Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 876–882
Marseille, 20-25 June 2022

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). Our results showed that FinBERT yields a 91.7% accuracy and 90.8% F1 score value, which outperforms all state-of-art models, including multilingual-BERT and CNN.

Keywords: Hate detection, Finnish hate, FinBERT, Finnish Hate dataset

1. Introduction

The proliferation of online social media platforms with millions of user-generated content every day has witnessed a substantial increase of online hate speech (HS), raising the concern of society, individuals, policymakers, and researchers.

(Brown, 2017) has defined hate speech as any textual or verbal practice that implicates issues of discrimination or violence against people regarding their race, ethnicity, nationality, religion, sexual orientation, and gender identity. According to (Anis and Maret, 2017) hate speech can occur in different linguistic styles and several acts like insulting, provocation, abusing, and aggression. However, according to (Chetty and Alathur, 2018), hate speech can be categorized into gendered, religious, and racist hate speech.

Especially, offensive language promotes discrimination based on gender, nationality, or race (Nockleby, 2000), which often leads to profound adverse effects, particularly for teenage groups, organizations, and society. Many internet companies offer criteria and generic guidelines that users must adhere to when posting content on certain sites to solve this problem. Simultaneously, they use manual annotators to detect offensive language and remove it accordingly. However, such a manual effort is invariably expensive, non-scalable, and non-sustainable, inspiring the benefit of a solution based on an automatic offensive language detection.

Prior work has studied detection of offensive language specially for English language in Twitter (Foong and Oussalah, 2017; Wiegand et al., 2018), FromSpring posts (Reynolds et al., 2011), Wikipedia comments and Facebook posts (Kumar et al., 2018). Besides, scholars examined various characteristics of offensive language such as the cyber aggression (Kumar et al., 2018), hate speech (Abderrouaf and Oussalah, 2019), abusive language (Mubarak et al., 2017), racism (Kwok and Wang, 2013) and offensive language (Wiegand et al., 2018).

Most of the previous work is based on English datasets where some of those works reported accuracy over 95% zhang2019hate (Jahan and Oussalah, 2021). This is due to adequate NLP resources for English language datasets. However, the lack of NLP resources for other languages affected detection accuracy compared to the English language. Although, several works have explored the issue in other languages as well, e.g., Arabic (Albadi et al., 2018; Refaee and Rieser, 2014), Italian (Nunes et al., 2018; Sigurbergsson and Derczynski, 2020), and Hindi (Bashar and Nayak, 2020). Recently Semeval-2020(Zampieri et al., 2020) organized a competition for hate speech detection from Twitter datasets and reported a detection accuracy of English(92%), Arabic(90%), Greek(85%), Danish(80%), and Turkish(82%). However, to the best of our knowledge, we have not found any good quality Finnish HS dataset to experiment with. Besides, it is also acknowledged that the Finnish language bears additional challenges for conventional parsers due to inherent characteristics of Finnish language because of flexible word order, unlimited compound building and a much richer inflection system.

In this respect, as part of our contribution to the Finnish hate speech detection challenge, we first used a Finnish annotated hate speech dataset that has not been used before in this domain. Then, we performed experiments using different machine learning models and applied various feature engineering strategies. Especially, we mostly focused on FinBERT model performance with the state-of-the-art CNN architecture. In addition, we compared the classifier performance using fasText and FinBERT applications in the deep learning embedding layer. In overall, the main contributions of this work are as follows:

(i) We constructed a new Finnish 10.7k hate-speech dataset. The dataset collection process and annotation guidelines are described in section2.

(ii) We compared the performance of FinBERT with other state-of-the-art approaches, namely, mBERT, CNN+fastText, Logistic regression (LR), Naive Bayes (NB), and RandomForest(RF). Here, FinBERT is a version

Statistics	Count
Number of Tokens	122,823
Vocabulary Size	23430
Number of Posts	10700
Average number of Tokens per post	1
Non-hate class	8914
Hate class	1786

Table 1: Statistic of Dataset.

