Multiword Expression Features for Automatic Hate Speech Detection

Nicolas Zampieri, Irina Illina, Dominique Fohr

Multiword Expression Features for Automatic Hate Speech Detection

Nicolas Zampieri, Irina Illina, and Dominique Fohr

University of Lorraine, CNRS, INRIA, Loria/ F-54000 Nancy, France
{firstname.lastname}@loria.fr

Abstract. The task of automatically detecting hate speech in social media is gaining more and more attention. Given the enormous volume of content posted daily, human monitoring of hate speech is unfeasible. In this work, we propose new word-level features for automatic hate speech detection (HSD): multiword expressions (MWEs). MWEs are lexical units greater than a word that have idiomatic and compositional meanings. We propose to integrate MWE features in a deep neural network-based HSD framework. Our baseline HSD system relies on Universal Sentence Encoder (USE). To incorporate MWE features, we create a three-branch deep neural network: one branch for USE, one for MWE categories, and one for MWE embeddings. We conduct experiments on two hate speech tweet corpora with different MWE categories and with two types of MWE embeddings, word2vec and BERT. Our experiments demonstrate that the proposed HSD system with MWE features significantly outperforms the baseline system in terms of macro-F1.

Keywords: Social media · Hate speech detection · Deep learning.

1 Introduction

Hate speech detection (HSD) is a difficult task both for humans and machines because hateful content is more than just keyword detection. Hatred may be implied, the sentence may be grammatically incorrect and the abbreviations and slangs may be numerous [12]. Recently, the use of machine learning methods for HSD has gained attention, as evidenced by these systems: [13, 8]. [9] performed a comparative study between machine learning models and concluded that the deep learning models are more accurate. Current HSD systems are based on natural language processing (NLP) advances and rely on deep neural networks (DNN) [11].

Finding the features that best represent the underlying hate speech phenomenon is challenging. Early works on automatic HSD used different word representations, such as a bag of words, surface forms, and character n-grams with machine learning classifiers [17]. The combination of features, such as n-grams, linguistic and syntactic turns out to be interesting as shown by [12].

In this paper, we focus our research on the automatic HSD in tweets using DNN. Our baseline system relies on Universal Sentence Embeddings (USE). We propose to enrich the baseline system using word-level features, called *multiword expressions* (MWEs) [14]. MWEs are a class of linguistic forms spanning conventional word boundaries that are both idiosyncratic and pervasive across different languages.[3] We believe that MWE modelling could help to reduce the ambiguity of tweets and lead to better detection of HS [16]. To the best of our

knowledge, MWE features have never been used in the framework of DNN-based automatic HSD. Our contribution is as follows. First, we extract different MWE categories and study their distribution in our tweet corpora. Secondly, we design a three-branch deep neural network to integrate MWE features. Finally, we experimentally demonstrate the ability of the proposed MWE-based HSD system to better detect hate speech: a statistically significant improvement is obtained compared to the baseline system. We experimented on two tweet corpora to show that our approach is domain-independent.

2 Proposed methodology

In this section, we describe the proposed HSD system based on MWE features. This system is composed of a three-branch DNN network and combines global feature computed at the sentence level (USE embeddings) and word-level features: MWE categories and word embeddings representing the words belonging to MWEs.

Universal sentence encoder provides sentence level embeddings. The USE model is trained on a variety of data sources and demonstrated strong transfer performance on a number of NLP tasks [2]. The HSD system based on USE obtained the best results at the SemEval2019 campaign (shared task 5) [8]. This power of USE motivated us to use it to design our system.

MWE features. A multiword expression is a group of words that are treated as a unit [14]. For example, the two MWEs *stand for* and *get out* have a meaning as a group, but have another meaning if the words are taken separately. MWEs include idioms, light verb constructions, verb-particle constructions, and many compounds. We think that adding information about MWE categories and semantic information from MWEs might help for the HSD task.

