IDENTIFICATION OF HATRED SPEECHES ON TWITTER.

[2]ADAMU HABU, [1]S. K ALARAMMA, [1]B. I YA'U, [1]M.A MUSA

[1]Abubakar Tafawa Balewa University, Nigeria. [2]University of St Andrews, Jack Cole Building, North Haugh, St Adrews, KY16 9SX, UK

Abstract

Freedom of speech has allowed people to freely communicate and social media platforms, such as Twitter offer users the opportunity to express their opinions and insights freely, there is a significant risk of users silencing each other based on prejudice by means of hateful Tweets. Hate speech is used more and more, to the point where it has become a serious problem invading these open spaces. Hate speech refers to the use of aggressive, violent or offensive language, targeting a specific group of people sharing a common property, whether this property is their gender (i.e., sexism), their ethnic group or race (i.e., racism) or their believes and religion. Since Twitter's public nature makes these messages more widely disseminated, it is important to aid in the detection of such messages, which may cause harm to targeted (groups of) users. . In this paper, we propose an approach to detect hate expressions on Twitter. Our approach is based on unigrams and patterns that are automatically collected from the training set. These patterns and unigrams are later used, among others, as features to train a machine learning algorithm. Our experiments on a dataset composed of 10,393 tweets show that our approach reaches an accuracy equal to 91.8% on detecting whether a tweet is hateful, offensive, or clean (ternary classification).

Keywords– Hate Speech Detection; Sentiment Analysis; Twitter.

INTRODUCTION

Social Media Networks (SMNs) are an ideal place for Internet users to keep in touch, share information about their daily activities and interests, publishing and accessing documents, photos and videos. They offer users a wide range of opportunities to communicate their thoughts and to share insights. Most social media platforms profile themselves as instrumental agents in promoting an Internet community in its most idealized form, namely as a space for uncensored, continuous discussion of any and all topics of interest to their users. However, the unrestricted nature of the debate

possibilities on these platforms entails an inherent risk due to the unpredictability of the users' discourse as such, they become an ideal plaza for proliferation of harmful information. Social media sites like Twitter maintain their base principle of freedom of expression and debate, but never to the expense of the well-being of their users. The underlying idea of their intolerance towards abusive and hateful behaviour is the importance of upholding a general atmosphere of safety, thereby ensuring all users feel sufficiently able to use Twitter in a productive way. Twitter is a famous platform for opinion and information sharing and this platform is mostly used before, during and after live events (Basile, 2019). To enforce their policy, Twitter, like many other social media sites, adheres to a varied strategy. They rely on user guidelines as well as the reactions of other Twitter users to disseminate the company policy. Users are able to report posts as containing hateful language, after which they are evaluated by a team of human evaluators before punitive action is undertaken towards the offending user. The human reporting and evaluation method works particularly well for instances in which the context of the Tweet largely determines its (non-) hateful nature. The Twitter policy therefore makes a distinction between "consensual" and "non-consensual" use of hateful terms, where the latter refers to actual hate speech and the former to jocular, friendly uses of offensive terms as a "means to reclaim terms that were historically used to demean individuals" (Davidson et al., 2017). It is especially these instances of consensual and covert offensive language, which pose the greatest challenges to automatic hate speech classification.

Definition of hate speech is often contested and it generally lies in a complex nexus with freedom of expression, individual, group and minority rights, along with concepts of equality, liberty and dignity. Hate speech has been defined as any form of communication which is intended to insult, intimidate or harass an individual or a group of individuals based on some characteristic (e.g., race, gender, sexual orientation, religion, nationality, etc.). Hate speech usually also expresses stereotypical assumptions about the target. Its degree of intensity can vary greatly, since its impact can range from causing offense and upsetting the target to threatening to harm or even kill the target. Davidson et al., (2017) have rightly advised researchers to not restrict themselves to the more extreme form of hate speech, which incites violence, since this would significantly decrease the amount of relevant data. Many shared tasks have been organized to tackle the challenge of hate speech detection on Twitter.