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 FinERT with FastText when used as feature embedding inputted to another classifier as in (Zhang and Luo, 2019). For this experiment, we have used CNN as a classifier and compared CNN+FinBERT and CNN+FastText. Section 2 describes the dataset development process, including corpus statistics, hate categories identification, annotator and annotation guidelines and disagreement handling. Section 3 illustrates the FinBERT classifier construction, including feature engineering and classifier architecture. The results are provided in Section 4. Finally, Section 5 draws the main findings and perspective works.

2. Dataset Development

This paper presents a new Finnish dataset of textual hate speech annotated at the sentence level. The content was extracted from Suomi24 corpus 2001–2017, VRT version 1.1 [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. First, we collected 5k posts from the original dataset by applying a set of profane words string matching (examples of profane words shown in Table 2). 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.

2.1. Corpus Statistics

The statistics of the collected dataset is summarized in Table 1. In total, it contains 10700 sentences.

Next, to identify hate-speech content from the collected dataset, we first highlight the categories of hate speech that are investigated in the subsequent analysis. This is detailed in the next subsection.

2.2. Hate Categories Identification

By definition, hate speech is a language that goes against groups based on specific characteristics; however, it may occur with different linguistic connotations, even in subtle forms (Fortuna and Nunes, 2018). In this paper, we identified five hate speech targets, which we describe and provide examples from the corpus as follows:

Racism / Racial Segregation: racism consists of an ideology of racial domination (Wolfe, 1999). However, (Clair and Denis, 2015) pointed out that racism could be considered biological or cultural superiority of one or more racial groups, such as skin color or physical difference. Example from corpus: 'Suurin osa rikoksista Suomessa ovat mustien tekemiä' -'Most of the crime done in Finland by Black people'.

Sexual: expressions with a sexual meaning or intention (e.g., 'Onko täällä kylmä vai onko sinulla ollut tissityötä?'- '"Is it cold in here, or have you had a b**b job?'). However, innocent sexual talk and sex educational conversion is considered differently (e.g., 'onko masturbaatio hyvä?'- 'is masturbation good?').

Xenophobia: expression primarily the form of discrimination manifested through biased actions and hates against foreigners (DE OLIVEIRA, 2020). An example: 'Niin monet Lähi-idän pakolaiset aiheuttavat ongelmia täällä' - 'Man so many refugees from the middle east are creating problems here'.

Religious Intolerance/fundamentalism: is consistently associated with high levels of intolerance and prejudice toward targeting specific religious groups (Altemeyer and Altemeyer, 1996). This is exemplified in the following post: 'Don't talk good things about them; I know Muslims and Christian. We don't need religion anymore'-'Älä puhu heistä hyviä asioita; Tunnen muslimeja ja kristittyjä. Emme tarvitse enää uskontoa'.

Homophobia: negative attitudes and feelings toward homosexuality or, in other words, people who are identified or perceived as being lesbian, gay, and bisexual. An example of this case from our corpus is: 'perheen täytyy olla pettynyt ur homo'-'family must be disappointed ur gay'.

Next, we describe the process of manual annotation, indicating whether a given post is a specific category of hate speech or not.

2.3. 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'. See examples shown in Table 3.

In our annotation, a sentence is considered a hate if it satisfies the following criteria drawn from the hate definition by (Brown, 2017; Anis and Maret, 2017; Chetty and Alathur, 2018): *deliberate attack directed towards a specific group of people or organization employing sexual attack, curse,*

[1] http://urn.fi/urn:nbn:fi:lb-2020021801

Type	**Words**	**English Trsnlation**
Offensive	Vittu	F**k
Offensive	Narttu	B**ch
Offensive	Pillua	P**sy
Offensive	jälkeenjäänyt	Retarded
Offensive	Homo	Gay
Offensive	Neekeri	Nigger
Offensive	Perse	A*s
Offensive	Pedot	Faggots
Swear words	Mene helvettiin	Go to Hell
Swear words	Tappaa sinut	kill you

Table 2: Example of profane words.

