Measuring the Impact of Sentiment for Hate Speech Detection on Twitter

Nina Bauwelinck and Els Lefever
LT^3, Language and Translation Technology Team
Ghent University, Belgium
Groot-Brittanniëlaan 45, 9000 Ghent
Email: nina.bauwelinck, els.lefever@ugent.be

***Abstract*—While 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. 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. Following current state of the art, we assume the usefulness of sentiment features for the detection of hate speech messages, which tend to exhibit a higher degree of negative polarity. Therefore, we investigate the impact of these sentiment features as well as Twitter-specific and hate speech features on the performance of a supervised classification method with Support Vector Machines (SVMs). The Twitter-specific features offer the best performance increase over our strong token n-gram baseline.**

Keywords–hate speech Detection; Sentiment Analysis; Twitter.

I. INTRODUCTION

Online platforms, such as social media networks and fora 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. 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. 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" [1]. It is especially these instances of consensual and covert offensive language, which pose the greatest challenges to automatic hate speech classification.

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. [2] 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. HatEval is one such task and has been organized by Basile et al. [3] in the context of SemEval-2019. Participating teams were asked to develop systems for the detection of hate speech against women and immigrants on Twitter, since these two groups are common targets of hateful messages online [3]. Two classification subtasks were proposed: (1) the main binary classification of the presence or absence of hate speech and (2) the fine-grained classification of hateful tweets in terms of the tweet's aggressiveness and the target of hate (individual or group).

This paper proposes a classification-based approach to hate speech detection and is an extension of previous research performed in the framework of the HatEval task [4]. The presented research is restricted to the main HatEval task, viz. the binary prediction of presence or absence of hate speech against women and immigrants. We perform a detailed analysis of the performance of a wide range of features and combinations of features, in order to get insights in the information sources that are most useful for the task of hate speech detection.

The paper is organized as follows. In Section II, we give a brief overview of the existing research and methodologies of hate speech detection on Twitter. Section III describes our experimental setup and reports on the specifics of the experiments we carried out and the different feature groups we used. In Section IV, we report on the results of our classifier incorporating different feature groups, perform a detailed error analysis and discuss possible improvements of the system. Section V concludes this paper.

II. 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 Support Vector Machines (SVMs)), as well as deep learning methods have been investigated and a wide range of features have been used to tackle the task [5]. Features typically utilized in the classification of hate

 ISBN: 978-1-61208-725-2

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 [5]. 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 [5]. 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 [2]. 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 [2].

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 [6]). However, customized hate lexicons are also constructed through the detection of language patterns in social media corpora [3]. Gitari et al. [7] 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 [8]) is used to provide generic context. Nobata et al. [9] 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. [10] use features derived from comment embeddings with neural language models as classification input, whereas Gao and Huang [11] used neural models to develop context-aware 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. [9] 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.

Finally, 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 [12], 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. [2], 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 [13] and detecting the use of othering language [14], 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.

This paper presents our contribution to the field of hate speech detection by developing a supervised classification method using Support Vector Machines (SVMs) with linguistic features inspired by the state of the art. We will investigate the classification performance impact of various feature groups and more specifically, the impact of sentiment features as opposed to lexical n-gram features. 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 [10].

III. Experimental Setup

The purpose of our experiments is to find out how well our framework is able to detect hate speech and to what measure sentiment features are able to improve the system performance in this task. To this end, we built various classifiers where different features and feature combinations were used. The task was approached as a supervised classification task and we applied the Library for Support Vector Machines (LIBSVM) [15] with the standard Radial Basis Function (RBF) kernel as the machine learning algorithm. In previous research [4], we performed a grid search to find the optimal hyperparameter settings for running the SVM on this type of data, resulting in a value of $c = 8.0$ and $g = 0.001953125$. In order to train and test the hate detection system, 5-fold Cross-validation was implemented, viz. the data is divided into 5 equal folds, allowing 80% of the data to run as training and 20% of the data as test within each fold.

A. Corpus

Our corpus consists of the English training data supplied in the HatEval shared task [3] of SemEval-2019. We conflated the development and training sets to make one large training set of 10,000 tweets. Half of these had the target "women", the other half "immigrants" [3]. The distribution of the labels is as follows: for the training set of 10,000 instances, 4210

 ISBN: 978-1-61208-725-2

are labeled as hateful (2000 of which are targeted towards immigrants, 2210 towards women) and 5790 are labeled as non-hateful (3000 targeted towards immigrants, 2790 towards women).