RELATED RESEARCH

The related research on hate speech detection on social media shows that most researchers consider the problem a supervised classification task. More traditional

machine learning algorithms (such as Random Forest), as well as deep learning methods have been investigated and a wide range of features have been used to tackle the task (Schmidt and Wiegand, 2017). Features typically utilized in the classification of hate speech include lexical surface level features like bag of words, unigrams and n-grams, which tend to perform quite well and provide a strong baseline. As is widely known, the automatic classification of User-Generated Content (UGC) poses a large amount of spelling variation problems. In order to capture as many language variants as possible of the offensive terms, character level n-grams are considered a vital feature (Schmidt and Wiegand, 2017). Surface-level features specific to Twitter have also been widely used, incorporating information, such as the occurrence and frequency of hashtags, mentions, URLs, retweets and tweet length (Hutto and Gilbert, 2015). Lexicon-based features consisting of "blacklists" of hateful and offensive terms are used to capture a variety of slurs and insults typical to hate speech messages. It has been shown that the more hateful racial and homophobic terms are present in a tweet, the more likely it is to be hate speech (Davidson et. Al., 2017). Syntactic information features like part-of-speech (POS) information and - on a deeper level – dependency relationships are also used to add linguistic information to the classifier. Specifically for the task of hate speech detection, the use of extra-linguistic features has been investigated. These features can be useful for the detection of the hateful intent behind the tweet, e.g., by considering the Twitter user's prior posting history and use of hateful terms. These also include information about the tweeter's ethnicity or gender, but this data is often unreliable or incomplete (Davidson et. al., 2017).

Sentiment analysis features have demonstrated their effectiveness in hate speech detection, based on the assumption that most instances of hate speech exhibit a higher degree of negative polarity than in cases where hate speech is not present. Such features can originate from external lexicons (in which case it is preferred that the lexicon be designed for the social media domain, such as VADER (Hutto and Gilbert, 2015)). However, customized hate lexicons are also constructed through the detection of language patterns in social media corpora (Basile, 2019). Gitari et al. (2015) have developed their own hate speech lexicon by using sentiment, subjectivity and semantic features. They then used this lexicon to develop a rule-based classifier for detecting hate speech.

Given the constraints in post length on a platform like Twitter, it is often difficult to determine whether a tweet truly contains hate speech. In order to supply the classifier with disambiguating contextual information, knowledge-based information (e.g., from ConceptNet Spear et. Al., 2017) is used to provide generic context. Nobata et al. (2016) utilize distributional semantics features, which relate to the immediate

context of tweets, resulting in such informative features as the preceding comments and the commenter's past behavior or comments. Djuric et al. (2015) use features derived from comment embeddings with neural language models as classification input, whereas Gao and Huang (2017) used neural models to develop contextaware models. It is evident that future research on hate speech detection would benefit greatly from the incorporation of more sophisticated contextual features.

As the state-of-the-art indicates, the task of hate speech detection is complicated by the characteristics of the social media data it is applied to. Nobata et al. (2016) consider the intrinsic noisiness of tweets as a helpful marker of hate speech and have developed features that capture different types of noise. As mentioned before, the spelling variation issue can hamper the performance of simple lexicon lookup features.

Two major issues remain as an obstacle to fully automated hate speech detection. On the one hand, there is the difficulty of detecting hateful speech whenever it is present in its more implicit form (Benicova et. al., 2018), for instance, when no offensive terms are present. On the other hand, the varied use of offensive language often leads to false positives, for example, as indicated by Davidson et al. (2017), when lyrics containing an offensive word are quoted, but more in general whenever a user is quoting someone else, often reporting on hate speech against their own person. This also includes all cases of what the Twitter policy on hateful conduct terms "consensual" use of hateful words. While the above overview and the current paper focus on the binary classification task of detecting the presence or absence of hate speech in tweets, it remains to be said that more and more researchers emphasize the importance of related sub-tasks, which offer up more fine-grained classification possibilities, especially for cases of implicit hate speech. Such tasks include detecting whether the hate is directed or generalized (Elsherif et al., 2018) and detecting the use of other language (Alorainy et al., 2019), which is a particularly salient feature for detecting hate speech against immigrants. It is important that such novel fine-grained classification methods continue to be investigated, since they show a lot of promise in capturing implicit hate speech when compared to traditional lexical "blacklist" methods.

As neural network approaches outperforms existing methods for text classification problems, this paper presents a contribution to the field of hate speech detection by developing a supervised classification method using a deep learning model, namely the Convolutional Neural Network (CNN) with linguistic features inspired by the state of the art. This classifier assigns each tweet to one of the categories of a Twitter dataset: hate, offensive language, and neither. The performance of this model has been tested using the accuracy, as well as looking at the precision, recall and F-score.