In our work, we focus on social media data. These textual data are very particular, may be grammatically incorrect and may contain abbreviations or spelling mistakes. For this type of data, there are no state-of-the-art approaches for MWE identification. A specific MWE identification system is required to parse MWEs in social media corpora. As the adaptation of an MWE identification system for a tweet corpus is a complex task and as it is not the goal of our paper, we decided to adopt a lexicon-based approach to annotate our corpora in terms of MWEs. We extract MWEs from the STREUSLE web corpus (English online reviews corpus), annotated in MWEs [15]. From this corpus, we create an MWE lexicon composed of 1855 MWEs which are classified into 20 lexical categories. Table 1 presents these categories with examples. Each tweet of our tweet corpora is lemmatized and parsed with the MWE lexicon. Our parser tags MWEs and takes into account the possible discontinuity of MWEs: we allow that one word, not belonging to the MWE, can be present between the words of the MWE. If, in a sentence, a word belongs to two MWEs, we tag this word with the longest MWE. We do not take into account spelling or grammatical mistakes. We add a special category for words not belonging to any MWE.

HSD system proposal. In this part, we describe our hate speech detection system using USE embeddings and MWE features. As USE is a feature at the sentence level and MWE features are at the word level, the architecture of our system is composed of a neural network with three branches: two branches are dedicated to the MWE features, the last one deals with USE features. Figure 1 shows the architecture of our system.

In the first branch, we associate to each word of the tweet the number of the MWE category (one-hot encoding). This branch is composed of 3 consecutive blocks of CNN (Conv1D) and

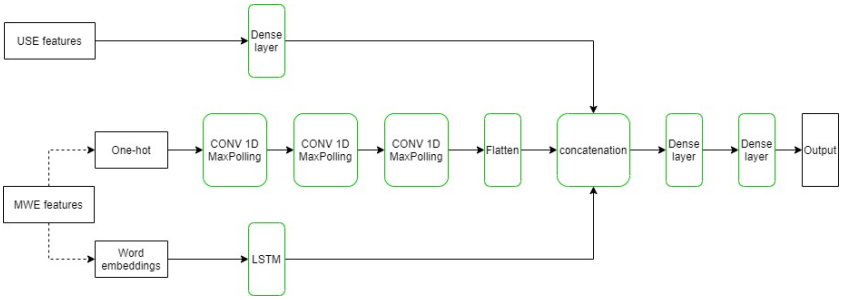

Fig. 1. Proposed hate speech detection system using USE and MWE features.

MaxPooling layers. Previous experiments with different DNN structures and the fast learning of CNN allow us to focus on this architecture. The second branch takes into account the semantic context of words composing MWE. If a given tweet has one or several MWEs, we associate a word embedding to each word composing these MWEs. We believe that the semantic meaning of MWEs is important to better understand and model them. This branch uses one LSTM layer. We propose to use two types of word embeddings: static where a given word has a single embedding in any context, or dynamic, where a given word can have different embeddings according to his long-term context. We experiment with word2vec and BERT embeddings [10, 4]. BERT uses tokens instead of words. Therefore, we use the embedding of each token composing the words of the MWEs. We think that using two branches to model MWEs allows us to take into account complementary information and provides an efficient way of combining different features for a more robust HSD system.

The last branch, USE embedding, supplies relevant semantic information at the sentence level. The three branches are concatenated and went through two dense layers to obtain the output. The output layer has as many neurons as the number of classes.

3 Experimental setup

3.1 Corpora

The different time frames of collection, the various sampling strategies, and the targets of abuse induce a significant shift in the data distribution and can give a performance variation on different datasets. We use two tweets corpora to show that our approach is domain-independent: the English corpus of SemEval2019 task 5 subTask A (called *HatEval* in the following) [1] and Founta corpora [5]. We study the influence of MWE features on the HatEval corpus, and we use the Founta corpus to confirm our results. Note that these corpora contain different numbers of classes and different percentages of hateful speech. We evaluate our models using the official evaluation script of SemEval shared task 5 [1] in terms of macro-F1. It is the average of the F1 scores of all classes.

HatEval corpus. In the HatEval corpus, the annotation of a tweet is a binary value indicating if HS is occurring against women or immigrants. The corpus contains 13k tweets. We

[1] https://github.com/msang/hateval/tree/master/SemEval2019-Task5/evaluation

Table 1. MWE categories with examples from STREUSLE corpus [15] and the number of occurrences of MWEs. The train set of HatEval. The column *Hateful (Non-hateful)* represents MWE occurences that appear only in hateful (non-hateful) tweets. The column *Both* represents MWE occurrences that appear in hateful and non-hateful tweets.