defamation, threat, gender, ethnicity, and identity. 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. For example, Sentence 3 from Table-1 does not contain any offensive words, though, by definition, it is very abusive and insulting to someone. Another example in Sentence 4 (Muslims are not terrorists), includes an offensive word 'Terrorist'; however, it includes a negation operator, making it a non-hate sentence. Therefore, with regards to HS, we decided to consider two aspects for its identification:

1. The target must be a group, an individual, or an organization.
2. The action, or more precisely the intention of the statement (Searle and Searle, 1969): this means that we must deal with a message that incites, spreads, promotes, or support violence or hatred towards the given target, or a statement that aims at dehumanizing, delegitimizing, hurting or intimidating the target.

To understand the action or intention of the speaker, the use of profane words plays an important role. This is defined as socially improper use of language, or may also be called offensive, cursing, swearing, or expletives. Table 2 shows examples of the 10 most frequent profane words extracted from the corpus.

Once labeled, 16.7% (1786) of the dataset was identified as hate, while the rest 83.3% (8914) was non-hate sentences. The details of the dataset collection will be released for the community on this GitHub page[2].

2.4. Inter Annotator Agreement

We used Krippendorff's alpha (α) (Krippendorff, 1970) to measure the inter-annotator agreement because of the nature of our annotation setup. This robust statistical measure accounts for incomplete data and, therefore, does not require every annotator to annotate every sentence.

$$\alpha = 1 - \frac{D_o}{D_e} \tag{1}$$

Here the alpha α is calculated by Equation (1), where (D_o) is the observed number of disagreements and (D_e) stands for the estimate of the likelihood of a disagreement occurring by chance.

We used nominal metrics to calculate annotator agreement. The range of α is between 0 and 1, $1 \geqslant \alpha \geqslant 0$. When α is 1, there is perfect agreement between annotators, and when α=0, the agreement is entirely due to chance. Our annotation produced an agreement reliability score of 0.92 using nominal metric .

Disagreement cases

Our inter-annotator agreement score was satisfactory (α = 0.92); however, some minor disagreements occurred. Here we talk about some problematic annotation examples and raised conflict between the two annotators.

1. 'Miksi mieluummin ajettu?' - 'Why do you prefer shaved?': Not sure whether the speaker means sexually shaved or not.
2. 'Luulen, että exälläsi on uusi tyttö' - 'I think ur ex has a new girl': This post doesn't consist of any hate/swear words; however, mentioning someone's 'Ex' might have the intention of defamation or insult or no intention at all. Therefore, it was complex to comprehend the intention of the speaker.
3. 'Haha hän on minun neekeri'-'Haha he's my nigga': The word nigger is an ethnic slur typically used against black people in the English language. However, the word 'nigga' is often used without any racist connotation.
4. 'Onko itsetyydytys mielestäsi hyödyllinen vai haitallinen?'-'Do you find the act of masturbation helpful or harmful?': Despite the fact that this sample contains offensive terms, the speaker's goal may be harmless, and the question may be asked for educational purposes.

3. Classifier Architecture and Feature Engineering

We have used five different models, namely BERT, CNN, NB, R, and LR. Two pre-trained models have been used for BERT: multilingual-BERT and FinBERT (both uncased). On the other hand, for CNN, two different setups were obtained with a similar architecture; one with non-contextual embeddings (fastText) and the other one with contextual embeddings (FinBERT). Below we present its technical description.

3.1. Preprocessing

For preprocessing, we mainly remove unidentified characters, symbols, numbers, mentioned tags, emojis tab tokens, URLs, etc. We have not performed removal of stop-words and stemming [3] since our initial test has shown a 0.5% decrease in overall accuracy and F1 score after removing of stop words and stemming.

[2] https://github.com/saroarjahan/FinBERT

[3] https://github.com/stopwords-iso/

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

3.2. BERT model

BERT is the Bidirectional Encoder Representations from Transformers. This seminal transformer-based language model applies an attention mechanism that enables learning contextual relations between words in a text sequence (Devlin et al., 2018). Two training strategies were used in our BERT model:

1. Masked-Language Modeling (MLM): where 15 % of the tokens in a sequence replaced (masked) for which the model learns to predict the original tokens, and

2. Next sentence prediction (NSP): here, the model receives two sentences as input, and the model learns whether the second sentence is a successor of the first sentence in the original document context.

