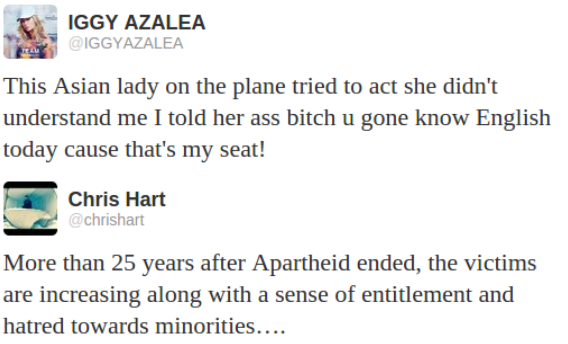

Figure 1: (Top) Tweet from 25 years-old Australian rapper Iggy Azalea. (Bottom) Tweet from senior South African economist Chris Hart.

To address these issues, we propose to focus on demographic features (age, gender, and location) of Twitter users to learn demographic word embeddings following the ideas of Bamman et al. (2014) for geographically situated language. The distributed representations learned from a large corpus containing a great proportion of racist tweets are then used to represent tweets in a straightforward way and serve as input to build an accurate supervised classification model for racist tweet detection. Different evaluations over a gold-standard dataset show that the demographic (age, gender, location) word embeddings methodology achieves F_1 score of 76.3 and significantly improves over classification performance of demographic-agnostic models.

2 Related Work

Despite prevalence and large impact of online hateful speech, there has been a lack of published works addressing this problem. The first studies on hatefuf speech detection are proposed by (Xu and Zhu, 2010; Kwok and Wang, 2013; Warner and Hirschberg, 2012) but with very limited scope and basic supervised machine learning techniques.

One of the very first attempts to propose a computational model that deals with offensive language in online communities is proposed by Xu and Zhu (2010). The authors propose a sentence-level rule-based semantic filtering approach, which utilizes grammatical relations among words to remove offensive contents in a sentence from Youtube comments. Although this may be a valuable work, its scope deviates from our specific goal, which aims at automatically detecting racist online messages.

The first identified studies that can directly match our objectives are proposed by Kwok and Wang (2013) and Warner and Hirschberg (2012). In Kwok and Wang (2013), the authors propose a naïve Bayes classifier based on unigram features to classify tweets as racist or non-racist. It is important to notice that the standard data sets of racist tweets were selected from Twitter accounts that were self-classified as racist or deemed racist through reputable news sources with regards to anti-Barack-Obama articles. Although this is the very first attempt to deal with Twitter posts, the final model is of very limited scope as it mainly deals with anti-black community racism. Similarly, (Warner and Hirschberg, 2012) propose a template-based strategy that exclusively focuses on the detection of anti-semitic posts from Yahoo!. In particular, some efforts were made to propose new feature sets but with unsuccessful results.

In Warner and Hirschberg (2012), the authors propose a similar idea focusing on the specific manifestation of anti-semitic posts from Yahoo! and xenophobic urls identified by the American Jewish Congress. In particular, they use a template-based strategy to generate features from the part-of-speech tagged corpus augmented with Brown clusters. Then, a model is generated for each type of feature template strategy, resulting in six SVM classifiers. Surprisingly, the smallest feature set comprised of only positive unigrams performed better than bigram and trigram features. Similarly to Kwok and Wang (2013), this study focuses on a specific community and is of limited scope. However, efforts were made to propose new features although with unsuccessful results.

Burnap and Williams (2016) is certainly the first study to propose a global understanding of hateful speech on Twitter including a wide range of protected characteristics such as race, disability, sexual orientation and religion. First, they build individual SVM classifiers for each of the four hateful-speech categories using combinations of feature types: ngrams (1 to 5), slur words and typed-dependencies. Overall, the highest F_1 scores are

achieved by combining ngrams and hateful terms, although the inclusion of typed dependencies reduces false positive rates. In a second step, the authors build a data-driven blended model where more than one protected characteristic is taken at a time and show that hateful speech is a domain dependent problem. The important finding of this work is the relevance of the simple dictionary lookup feature of slur words.

In Waseem and Hovy (2016), the authors study the importance of demographic features (gender and location) for hateful speech (racism, sexism and others) classification from Twitter as well as propose to deal with data sparseness by substituting word ngrams with character ngram (1 to 4) features. 10-fold cross validation results with a logistic regression show that the demographic features (individually or combined) do to not lead to any improvement, while character-based classification outperforms by at least 5 points the F_1 score of word-based classification. Although non-conclusive results are obtained with demographic features, we deeply believe that these contextual features can lead to improvements.