For preprocessing the data, the Twitter-specific module tweetokenize was used [16]. 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.

B. Information Sources

We aimed to develop a rich feature set that focused on lexical information, supplied with linguistic features. This featurization pipeline is based on work in cyberbullying detection and analysis [17]. Following similar research [5], we added surface-level Twitter-specific features to capture the use of hashtags, mentions and URLs. In order to investigate the performance impact of sentiment information and following the assumption that hate speech exhibits a higher degree of negative polarity, we added several general purpose sentiment lexicons as well as Linguistic Inquiry and Word Count (LIWC) [18] for capturing psychometric information. Additionally, we also used a sentiment polarity lexicon for emojis, the Emoji Sentiment Ranking lexicon [19], consisting of the 751 most commonly used emojis and their sentiment polarity score (positive, negative, neutral). Even though emojis tend to be more common in tweets than smileys consisting of only typographical characters, we also included a sentiment polarity lexicon for smileys. Finally, we developed a set of features specific to hate speech, comprising a lexicon look-up of profanity words, a feature to capture self-referential use of commonly used offensive words and a feature capturing the combination of a mention and a profanity word present in the tweet.

1) Linguistic features:

- **Token:** token unigrams, bigrams and trigrams.
- **Char:** character bigrams, trigrams and fourgrams.
- **Linguistic:** binary lexicon look-up features for the following types of linguistic information:
  - Allness: presence of allness word (e.g., "always", "everybody").
  - Diminishers: presence of diminisher word (e.g., "almost", "meh", "little").
  - Intensifiers: presence of intensifier word (e.g., "as fuck", "awful").
  - Negations: presence of negation word (e.g., "none", "nah", "nobody").
  - Imperative: presence of imperative mood.
  - Person-Alternation: if the instance contains references to both first and second person pronouns (e.g., "my" and "ur").
  - Names: presence of a proper noun.

2) Sentiment features:

- **Sentiment Lexicons (SL):** ratio of positive, negative and objective lexicon entry matches vs. all matches for the four sentiment polarity lexicons listed below; as well as the polarity sum of all matches in document:
  - AFINN-111 [20]: 2,477 English terms with sentiment score of -5 to 5.
  - Multi-perspective Question Answering (MPQA) opinion corpus [21]: 8,222 English terms with four sentiment score labels (positive, negative, both, neutral).
  - General Inquirer [22]: 3,644 English terms with sentiment score labels (positive, negative).
  - Hu and Liu Opinion Lexicon [23]: 13,202 English terms with sentiment score labels (positive, negative).
- **Linguistic Inquiry and Word Count (LIWC) Psychometric Features:** relative frequency of 64 psychometric categories in the 2001 version of the Linguistic Inquiry and Word Count dictionary [18].
- **Smiley and Emoji Sentiment Lexicons:** ratio of positive, negative and objective lexicon entry matches vs. all matches for two sentiment polarity lexicons (one containing 125 typographic smileys; the other one being the Emoji Sentiment Ranking lexicon [19], consisting of the 751 most commonly used emojis) as well as the sum of all matches in the tweet.

3) Twitter-Specific Features:

- **Hashtag, URL, Mention:** binary feature recording the presence of a hashtag, URL and @-Mention and a count feature making the sum of all hashtags present.

4) Hate Speech Features:

- **Profanity Lexicon:** counts exact matches with a lexicon containing 2,315 single and multiword expressions commonly used as slurs.
- **Self-Referential:** binary feature recording the presence of both a first person pronoun (singular and plural) and a common profanity word (e.g., "I" and "bitch").
- **Mention and Profanity:** binary feature recording the presence of both an @-mention and a common profanity word in the tweet.

IV. Results and Analysis

In this section, we present the results of our experiments with 5-fold Cross-Validation on the training data. We start by discussing the global scores and then we discuss some of the features in isolation, focusing on the outliers. We conduct a brief error analysis, observing some of the trends and instances where classification performance was particularly good and bad. We end this section by making some suggestions towards possible improvements on our current features.

We experimented with the different feature groups and individual features described in Section III in order to get a comprehensive overview of the precise impact of each feature addition on the performance of our hate speech classifier. The scores of our systems overall indicate good performance, since there are no massive outliers and none of our systems score lower than 57.91% micro-averaged F-score. The best performing system is TWIT-2, which utilizes token n-gram and twitter-specific (hashtag, URL, mention) information. The system combining all features performs well, with an average F-score of 78.11%, making it the third best out of all of the systems we trained.

