Finnish Hate-Speech Detection on Social Media Using CNN and FinBERT

Md Saroar Jahan, Mourad Oussalah, Nabil Arhab
University Of Oulu, CMVS, BP 4500, 90014 Finland
{Md.Jahan, Mourad.Oussalah,nabil.arhab}@oulu.fi

Abstract

There has been a lot of research in identifying hate posts from social media because of their detrimental effects on both individuals and society. The majority of this research has concentrated on English, although one notices the emergence of multilingual detection tools such as multilingual-BERT (mBERT). However, there is a lack of hate speech datasets compared to English, and a multilingual pre-trained model often contains fewer tokens for other languages. This paper attempts to contribute to hate speech identification in Finnish by constructing a new hate speech dataset that is collected from a popular forum (Suomi24). Furthermore, we have experimented with FinBERT pre-trained model performance for Finnish hate speech detection compared to state-of-the-art mBERT and other practices. In addition, we tested the performance of FinBERT compared to fastText as embedding, which employed with Convolution Neural Network (CNN).

Objectives

In overall, the main contributions of this work are as follows:

(i) We constructed a new Finnish 10.7k hate-speech dataset.

(ii) We compared the performance of FinBERT with other state-of-the-art approaches, namely, mBERT, CNN + fastText. Here, FinBERT is a version of Google's BERT deep transfer learning model for the Finnish language. The model can be fine-tuned to achieve state-of-the-art results for various Finnish natural language processing tasks.

(iii) Finally, we compared the performance of the proposed FinBERT with fastText when used as feature embedding inputted to another classifier as in [1]. For this experiment, we have used CNN as a classifier and compared CNN+FinBERT and CNN+fastText.

Dataset Development

The dataset content was extracted from Suomi24 corpus 2001–2017, VRT version 1.1. The corpus contains all the texts available in the discussion forums of the Suomi24 online social networking website from 1 January 2001 to 31 December 2017. The original dataset contains more than 35 million sentences covering diverse topics; therefore, we collected a subset of the original dataset and annotated it.

Dataset Devlopment

First, we collected 5k posts from the original dataset by applying a set of profane words string matching. Filtering with profane words increases the chances of hate speech in the sentence. However, since it is more realistic to have non-hate speech in the dataset, the rest of the subset was collected randomly from the original dataset, which has a negligible amount (0.93%) of hate sentences. Our collected dataset's total size is 10.7k, which does not include any noise comments or statements presenting only emoticons or numbers.

Annotator and Annotation Guidelines: The annotation involves identifying whether each sentence contains a hate speech or not. In this study, all the annotators together created and discussed the guidelines to ensure all participants had the same understanding of hate speech. Two independent labelers (who have knowledge in this field and completed a master's thesis on hate speech detection and NLP) have been employed separately for annotation to avoid bias. While, a third one (a senior research fellow who completed his Ph.D. in this field) is called upon whenever a disagreement between the two arises (total disagreement 197). If a sentence includes a hate (regardless of the category of the hate it belongs to), it is given a label '1'; otherwise, it is assigned '0'. However, it must be emphasized that the presence or absence of offensive words in a sentence cannot systematically be considered sufficient evidence to confirm the existence of hate or not-hate.

Type	Words	English Trsnlation
Offensive	Vittu	F**k
Offensive	Narttu	B**ch
Offensive	Pillua	P**sy

Table1: Example of profane words.

Sentence	Translation	Label
1. vittu mik tollo	F**k mik tollo	1
2. Nyt on kovaa setti	Now there is a hard set	0
3. Kuinka lyhyet ovat lyhyimmät shortsisi, päätän sen pienen	How short is your shortest shorts, I head its small	1
4. Muslimit eivät ole pahoja	Muslims are not bad bro	0

Table 2: Labelling example from the original dataset with English translation. Label 1 refers to hate, and 0 refers to non-hate.