While all these efforts are of great importance, most works point at the drawbacks of discrete representations of words and texts. This especially holds in the context of the racist language where offenders often use simple yet effective tricks to obfuscate their comments and make them more difficult for automatic detection (such as replacing or removing characters of offensive words), while still keeping the intent clear to a human reader. For that purpose, including continuous distributed representations of words and texts may lead to improved classification results.

Following this assumption, two studies have been proposed with different strategies. Tulkens et al. (2016) present a dictionary-based approach to racism detection in Dutch social media comments following the findings of Burnap and Williams (2016). In particular, they broaden the coverage of the categories in their dictionaries by performing a dictionary expansion strategy that uses a word embedding (skip-gram model) obtained from a general-purpose 3.9 billion words Dutch corpus. For instance, they show that the entry "mohammed"[1] can be expanded with "mohamed", "mohammad" and "muhammed". The SVM classification results show that the automated expansion of the dictionary slightly boosts the performance, but with no statistical significance. To our point of view, this may be due to the non-specificity of the corpus used to produce the word embeddings.

Indeed, more successful results are obtained by Djuric et al. (2015), who propose to build specific paragraph embeddings for hateful speech. In particular, they show a two-step method. First, paragraph2vec (Le and Mikolov, 2014) is used for joint modeling of comments and words from Yahoo! comments with the bag-of-words neural language model. Then, these embeddings are used to train a binary classifier to distinguish between hateful and clean comments. Results of the logistic regression for 5-fold cross-validation show improved performance of the continuous representation when compared to discrete ngram features.

In this paper, we propose to study further the initial "unsuccessful" idea of Waseem and Hovy (2016) of taking into account demographic factors for racism classification in Twitter. Indeed, recent linguistic studies have shown that the racist language can differ from place to place (Chaudhry, 2015). Moreover, we deeply think that the specificity of racist language can be better modeled by continuous spaces as shown in Djuric et al. (2015).

As a consequence, we propose to build demographic word embeddings following the initial findings of Bamman et al. (2014) (location) and Hovy (2015) (age and gender) for discrete features. These embeddings are computed on a corpus specifically built for racism detection and obtained by massive gathering of tweets containing slur words listed in different referenced sources such as the racial slur database.[2] Then, the low-space distributed continuous representations are used as direct input of binary (racist vs. non racist) SVM classifiers, which are compared to concurrent approaches over a gold standard data set.

3 Data Sets

In this section, we first detail the process that consists in gathering a huge quantity of potentially racist and non-racist tweets from which the demographic word embeddings are computed. Then, we present the manual annotation process defined to create a gold standard data set and used to evaluate the proposed classification strategy.

[1] The study specifically focuses on anti-Islamic racism.

[2] http://www.rsdb.org/ Last accessed on 10-04-2017.

3.1 Unlabeled Data Set

The process starts by crawling English tweets using the Twitter streaming API[3] for a period of three months (05/02/2015 to 05/05/2015). From this set, potentially racist tweets are collected by selecting all those that contain at least one racist candidate keyword from (1) a collection of 3000 racial and ethnic slurs or (2) a set of racially motivated phrases. In particular, the first list of keywords is compiled from the Wikipedia list of ethnic slurs[4] and the racial slur database.[5] The second list of racist phrases is produced by combining a general purpose insult with the name of a given ethnicity, giving rise to slurs such as "dirty jew" or "russian pig". For that particular purpose, we collected 70 common insulting modifiers such as "dog", "filthy", "honky", "redneck", or "rat". This process allowed to gather about 17.2 million potentially racist tweets from 1.8 million users.

In parallel, a random sample ($\approx 1\%$ of all retrieved tweets) of potentially clean (i.e. that do not contain racial and ethnic slurs) tweets is collected resulting in 41.3 million tweets from 4.6 million users all over the world.

In addition to the message conveyed in a given tweet, some information is stored for all tweets, such as location, time and user's profile. However, this information is not accessible for most tweets. As a consequence, we propose different text processing methodologies to compute the three demographic variables: age, gender and location.