Overall, our systems score better on the NOT (not hate speech) label than on the HS (hate speech) label. For the

 ISBN: 978-1-61208-725-2

linguistic feature groups, it can be noted that the combination of token and character n-grams works well with linguistic information (LING-3: 77.94% Avg. F1). Additionally, this system also has one of the highest F1-scores for the HS label out of all the systems. The sentiment feature groups perform poorly when compared to the other groups, since most of them overgenerate on the HS label (SENT-1, SENT-3, SENT-5, SENT-7, SENT-9). We assume that the fact that information related to sentiment is often omnipresent in tweets labeled as HS and not just those labeled as NOT, leads the classifier to consider too many instances as HS. SENT-5 presents us with the most severe case of overgenerization. This is due to a similar case of smileys and (especially) emojis being present in HS tweets as much, if not more than in non-hateful tweets.

It can be noted that the addition of token n-grams seems to balance out the scores considerably, especially for the sentiment features group. Therefore, it is no surprise that the highest scoring sentiment-informed system combines token n-gram information with sentiment features (SENT-10, with avg. F1-score of 78.05%). The LIWC feature on its own (SENT-3) overgenerates more over the HS label than the Sentiment Lexicons as a separate feature (SENT-1).

Aside from the combination system (ALL), the TWIT-2 system is the highest scoring system for the HS label (73.0% F-score) and also for the NOT label (82.3% F-score, surpassing the score of our ALL system). It makes sense that this system performs quite well, since a glance at the training data confirms that the presence of a mention is usually indicative of a hateful tweet (it assumes a target is being addressed).

(1) USERNAME You're a vapid whore; one day you'll be ugly and begging for dick scraps

Additionally, lots of the hate speech tweets targeting immigrants abound in hashtags, URLs and mentions.

(2) ey #Democrats Obama agreed wth USERNAME on ILLEGAL #Immigration Now Democrats Stop Whining and Lying and Pass a BILL Your jobs depend on it #ElectionDay #RedNationRising #Trump #MAGA #GOP USERNAME URL

Overgenerization does not solely occur for the sentiment features, however. The combination of all hate speech features, without any token n-grams (namely, HATE-4) also overgenerates for HS. For the remaining hate speech features, the addition of token n-grams once more has a balancing effect on the scores.

Having discussed the global scores for all sentiment groups, we examine some of the errors made by our best performing system (TWIT-2, with an avg. F-score of 78.59%). We observe some trends in the tweets misclassified by TWIT-2 as not containing hate speech. First of all, it is clear the Mention feature is detected in both HS and NOT-labelled instances. This is illustrated in (3), where the combination of the offensive word "cunt" and the double mention does not determine the label to be HS.

(3) USERNAME USERNAME you cunt.

Secondly, the presence of a URL is also characteristic of both labels. As illustrated in (4) and (5), this feature is often present in tweets where context is important in order to arrive at the correct classification.

(4) 30 seconds after you ' re done fucking the attitude out of her URL

(5) How basic bitches wash away their weekend sins and mistakes URL

Since a lot depends on the content that the link is referring to, it would be necessary to expand this feature to exploit this information to the full.

Thirdly, the hashtag sum feature is probably the most informative of our Twitter-specific features, since a lot of the tweets in our training data labelled as HS contain a large number of them. These are predominantly hateful tweets targeting immigrants, containing both official and unofficial campaign slogans from British ("#VoteLeave", "#Brexit") and American politics ("#EndDACA", "#DrainTheSwamp", "#BuildThatWall"); as well as hashtags like "#IllegalAlien" and aggressive imperatives, such as "#SendThemBack", "#DeportThemAll", #LockThemUp" and "#StopTheInvasion" as in (6).

(6) the cubans never assimilated in miami. thats why I left. #ThirdWorldCountry #StopTheInvasion

In order to assess the performance of all the other feature groups we experimented with, we will discuss the main trends we noted in the misclassifications made by our ALL system (where the gold standard has the label HS and our system predicted NOT), containing all feature information. Concerning our hate speech features, we observe a number of errors related to specific offensive terms, which were not present in our profanity lexicon (HATE-1), for example the term "rapefugee" as in (7) and "roachingfugee" as in (8).

(7) USERNAME He's not a refugee, he's a RapeFugee!!!Past time to PURGE the West

(8) Absurdity! The Swedes overwhelmingly voted for Democracy, Freedom, Human Rights, and the Nectar of Rapefugee welfare! I'm so moved I'm willing to fund 10 of the local Roachingfugees to fuck off there never return. I hope the Swedish Gormint funds me in this grand undertaking.

