The binary trio at SemEval-2019 Task 5: Multitarget Hate Speech Detection in Tweets

Patricia Chiril
IRIT
Toulouse University
patricia.chiril@irit.fr

Farah Benamara
IRIT, CNRS
Toulouse University
benamara@irit.fr

Véronique Moriceau
LIMSI-CNRS
Univ. Paris-Sud
moriceau@limsi.fr

Abhishek Kumar
Indian Institute of Science
abhishekkumar12@iisc.ac.in

Abstract

The massive growth of user-generated web content through blogs, online forums and most notably, social media networks, led to a large spreading of hatred or abusive messages which have to be moderated. This paper proposes a supervised approach to hate speech detection towards immigrants and women in English tweets. Several models have been developed ranging from feature-engineering approaches to neural ones. We also carried out a detailed error analysis to show main causes of misclassification.

1 Motivation

Social media networks such as Facebook, Twitter, blogs and forums, have become a space where users are free to relate events, personal experiences, but also opinions and sentiments about products, events or other people. This massive growth of user generated web content, along with the interactivity and anonymity the internet provides, may lead to a large spreading of hatred or abusive messages which have to be moderated.

In spite of no universally accepted definition of hate speech and the way it differs from offensive language, there are some common elements that seem to arise. In particular, these messages may express threats, harassment, intimidation or ”disparage a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic” (Nockleby, 2000).

In this paper, we focus on automatic hate speech detection towards *two different targets* – immigrants and women and we propose several multi-target hate speech detection systems. The task is performed over a collection of English tweets annotated as conveying hate speech against both immigrants and women, as part of HateEval@SemEval2019 (Basile et al., 2019). The first challenge involves building a binary classifier able to determine whether a tweet with a given target (women or immigrants) is hateful or not hateful. For this, we propose both features-based models (relying on both language-dependent and language independent features) and a neural model. We also performed a detailed error analysis to identify main causes of misclassification. Our analysis shows that errors come from several factors, which show the complexity of the task: the presence of irony and sarcasm, the lack of context and implicit hate speech. We also identified tweets for which we question the original label when taking into account the class definition.

The paper is organized as follows. Section 2 briefly presents the current state of the art. Section 3 describes our data and models, while Section 4 analyses the experiments we carried out on multi-target detection. We conclude by providing some perspectives for future work.

2 Related work

Hateful speech can be expressed at different linguistic granularity levels going from lexical to discursive (Cameron, 1992). Both sexism and racism can be expressed explicitly or implicitly (see the following tweets from our data) using different pragmatic devices, including:

- Negative opinion, abusive message: *Stop tweeting about football. You're a girl and you opinion doesn't count. #WomenSuck.*
- Stereotype: *Illegals are dumping their kids heres o they can get welfare, aid and U.S School Ripping off U.S Taxpayers #SendThemBack ! Stop Alowing illegals to Abuse the Taxpayer #Immigration.*
- Humor, irony, sarcasm: *Where is this? Brazil? Uganda? Sudan? Nope, it is France.*

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 489–493
Minneapolis, Minnesota, USA, June 6–7, 2019. ©2019 Association for Computational Linguistics

> *Got to love that cultural enrichment thing going on. #openborders #refugeesnotwelcome #slums.*

For most of the harassment and hate speech detection tasks, the classifiers still rely on supervised learning, and when creating a new classifier, one may directly feed different types of features to the classical algorithms (Naive Bayes, Logistic Regression, Random Forest, SVM) or use deep learning methods that will automatically learn abstract features from data instances. Due to the noise present in the data (especially on social media), many authors choose to combine n-grams (due to their high prediction rate) with a large selection of additional features: linguistic features that take into consideration the POS information, dependency relations (long-distance relationship in between words), or word embeddings, which have the advantage of having similar vector representations for different, but semantically similar words. Several approaches incorporate sentiment analysis as a supplementary classification step, assuming that generally negative sentiment relates to a hateful message (Dinakar et al., 2012; Sood et al., 2012).

Although within the Automatic Misogyny Identification shared task at IberEval 2018 the best results were obtained with Support Vector Machine models with different feature configurations, there are also a few notable neural networks techniques deployed in order to detect hate speech in tweets that outperform the existing models: in (Badjatiya et al., 2017) the authors used three methods (Convolutional Neural Network (CNN), Long short-term memory and FastText) combined with either random or GloVe word embeddings. In (Zhang and Luo, 2018) the authors implemented two deep neural network models (CNN + Gated Recurrent Unit layer and CNN + modified CNN layers for feature extraction) in order to classify social media text as racist, sexist, or non-hateful.