In particular, age and gender are predicted using the text-based models described in Sap et al. (2014). These models are trained on data from over 70,000 Facebook users and report an accuracy rate of 91.9% for gender (resp. 88.9% on Twitter data) and a Pearson correlation coefficient of 0.84 for age prediction. For location, we use the geo-coding scheme proposed in Chen and Neill (2014), which is based on three major rules with priorities.[6] For each tweet, (1) we search for a location mention from GeoNames[7] in the text message, then (2), we verify if the user enabled the geo-coding function of his/her device, and (3) we look for location information from the stored user's profile. The first location information identified is then returned as the geographic location of the tweet. Note that the returned location information is at country-level.

Finally, similarly to Hovy (2015), we lowercase tweet texts, remove special characters, replace numbers with a 0, and join frequent bigrams with an underscore to form a single token.

Table 1 presents the distribution of the obtained corpus broken down by the contextual variables.

Country	Tweets	Users	Users Demographics			
			U. 35	O. 35	M	F
Potentially racist tweets						
USA	3.9M	124.2k	54.7k	69.5k	79.9k	44.3k
India	3.4M	132k	89.7k	42.3k	93.8k	38.2k
UK	2.8M	105k	55.8k	49.2k	61.3k	43.7k
Canada	1.3M	96k	44.7k	51.3k	52.8k	43.2k
Japan	1.1M	98.4k	50.6k	44.8k	49.8k	45.6k
Potentially clean tweets						
USA	11.8M	259.4k	121.6k	137.8k	137.4k	121.9k
UK	8.3M	184.5k	98.3k	86.2k	94.1k	90.4k
India	6.4M	218.4k	136.2k	82.2k	145.6k	72.8k
Japan	2.3M	102.5k	54.8k	47.7k	57.3k	45.2k
Indon.	2.1M	109.4k	63.2k	46.2k	71.9k	37.5k

Table 1: Top five countries with most tweets in the unlabeled data sets. Tweet users are broken down by age - (i) Under 35 (U. 35); (ii) Over 35 (O. 35) - and gender - (i) Male (M); (ii) Female (F).

3.2 Classification Data Set

It is anticipated that recognizing racist tweets can be difficult. Certainly, the use of ethnic and racial slurs is clearly not always hateful. A number of studies have examined how the slur "nigger" has been appropriated by African Americans as a way of actively rejecting the connotations it carries, e.g. for comedic purposes, a status symbol, a shorthand term expressing familiarity among friends, or even forgetting what the term ever denoted in the first place (Anderson and Lepore, 2013). Therefore, tweets that merely contain these slurs may not be always offensive.

In order to build a gold-standard data set that allows the evaluation of classifiers learned to detect "true" racist tweets, we adopt an approach, which depends on the assessments made by users of a crowdsourcing platform. For that purpose, we randomly sample 4 sub-collections from the collection of potentially racist tweets gathered in section 3.1, namely two for gender (Male and Female), and two for age (Under 35 and Over 35). The same strategy is followed to select 4 sub-

[3] https://dev.twitter.com/streaming/overview. Last accessed on 10-04-2017.

[4] https://en.wikipedia.org/wiki/List_of_ethnic_slurs. Last accessed on 10-04-2017.

[5] http://www.rsdb.org/. Last accessed on 10-04-2017.

[6] This methodology is clearly prone to error. But, improving this pre-processing step is out of the scope of this paper.

[7] http://www.geonames.org/. Last accessed on 10-04-2017.

Tweet	R1	R2	R3	Maj.V
Dog, this nigga does not stop staring at me in the gym. Dickface got a staring problem!	Y	Y	Y	Y
The after effect of being a wigger. http:/***	N	N	N	N
@*** i told yo racist ass to stop callin me a niger. Dumb white boy!	Y	N	Y	Y
Who cares where they were born, camel breath, they call themselves Israelis and Jews.	Un	Y	Un	Un
Radical Islam on the rise in Indonesia. http://****.	N	Un	Un	Un
So I can see iphone emojis now, soooo coolie!!!!	N	N	Y	N

Table 2: Examples of tweets annotated for racism: Yes (Y), No (N), and Unsure (Un.). R1, R2, R3: judgements from each rater. Maj.V: choice from majority voting.

Country	Number of Tweets				
	Total	U. 35	O. 35	M	F
Racist tweets					
USA	1036	387	649	572	464
UK	728	346	382	382	346
India	587	413	174	445	142
Japan	485	268	217	289	196
Canada	431	198	233	256	175
Clean tweets					
India	1132	712	420	694	438
USA	956	438	518	534	422
UK	762	352	410	387	375
Canada	631	309	322	336	295
Japan	610	294	316	327	283

Table 3: Top five countries with most tweets in the classification data set. Tweets are broken down by age and gender.

collections from the set of potentially clean tweets.

