IIT Bombay at HASOC 2019: Supervised Hate Speech and Offensive Content Detection in Indo-European Languages

Urmi Saha¹, Abhijeet Dubey², and Pushpak Bhattacharyya¹

¹ Indian Institute of Technology, Bombay ² Apple Inc. {urmisaha, abhijeetd, pb}@cse.iitb.ac.in

Abstract. Text classification is a classical problem in NLP and has impactful applications. An essential business application is hate speech detection from online data. With the enormous amount of social media data getting generated continuously across the world, detection of hate speech is considered a very challenging task in NLP. In this paper, we describe our approaches for three shared tasks on hate speech and offensive content identification in Indo-European languages (Mandl et al. [9]). We describe statistical machine learning-based approaches as well as deep learning-based approaches and present their comparisons. We observe that convolutional neural networks perform quite well in the classification task.

Keywords: machine learning \cdot neural networks \cdot hate speech \cdot feature engineering \cdot word embeddings.

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). FIRE 2019, 12-15 December 2019, Kolkata, India. U. Saha et al.

1 Introduction

Social media has become an important communication medium today. Social media technology enables a piece of information to be spread quickly. With the exponential growth of social media users, public platforms are often used to express satisfaction or grievance regarding any product, service or experience. A huge amount of such data is generated continuously across the world. Unfortunately, these data often contain offensive words, which can be considered as *hate speech*. The anonymity and mobility provided by the social media platforms have made the breeding and spread of hate speech (Zhang et al. [17]) - eventually leading to cybercrime.

The term 'hate speech' was formally defined as 'any communication that disparages a person or a group on the basis of some characteristics (to be referred to as types of hate or hate classes) such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics (Nockleby et al. [11]). In many countries, including the United Kingdom, Canada, and France, there are laws prohibiting hate speech, which tends to be defined as speech that targets minority groups in a way that could promote violence or social disorder (Davidson et al. [5]). Constructing such countermeasures for online speech requires as first step, correct identification of hate speech. Therefore, analyzing the quality of this the huge amount of social media data has found its importance in many NLP tasks. Hate speech detection is critical for applications like controversial event extraction, building AI chatbots, content recommendation, and sentiment analysis (Badjatiya et al. [3]).

In this paper, we use tweets in English, Hindi and German language and perform three classification tasks on them:

- Sub-task A: Hate and Offensive, and Non Hate-Offensive
- Sub-task B: Hate speech, Offensive, and Profane
- Sub-task C: Targeted Insult, and Untargeted

We perform the above classification task using both statistical machine learning methods as well as deep learning methods - we use SVM and CNN, respectively. We perform feature engineering on our datasets before creating feature vectors for SVM. For CNN, we create embedding vectors using different word embeddings are use them in our model.

The rest of the paper is organized as follows. We discuss the related work in detail in the next section. Next, we describe our approaches in detail in Section 3. Then, we outline the experimental setup in Section 4 and present the results of our experiments in Section 5. Finally, we conclude the paper and discuss future work in Section 6.

2 Related Work

Traditional machine learning methods have performed quite well in classification tasks. There has been extensive work on classification tasks with feature engineering. Nobata et al.[10] experimented with several n-gram-based, syntactic, and distributional semantic features and showed that character n-grams contribute most for an online gradient descent learner. Sood et al. [13] trained several Support Vector Machine classifiers. They showed that that classification performance keeps improving with increased datasets, but not as rapidly after the data size had passed 1,500 items.

With the constant generation of huge amounts of data, neural networks are starting to take over statistical machine learning models. Gehrmann et al. [7] compare rule-based and deep learning-based models on ten different phenotyping tasks and show that CNNs outperform other phenotyping algorithms in all of them.

Word embeddings in neural networks are quite influential for classification tasks. Akhtar et al. [1] show that sentiment embedded vectors make deep learning architectures highly efficient. Recently, Xu et al. [16] present a CrossNet model based on context encoding layer, which learns from a source to analyze an unseen similar destination target. Djuric et al. [6] in their work describe how low-dimensional representations of comments can be learnt using neural language models and can be fed into a classification algorithm. Similarly, in our work, we use domain-specific word embeddings trained on tweets for our convolutional neural network.

Hate speech detection in the English language has been an extensive area of research. Xiang et al. [15] created offensive language topic clusters using Logistic Regression over a set of 860,071 tweets and supplemented with a dictionary of 339 offensive words. Wulczyn et al. [14] illustrates a method that combines crowdsourcing and machine learning to analyze personal attacks at scale. Besides English, substantial work has been done on hate speech detection in textual data of other languages too. Alfina et al. [2] detect hate speech in the Indonesian language. Kamble et al. [8] use domain-specific embeddings for hate speech detection in English-Hindi code-mixed tweets. Our task involves hate speech and offensive content detection in Hindi and German tweets, besides English tweets. We use these domain-specific embeddings for our convolutional neural network model.

