LT3 at SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter (hatEval)

Nina Bauwelinck, Gilles Jacobs, Véronique Hoste and Els Lefever
LT3, Language and Translation Technology Team
Department of Translation, Interpreting and Communication – Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
firstname.lastname@ugent.be, gillesm.jacobs@ugent.be

Abstract

This paper describes our contribution to the SemEval-2019 Task 5 on the detection of hate speech against immigrants and women in Twitter (hatEval). We considered a supervised classification-based approach to detect hate speech in English tweets, which combines a variety of standard lexical and syntactic features with specific features for capturing offensive language. Our experimental results show good classification performance on the training data, but a considerable drop in recall on the held-out test set.

1 Introduction

The exponential growth of social media such as Twitter, Facebook, Youtube and community forums has created a variety of novel ways for all of us to communicate with each other, but this opportunity to freely communicate online has unfortunately also given a forum to people who want to denigrate others because of their race, colour, gender, sexual orientation, religion, etc. While there has been an increasing interest in automatic hate speech detection in social media, the problem is far from solved, partly due to the low consensus on what exactly constitutes hate speech, how it relates to offensive language and bullying and thus the low reliability of hate speech annotations (Ross et al., 2017). Davidson et al. (2017) for example observe that their classifications of hate speech tend to reflect their own subjective biases: while racist and homophobic insults are considered hateful, they tend to see sexist language as merely offensive. When we consider the different approaches that address hate speech, we can observe that -apart from simple methodologies that rely on lookup in a dictionary of hateful terms (Tulkens et al., 2016) - most methods cast the problem as a supervised classification task either using a more standard machine learning approach or deep learning methods (Pitsilis et al., 2018). This was also the approach we took for our hate speech detection system.

We participated for both subtasks proposed for English for Task 5 (Basile et al., 2019), being TASK A, which was defined as a binary classification task where systems have to predict whether a tweet with a given target (women or immigrants) is hateful or not hateful, and TASK B, where systems are asked first to classify hateful tweets as aggressive or not aggressive, and second to identify the target harassed as individual or generic (i.e. single human or group).

2 System Description

We designed a cascaded classification-based approach, where a first classifier categorizes a tweet as being hateful or not, while in a second step the hateful tweets are classified as (a) being aggressive or not, and (b) the target as being individual or generic. For the second step we built separate classifiers for both subtasks (a) and (b).

2.1 Preprocessing

We applied the Twitter-specific tweetokenize (Suttles, 2013)[1] module for tokenization and preprocessing. With this module, we were able to ensure all unicode emojis would be preserved. It also allowed us to add emoticons to the module's lexicon, to avoid splitting them up. We used the module with standard settings: allcaps were maintained; @mentions were replaced by "USERNAME"; urls replaced by "URL". We decided to preserve hashtags and numbers (but replacing phonenumbers and times by "PHONENUMBER" and "TIME", respectively), as well as quotes and stopwords. Finally, we applied a function to tokenize

[1] https://github.com/jaredks/tweetokenize

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 436–440

the hashtags, but this proved insufficient, as it did not tokenize camelcased hashtags correctly.

2.2 Featurization

We aimed to develop a rich feature set that focused on lexical information with some syntactic and non-linguistic features included. This featurization pipeline is based on work in cyberbullying detection and analysis (Van Hee et al., 2018). The whole set of features listed below was used to build all three classifiers. We did not apply any feature selection.

Bag-of-words features: We included binary token unigrams, bigrams and trigrams, along with character trigrams and fourgrams. The latter provide robustness to the spelling variation typically found in social media.

Lexicon features: We computed positive and negative opinion word ratio and overall post sentiment using both the MPQA (Wilson et al., 2005) and Hu and Liu's (Hu and Liu, 2004) opinion lexicons. We added positive, negative and neutral emoji counts based on the BOUNCE emoji sentiment lexicon (Kökciyan et al., 2013). We also included the relative frequency of all 64 psychometric categories in the Linguistic Inquiry and Word Count (LIWC) dictionary (Pennebaker et al., 2007). Furthermore, we included diminisher, intensifier, negation, and "allness" lexicons which relate to a negative mindset in the context of suicidality research (Osgood and Walker, 1959; Gottschalk and Gleser, 1960; Shapero, 2011) as well as a proper name gazetteer.

Syntactic features: Two binary features were implemented indicating whether the imperative mood was used in a post and whether person alternation occurred (i.e. combinations of first and second person pronouns).