The samples are equally distributed among male, female, and the two age groups, and the geographic distributions (at the country level) in the 8 sub-collections are in line with the distributions over the whole corpus of tweets.

Note that while the effect of age on language use is undisputed (Rickford and Price, 2013), providing a clear cut-off is hard. We therefore used age ranges that result in roughly equally sized data sets for both groups in the overall corpus.

Each sub-collection consists of 1000 tweets that are uploaded to the crowdsourcing service of the CrowdFlower platform[8] for annotation. In particular, each tweet is represented by its text, location information, user's age and gender, and a multiple choice question is asked to the annotators to decide whether the tweet has indeed a racist intent or not. The available answers are "Yes", "No", and "Unsure". Each tweet is annotated by at least 3 annotators. Each annotator requires to be an English speaker and preferably in the same country as the origin of the tweet.

Out of the 8000 tweets that were judged, 7358 received a majority of "Yes" or "No" votes. The remainder were less determinant with the addition of "Unsure" votes. Of these 7358 tweets, 3267 (44%) had a majority of votes confirming they were racist, with the remainder 4091 (56%) considered as not racially motivated tweets. Table 2 shows some examples of tweets with their annotations and Table 3 summarizes the classification data set by demographic variable.

This data set can be thought as a hard test for classifiers as non racist tweets may contain slurs unlike most works so far, which assess their models based on the hypothesis that non racist tweets usually contain general vocabulary and do not exhibit any critical content. In parallel, we acknowledge that some racist tweets may not contain any slur such as illustrated in Figure 1. Future work will definitely have to deal with the integration of racist tweets that do not present any direct racist mention, in the classification data set.

4 Methodology

Following the self-taught learning paradigm (Raina et al., 2007), we first construct demographic word embeddings from the unlabeled data described in section 3.1. Then, the obtained context-aware high-level low-dimension features form a succinct input representation for the specific task of racist tweet detection.

4.1 Demographic Word Embeddings

Djuric et al. (2015) were the first to propose a self-taught learning strategy in the context of hateful speech detection, where they simultaneously learn low-dimension representations of documents and words in a common vector space. However, contextual/demographic characteristics were not taken into account.

A simple way to take demographic variables into account when building word embeddings is to train individual models on each variable value (e.g. a "male" model is obtained by using data

[8] http://www.crowdflower.com/. Last accessed on 10-04-2017.

only from male users). This has been the strategy followed by Hovy (2015) for age and gender in the context of sentiment classification, topic identification, and author attribute identification.[9]

A more sophisticated model has been proposed in Bamman et al. (2014), which defines a joint parameterization over all variable values in the data. In this specific case, they study geographic variables. As such, information from data originating in some location can influence the representations learned for other locations. A joint model has three a priori advantages over independent models: (i) sharing data across variable values encourages representations across those values to be similar; (ii) such sharing can mitigate data sparseness for less-witnessed variables; and (iii) all representations are guaranteed to be in the same vector space and can therefore be compared to each other.

In this work, we propose to follow the work of Bamman et al. (2014) and introduce a set of demographic features, namely Age (Under 35 and Over 35), Gender (Male and Female) and Location (top 20 countries in the unlabeled data set[10]).

This model corresponds to an extension of the skip-gram language model proposed in (Mikolov et al., 2013). Formally, given an input word w_i in a vocabulary V and a set of demographic variables A, the objective is to maximize the average data log-likelihood given in equation (1), where c is the length of the word context.

$$\mathcal{L} = \frac{1}{|V|}\sum_{i=1}^{|V|}\sum_{c\in\{-1,1\}}\log \Pr(w_{i+c}|w_i) + \frac{1}{|A|}\sum_{a=1}^{|A|}\frac{1}{|V_a|}\sum_{i=1}^{|V_a|}\sum_{c\in\{-1,1\}}\log \Pr(w_{i+c}|w_i^a) \quad (1)$$

So, while any word has a common low-dimension representation that is invoked for every instance of its use (regardless of its demographic context), the word embedding specific to a given demographic variable indicates how that common representation should shift in the k-dimension space when used in this special context.

In terms of implementation, back-propagation functions as in the standard skip-gram language model, with gradient updates for each training example and computation is speeded up using the hierarchical softmax function.

[9] Thus, outside our application scope.

[10] Remaining countries did not have reasonable amount of tweets to learn embeddings.

4.2 Racist Tweet Classification

For the classification process, we use linear Support Vector Machines (SVM) models that take as input a feature representation of tweets based on demographic word embeddings.