3 Approaches

We implement three different approaches for the given task:

- Support Vector Machine without feature engineering
- Support Vector Machine with feature engineering
- Convolutional Neural Network

3.1 SVM without feature engineering

In this approach, we first clean the given tweets by the following steps:

- 1. removing blank rows if any
- 2. replacing any digit with 0

U. Saha et al.

- 3. modifying URLs to <url>
- 4. changing text to lowercase
- 5. tokenizing each tweet into words
- 6. removing stop words and performing stemming

After the data is cleaned, the tweets are encoded as feature vectors and are directly used as input to our SVM. Results of this method are shown in Section 5.

3.2 SVM with feature engineering

In this approach, we perform feature engineering before creating feature vectors for SVM. We select a handful of features which carry some relevant information about a tweet. For example, if the user uses one or more *angry* emoticons in a tweet, the tweet is more likely to be carrying the hatred emotion towards something.

We select the following features and include them in our feature vector:

- emoticons: we create a dictionary of happy, sad, anger, fear, surprise, disgust, others emoticons and count the number of each of them used in a tweet
- hashtags: we extract out the words used in hashtags in a tweet. Users often summarize their opinion through a hashtag. This can help find out the emotion expressed in a tweet.
- intensifiers: words like *exceptionally*, *incredibly*, *awful*, *insanely*, *etc* are often used to emphasize on some descriptive word. We call them intensifiers.
- negations: words like *never*, no. nothing, nowhere, etc are used to thwart the meaning of a piece of text.
- hate words: we use a list of hate words and find their occurrences in a tweet

Other features include char n-grams, word n-grams, etc

3.3 Convolutional Neural Network

In the deep neural network approach, we tokenize each input sentence and find the embedding of each word. Our hate speech dataset consists of tweets in three different languages. We use domain-specific word embeddings which are trained on Twitter data.

We pass the embeddings through our convolutional layers with multiple filter widths and feature maps. After each convolutional layer, we perform max-overtime pooling before passing them through a final fully connected layer with softmax output. We train this model by minimizing the categorical cross-entropy loss.

4 Experimental Setup

For statistical machine learning-based approaches, we use **SVM** with RBF kernel and c = 1.0 using grid-search and **Random-forest** with number of estimators = 50. We use nltk³ libraries for all data processing steps in our SVM model.

For our deep learning-based approaches, we use CNN. We have 2 convolutional layers with total 128 filters with size 5 and max-pooling of 5 and 30. We make our CNN deeper by using multiple filters - 3, 4, and 5. We use different word embeddings for each language which we feed into our model as input.

- English: Used domain specific word embeddings trained on Twitter domain (Kamble et al. [8].)
- **Hindi:** Embeddings trained on Hindi corpus available in CFILT⁴.
- German: Embeddings from Europarl trained with FastText.

FastText considers sub-word embedding. It is helpful in our case, as tweets are often informal and sub-word information should be given importance to extract the semantics of tweets.

We experiment with word embeddings of different dimensions and find that 100 dimensional word embeddings perform the best.

5 Results

Table 1 shows results in different metrics for English dataset when tried with the basic SVM method. The class labels to be predicted are mapped to numbers in the following way:

- Task 1: 0 HOF, 1 NOT
- Task 2: 0 PRFN, 1 OFFN, 2 HATE, 3: NONE
- Task 3: 0 TIN, 1 UNT, 2 NONE

Table 2 shows results in different metrics for English dataset when tried with the advanced SVM method.

6 Conclusion and Future Work

In this paper, we present our approaches for the three tasks of HASOC2019. We describe feature engineering for our statistical machine learning-based approaches. We also use a list of hate words for each of the three languages and use them in feature engineering for creating feature vectors of our model. We also describe a deep learning-based approach. In this approach, we use domain-specific word embeddings and find that word embeddings trained on a particular domain performs better for a text from the same domain. We also see improvement in

³ https://www.nltk.org/

⁴ http://www.cfilt.iitb.ac.in/

U. Saha et al.

	Task 1			Task 2				Task 3				
	Р	R	\mathbf{F}	\mathbf{S}	Р	\mathbf{R}	\mathbf{F}	\mathbf{S}	Р	\mathbf{R}	\mathbf{F}	\mathbf{S}
0	0.67	0.31	0.43	679	0.40	0.06	0.10	350	0.66	0.94	0.78	1077
1	0.68	0.91	0.77	1077	0.65	0.96	0.78	1077	0.67	0.25	0.36	615
2					0.47	0.07	0.12	125	0.00	0.00	0.00	64
3					0.78	0.34	0.47	204				
accuracy			0.68	1756			0.65	1756			0.66	1756
macro avg	0.67	0.61	0.60	1756	0.57	0.36	0.37	1756	0.44	0.40	0.38	1756
weighted avg	0.68	0.68	0.64	1756	0.60	0.65	0.56	1756	0.64	0.66	0.60	1756