3 Multitarget hate speech detection systems

Automatically labelling tweets as hateful or not hateful is a challenging task because the language of tweets is full of grammatically and/or syntactic errors, it lacks conversational context, might consist of only one or a few words and because they can be indirectly hateful (by employing techniques such as sarcasm, satire or irony) it makes the task of text-based feature extraction difficult.

3.1 Data

Our data comes from two corpora. The first one, is an already existing corpus containing English tweets annotated for hate speech against immigrants and women, as part of the HatEval task at SemEval2019 (Basile et al., 2019). The second one was created as a result of the conclusions we had drawn after analyzing the data, i.e. we observed that for most of the tweets, even though the message appeared to be positive, just by having a certain hashtag used, it becomes negative. The hashtag importance is also supported by a simple experiment that includes in the pre-processing step hashtag removal. This leads to a decrease in accuracy by 4% and F-score by 5%. Thus we created a new dataset (DATASET++) by collecting the most used hashtags (we used scrape-twitter[1]) in both hateful (#buildThatWall) and non-hateful tweets (#refugees), as well as the most used hashtags in the misclassified tweets[2].

Table 1 shows the distribution of the tweets for the task of hate speech detection.

Task	#hate	#nonHate	Total
DATASET	5 512	7 559	13 071
DATASET++	17 989	21 921	39 909

Table 1: Tweet distribution in the corpora

For the task at hand, several models have been built, all tested using 10-cross-validation. In the next sections, we detail our models and then provide our results.

3.2 Models

Baseline (B)**.** In all the experiments, we used Bag of Words (BoW) model as lexical features. Due to the noise in the data, we performed standard text pre-processing by removing user mentions, URLs, RT, stop words, degraded stop words and the words containing less than 3 characters, and we stemmed all the remaining words by using the Snowball Stemmer[3].

Feature-based models. We experimented with several state of the art features that have shown to

[1] https://www.npmjs.com/package/scrape-twitter

[2] #maga, #usa, #trump, #sendThemBack, #immigration, #noDaca, #deportThemAll, #meToo, #stopTheInvasion, #illegalAliens, #apathyKills, #withImmigrants

[3] http://snowballstem.org

be useful in hate speech detection and we relied on a manually built emoji lexicon that contains 1 644 emojis along with their polarity. We also tested whether by identifying the users opinion we can better classify his attitude as hateful or non-hateful by making use of HurtLex (a multilingual hate word lexicon divided in 17 categories) (Bassignana et al., 2018) and a lexicon containing 1 818 profanity English words created by combining a manually built offensive words list, the noswearing dictionary [4] and an offensive word list[5].

We experimented with several combinations of the features above and we used the best performing ones for training four classifiers:

- C_1 : combines the length of the tweet with the number of words in the HurtLex lexicon with a Baseline architecture
- C_2 : combines the number of words in the offensive lexicon, the number of positive and negative emojis and emoticons and the presence of URLs with a Baseline architecture, but applied on the extended dataset
- C_3 : combines the number of words in the offensive lexicon, the number of positive and negative emojis and emoticons and performs linear dimensionality reduction by means of truncated Singular Value Decomposition and used Random Forest only for intermediate classification, whose output were then combined and passed onto a final Extreme Gradient Booster classifier
- C_4 : the same as C_3 but applied on the extended dataset

Neural model. The last model (C_5) used a Bidirectional LSTM with an attention mechanism. For the task at hand, we used pre-trained on tweets Glove embeddings (Pennington et al., 2014).

4 Results

We tried several machine learning algorithms in order to evaluate and select the best performing one. Hereby, the hate speech system baseline is a Random Forest classifier. Table 2 shows how the experiments were set up and presents the results in terms of accuracy (A), macro-averaged F-score (F), precision (P) and recall (R). For each of the systems we present the results obtained on 10-cross validation (using the provided train, trial and dev datasets) and the official results.

Among the five systems, C_2 represents our best performing one during the development phase [6], while C_5 performed best in the evaluation phase.

Due to a significant decrease in both accuracy and F-score on the official test data, we also investigated the influence of the data distribution in the train and test datasets. The results obtained after shuffling and re-splitting the data (while keeping the original distribution of the tweets) are also presented in Table 2. It is important to realize that these results were obtained by using a train-test configuration on a random test, not by using cross validation. These results are comparable to the ones obtained during the development phase.

As we encountered a significant decrease in the system's performance in the official test, we decided to conduct a deeper analysis in order to identify the main causes of errors.