On the one hand, each word of a given tweet is represented by its joint embedding, which is the concatenation of its common low-dimension representation of dimension k and each low-dimension representation of its specific demographic embeddings, each one also of dimension k. For example, if a given word appears in a tweet issued by a male user of 45 years-old in the USA, its representation will be the concatenation of its common embedding with its 3 specific embeddings computed from the active variables Age=Over 35, Gender=Male and Location=USA. In this particular case, each word will be represented by a vector of $4k$ continuous values.

On the other hand, we need to represent tweets with variable lengths based only upon the concatenated embeddings of the words they contain. For that purpose, we follow the same strategy as proposed in Hovy (2015). For each learning instance (i.e. tweet), we collect 5 N-dimensional statistics (i.e. minimum, maximum, mean representation, as well as one standard deviation above and below the mean) over a $N \times t$ input matrix, where N is the dimensionality of the concatenated embeddings (i.e. $N = |A^*| \times k$, where $|A^*|$ is the number of active variables and k is the dimension of each individual embedding), and t is the sentence length in words. We then concatenate those 5 N-dimension vectors to a $(5N)$-dimension vector to represent each learning instance in a single format and feed it to the SVM classifiers.

Note that taking the maximum and minimum across all embedding dimensions is equivalent to representing the exterior surface of the instance manifold (i.e. the volume in the embedding space within which all words in the instance reside). And adding the mean and standard deviation acts as the density per-dimension within the manifold.

We are aware that different tweet representations could have been tested, namely those based on neural sequence modeling. In particular, future work will aim to adapt models such as paragraph2vec (Le and Mikolov, 2014) or LSTM (Tai et al., 2015) to demographic embeddings.

5 Evaluation

5.1 Experimental Setups

The demographic word embeddings were computed based on the implementation provided by Bamman et al. (2014)[11] with $k = 100$ for the low representation dimension, while demographic-agnostic models were built using the word2vec[12] implementation for the same size of k. In particular, we built 8 different embeddings i.e. one demographic-agnostic embedding based on the unlabeled data set presented in section 3.1 without any demographic variable and 7 others with different demographic variables combinations such as Age, Gender, Location, Age+Gender, Age+Location, Gender+Location and Age+Gender+Location. With respect to the discrete representation of tweets, we proposed 6 different configurations using either word unigrams and/or bigrams, or character ngrams (n=1 to 4). In this context, the demographic information was included as binary variables for Gender and Age, and a n-ary variable for Location. One final model was proposed that joins demographic-agnostic word embeddings with demographic discrete variables. As a consequence, 15 different input settings were tested for classification as shown in the first column of Table 4, using the linear SVM model implemented in the Weka[13] platform with standard parameters settings[14].

As for evaluation, we proposed to split the gold-standard data set of 7358 tweets described in section 3.2 into a training data set of 5149 tweets (70%) and a test data set of 2209 tweets (30%). In particular, the test data set is equally distributed among male, female, the two age groups, and the top 20 countries with most tweets in our collected unlabeled data set. In order to evaluate the capacity of the classification models to generalize over the original data distribution, we also performed 10-fold cross-validation for all the settings. Moreover, we analyzed the effect of the training data size for the classification purposes.

[11] https://github.com/dbamman/geoSGLM. Last accessed on 10-04-2017.

[12] https://code.google.com/archive/p/word2vec/. Last accessed on 03-10-2017.

[13] http://www.cs.waikato.ac.nz/ml/weka/. Last accessed on 10-04-2017.

[14] The magnitude of the improvements could be improved by tuning the parameters over a development set. But this remains out of the scope of this paper.

5.2 Quantitative Results

The classification results over the annotated gold standard test data presented in section 3.2 are given in Table 4 for all 15 different configurations.

Tweet Representation	Prec.	Rec.	F_1
Uni.	58.2	54.2	56.1
Uni. + Age + Loc. + Gen.	60.1	58.3	59.0
Uni. + Bi.	58.9	57.2	58.0
Uni. + Bi. + Age + Loc. + Gen.	61.8	60.2	60.9
Ch. ngrams	60.6	62.2	61.3
Ch. ngrams + Age + Loc. + Gen.	60.3	61.8	61.0
W.2V.	67.3	66.4	66.8
W.2V. + Age + Loc. + Gen.	70.3	70.7	70.5
Demo. W.2V. (Age)	72.3	71.5	71.9
Demo. W.2V. (Gen.)	68.7	67.5	68.0
Demo. W.2V. (Loc.)	73.6	73.1	73.4
Demo. W.2V. (Age + Gen.)	72.7	72.1	72.4
Demo. W.2V. (Age + Loc.)	75.3	**76.2**	75.8
Demo. W.2V. (Gen. + Loc.)	74.0	73.4	73.7
Demo. W.2V. (Age + Gen. + Loc.)	**76.4**	76.1	**76.3**