	MWE categories	Examples	Hateful	Non-hateful	Both
MWE5	Adjective	*dead on*	9	8	255
	Adverb	*once again*	1	5	194
	Discourse	*thank you*	12	15	401
	Nominal	*tax payer*	25	36	189
	Adposition phrase (idiomatic)	*on the phone*	9	36	134
VMWE5	Inherently adpositional verb	*stand for*	11	21	447
	Full light verb construction	*have option*	9	10	36
	Verbal idioms	*Give a crap*	14	24	384
	Full verb-particle construction	*take off*	11	20	387
	Semi verb-particle construction	*walk out*	6	18	153
	Auxiliary	*be suppose to*	4	0	475
	Coordinating conjunction	*and yet*	1	0	8
	Determiner	*a lot*	1	2	242
	Infinitive marker	*to eat*	0	0	12
	Adposition	*apart from*	3	13	573
	Non-possessive pronoun	*my self*	0	3	11
	Subordinating conjunction	*even if*	0	0	28
	Cause light verb construction	*give liberty*	1	0	0
	Symbol	*A+*	0	0	0
	Interjection	*lo and behold*	0	0	0

use standard corpus partition in training, development, and test set with 9k, 1k, and 3k tweets respectively. Each set contains around 42% of hateful tweets. The vocabulary size of the corpus is 66k words.

We apply the following pre-processing for each tweet: we remove mentions, hashtags, and URLs. We keep the case unchanged. We use this pre-processing because the systems using this pre-processing obtained the best results at the SemEval2019 task 5.

For train and development sets, we keep only tweets that contain at least two words. Thus, we obtain 8967 tweets for the training set and 998 tweets for the development set. We split the training part into two subsets, the first one (8003 tweets) to train the models, and the second one (965 tweets) for model validation. In the test set, we keep all tweets after pre-processing, even empty tweets. We tag empty tweets as non-hateful.

Founta corpus contains 100k tweets annotated with normal, abusive, hateful, and spam labels. Our experiments focus on HSD, so we decided to remove spams and we keep around 86k tweets. The vocabulary size of the corpus is 132k words. We apply the same pre-processing as for the HatEval corpus. We divide the Founta corpus into 3 sets: train, development, and test with 60%, 20%, and 20% respectively. As for the HatEval corpus, we use a small part of training as the validation part. Each set contains about 62%, 31%, and 6% of normal, abusive, and hateful tweets.

3.2 System parameters

Our baseline system utilizes only USE features and corresponds to figure 1 without MWE branches. The system proposed in this article uses USE and the MWE features as presented in figure 1[2]. For the USE embedding, we use the pre-trained model provided by google[3] (space dimension is 512) without fine-tuning.

We tag the MWE of each tweet using the lexicon, presented in the section 2. If an MWE is found, we put the corresponding MWE category for all words of the MWE. To perform fine-grained analysis, we decided to select MWE categories that have more than 50 occurrences (arbitrary value) and occurrences appear less than 97% in hate and non-hate tweets at the same time. We obtain 10 MWE categories: called MWE5 and VMWE5 which are respectively the first and second part of Table 1. VMWE5 is composed of Verbal MWE categories and MWE5 with the rest of the categories. The training part of the HatEval corpus contains 1551 occurrences of VMWE5 and 1329 occurrences of MWE5. During our experiments, we experiment with all MWE categories presented in Table 1 (containing 19 categories: 18 categories, and a special category for words not belonging to any MWE) and with the combination of VMWE5 and MWE5 (10 MWE categories and a special category).

Concerning the MWE one-hot branch of the proposed system, we set the number of filters to 32, 16, and 8 for the 3 Conv1D layers. The kernel size of each CNN is set to 3. For the MWE word embedding branch, we set the LSTM layer to 192 neurons. For BERT embedding, we use pre-trained uncased BERT model from [4] (embedding dimension is 768). The BERT embeddings are extracted from the last layer of this model. For word2vec embedding, we use the pre-trained embedding of [7]. This model is trained on a large tweet corpus (embedding dimension is 400). In our systems, each dense layer contains 256 neurons.