Table 1. Results of all three tasks on English dataset using SVM model

Table 2. Results of advanced SVM model for Task 1 on English dataset

	Ρ	\mathbf{R}	\mathbf{F}	\mathbf{S}
HOF	0.21	0.95	0.34	540
NOT	0.94	0.17	0.28	2386
avg / total	0.80	0.31	0.29	2926

the performance of text classification by convolutional neural networks on using domain-specific word embeddings.

In this paper, we show that statistical machine learning-based methods perform well for classification tasks when the dataset is not very huge. However, with the constant generation of data over social media platforms, such tasks often need to cater to a wide variety with fast processing. Deep learning-based methods perform better on such massive data.

Our work is an effort to improve the existing methodology of detecting hate speech in social media comments. With the increase in social media usage by people across the world, hate speech is getting generated and propagated at high speed. Due to the massive scale of the web, methods that automatically detect hate speech are required (Schmidt et al. [12]). Both Facebook and Twitter have responded to criticism for not doing enough to prevent hate speech on their sites by instituting policies to prohibit the use of their platforms for attacks on people based on characteristics like race, ethnicity, gender, and sexual orientation, or threats of violence towards others (Bird et al. [4]). Hate speech detection is an urgent problem to solve to avoid an increase in cybercrime.

For future work, we will add more feature engineering and compare results with the current model. We will run our deep neural network on a larger dataset, using domain-specific word embeddings and observe its performance.

References

1. Akhtar, M.S., Kumar, A., Ekbal, A., Bhattacharyya, P.: A hybrid deep learning architecture for sentiment analysis. In: Proceedings of COLING 2016, the 26th

International Conference on Computational Linguistics: Technical Papers. pp. 482–493 (2016)

- Alfina, I., Mulia, R., Fanany, M.I., Ekanata, Y.: Hate speech detection in the indonesian language: A dataset and preliminary study. In: 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS). pp. 233–238. IEEE (2017)
- Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for hate speech detection in tweets. In: Proceedings of the 26th International Conference on World Wide Web Companion. pp. 759–760. International World Wide Web Conferences Steering Committee (2017)
- 4. Bird, S., Klein, E., Loper, E.: Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc." (2009)
- Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Eleventh international aaai conference on web and social media (2017)
- Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., Bhamidipati, N.: Hate speech detection with comment embeddings. In: Proceedings of the 24th international conference on world wide web. pp. 29–30. ACM (2015)
- Gehrmann, S., Dernoncourt, F., Li, Y., Carlson, E.T., Wu, J.T., Welt, J., Foote Jr, J., Moseley, E.T., Grant, D.W., Tyler, P.D., et al.: Comparing rule-based and deep learning models for patient phenotyping. arXiv preprint arXiv:1703.08705 (2017)
- 8. Kamble, S., Joshi, A.: Hate speech detection from code-mixed hindi-english tweets using deep learning models. arXiv preprint arXiv:1811.05145 (2018)
- Modha, S., Mandl, T., Majumder, P., Patel, D.: Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages. In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)
- Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y.: Abusive language detection in online user content. In: Proceedings of the 25th international conference on world wide web. pp. 145–153. International World Wide Web Conferences Steering Committee (2016)
- Nockleby, J.T.: Hate speech. Encyclopedia of the American constitution 3, 1277– 1279 (2000)
- Schmidt, A., 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. pp. 1–10 (2017)
- Sood, S.O., Churchill, E.F., Antin, J.: Automatic identification of personal insults on social news sites. Journal of the American Society for Information Science and Technology 63(2), 270–285 (2012)
- Wulczyn, E., Thain, N., Dixon, L.: Ex machina: Personal attacks seen at scale. In: Proceedings of the 26th International Conference on World Wide Web. pp. 1391– 1399. International World Wide Web Conferences Steering Committee (2017)
- Xiang, G., Fan, B., Wang, L., Hong, J., Rose, C.: Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In: Proceedings of the 21st ACM international conference on Information and knowledge management. pp. 1980–1984. ACM (2012)
- Xu, C., Paris, C., Nepal, S., Sparks, R.: Cross-target stance classification with self-attention networks. arXiv preprint arXiv:1805.06593 (2018)
- 17. Zhang, Z., Luo, L.: Hate speech detection: A solved problem? the challenging case of long tail on twitter. Semantic Web (Preprint), 1–21 (2018)