Table 4: Precision (Prec.), Recall (Rec.), and F_1 Score (F_1) for racist tweet detection over the test data set. Demographic variables Location and Gender are represented by Loc. and Gen., respectively. W.2V. stands for word2vec model learned over the racist data set.

The first finding is that demographic-aware models perform better than the demographic-agnostic ones across almost all configurations. This improvement raises at 5.3% of F_1 on average with a maximum increase of 9.5%. The only contradictory case is when character ngrams (n=1 to 4) are used as text representation. In that specific configuration, the inclusion of demographic variables has no impact on the results, confirming previous results of Waseem and Hovy (2016). However, using character ngrams improves over a word ngrams representation as noted by WaseemH16. But, it fails to benefit from the introduction of demographic variables unlike the word ngrams representation which is boosted by the introduction of contextual variables. Note that this result was not reported in Waseem and Hovy (2016). Furthermore, we computed the exact same best configuration of Waseem and Hovy (2016), i.e. logistic regression over character ngrams with the Gender variable, and a F_1 score of 60.7% was achieved, which is comparable to our results with SVM reaching F_1 score of 61.0%.

The second important result is that the model learned over continuous feature representations of tweets from general embeddings in combination

with features containing demographic information namely Age, Gender, Location (row 8 in Table 4) outperforms all demographic-aware configurations based on discrete text representations (i.e. unigrams, bigrams or character ngrams) by a minimum (resp. maximum) margin of 9.5% (resp. 11.5%) of F_1 score. This result particularly shows that word semantics is better captured by word embeddings than classical text representations.

The third result supports our initial hypothesis that demographic word embeddings can improve the task of racist tweet classification. Indeed, all demographic-aware embeddings improve over the demographic-agnostic embeddings, even boosted by discrete demographic variables, to the only exception of the gender-aware model (row 10 in Table 4). In particular, a maximum difference is obtained by the demographic word embedding that counts with Age, Gender and Location variables at levels of 9.5% of F_1 score, when compared to demographic-agnostic word embeddings.

Finally, although all demographic variables improve on classification, the Gender variable seems to have the smallest impact on the overall results. Indeed, at each inclusion it slightly improves results, while Location is the most productive variable, as hypothesized by Chaudhry (2015). In particular, note that the demographic-aware continuous model with the single Gender variable is outperformed by the model built on general word embeddings increased by discrete context variable (row 8 in Table 4).

Tweet Representation	Prec.	Rec.	F_1
Uni.	65.3	61.5	63.3
Uni. + Age + Loc. + Gen.	66.7	62.4	64.4
Uni. + Bi.	65.8	60.7	63.1
Uni. + Bi. + Age + Loc. + Gen.	66.2	62.3	64.1
Ch. ngrams	64.2	65.6	64.8
Ch. ngrams + Age + Loc. + Gen.	65.0	67.2	66.0
W.2V.	71.2	69.5	70.3
W.2V. + Age + Loc. + Gen.	73.5	72.8	73.1
Demo. W.2V. (Age)	75.3	74.5	74.8
Demo. W.2V. (Gen.)	72.3	71.2	71.7
Demo. W.2V. (Loc.)	75.4	76.3	75.8
Demo. W.2V. (Age + Gen.)	75.1	74.4	74.7
Demo. W.2V. (Age + Loc.)	78.2	**78.8**	78.5
Demo. W.2V. (Gen. + Loc.)	77.6	74.2	75.8
Demo. W.2V. (Age + Gen. + Loc.)	**79.0**	77.4	**77.1**

Table 5: Precision (Prec.), Recall (Rec.), and F_1 Score (F_1) for 10-fold cross-validation.

Results of the 10-fold cross-validation are presented in Table 5 and show similar conclusions but following the original distribution of the learning data sets, oppositely to the first experiment, where we forced the test data to be balanced. The only small difference is that in these conditions the character ngrams representation seems to take advantage of the introduction of contextual variables reaching the best results in terms of discrete text representation.