For each system configuration, we train 9 models with different random initialization. We select the model that obtains the best result on the development set to make predictions on the test set.

4 MWE statistics

We first analyze the distribution of the MWEs in our corpora. We observe that about 25% of the HatEval training tweets contain at least one MWE and so the presence of MWE can influence the HSD performance.

As a further investigation, we analyze MWEs appearing per MWE category and for hate/non-hate classes. In the training set of the HatEval corpus our parser, described in section 2, annotated 4257 MWEs. Table 1 shows MWEs that appear only in hateful or non-hateful tweets or both in HatEval training part. We observe that some MWE categories, as *symbol* and *interjection*, do not appear in HatEval training set. We decided to not use these two categories in our experiments. Most of the categories appear in hateful and non-hateful tweets. For the majority of MWE categories, there are MWEs that occur only in hateful speech and MWEs that occur only in non-hateful tweets.

Finally, we analyze the statistics of each MWE category for hate and non-hate classes. As in HatEval the classes are almost balanced, there is no bias due to imbalanced classes.

[2] https://github.com/zamp13/MWE-HSD

[3] https://tfhub.dev/google/universal-sentence-encoder-large/3

We observe that there are no MWE categories used only in the hateful speech or only in the non-hateful speech excepted for the *cause light verb construction* category, but this category is underrepresented. We note that there is a difference between the use of MWEs in the hateful and the non-hateful tweets: MWEs are used more often in non-hateful speech. These observations reinforce our idea that MWE features can be useful for hate speech detection.

5 Experimental results

The goal of our experiments is to study the impact of MWEs on automatic hate speech detection for two different corpora: HatEval and Founta. We carried out experiments with the different groups of MWE categories: MWEall, including all MWE categories, and the combination of VMWE5 and MWE5.

Table 2 displays the macro-F1 on HatEval and Founta test sets. Our baseline system without MWE features, called *USE* in Table 2, achieves a 65.7% macro-F1 score on HatEval test set. Using MWE features with word2vec or BERT embeddings, the system proposed in this paper performs better than the baseline. For instance, on HatEval, MWEall with BERT embedding configuration achieves the **best result** with 66.8% of macro-F1. Regarding Founta corpus, we observe a similar result improvement: the baseline system achieves 72.2% and systems with MWE features obtain scores ranging from 72.4% to 73.0% of macro-F1. It is important to note that according to a matched pair test in terms of accuracy with 5% risk [6], the systems using MWE features and word2vec or BERT embeddings *significantly* outperform the baseline system on the two corpora. Finally, the proposed system with MWEall and BERT embedding for HatEval outperforms the state-of-the-art system FERMI submitted at HatEval competition (SemEval task 5): 66.8% for our system versus 65% for FERMI of macro-F1 [8].

To analyze further MWE features, we experiment with different groups of MWE categories: VMWE5, MWE5, and MWEall. Preliminary experiments with the two-branch system with USE and word embeddings branches only gave a marginal improvement compared to the baseline system. Using the three-branch neural network with only VMWE5 or MWE5 instead of MWEall seems to be interesting only for word2vec embedding. With BERT embedding it is better to use MWEall categories. Finally, the use of all MWEs could be helpful rather than the use of a subgroup of MWE categories. Comparing word2vec and BERT embeddings, dynamic word embedding performs slightly better than the static one, however, the difference is not significant.

We compare the confusion matrices of two systems: the baseline system and the proposed one with MWEall and BERT embeddings. On the HatEval, the proposed system classifies better non-hateful tweets than the baseline system. In contrast, on Founta our system is more accurate to classify hateful tweets. We think that the balance between the classes plays an important role: in the case of HatEval corpus, the classes are balanced, in the case of Founta, the classes are unbalanced.

To perform a deeper analysis, we focus our observations on only the tweets from the test sets containing at least one MWE: 758 tweets from the HatEval test set and 3508 tweets from the Founta test set. Indeed, according to section 4, there is about 25% of tweets containing MWEs. The second part of table 2 shows that the results are consistent with those observed previously in this section, and the obtained improvement is more important.