Following the assumption that hate speech typically exhibits a higher degree of negative polarity, we anticipate that adding sentiment information will improve performance. We believe adding sentiment features will help to capture more implicitly hateful tweets, which may help in the detection of tweets which have been 'edited' by offenders to ensure their messages can slip through the net of current automated hate speech detection methods (Djuric, 2015).

3. EXPERIMENTAL SETUP

A. Data

Tweets were scrapped from Twitter using the Twitter Python API which uses the advanced search option of twitter. We have mined the tweets by selecting certain hashtags and keywords from politics, religion, tribe, and public protests, riots, etc., which have a good propensity for the presence of hate speech. We retrieved 788,231 tweets from Twitter in json format, which consists of information such as timestamp, URL, text, user, re-tweets, replies, full name, id and likes. An extensive processing was carried out to remove all the noisy tweets. The dataset features 10,393 English tweets that has been classified into three classes:

1. *Hate:* tweet contains hate speech
2. *Offensive:* tweet contains offensive language but no hate speech
3. *Neither:* tweet does not contain hate speech nor offensive language

Class	**# of Tweets**	
Hate	879	
Offensive	7,511	
Neither	2,003	
Total	10,393	

Table 1.1: Summary classes in the Dataset

The distribution of the tweets across the three classes is shown in Table 1.1. These numbers indicate that approximately 8% of the tweets contain hate speech, while the majority of the tweets (72%) contains offensive language. This means that the number of tweets belonging to the three classes is quite skewed, leading to an unbalanced dataset. The size of the dataset is rather small, but will still be used for this research.

B. Preprocessing

For preprocessing the data, the Twitter-specific module tweetokenize was used (Suttles, 2019). This module took care of tokenization and converted all mentions, numbers and URLs by placeholder tags. We applied an additional function to

tokenize hashtags that was able to capture camelcased hashtags correctly. Since we created external lexicons for our emoji and smiley sentiment features, we also replaced all emojis in the data with a placeholder ('emoji') followed by the Unicode code of the emoji (e.g., 'emoji0001f194'), to ensure our featurizer would be able to recognize its presence in the document.

C. Experimental setup

The experiments have been done in Python. Here the neural network and machine learning libraries are utilized. To be more specific, for the training of the model, the Keras library with Tensorflow back-end has been used. This subsection describes the setup of the experiments that assessed the performance of the convolutional neural network.

- **Data splitting.** For this research paper it has been decided to split the data into the three separate sets, instead of using cross-validation. The reason for this is the short time span of this research. Training a CNN is computationally expensive and with the use of cross-validation it would take too much time. Thus, the division has split the dataset in training, validation, and test set of 50%, 30%, and 20% respectively. The rationale behind this ratio comes from the size of the whole dataset. To evaluate the model and refine the parameters, a larger validation set is needed.
- **Architecture.** The convolutional networks are commonly made up of only three layer types: convolution, pooling and fully-connected. Several studies have shown that it is very uncommon to have resources to successfully train a full convolutional neural network from scratch. The CNN model that has been used on the dataset is the architecture used by Zhang and Luo, (2019), which consists of a non-linear convolution layer, max-pooling layer, and softmax layer. The architecture of is used for a sentence classification, which is why it is expected that a big part of the model is transferable to this hate speech problem.
- **Input + word embedding.** The input of this model is a preprocessed tweet, which is treated as a sequence of words. To set the weight of the embedding layer, this work used the publicly available word2Vec word embedding with 300 dimensions pre-trained on the 3-billion-word from Google News with a skip-gram model.
- **Hyperparameters.** The model parameters will be based on default values or on (empirical) findings that have been reported earlier. However, the batch size and epoches are derived from the training model, which will be explained later. It should be noted that these values may not lead to optimal results, as

each setting is dependent of the data. Nonetheless, this work will show that using these parameters leads to obtaining good performances even without tuning the parameters.