Note that all improvements in Table 4 and Table 5 are statistically significant. In particular, we adopted a bootstrap-sampling test similarly to Hovy (2015) with a standard cutoff of $p < 0.05$.

Effect of Training Data Size: The size of the training data set is an important concern in supervised learning methods as lots of manual efforts are required to tag learning instances. Thus, we want to evaluate the impact of the training data set size on the performances of two different word embeddings configurations: (1) W.2V. + Age + Gen. + Loc. (Baseline) and (2) Demo. W.2V. (Age + Gen. + Loc.).

For that purpose, we randomly select $d\%$ (by steps of 20%) of the training data set to train the classifiers and test them on the test data set. Note that for each $d\%$, we generate the training set 20 times and the averaged performance is recorded. F_1 scores for both approaches over the test data are presented in Table 6.

Data size	Baseline	Demo. W.2V. (Age+Gen.+Loc.)
1k	63.4	68.3
2k	65.2	70.2
3k	68.5	73.8
4k	69.9	75.5
5.1k (all)	70.5	76.3

Table 6: Classification performances (F_1 score) with different sizes of training data.

Results show that our proposed framework performs consistently better than its counterpart. In particular, results show that with 3k training examples better results can be obtained by our approach than relying on 5.1k training examples for the state-of-the-art supervised machine learning approach (Baseline) based on common word embeddings for racist tweet detection.

5.3 Qualitative Results

In order to better understand figures given in section 5.2, we examined different qualitative criteria.

Error Analysis We performed a manual error analysis of the instances where our best performing configuration and manual annotation differed. We noticed that some of the tweets were difficult to classify because of their ambiguity with respect to racism classification. For example, tweets such as "*adam 15 boy prob bi whitey is an irl egg*" or "*@... but the pakis are still trying to get across the border*" could be "racist" tweets, but without the context it is difficult to judge even for human. Part of our future work will be to extract the context (e.g. previous tweets, threads) of tweets and use it for (i) annotation purposes and (2) classification issues.

Lexical Distribution. Table 7 represents the top ten most frequent racial slurs occurring in automatically tagged racist tweets broken down by Age and Gender. Results show that majority of racial slurs discuss about Islam. Considering that many tweets are issued from the USA or India rather than other countries, it is not surprising that several of the terms are in line with the Indian and American political discourses orientations.

Overall	U.35	O.35	M	F
islam	nigger	islam	islam	nigga
nigger	muslims	hebe	muslims	islam
muslims	paki	muslims	paki	nigger
white boy	islam	white boy	white boy	desi
mohammed	prophet	nigger	chinki	mohammed
pedophile	bihari	negro	mohammed	pedophile
paki	white boy	pedophile	nigger	whitey
prophet	pedophile	whitey	whitey	muslims
whitey	gook	paki	bihari	hebe
bihari	whitey	coon	mallu	coon

Table 7: Ten most frequently occurring racial slurs in racist tweets broken down by age and gender.

Demographic Distribution. The distributions by Age and Gender of the automatically tagged racist tweets are presented in Table 8. It can be seen that the Gender distribution is skewed towards men. This goes in line with an earlier study made by (Roberts et al., 2013) who found that the majority (more than 70%) of the offenders of hate crimes were men. The results also demonstrate that people under the age of 35 years-old seem to be more racist than people with the age over 35. Note that 40.2% of the Twitter users are less than 35 years old[15], which indicates a clear bias towards racism by youngers. We further analyzed the racist tweets and found that roughly 35% of total racist offenders are aged under 25. So, it seems that younger adults are more likely to be involved in racist offences than older adults.

Variable	Value	% Racist tweets
Gender	Men	64.7%
	Women	35.3%
Age	Under 35	67.2%
	Over 35	32.8%

Table 8: Distribution of tweets by Gen. and Age.

Usage Patterns of Racial Slurs. Finally, we analyzed the set of automatically tagged racist tweets and categorized the patterns of usage of racial slurs into four main categories:
- *Group Demarcation:* The user's intention is to demarcate group boundaries (people as in or out) as in the follwoing example "*Dirty Jews, Im Hitler, Ill kill the Jews.*".
- *Directed Attack:* Here, an attack is directed at an individual or group known personally to the sender. Tweets in this pattern use racial/ethnic slurs directly such as in "@ *** *why you rotten nigger bitch, how dare you.*".
- *Unnecessary Use:* In this case, the main discourse of the tweet may not be race or ethnicity but rather the use of a given slur in an offhand or casual fashion, e.g. "*@: disgusting Indian shops that charge you for paying by card.*" ;
- *Ideological:* Here, authors consciously use racial/ethnic slurs within a political statement that would justify an action in the real world, such as in "*Leftist have no right view. They have agenda for Paki and muslim. Support BJP*".