Table 2. The first part represents F1 and macro-F1 scores (%) on *HatEval* and *Founta* test sets. The second part represents F1 and macro-F1 scores (%) on tweets containing at least one MWE in *HatEval* and *Founta* test sets.

Features	HatEval			Founta			
	F1		Macro-F1	F1			Macro-F1
	Hateful	Non-hate		Norm	Abus	Hate	
All test set							
USE	64.9	66.4	65.7	94.2	87.8	34.6	72.2
USE, MWEall, word2vec	64.5	68.2	66.3	93.8	86.9	36.5	72.4
USE, VMWE5, MWE5, word2vec	66.1	67.0	66.5	93.9	87.1	37.2	72.7
USE, MWEall, BERT	64.2	69.4	**66.8**	94.0	87.1	37.5	72.9
USE, VMWE5, MWE5, BERT	64.8	68.2	66.5	93.8	86.9	38.2	**73.0**
Tweets containing at least one MWE							
USE	67.8	62.3	65.0	91.1	94.1	41.6	75.6
USE, MWEall, word2vec	71.7	61.4	66.6	91.4	86.9	44.6	**76.5**
USE, MWEall, BERT	73.9	61.3	**67.6**	90.9	94	43.3	76.1

6 Conclusion

In this work, we explored a new way to design a HSD system for short texts, like tweets. We proposed to add new features to our DNN-based detection system: mutliword expression features. We integrated MWE features in a USE-based neural network thanks to a neural network of three branches. The results were validated on two tweet corpora: HatEval and Founta. The models we proposed yielded significant improvements in macro-F1 over the baseline system (USE system). Furthermore, on HatEval corpus, the proposed system with MWEall categories and BERT embedding significantly outperformed the state-of-the-art system FERMI ranked first at the SemEval2019 shared task 5. These results showed that MWE features allow to enrich our baseline system. The proposed approach can be adapted to other NLP tasks, like sentiment analysis or automatic translation.

References

1. Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F.M., Rosso, P., Sanguinetti, M.: 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. pp. 54–63. ACL (2019)
2. Cer, D., Yang, Y., Kong, S.y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., Kurzweil, R.: Universal sentence encoder for English. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 169–174. ACL (2018)
3. Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., Todirascu, A.: Multiword expression processing: A Survey. Computational Linguistics pp. 837–892 (2017)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT (2019)
5. Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., Kourtellis, N.: Large scale crowdsourcing and characterization of twitter abusive behavior (2018)

6. Gillick, L., Cox, S.J.: Some statistical issues in the comparison of speech recognition algorithms. In: International Conference on Acoustics, Speech, and Signal Processing,. pp. 532–535 vol.1 (1989)
7. Godin, F.: Improving and Interpreting Neural Networks for Word-Level Prediction Tasks in Natural Language Processing. Ph.D. thesis, Ghent University, Belgium (2019)
8. Indurthi, V., Syed, B., Shrivastava, M., Chakravartula, N., Gupta, M., Varma, V.: FERMI at SemEval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in Twitter. In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 70–74. ACL (2019)
9. Lee, Y., Yoon, S., Jung, K.: Comparative studies of detecting abusive language on twitter. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2). p. 101–106 (2018)
10. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: ICLR Workshop Papers (2013)
11. Mozafari, M., Farahbakhsh, R., Crespi, N.: A bert-based transfer learning approach for hate speech detection in online social media (2019)
12. 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. p. 145–153 (2016)
13. Pamungkas, E.W., Cignarella, A.T., Basile, V., Patti, V.: Automatic identification of misogyny in english and italian tweets at evalita 2018 with a multilingual hate lexicon. In: EVALITA@CLiC-it (2018)
14. Sag, I., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword expressions: A pain in the neck for nlp. In: Proceedings of CICLING-2002. pp. 1–15 (2002)
15. Schneider, N., Smith, N.A.: A corpus and model integrating multiword expressions and supersenses. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1537–1547. ACL (2015)
16. Stanković, R., Mitrović, J., Jokić, D., Krstev, C.: Multi-word expressions for abusive speech detection in Serbian. In: Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons. pp. 74–84. ACL (2020)
17. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In: Proceedings of the NAACL Student Research Workshop. pp. 88–93. ACL (2016)