Classifier	Accuracy	F1
NB + W.Level TF-IDF	81.5	74.6
NB + CharLevel Vector TF-IDF	81	74
LR + Word Level TF-IDF	82.3	77.5
LR+ Char Level Vector TF-IDF	82	77
RF + Word Level TF-IDF	82	80
RF+ Char Level Vector TF-IDF	83	77
fastText	59.5	56.4
CNN + fastText Word Embedding	90	89.7
CNN + FinBERT as embedding	91.3	90
FinBERT (bert-base-finnish-uncased-v1)	**91.7**	**90.8**
BERT-multilingual-uncased	88.2	85.3

Table 3: Classifier Accuracy (%) and F1 scores (%) for Finnish hate speech detection. Best in bold.

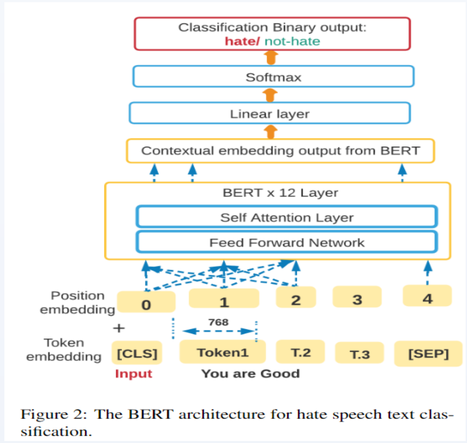

Figure 2: The BERT architecture for hate speech text classification.

Once our data was prepossessed, we performed the binary hate speech classification using 70% training, 10%validation, and 20% test samples for all models.

RESULTS

All five classifiers, FinBERT outperformed all other classifiers, indicating that the suggested FinBERT contextual model works better than other deep learning (CNN) and non-deep learning models (NB, LR, and RF). However, CNN showed a close performance as CNN with FinBERT 91.3% accuracy and 90% F1 score. These results indicate that NLP based hate speech detection contextual model is preferable to deep learning as word-embeddings features compared to non-contextual word embeddings like fastText. Since we have used pre-trained word embedding and provided the best accuracy and F1 scores, we assume that pre-trained word embeddings could be a reliable choice in this case. Therefore, we experimented with both FinBERT and fastText word-embedding as a feature.

Our experiment showed that FinBERT has 91.3% accuracy and 90% F1, which is 1.3% and .5% better in terms of accuracy and F1 scores compared to fastText embedding. Comparing FinBERT and BERT-multilingual, FinBERT outperformed 3.5% in accuracy and 5% in F1 score. This low performance of mBERT can be explained since mBERThas trained over 102 languages; however, it has only 3% Finnish text. Otherhand, FinBERT pre-trained over 3 billion tokens

Conclusion

This paper introduced a new Finnish hate speech annotated dataset and experimented with BERT, CNN, and non-deep learning classifiers for hate-speech detection. To the best of our knowledge, this work is the first application of BERT for hate speech detection in the Finnish language. In all cases, FinBERT has performed outstandingly to detect hate speech compared to the CNN+fastText as we anticipated. In addition, this experiment shows the effectiveness of contextual models' performance over the non-contextual model. For example, when FinBERT contextual embedding was applied with CNN, it offered better performance compared to CNN with fastText (Non-contextual embeddings). Furthermore, FinBERT performed much better than NB, LR, and BERT-multilingual models. Our findings showed that FinBERT yields 91.7% accuracy and 90.8% F1 scores, which is better than all other learning models and features. In the future, we would like to experiment with a larger dataset and solely work on improving the deep learning method for Finnish hate speech detection.

References

[1] Zhang, Z. and Luo, L. (2019). Hate speech detection: A solved problem? the challenging case of long tail on twitter. Semantic Web, 10(5):925–945

[2] Searle, J. R. and Searle, J. R. (1969). Speech acts: An essay in the philosophy of language, volume 626. Cambridge university press.

Acknowledgements

This work was partially supported by EU Project YougRes on youth polarization \& radicalization (ID: 823701) and COST Action NexusLinguarum – "European network for Web-centered linguistic data science" (CA18209), which are gratefully acknowledged