6 Conclusions

In this paper, we proposed to build demographic-aware word embeddings for the classification of racist tweets. We showed that such models outperform strategies based on discrete text representations and demographic-agnostic word embeddings. However, overall performance ($F_1 = 76.3\%$) is still insufficient confirming that hateful speech detection is a hard task. To improve initial results, future works will aim to (1) incorporate additional context (such as connected tweets) to leverage tweet ambiguity, (2) build demographic-aware sequence embeddings to better capture text semantics, (3) combine both discrete and continuous representations to build semi-supervised models, as weak detection of racist tweets in large amount is possible and (4) test new character-based word embeddings (Bojanowski et al., 2017).

[15] http://goo.gl/qHlIQq. Last Accessed on 6-06-2017.

7 Acknowledgment

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

Luvell Anderson and Ernie Lepore. 2013. Slurring words 1. *Noûs* 47(1):25–48.

David Bamman, Chris Dyer, and Noah A. Smith. 2014. Distributed representations of geographically situated language. In *Proceedings of the Association for Computational Linguistics (ACL 2014)*. pages 828–834.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *TACL* 5:135–146.

Pete Burnap and Matthew L. Williams. 2016. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Science* 5(1):11.

Irfan Chaudhry. 2015. hashtagging hate: Using twitter to track racism online. *First Monday* 20(2).

Feng Chen and Daniel B. Neill. 2014. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pages 1166–1175.

William A. Darity Jr. 2003. Employment discrimination, segregation, and health. *American Journal of Public Health* 93(2):226–231.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web (WWW 2015)*. pages 29–30.

Katharine Gelber and Adrienne Sarah Ackary Stone. 2007. *Hate Speech and Freedom of Speech in Australia*, volume 2118. Federation Press.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the Association for Computational Linguistics (ACL 2015)*. pages 752–762.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI 2013)*. pages 1621–1622.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*. pages 1188–1196.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781.

Andre Oboler and Karen Connelly. 2014. Hate speech: A quality of service challenge. In *Proceedings of IEEE Conference on E-Learning, E-Management and E-Services (IC3E 2014)*. pages 117–121.

Yin C. Paradies. 2006a. Defining, conceptualizing and characterizing racism in health research. *Critical Public Health* 16(2):143–157.

Yin C. Paradies. 2006b. A systematic review of empirical research on self-reported racism and health. *International Journal of Epidemiology* 35(4):888–901.

Naomi Priest, Yin C. Paradies, Brigid Trenerry, Mandy Truong, Saffron Karlsen, and Yvonne Kelly. 2013. A systematic review of studies examining the relationship between reported racism and health and well-being for children and young people. *Social Science & Medicine* 95:115–127.

Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y Ng. 2007. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th International Conference on Machine Learning (ICML 2007)*. ACM, pages 759–766.

John Rickford and Mackenzie Price. 2013. Girlz ii women: Age-grading, language change and stylistic variation. *Journal of Sociolinguistics* 17(2):143–179.

Colin Roberts, Martin Innes, Matthew Leighton Williams, Jasmin Tregidga, and David Gadd. 2013. *Understanding who commits hate crimes and why they do it*. Welsh Government.

Maarten Sap, Gregory J. Park, Johannes C. Eichstaedt, Margaret L. Kern, David Stillwell, Michal Kosinski, Lyle H. Ungar, and H. Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*. pages 1146–1151.

Leandro Silva, Mainack Modal, Denzil Correa, Fabricio Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *Proceedings of the International AAAI Conference on Web-Blogs and Social Media (ICWSM 2016)*.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*

and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, (ACL/IJCNLP 2015). pages 1556–1566.

Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. A dictionary-based approach to racism detection in dutch social media. *CoRR* abs/1608.08738.

Brendesha M. Tynes, Michael T. Giang, David R. Williams, and Geneene N. Thompson. 2008. Online racial discrimination and psychological adjustment among adolescents. *Journal of Adolescent Health* 43(6):565–569.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media (LSM 2012) of the Association for Computational Linguistics (ACL 2012)*. pages 19–26.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the Student Research Workshop of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL 2016)*. pages 88–93.

Zhi Xu and Sencun Zhu. 2010. Filtering offensive language in online communities using grammatical relations. In *Proceedings of the 7th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS 2010)*. pages 1–10.