FinBERT is a BERT language model further trained on 1 million training steps on over 3 billion tokens (24B characters) of Finnish text drawn from news, online discussion, and internet crawls. We have used FinBERT uncased version [4]. By contrast, multilingual-BERT pre-trained on the top 102 languages with the largest Wikipedia using a masked language modeling (MLM) objective, the Finnish Wikipedia text covers only approximately 3% (Devlin et al., 2018).

3.3. Feature Engineering

A set of features have been employed and evaluated for Finnish hate speech detection.

TF-IDF. The term frequency (TF) accounts for the absolute frequency of the tokens in the corpus. The TF-IDF considers the rate of occurrence of each token weighted by its inverse document frequency in the corpus. The latter reflects how important an individual token is to a document in the database.

n-grams. Word n-grams consider a sequence of n words and, therefore, allow to account for words ordering, unlike unigram. We have used three different combinations of TF-IDF: Word-level, n-gram word level (for N=2, 3), and n-gram Character level (for N=3, 4). We restricted to 5000 features for each type to avoid the computational cost.

Word Embeddings Features. Word embedding maps each token to a vector of real numbers aiming to quantify and categorize the semantic similarities between linguistic terms based on their distributional properties in a large corpus using machine learning or related dimensional reduction techniques. We used the pre-trained word embedding; namely, fastText [5] and FinBERT.

3.4. Classification Architecture

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

CNN-fastText Model Structure: We adopted (Kim, 2014) CNN architecture, where the input layer is represented by a concatenation of the words forming the post (up to 70 words), except that each word is now represented by its fastText embedding representation with a 300 embedding vector. A convolution 1D operation with a kernel size of 3 was used with a max-over-time pooling operation over the feature map with a layer dense 50. Dropout on the penultimate layer with a constraint on the l2-norm of the weight vector was used for regularization. Fig. 1 illustrates our CNN architecture.

CNN-BERT Model Structure: The first part is where FinBERT is used, in which the text is passed through 12 layers of self-attention to obtain contextualized vector representations. Here the input layer is represented by a concatenation of the words forming the post (up to 120 words). The other part being CNN, which was used as a classifier. It uses the same architecture as fastText with CNN; however, it produces embedding representation with a 768 embedding vector for each word, unlike fastText, which used 300. The details of the implementation are reported on the GitHub page of the project[6].

BERT Model Structure: We used Huggingface Transformers (Wolf et al., 2019) library for implementing the classifiers. We fine-tuned different transformer training data using 70% corresponding training data. The following models were tested: FinBERT (Bert-base-finnish-uncased) and multilingual-BERT (mBERT uncased). Each model was fine-tuned for 6 epochs with a learning rate of 5e-6, the maximum input sequence length of 128, and batch size 4. After each epoch, the model was evaluated on the test set. Fig. 2 illustrates our BERT architecture.

In addition to the designed CNN and BERT architecture,

[4] `https://huggingface.co/TurkuNLP/bert-base-finnish-uncased-v1`

[5] https://fasttext.cc/docs/en/crawl-vectors.html (accessed 30.12.2021)