- **CNN.** The dimension of the word vector is set to d = 100 at first and thus the embedding layer passes an input feature space that has a 3-dimensional tensor of shape (None, 100, 300) . The output of this layer is then fed into a 2D convolutional layer with filter layers of 3, 4, and 5, each having a 100 feature map. The rectified linear unit function is used for activation. In order to build the convolutional layers followed by max-pooling, it is needed to convert the output of the embedding in a 4-dimensional tensor. The layer of the shape then becomes (None, 100, 300, 1). As this network deals with filters of different sizes, each filter has its own layer, which is then merged into one feature vector. The filters are then slided over the sentences without padding the edges. For example, the filter that slides over 3 words gives a tensor of shape (None, 944, 1, 100) . This input feature space is then further down-sampled by a max pooling layer with a stride size of 1, which gives a tensor of shape (None, 1, 1, 100). Once all the output tensor from the pooled layers are created, each filter size is then combines into one long feature vector. Then, the activations are dropped randomly with the probability p = 0.5 and the dimension is flatten when possible. At last, the output feature vector is then taken as input in a softmax layer. This layer then predicts the probability distribution over all possible classes.
- **Optimization.** To train the model, the Adam algorithm has been used. Furthermore, default parameters given in Djuric, (2015) are used and if necessary are adjusted to get a better performance. The batch size is set to 64 and the model will be trained with 10 epochs.

RESULTS AND ANALYSIS

In this section, we present the results of our experiments on the training data. Before finalizing the model, it is important to see how the model performs by using the validation set. This will give an indication whether the parameters needs to be tuned in for a better performance. Fig. 1.1 and Fig. 1.2 give an overview of how the classifier with the given parameters (as discussed previously) has performed. Initially, a learning rate of $\alpha = 10^{-3}$, batch size of $\beta = 64$ and a dropout probability of p = 0.5 have been applied. A plot of the loss function is given in Fig 1.1.

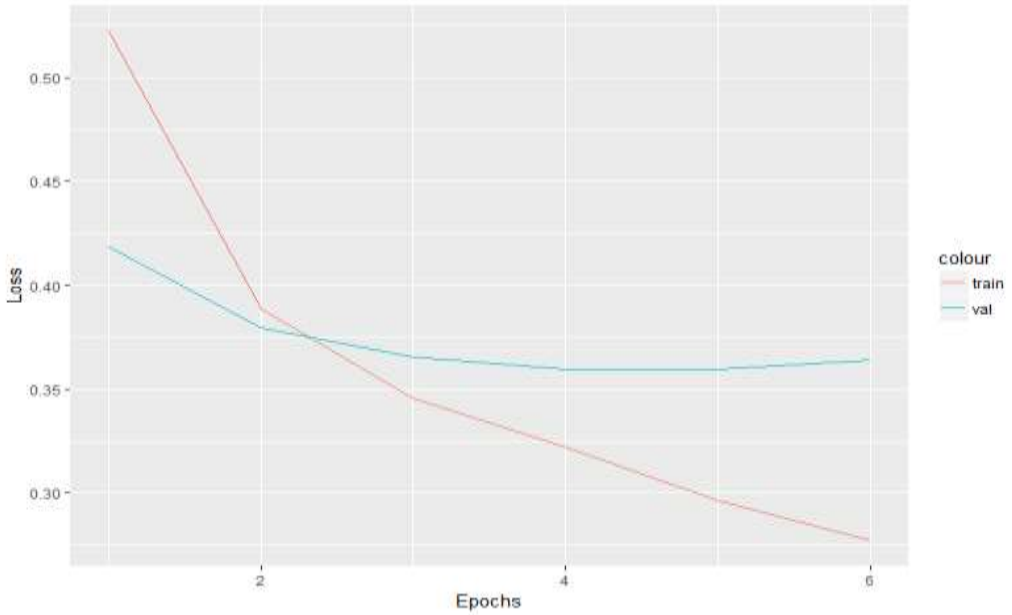

Figure 1.1: The loss for the initial model.

