Using Sentence Embeddings to identify Hate Speech against Immigrants and Women on Twitter

by

Vijaysaradhi Indurthi, Bakhtiyar Syed, Manish Shrivastava, Nikhil Chakravartula, Manish Gupta, Vasudeva Varma

in

13th International Workshop on Semantic Evaluation
(*SemEval-2019*)

Minneapolis, USA

Report No: IIIT/TR/2019/-1

Centre for Search and Information Extraction Lab
International Institute of Information Technology
Hyderabad - 500 032, INDIA
June 2019

Fermi at SemEval-2019 Task 5: Using Sentence Embeddings to identify Hate Speech against Immigrants and Women on Twitter

Vijayasaradhi Indurthi[1,3], Bakhtiyar Syed[1], Manish Shrivastava[1]
Nikhil Chakravartula[3], Manish Gupta[1,2], Vasudeva Varma[1]
[1] IIIT Hyderabad, India
[2] Microsoft, India
[3] Teradata, India
[1]{vijaya.saradhi, syed.b}@research.iiit.ac.in
[1]{m.shrivastava, manish.gupta, vv}@iiit.ac.in
[2]gmanish@microsoft.com [3]nikhil.chakravartula@teradata.com

Abstract

This paper describes our system (Fermi) for Task 5 of SemEval-2019: HatEval: Multilingual Detection of Hate Speech Against Immigrants and Women on Twitter. We participated in the subtask A for English and ranked first in the evaluation on the test set. We evaluate the quality of multiple sentence embeddings and explore multiple training models to evaluate the performance of simple yet effective embedding-ML combination algorithms. Our team - **Fermi**'s model achieved an accuracy of 65.00% for English language in task A. Our models, which use pretrained Universal Encoder sentence embeddings for transforming the input and SVM (with RBF kernel) for classification, scored first position (among 68) in the leaderboard on the test set for Subtask A in English language. In this paper we provide a detailed description of the approach, as well as the results obtained in the task.

1 Introduction

Microblogging platforms like Twitter provide channels to exchange ideas using short messages called tweets. While such a platform can be used for constructive ideas, a small group of people can propagate their notions including hatred against an individual, or a group or a race to the entire world in a few seconds. This necessitates the need to come up with computational methods to identify hate speech in user generated content.

Using computational methods to identify offense, aggression and hate speech in user generated content has been gaining attention in the recent years as evidenced in (Waseem et al., 2017; Davidson et al., 2017; Malmasi and Zampieri, 2017; Kumar et al., 2018) and workshops such as Abusive Language Workshop (ALW) [1] and Workshop on Trolling, Aggression and Cyberbullying (TRAC) [2].

2 Related Work

In this section we briefly describe other work in this area.

A few of the early works related to hate speech detection employed the use of features like bag of words, word and character n-grams with relatively off-the-shelf machine learning classifiers for detection (Dinakar et al., 2011; Waseem and Hovy, 2016; Nobata et al., 2016). Deep learning methods for hate speech detection were used by Badjatiya et al. (2017) wherein the authors experimented with a combination of multiple deep learning architectures along with randomly initialized word embeddings learned by Long Short Term Memory (LSTM) models.

Papers published in the last two years include the surveys by (Schmidt and Wiegand, 2017) and (Fortuna and Nunes, 2018), the paper by (Davidson et al., 2017) which presented the Hate Speech Detection dataset used in (Malmasi and Zampieri, 2017) and a few other recent papers such as (ElSherief et al., 2018; Gambäck and Sikdar, 2017; Zhang et al., 2018).

A proposal of typology of abusive language sub-tasks is presented in (Waseem et al., 2017). For studies on languages other than English see (Su et al., 2017) on Chinese and (Fišer et al., 2017) on Slovene. Finally, for recent discussion on identifying profanity versus hate speech see (Malmasi and Zampieri, 2018). This work highlighted the challenges of distinguishing between profanity, and threatening language which may not actually contain profane language.

Some of the similar and related previous workshops are Text Analytics for Cybersecurity and

[1] https://sites.google.com/view/alw2018

[2] https://sites.google.com/view/trac1

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 70–74
Minneapolis, Minnesota, USA, June 6–7, 2019.

Online Safety (TA-COS) [3], Abusive Language Workshop [4], and TRAC [5]. Related shared tasks include GermEval (Wiegand et al., 2018) and TRAC (Kumar et al., 2018).

3 Methodology

In this paper, we make use of several word embedding and sentence embedding methods.