Such cases can easily be captured in future by sufficiently expanding the profanity lexicon. However, determining how far such a lexicon needs to be expanded is not straightforward, since many creative insults appear with words less commonly considered to be offensive (as is the case in (9)).

(9) USERNAME USERNAME Tina, you willfully ignorant somnabulist kunt, have a beer

Our 'Self-Referential' hate speech feature was introduced in the assumption that it would increase the classification performance on instances, which contained a type of "consensual" use of offensive terms, namely when the Twitter users are referring to themselves by means of an offensive word in a self-deprecating manner. However, all of these instances were incorrectly classified by our system as hate speech, e.g., (10):

(10) I'm such a little pussy ass bitch on my period what tha fook

TABLE I. PRECISION, RECALL AND F-SCORES FOR THE HS (hate speech) AND NOT (NOT hate speech) LABEL, AND THE MICRO-AVERAGED F-SCORE (%). RESULTS OF 5-FOLD CROSS-VALIDATION EXPERIMENTS ON THE TRAINING SET.

Feature Group	Features	P_HS	R_HS	F_HS	P_NOT	R_NOT	F_NOT	AVG_F-score
Lexical Features								
LING-1	Token	65.7	**77.9**	71.3	86.5	77.6	81.8	77.71
LING-2	Char	69.0	75.9	72.2	84.0	78.8	81.4	77.68
LING-3	Token + Char + linguistic	69.5	76.0	72.6	84.1	**79.1**	81.5	77.94
Sentiment Features								
SENT-1	Sentiment Lexicons (SL)	28.5	62.8	39.2	87.7	62.8	73.2	62.76
SENT-2	Token + SL	67.9	77.2	72.2	85.5	78.5	81.8	78.04
SENT-3	LIWC	3.6	73.4	6.9	99.1	58.6	73.6	58.86
SENT-4	Token + LIWC	66.7	77.8	71.8	86.1	78.0	81.9	77.93
SENT-5	Smiley and Emoji	0.1	62.5	0.2	**99.9**	57.9	73.3	57.91
SENT-6	Token + Smiley and Emoji	65.4	77.7	71.1	86.4	77.5	81.7	77.55
SENT-7	SL + Smiley and Emoji	28.2	63.3	39.0	88.1	62.8	73.3	62.87
SENT-8	Token + SL + Smiley and Emoji	67.6	77.1	72.0	85.4	78.4	81.7	77.89
SENT-9	SL + Smiley and Emoji + LIWC	28.5	63.9	39.4	88.3	62.9	73.5	63.11
SENT-10	Token + SL + Smiley and Emoji + LIWC	68.1	77.1	72.3	85.3	78.6	81.8	78.05
Twitter-Specific Features								
TWIT-1	Hashtag, URL, Mention	**73.2**	56.5	63.8	59.0	75.2	66.1	64.95
TWIT-2	Token + Hashtag, URL, Mention	68.6	**77.9**	73.0	85.8	79.0	**82.3**	**78.59**
TWIT-3	Token + Hashtag, URL, Mention + Smiley and Emoji	68.2	77.5	72.6	85.6	78.7	82.0	78.27
Hate Speech Features								
HATE-1	Token + Profanity Lexicon	66.6	77.7	71.7	86.0	78.0	81.8	77.83
HATE-2	Token + Self-Referential	65.6	77.7	71.2	86.3	77.5	81.7	77.59
HATE-3	Token + Mention and Profanity	65.4	77.6	71.0	86.3	77.4	81.6	77.48
HATE-4	Profanity Lexicon + Self-Referential + Mention and Profanity	16.3	69.0	26.4	94.7	60.9	74.1	61.68
HATE-5	Token + Profanity Lexicon + Self-Referential + Mention and Profanity	66.7	77.4	71.7	85.8	78.0	81.7	77.78
All Features								
ALL	All Lexical + Sentiment + Twitter-Specific + Hate Speech Features	71.1	75.5	**73.2**	83.2	79.8	81.5	78.11

There are also instances in which it contributes to the misclassification as NOT hate speech of examples, such as (11):

(11) USERNAME USERNAME USERNAME USERNAME and I thought I was a bitch, but you are, well the worst cunt ever.

This feature did perform well in the classification of tweets as NOT hate speech, in which the user is reporting having experienced hate speech against their person (12) or someone close to them (13).