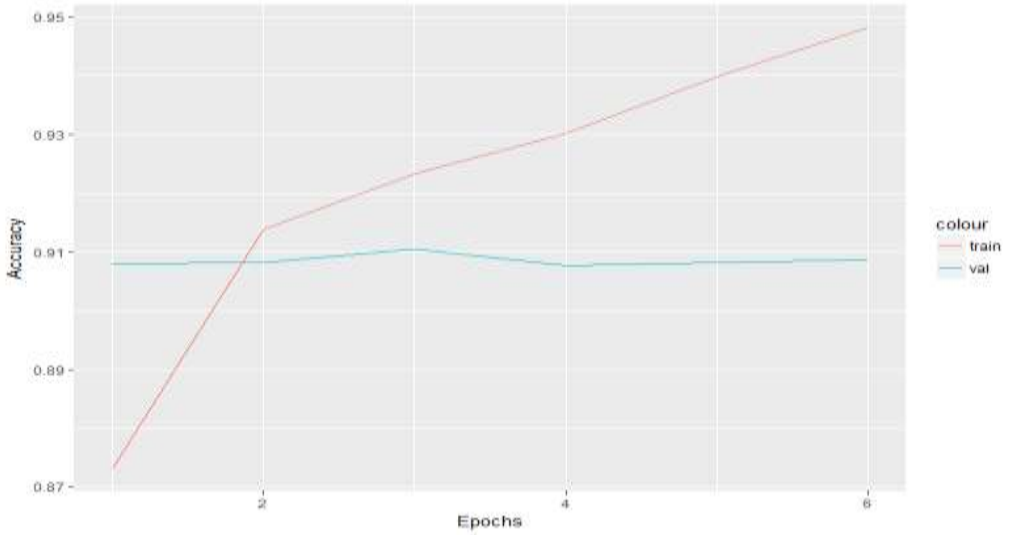

Figure 1.2: The accuracy for the initial model.

Fig. 1.2 shows the accuracy performance of the model. This shows whether the classifier is overfitting the training data. Looking at both plots, it can be seen that the model tends to overfit the training data. Moreover, these plots indicate that the learning rate may be decreased. Thus, based on these observations, the training model has been retrained with the same batch size and probability, but with a lower learning rate of $\alpha = 10^{-4}$. The results of this attempt are given in Fig. 1.3 and Fig. 1.4. Here it can be seen that the model did benefit from the lower learning rate. It can also be seen that the model overfits the training data a bit. However, the

performance of the accuracy for the training and validation do not differ too much. Therefore, a little overfitting of the training is somewhat affordable.

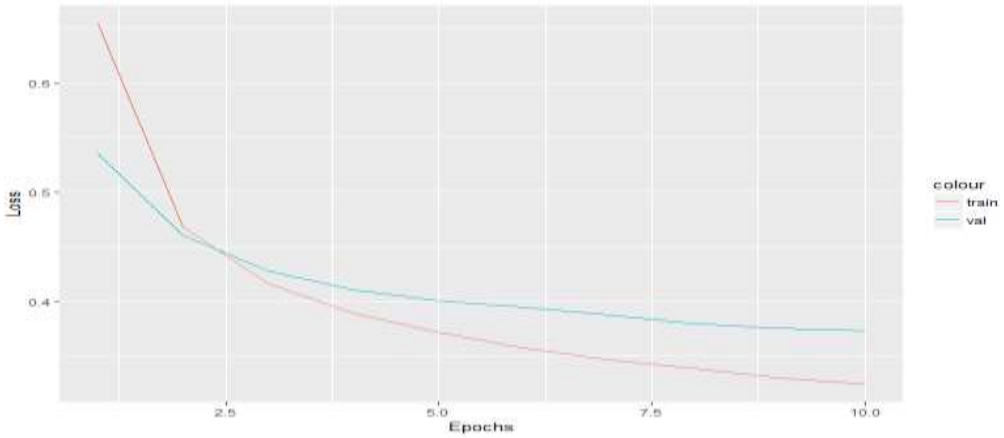

Figure 1.3: The loss for the after the changes

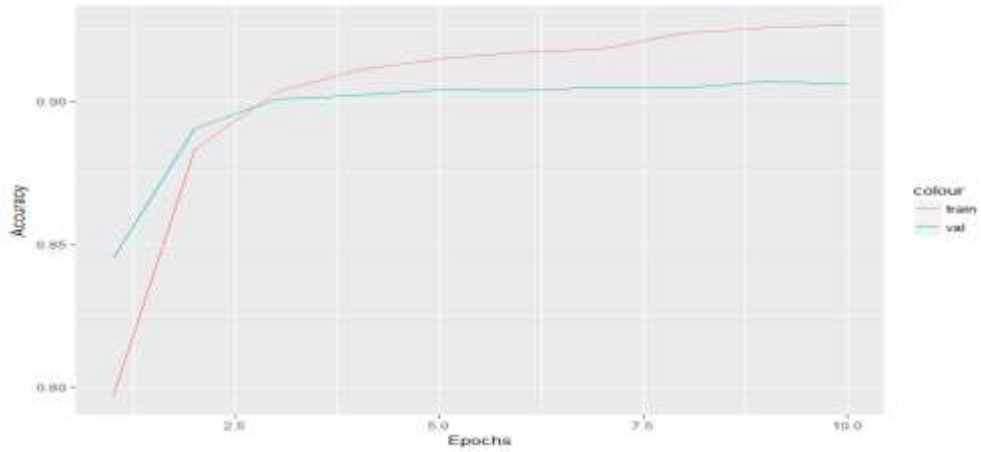

Figure 1.4: The accuracy for after the changes.