2.3 Experimental Setup

As mentioned in Section 2, we built three different classifiers to tackle the various subtasks: (1) determine whether a tweet is hateful or not, (2) for tweets classified as hateful, determine whether the target is individual or generic and (3) for tweets classified as hateful, determine whether the tweet is aggressive or not. As the classification algorithm we used LIBSVM (Chang and Lin, 2011) with linear kernel. For each classification task, we performed a grid search to find the optimal cost parameter using 5-fold cross-validation (CV) on the training data. The resulting hyperparameter ($c = 0,03125$) was applied in four different experimental setups: LIBSVM with RBF kernel not normalized, RBF kernel normalized, linear kernel not normalized and linear kernel normalized. These experiments revealed the setup with the linear kernel using normalized data as the best performing system. Table 1 presents the 5-fold CV results for this system on the training set. The experimental results on the training data show good detection results of hate speech (F-score of 71.7% on the positive class), very good results for the classification of the target as being generic or individual (average F-score of 87.5%) and lower classification performance for the classification of aggressive tweets (average F-score of 66.2%).

2.4 Competition Results and Analysis

Table 2 lists the results of our optimized linear kernel system for Task 1, whereas Table 3 shows the results for all three subtasks. As is clear from the results for task 1, our system undergenerates, resulting in a recall of only 6.8%, as opposed to 74.3% on the training data.

Upon closer inspection of the classification errors made by our system for task 1 (hate speech or not), it is possible to distinguish a few broad error categories. In order to reveal possible causes for our low recall, we will focus on hate speech that was not detected by the classifier.

2.4.1 HS Towards Women

Prediction: NOT, Gold: HS. Firstly, we consider the tweets targeted at women. For most of the tweets misclassified by our system as not containing hate speech, it is unclear which features of the text may have caused confusion, as is illustrated by example 1:

(1) That bitch on the spoiled whore list Thought she could play games with Drake but forfeit

Similar examples include at least one offensive term and are clear instances of hate speech. There are, however, some tweets containing some elements which may explain the erroneous classification.

Insults which contain a metaphorical element (such as "meat" in the example below) are hard to classify for our system:

(2) Open your mouth & take the meat like a hoe you bitch ass

	#instances	Recall	Precision	F-score
	TASK 1: hateful or not			
Hateful	4210	74.3	69.2	71.7
Not Hateful	5790	78.7	82.6	80.6
Macro Avg	10000	76.5	75.9	76.2
	TASK 2: individual or generic			
Individual	1560	83.4	82.8	83.1
Generic	2650	89.9	90.3	90.1
Macro Avg	4210	86.7	86.6	86.6
	TASK 3: aggressive or not			
Aggressive	1763	61.1	53.0	56.8
Not Aggresive	2447	69.1	75.7	72.2
Macro Avg	4210	65.1	64.4	64.5

Table 1: Cross-validation results on the training data with the linear kernel with optimised hyperparameter settings.

	Recall	Precision	F-score
	TASK 1: hateful or not		
Hateful	6.8	55.9	12.1
Not Hateful	96.1	58.6	72.8
Macro Avg	51.4	57.3	42.5
Accuracy	58.5		

Table 2: Final results task 1 of best system (optimized linear kernel with normalization).

	Recall	Precision	F-score
	TASK 1: hateful or not		
Macro Avg	51.4	57.3	42.5
Accuracy	58.5		
	TASK 2: individual or generic		
Macro Avg	53.4	60.9	52.9
Accuracy	81.3		
	TASK 3: aggressive or not		
Macro Avg	50.1	90.1	44.7
Accuracy	80.2		

Table 3: Final results of best system (optimized linear kernel with normalization).

Finally, we have noticed that tweets containing self-directed or self-referential insults are often misclassified as not containing hate speech:

(3) @USERNAME 1o million, one cent less, i am a liei8ng son of a bitch and my mom is a whore

Some of the errors can be explained by the ambiguity inherent in the tweet in question. First of all, our system has some difficulty distinguishing between genuine insults and jocular insults between friends. For example, this tweet has been classified as not containing hate speech:

(4) @USERNAME you hoe you need to be checked for aids bitch.

On the other hand, many tweets have been correctly classified as not containing hate speech. These usually contain positive words like ”love” and ”friend”:

(5) Lana i love you bitch. Put that flag back up hoe [USA flag emoji] #lustfoflife

There are also some edge cases, where there is ambiguity as to the intent of the tweeter. It makes sense that our undergenerating system is not preferred in cases such as these:

(6) @USERNAME Bitch RT me one more time & not answer my text [face with steam from nose]

A second error type of ambiguous tweets is that of the tweet containing an intensifier, which is usually ”bitch”, for example:

(7) i’m so salty i tried to go in a store with no shoes on and the ladies like ”you have to have shoes” BITCH IM NOT VCARRYING A DISEASE

A third type of ambiguity occurs when there is some kind of reference to or citation of an existing text containing offensive language. For instance, Tweets sometimes contain references to well-known vines. Some of these are correctly classified by our system (first example, classified

as NOT HS), while others are not (second example, misclassified as NOT HS):

(8) I love you...bitch. I ain't never gonna stop loving you...bitch. You're my best friend in the whole wide world and I can't wait to hug you

(9) when ugly hoes try to come for you .... Bitch DISGUSTING

A fourth category of errors caused by ambiguity occurs whenever the tweeter is making a general remark, characterized by the use of plural form ("bitches", "niggas"). These tweets usually begin with a phrase indicating the sentiment of the speaker ("I'm tired of") and can contain an intensifier.