5 Discussion

Error analysis shows that in the misclassification of hateful instances intervene several factors: the presence of off-topic tweets, the lack of context (as some words that trigger hate in certain contexts may have different connotations in others) and implicit hate speech that employs stereotypes or metaphors in order to convey hatred.

Although the results of the system employed on the extended dataset seemed promising, we couldn't see any improvement on the official test dataset. This might be as a result of not having any information on the actual distribution of the tweets (the number of tweets that convey hate towards immigrants and the number of tweets that convey hate towards women, information that might have been useful when extending the dataset), neither on the way the annotation was done and our definition of hate speech (and the way it differs from offensive language) might have been different. We also identified tweets for which we question the original label when taking into account the class definition. Below, we have provided some examples.

Example 1: The first tweet (annotated as not hateful), containing the users opinion on Poland

[4] https://www.noswearing.com/dictionary

[5] http://www.cs.cmu.edu/ biglou/resources/bad-words.txt

[6] As only the last submitted system was taken into consideration, this also corresponds to the system used for the official ranking

	10-cross validation results				Official results				Train-test configuration on a random test			
	A	F	P	R	A	F	P	R	A	F	P	R
B	0.772	0.762	0.764	0.669	–	–	–	–	0.756	0.747	0.730	0.671
C_1	0.733	0.747	0.623	0.716	0.461	0.401	0.566	0.523	0.757	0.747	0.732	0.666
C_2*	**0.934**	**0.938**	**0.917**	**0.934**	0.437	0.351	0.540	0.508	**0.856**	**0.824**	**0.844**	**0.805**
C_3	0.711	0.707	0.607	0.700	0.450	0.393	0.532	0.513	0.750	0.743	0.701	0.704
C_4	0.924	0.927	0.904	0.923	0.448	0.388	0.530	0.511	0.835	0.820	0.829	0.813
C_5	0.736	0.727	0.709	0.646	**0.5**	**0.471**	**0.581**	**0.551**	0.793	0.779	0.830	0.773

Table 2: Hate speech detection results. Best results are presented in bold font while * marks the system submitted for the official ranking.

receiving immigrants, seems more hateful than the second tweet (hateful), in which the user depicts a series of events.

- If Germans want rapefugees [7] they can keep them. Poland will not accept a single rapefugee. Not even one!!! Poland remains proud and firm! https://t.co/VHoFrxFjB1
- GERAMNY: African rapefugee climbs into house, steel expensive goods, rook a knife and abuse a girl of the family. https://t.co/fVXhWMYcE8

Example 2: Although in the first tweet (not hateful) the user talks about Donald Trump, which doesn't fit in the targeted categories, the annotation raises problems when trying to classify tweets such as the second one (annotated as hateful).

- I love my religious brothers and sisters, but @realDonaldTrump, FUCK YOU, YOU'RE NOT EVEN A REAL THEOCRAT YOU FAT USLESS BITCH.
- @menzemerized_ Worse i have proof. A picture i took of you and one you took of me on the same night. Useless ungreatful kunt!

6 Conclusion

This paper proposed several models that can be used in order to identify messages that convey hate towards women and immigrants, incorporating a variety of features for capturing offensive language. Our results revealed good classication performance on the training dataset, but a lower performance on the evaluation data, with a notable decrease in both accuracy and F-score. Error analysis shows that this decrease is mainly due to the lack of context to infer hateful intents, and the way hate speech was defined in the manual annotation of the dataset.

As the meaning of a message might change in different contexts (as it can be highly dependent on knowledge about the world), in our future work we plan on studying ways to retrieve contextual information.

Acknowledgments

This work has been funded by Maison des Sciences de l'Homme et de la Société de Toulouse under the project AMeSexTo.

References

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee.

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.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, pages 1–6. CEUR-WS.

Deborah Cameron. 1992. *Feminism and Linguistic Theory.* Palgrave Macmillan.

Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):18.

[7] Accordind to Urban Dictionary, the term rapefugee is usually used when referring to the Muslim refugees coming into Europe in a derogatory way, as refugees are perceived as being more likely to raping people.

John T. Nockleby. 2000. Hate speech. In *Encyclopedia of the American Constitution (2nd ed., edited by Leonard W. Levy, Kenneth L. Karst et al.*, pages 1277–1279.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Sara Owsley Sood, Elizabeth F Churchill, and Judd Antin. 2012. Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology*, 63(2):270–285.

Ziqi Zhang and Lei Luo. 2018. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *arXiv preprint arXiv:1803.03662*.