It should also be noted again that this model has been tested on a fairly small dataset. Therefore, to reduce overfitting, it is probably desirable to collect more data.

Final results. The parameters described above results in the model predicting each category with 91.8% accuracy and a loss of 36%. The final model gives an overall precision of 0.92, a recall of 0.88 and a F1-score of 0.92. Looking at Fig.1.2, it can be observed that the model overall did not identify many tweets as hate-speech tweets: almost 80% of the hate class is misclassified. This may be due to insufficient and unbalanced training data, as mentioned in Section 3 (A). This leads the model to be biased towards classifying tweets as offensive. Furthermore, the model also incorrectly identified some non-hate speech as hate speech. However it did perform

better in identifying the offensive class. This is because the number of tweets that are categorized as offensive are larger than the other two categories.

CONCLUSION AND FUTURE WORK

We have experimented with a deep learning approach incorporating different informative features for the task of hate speech detection on Twitter. Our model employed a varied feature space, ranging from linguistic information, sentiment and Twitter-specific features, to hate speech specific features. To enable successful execution of the research it was first necessary to understand what hate speech is. To accomplishing this, an overview of this topic has been conducted. Here it can be concluded that hate speech has several definition, all coming from different platforms. Hate speech detection is a classification-related task, and that is why further literature was reviewed to understand the idea behind Sentiment Analysis and the application of its various techniques. Previous work showed that deep learning models improve the state-of-art approaches within hate speech classification tasks. Therefore, a deep learning method, namely a Convolutional Neural Network (CNN), has been applied on a Twitter dataset. This data contains tweets annotated with three labels: hate, offensive language and neither. Our best model used Twitter specific features (hashtag, URL, mention) (Avg. 78.59% Fscore) and was an expansion on our token n-gram baseline (Avg. 77.71% F-score), which appeared to be a very strong baseline, as is the case for many related Natural Language Processing tasks. The sentiment features we added ended up over generating on the hate speech label, but when combined with our baseline, the scores evened out. The detailed error analysis we performed on our best and combined systems has made us reflect more generally on the biases related to the tasks of hate speech detection and the use of offensive language on social media like Twitter. Aside from the subjective biases impacting the annotations of different types of hate speech Davidson et al., (2017), it is useful to consider the research bias in hate speech detection identified by Zhang and Luo, (2018). According to these authors, the problem of hate speech detection is often viewed starting from the same research question, namely: how can we improve the system to ensure that non-hateful instances do not get classified as hateful? This leads to evaluations, which are biased towards the detection of non-hateful messages, rather than hateful ones (Zhang and Luo, 2018). It is interesting to consider how this perspective is indicative of a different focus on the use(fulness) of social media. On the one hand, the principle of freedom of expression seems to lie at the root of the bias towards detecting non-hateful tweets, since the positively evaluated detection systems are those which would not result in users innocent of the use of hate speech to be banned or to receive a warning for their “consensual” use of

offensive terms. On the other hand, system evaluations which are biased towards detecting hateful tweets seem driven by another guiding principle of social media platforms, i.e., the need to maintain the assurance of a safe space for its users. We agree with Zhang and Luo, (2018) that the second perspective is perhaps the more urgent of the two in the context of hate speech detection, but it is our opinion that other related tasks, such as detecting offensive messages would benefit more from the first perspective.

Future work direction:

- *Error analysis.* As seen in Fig. 1.2 there are a lot of tweets that have been misclassified, namely in classifying hate speech. An error analysis could therefore help to provide insights about the performance of the model. For example, when examining the wrong predictions of the hate class, this could help with clarifying why this class is so hard to predict. It would be interesting to see if and how certain terms are useful for distinguishing between hate speech and offensive language.
- *Data*. There are three ways of showing hate on Twitter: directly to a person or group, in a conversation between people, and randomly to nobody in particular. In future work, when looking at hate speech, it can also be interesting to look at and distinguish between the three ways that people show hatred on Twitter. Future work can also focus on the individual characteristics and motivation of a user for example. And of course, if possible it would be better to collect more data to make a more accurate distinction between the model performances.
- *Ensemble.* Previous works have used ensemble methods and achieved success in classification tasks. Therefore, it is expected that when these methods are used, they can further improve the results that has been obtained.