[6] `https://github.com/saroarjahan/FinBERT`

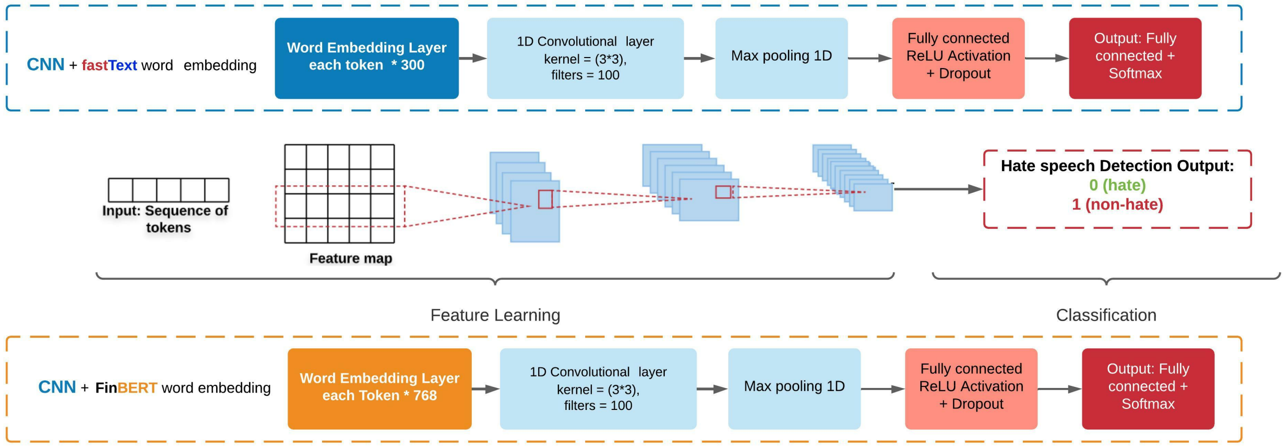
CNN + fastText word embedding
Word Embedding Layer each token * 300
1D Convolutional layer kernel = (3*3), filters = 100
Max pooling 1D
Fully connected ReLU Activation + Dropout
Output: Fully connected + Softmax
Input: Sequence of tokens
Feature map
Hate speech Detection Output:
0 (hate)
1 (non-hate)
Feature Learning
Classification
CNN + FinBERT word embedding
Word Embedding Layer each Token * 768
1D Convolutional layer kernel = (3*3), filters = 100
Max pooling 1D
Fully connected ReLU Activation + Dropout
Output: Fully connected + Softmax

Figure 1: The architecture of hate speech detection CNN with fastText (Top architecture), and CNN with FinBERT (Bottom architecture).

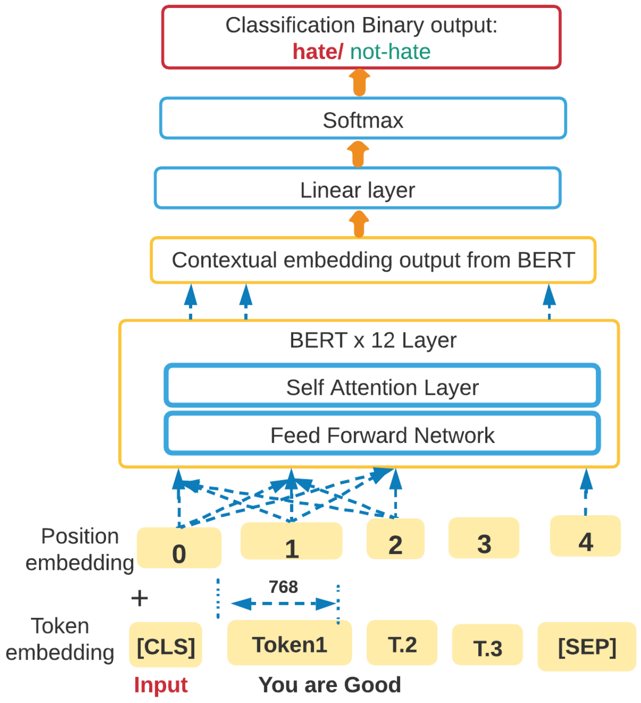

Figure 2: The BERT architecture for hate speech text classification.

we have used three non-deep-learning algorithms: Logistic regression (LR), Naive Bayes (NB), and Random Forest (RF). As features, we have used Tf-IDF word-level and n-gram character levels with NB, LR, and RF models. Furthermore, we have used FinBERT word embedding with this non-deep learning as well to compare potential improvements.

4. Results

For classification purpose, we randomly shuffled and divided the entire collected dataset into three parts: the training set consisted of 6700 sentences, the validation set consisted of 1000 sentences, and the testing set consisted of 3000 sentences.