3.1 Word Embeddings

Word embeddings have been widely used in modern Natural Language Processing applications as they provide vector representation of words. They capture the semantic properties of words and the linguistic relationship between them. These word embeddings have improved the performance of many downstream tasks across many domains like text classification, machine comprehension etc. (Camacho-Collados and Pilehvar, 2018). Multiple ways of generating word embeddings exist, such as Neural Probabilistic Language Model (Bengio et al., 2003), Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and more recently ELMo (Peters et al., 2018).

These word embeddings rely on the distributional linguistic hypothesis. They differ in the way they capture the meaning of the words or the way they are trained. Each word embedding captures a different set of semantic attributes which may or may not be captured by other word embeddings. In general, it is difficult to predict the relative performance of these word embeddings on downstream tasks. The choice of which word embeddings should be used for a given downstream task depends on experimentation and evaluation.

3.2 Sentence Embeddings

While word embeddings can produce representations for words which can capture the linguistic properties and the semantics of the words, the idea of representing sentences as vectors is an important and open research problem (Conneau et al., 2017).

Finding a universal representation of a sentence which works with a variety of downstream tasks is the major goal of many sentence embedding techniques. A common approach of obtaining a sentence representation using word embeddings is by the simple and naïve way of using the simple arithmetic mean of all the embeddings of the words present in the sentence. Smooth inverse frequency, which uses weighted averages and modifies it using Singular Value Decomposition (SVD), has been a strong contender as a baseline over traditional averaging technique (Arora et al., 2016). Other sentence embedding techniques include p-means (Rücklé et al., 2018), InferSent (Conneau et al., 2017), SkipThought (Kiros et al., 2015), Universal Encoder (Cer et al., 2018).

Task A (Hate speech detection) is a two-class classification where systems have to predict whether a tweet in English or in Spanish with a given target (women or immigrants) is hateful or not hateful. TASK B (Aggressive behavior and Target Classification) is a two-class classification where systems have to classify hateful tweets (e.g., tweets where Hate Speech against women or immigrants has been identified) as aggressive or not aggressive, and second to identify the target harassed as individual or generic (i.e. single human or group).

We formulate sub-task A of HatEval as a text classification tasks. In this paper, we evaluate various pre-trained sentence embeddings for identifying the offense, hate and aggression. We train multiple models using different machine learning algorithms to evaluate the efficacy of each of the pre-trained sentence embeddings for the downstream task. We observe that there is a class label imbalance in the dataset. To prevent any bias induced due to imbalanced classes, we process the transformed training dataset using SMOTE (Chawla et al., 2002) which synthetically oversamples data and ensures that all the classes have same number of instances.

In the following, we discuss various popular sentence embedding methods in brief.

- InferSent (Conneau et al., 2017) is a set of embeddings proposed by Facebook. InferSent embeddings have been trained using the popular language inference corpus. Given two sentences the model is trained to infer whether they are a contradiction, a neutral pairing, or an entailment. The output is an embedding of 4096 dimensions.

- Concatenated Power Mean Word Embedding (Rücklé et al., 2018) generalizes the concept of average word embeddings to power mean word embeddings. The concatenation of different types of power mean word embeddings

[3] http://ta-cos.org/
[4] https://sites.google.com/site/alw2018
[5] https://sites.google.com/view/trac1

Model	LR		RF		SVM-RBF		XGB	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
InferSent	64.26	64.34	63.96	62.45	57.13	41.54	**71.18**	**71.21**
Concat-p mean	63.35	63.43	67.17	65.83	63.86	60.98	71.08	70.67
Lexical Vectors	67.27	66.61	67.97	67.09	58.53	46.03	67.87	68.31
Universal Encoder	70.58	70.63	70.48	70.05	57.13	41.54	64.26	64.34
ELMo	69.68	69.78	65.96	65.12	68.37	68.44	66.57	66.59
NNLM	66.57	66.46	64.36	62.83	65.56	63.88	66.37	65.74

Table 1: *Dev* Set Accuracy and Macro-F1 scores(in percentage) for **Sub-Task A- English.**

considerably closes the gap to state-of-the-art methods mono-lingually and substantially outperforms many complex techniques cross-lingually.

- Lexical Vectors (Salle and Villavicencio, 2018) is another word embedding similar to fastText with slightly modified objective. FastText (Bojanowski et al., 2016) is another word embedding model which incorporates character n-grams into the skipgram model of Word2Vec and considers the sub-word information.