(10) I'm tired of bitches wearin no clothes then got the nerve to say niggas is insecure like bitch I'm tryn help you not look like a hoe

We would like to finish this section with some general remarks on this part of the task, with the focus on HS targeted towards women. After close inspection of the results, we noted the difficulty of annotating such data. First of all, tweets with jocular intent are not always easy to distinguish. We did observe that tweets containing references to birthdays tended not to be hate speech, for example:

(11) @USERNAME happy birthday bitch! Hope you have a fab day you hoe n get on lifts soon [car emoji] URL

Second, the truly ambiguous tweets, those which confuse both man and machine, appear to be those in which there is potential ironic intent. This is mostly the case whenever there is overexaggeration (the reaction to the colleague's question is disproportionate to the topic of tie dye t-shirts):

(12) This bitch at work had the nerve to ask why are all y'all wearing tye die.. because u lil whore it's lit

Third, lots of tweets include references to real-life situations (e.g. someone quoting an insult they have received); or to memes (like the "Me: X, Also Me: X" meme in the second example):

(13) Fat whore!!! Ugly bitch!!! Me: URL

2.4.2 HS Towards Immigrants

Prediction: NOT, Gold: HS. Once again, our system undergenerates on detecting hate speech targeted at immigrants. First of all, many tweets contained the offensive term "shithole". The fact that this is a compound term may explain why our classifier was unable to pick up on "shit" as offensive:

(14) @JoShawTx a_meluzzi #BanMuslims #BanIslam #SendThemBack to their shitholes!! #CloseTheBorders

Our classifier was not able to pick up on hate speech in tweets containing a metaphorical meaning. For example:

(15) Bloody foreign hornets coming over here killing our bees. #brexitmeansbrexit #sendthemhome #controlourborders

Finally, it is clear that this part of the task requires a classifier which is better able to pick up on derogatory use of the terms "invasion/invader", "illegal alien" and is able to deal with contracted insults like "rapefugee":

(16) @ShadilayAlways Illegal invaders that deserve to be deported, as usual. #DeportThemAll #BuildTheWall #KAG #SupportICE

(17) @AhsanNasir5 @enough14 @AZeckenbiss Gtfo off europe. Here is no place for you, you rapefugee.

It turns out our system was only able to correctly classify the tweet as containing hate speech when it also contained other offensive words:

(18) This fuck is a weak cunt and puy his fellow country men in a bad way with the rapefugee invasion. Deport THEM ALL.

3 Conclusion

We applied a supervised classification-based approach to the task of hate speech detection against women and immigrants, incorporating a variety of standard lexical and syntactic information with specific features for capturing offensive language. Our results revealed good classification performance on the training data, but a lower performance on the evaluation set, with a notable drop in recall for all subtasks. A detailed error analysis revealed a number of tendencies, but inherently

ambiguous tweets (irony, references to memes and vines, etc.) appeared to be the major cause of classification errors for hate speech towards women. Hate speech against immigrants seems to be characterized by compound terms containing offensive parts (e.g. "rapefugee", "shitholes").

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Rangel, Paolo Rosso, and Manuela Sanguinetti. 2019. 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.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27:1–27:27. ISSN: 2157-6904.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of ICWSM*.

Louis Gottschalk and Goldine Gleser. 1960. An analysis of the verbal content of suicide notes. *British Journal of Medical Psychology*, 33(3):195–204.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Nadin Kökciyan, Arda Çelebi, Arzucan Özgür, and Suzan Üsküdarl. 2013. BOUNCE: Sentiment Classification in Twitter using Rich Feature Sets. In *Second Joint Conference on Lexical and Computational Semantics (*SEM): Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 554–561, Atlanta, Georgia, USA. Association for Computational Linguistics.

Charles Osgood and Evelyn Walker. 1959. Motivation and language behavior: A content analysis of suicide notes. *The Journal of Abnormal and Social Psychology*, 59(1):58.

James Pennebaker, Roger Booth, and Martha Francis. 2007. Liwc2007: Linguistic inquiry and word count. *Austin, Texas: liwc. net*.

Georgios Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Effective hate-speech detection in twitter data using recurrent neural networks. *Applied Intelligence*.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis.

Jess Jann Shapero. 2011. *The language of suicide notes*. Ph.D. thesis, University of Birmingham.

Jared Suttles. 2013. tweetokenize. *GitHub repository*.

Stephan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. A dictionary-based approach to racism detection in dutch social media. *Proceedings of the First Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS 2016)*.

Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and Veronique Hoste. 2018. Automatic detection of cyberbullying in social media text. *PLOS ONE*, 13:22.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.

—