Table 4: Classifier Accuracy (%) and F1 scores (%) for Finnish hate speech detection. Best in bold.

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

The results of the binary classification of the Finnish hate speech dataset summarized in Table 4 shows classifier accuracy and F1 score for all four types of classifiers.

The result shows that fastText as a classifier yielded only 59.5% accuracy and F1 56.4% scores. In contrast, fastText as word embedding with CNN yielded 90% accuracy and an F1 score of 89.7%. This outcome motivates the use of fastText as a word embedding for this particular hate speech domain.

Among 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.

Among non-deep learning models, LR models outperformed, including NB and RF, in terms of accuracy and F1 scores. In all cases, NB performed lower than others. On the other hand, RF classifiers performed 1% better than NB However, none of these non-deep learning performed as good as CNN or FinBERT.

Strictly speaking, the Suomi24 dataset and fastText/FinBERT have not been used for this domain-specific task. Therefore, it was not possible to compare our work with any previous work for this Finnish dataset.

5. 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.

6. 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.

7. Bibliographical References

Abderrouaf, C. and Oussalah, M. (2019). On online hate speech detection. effects of negated data construction. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 5595–5602. IEEE.

Albadi, N., Kurdi, M., and Mishra, S. (2018). Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 69–76. IEEE.

Altemeyer, R. A. and Altemeyer, B. (1996). *The authoritarian specter*. Harvard University Press.

Anis, M. and Maret, U. (2017). Hatespeech in arabic language. In *International Conference on Media Studies*.

Bashar, M. A. and Nayak, R. (2020). Qutnocturnal@ hasoc'19: Cnn for hate speech and offensive content identification in hindi language. *arXiv preprint arXiv:2008.12448*.

Brown, A. (2017). What is hate speech? part 1: The myth of hate. *Law and Philosophy*, 36(4):419–468.

Chetty, N. and Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and violent behavior*, 40:108–118.

Clair, M. and Denis, J. S. (2015). Sociology of racism. international encyclopedia of the social and behavioral sciences 2nd.

DE OLIVEIRA, L. M. (2020). Imigrantes, xenofobia e racismo: uma análise de conflitos em escolas municipais de são paulo.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Foong, Y. J. and Oussalah, M. (2017). Cyberbullying system detection and analysis. In *2017 European Intelligence and Security Informatics Conference (EISIC)*, pages 40–46. IEEE.

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Jahan, M. S. and Oussalah, M. (2021). A systematic review of hate speech automatic detection using natural language processing. *arXiv preprint arXiv:2106.00742*.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2018). Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11.

Kwok, I. and Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. In *Twenty-seventh AAAI conference on artificial intelligence*.

Mubarak, H., Darwish, K., and Magdy, W. (2017). Abusive language detection on arabic social media. In *Proceedings of the first workshop on abusive language online*, pages 52–56.

Nockleby, J. T. (2000). Hate speech. *Encyclopedia of the American constitution*, 3(2):1277–1279.

Nunes, S., Fortuna, P., and Bonavita, I. (2018). Merging datasets for hate speech classification in italian.

Refaee, E. and Rieser, V. (2014). An arabic twitter corpus for subjectivity and sentiment analysis. In *LREC*, pages 2268–2273.

Reynolds, K., Kontostathis, A., and Edwards, L. (2011). Using machine learning to detect cyberbullying. In *2011*

10th International Conference on Machine learning and applications and workshops, volume 2, pages 241–244. IEEE.

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

Sigurbergsson, G. I. and Derczynski, L. (2020). Offensive Language and Hate Speech Detection for Danish. In *Proceedings of the 12th Language Resources and Evaluation Conference*. ELRA.

Wiegand, M., Siegel, M., and Ruppenhofer, J. (2018). Overview of the germeval 2018 shared task on the identification of offensive language.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Wolfe, A. (1999). The bridge over the racial divide: Rising inequality and coalition politics. *The Journal of Blacks in Higher Education*, (26):127.

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., and Çöltekin, Ç. (2020). Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *arXiv preprint arXiv:2006.07235*.

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.