REFERENCES

Davidson, T., Warmsley, D., Macy, M., and Weber, I. "Automated Hate Speech Detection and the Problem of Offensive Language," CoRR, vol. abs/1703.04009, 2017, P. 512–515. [Online]. Available:http://arxiv.org/abs/1703.04009

Basile, V. "Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter," in Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019). Association for Computational Linguistics, 2019, P. 54–63.

Bauwelinck, N., Jacobs, G., Hoste, V., and Lefever, E. "Lt3 at semeval-2019 task 5 : multilingual detection of hate speech against immigrants and women in twitter (hateval)," in Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019). Association for Computational Linguistics, 2019, p. 5.

Schmidt A. and Wiegand, M. "A Survey on Hate Speech Detection using Natural Language Processing," in Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, 2017, P. 1–10. [Online] Available https: //doi.org/10.18653/v1/W17-1101

Hutto C. J. and Gilbert, E."VADER: A parsimonious rule-based model for sentiment analysis of social media text," in Proceedings of the 8th AAAI conference on weblogs and social media (ICWSM), 2014, P. 216–225.

Gitari, N., Zuping, Z., Damien, H. and Long, J. "A lexicon-based approach for hate speech detection," International Journal of Multimedia and Ubiquitous Engineering, vol. 10(4), 2015, P. 215–230.

Speer, R., Chin, J. and Havasi, C. "Conceptnet 5.5: An open multilingual graph of general knowledge." in Proceedings of the 31st international artificial intelligence research society conference (AAAI 31), 2017, P. 4444–4451.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y. and Chang, Y. "Abusive language detection in online user content," in Proceedings of the 25th International Conference on World Wide Web, 2016, P. 145–153.

Djuric N., "Hate speech detection with comment embeddings," in Proceedings of the 24th International Conference on World Wide Web. ACM, 2015, P 29–30.

Gao, L and Huang, R. "Detecting Online Hate Speech Using Context Aware Models," CoRR, vol. abs/1710.07395, 2017. [Online]. Available: http://arxiv.org/abs/1710.07395

Benikova, D., Wojatzki, M., and Zesch, T. "What Does This Imply? Examining the Impact of Implicitness on the Perception of Hate Speech," in Language Technologies for the Challenges of the Digital Age, G. Rehm and T. Declerck, Eds., 2018, P. 171–179.

ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., and Belding-Royer, E. "Peer to peer hate: Hate speech instigators and their targets," in Proceedings of the 12th international AAAI conference on weblogs and social media (ICWSM), 2018, P. 52–61.

Alorainy, W., Burnap, P., Liu, H. and Williams, M. "'The Enemy Among Us': Detecting Cyber Hate Speech with Threats-based Othering Language Embeddings," ACM Transactions on the Web, 9(4), P. 1–26.

Chang, C-C. and Lin, C. J. "LIBSVM: A Library for Support Vector Machines," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 2, no. 3, 2011, P. 27:1–27:27, ISSN: 2157-6904.

Suttles, J. tweetokenize. https://github.com/jaredks/tweetokenize [accessed: 2019-06-02]. (2013)

Van Hee C. "Automatic detection of cyberbullying in social media text," PLOS ONE, vol. 13, no. 10, 2018, P. 1–22. [Online].Available: http://dx.doi.org/10.1371/journal.pone.0203794

Tausczik, P. J. "The psychological meaning of words: Liwc and computerized text analysis methods," Journal of Language and Social Psychology, vol. 29, no. 1, 2010, P. 24–54. [Online]. Available: https://doi.org/10.1177/0261927X0935167

Novak, P., Smailovi, J., Sluban, B. and Mozeti I., "Sentiment of emojis," PLOS ONE, vol. 10, no. 12, 2015, P. 1–22. [Online]. Available: https://doi.org/10.1371/journal.pone.0144296

Stone P. J and Hunt, E. B. "A computer approach to content analysis: studies using the general inquirer system," in Proceedings of the AFIPS, 1963, P. 241–256.

Hu M. and Liu, B. "Mining and summarizing customer reviews," in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004, P. 168–177.

Zhang Z., and Luo, L. "Hate speech detection: A solved problem?the challenging case of long tail on twitter," Semantic Web, vol. 1, no. 0, 2018, P. 1–51.