(12) This is besides the catcalls, the hey babys and calling me a slut and bitch when I wasnt interested. Fuck this shit.

(13) Who the fuck is calling my girl a whore and a bitch? She hasn't even been at that school for a whole month, people are fucking stupid

The self-referential feature has helped to correctly classify tweets containing consensual use of offensive terms in very obviously non-hateful contexts, such as birthdays (14):

(14) ::smirking face emoji:: ::dancer emoji:: ::rose emoji:: ::revolving hearts emoji:: happiest of birthdays to the main hoe I hope you have a wonderful day angel USERNAME

We also believe that the Self-Referential feature might help to capture instances containing (rap) lyrics [2], since these are often sung from a first person perspective.

Finally, even though our sentiment features led to over-generation on the HS label, our combined system (ALL) also misclassified examples like (15). This leads us to recommend the use of a feature capturing the combination of strongly negative words like "curse' and "cunt" with smileys with positive polarity as being indicative of aggression.

(15) I saw a women driving whilst speaking on the phone I cursed that cunt :)

In order to increase the effectiveness of our lexicon look-up features, it would be useful to include bigram lexicons as well. For our sentiment features, we used general purpose sentiment lexicons. However, for future experiments, it would be useful to try out lexicons, which are more sensitive to sentiment expressions in social media contexts, such as the Valence Aware Dictionary and sEntiment Reasoner (VADER) [6]. Furthermore, our system currently lacks syntactic information. We tried to mitigate this by adding lexicons containing linguistic information, like the presence of 'allness' and 'diminisher' words, but in future research we would like to incorporate more syntactic features, like POS and dependency relationships. As illustrated by example (16), which was misclassified as NOT hate speech, the performance of our combined system would benefit from begin able to capture long-range dependencies:

(16) Anyone would think that Julie Hartley-Brewer is jealous of rape victims, the way the vacuous, malicious cunt is going on about it.

This syntactic information could also improve our performance on examples of self-referential and "consensual" use of offensive terms (cfr. Supra for (10)). Finally, we can conclude that our system managed to classify the more violence-inciting instances of hate speech correctly ((17), (18)).

(17) I'd punch that cunt in the fuckin mouth. Idgaf if she's a woman URL

(18) USERNAME Its time to give hunting licenses to border homes and make it year round open season on this invading force! Free guns and ammo to

those willing to defend our borders!!! #BuildThatWall #BuildThatDamnWallNow

V. Conclusion and Future Work

We have experimented with a supervised machine learning approach incorporating different informative features for the task of hate speech detection on Twitter, building upon our previous participation to the HatEval task of SemEval-2019. Our model employed a varied feature space, ranging from linguistic information, sentiment and Twitter-specific features, to hate speech specific features. Our best model used Twitter-specific features (hashtag, URL, mention) (Avg. 78.59% F-score) 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 overgenerating 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 [2], it is useful to consider the research bias in hate speech detection identified by Zhang and Luo [24]. 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 [24]. 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 [24] 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.

References

[1] "Hateful conduct policy," 2019, URL: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy [accessed: 2019-06-02].

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

[3] V. Basile et al., "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, pp. 54–63.

[4] N. Bauwelinck, G. Jacobs, V. Hoste, and E. Lefever, "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.

[5] A. Schmidt and M. Wiegand, "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, pp. 1–10. [Online]. Available: https://doi.org/10.18653/v1/W17-1101

[6] C. J. Hutto and E. Gilbert, "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, pp. 216–225.

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

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

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

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

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

[12] D. Benikova, M. Wojatzki, and T. Zesch, "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, pp. 171–179.

[13] M. ElSherief, S. Nilizadeh, D. Nguyen, G. Vigna, and E. Belding-Royer, "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, pp. 52–61.

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

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

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

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

[18] P. J. Tausczik, Y.R., "The psychological meaning of words: Liwc and computerized text analysis methods," Journal of Language and Social Psychology, vol. 29, no. 1, 2010, pp. 24–54. [Online]. Available: https://doi.org/10.1177/0261927X09351676

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

[20] F. Nielsen, "A new anew: Evaluation of a word list for sentiment analysis in microblogs," in Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages 718 in CEUR Workshop Proceedings, 2011.

[21] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in Proceedings of the conference on human language technology and empirical methods in natural language processing. ACL, 2005, pp. 347–354.

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

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

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

 ISBN: 978-1-61208-725-2