- The Universal Sentence Encoder (Cer et al., 2018) encodes text into high dimensional vectors. The model is trained and optimized for greater-than-word length text, such as sentences, phrases or short paragraphs. It is trained on a variety of data sources and a variety of tasks with the aim of dynamically accommodating a wide variety of natural language understanding tasks. The input is variable length English text and the output is a 512 dimensional vector.

- Deep Contextualized Word Representations (ELMo) (Peters et al., 2018) use language models to get the embeddings for individual words. The entire sentence or paragraph is taken into consideration while calculating these embedding representations. ELMo uses a pre-trained bi-directional LSTM language model. For the input supplied, the ELMo architecture extracts the hidden state of each layer. A weighted sum is computed of the hidden states to obtain an embedding for each sentence.

Using each of the sentence embeddings we have mentioned above, we seek to evaluate how each of them performs when the vector representations are supplied for classification with various off-the-shelf machine learning algorithms. For each of the evaluation tasks, we perform experiments using each of the sentence embeddings mentioned above and show our classification performance on the *dev* set given by the task organizers.

Using each of the sentence embeddings we have mentioned above, we seek to evaluate how each of them performs when the vector representations are supplied for classification with various off-the-shelf machine learning algorithms. For each of the evaluation tasks, we perform experiments using each of the sentence embeddings mentioned above and show our classification performance on the *dev* set given by the task organizers.

4 Dataset

The data collection methods used to compile the dataset used in HatEval is described in (Basile et al., 2019). We did not use any external datasets to augment the data for training our models.

5 Results and Analysis

The official *test* set results scored on CodaLab have been presented below in Table 2.

System	F1 (macro)	Accuracy
Universal Encoder	0.65	0.65

Table 2: Results for Sub-task A using Universal Encoder Sentence embeddings with SVM classifier using RBF kernel.

Our results on the different algorithms from the ones stated above have been mentioned henceforth and described in Table 1.

As described in Table 1 the dev set macro-averaged F-1 and accuracy is given for the task A-English.

We notice the best performance for task A in English on the official test set was bagged by the model which used pretrained Universal sentence embeddings using SVM with RBF kernel. However, pretrained Infersent embeddings along with XGBoost algorithm outperformed every other combination on the dev test. This can be probably due to the difference between the distributions in the dev and the official test sets.

Overall, this work shows how different set of pretrained embeddings trained from different state-of-the-art architectures and methods when used with simple machine learning classifiers perform very well in the classification task of categorizing text as offensive or not.

6 Conclusions and Future Work

It is also important to note that the experiments are performed using the default parameters, so there is much scope for improvement with a lot of fine-tuning, which we plan on considering for future research purposes. Further, we can explore augmenting data from other similar shared tasks to achieve better performance.

References

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2016. A simple but tough-to-beat baseline for sentence embeddings.

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.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *The Social Mobile Web*, pages 11–17.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media. *arXiv preprint arXiv:1804.04257*.

Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 2017. Legal Framework, Dataset and Annotation Schema for Socially Unacceptable On-line Discourse Practices in Slovene. In *Proceedings of the Workshop Workshop on Abusive Language Online (ALW)*, Vancouver, Canada.

Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys (CSUR)*, 51(4):85.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using Convolutional Neural Networks to Classify Hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking Aggression Identification in Social Media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbulling (TRAC)*, Santa Fe, USA.

Shervin Malmasi and Marcos Zampieri. 2017. Detecting Hate Speech in Social Media. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 467–472.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in Discriminating Profanity from Hate Speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30:1–16.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153. International World Wide Web Conferences Steering Committee.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Andreas Rücklé, Steffen Eger, Maxime Peyrard, and Iryna Gurevych. 2018. Concatenated p-mean word embeddings as universal cross-lingual sentence representations. *arXiv preprint arXiv:1803.01400*.

Alexandre Salle and Aline Villavicencio. 2018. Incorporating subword information into matrix factorization word embeddings. *arXiv preprint arXiv:1805.03710*.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection Using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Association for Computational Linguistics*, pages 1–10, Valencia, Spain.

Huei-Po Su, Chen-Jie Huang, Hao-Tsung Chang, and Chuan-Jie Lin. 2017. Rephrasing Profanity in Chinese Text. In *Proceedings of the Workshop Workshop on Abusive Language Online (ALW)*, Vancouver, Canada.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Langauge Online*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *SRW@HLT-NAACL*.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of GermEval*.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. In *Lecture Notes in Computer Science*. Springer Verlag.