Fermi at SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media using Sentence Embeddings

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

Abstract

This paper describes our system (Fermi) for Task 6: OffensEval: Identifying and Categorizing Offensive Language in Social Media of SemEval-2019. We participated in all the three sub-tasks within Task 6. We evaluate multiple sentence embeddings in conjunction with various supervised machine learning algorithms and evaluate the performance of simple yet effective embedding-ML combination algorithms. Our team (**Fermi**)'s model achieved an F1-score of 64.40%, 62.00% and 62.60% for sub-task A, B and C respectively on the official leaderboard. Our model for sub-task C which uses pretrained ELMo embeddings for transforming the input and uses SVM (RBF kernel) for training, scored third position on the official leaderboard.

Through the paper we provide a detailed description of the approach, as well as the results obtained for the task.

1 Introduction

Social media provides anonymity which can be misused to target offensive comments to targeted parties. Users may engage in generating offensive content on social media which may show aggressive behaviour and may also include hate speech. As a result, it is imperative for social media platforms to invest heavily in creating solutions which can identify offensive language and to prevent such behaviour on social media.

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.

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; Badjatiya et al., 2017).

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 vs. 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 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

3.1 Word Embeddings

Word embeddings have been widely used in modern Natural Language Processing applications as

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

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 611–616
Minneapolis, Minnesota, USA, June 6–7, 2019.

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).

We formulate each of the sub-tasks of OffensEval as a text classification task. 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 sub-tasks as defined in this task. 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 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

Model	LR		RF		SVM-RBF		XGB	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
InferSent	70.32	70.46	70.77	67.26	65.45	60.12	75.52	74.21
Concat-p	69.82	69.95	71.60	70.68	70.37	71.11	75.41	75.23
Lexical Vectors	82.80	82.11	74.42	81.55	79.3	68.3	81.87	81.92
Universal Encoder	74.57	71.07	58.52	74.90	69.67	56.43	75.44	71.37
ELMo	80.00	78.72	73.54	85.20	82.66	73.44	83.27	80.90

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

Model	LR		RF		SVM-RBF		XGB	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
InferSent	82.98	80.47	82.29	82.00	80.49	84.02	85.30	83.99
Concat-p	83.17	82.13	80.29	83.64	80.37	82.39	85.17	84.14
Lexical Vectors	76.80	74.16	77.47	81.30	79.3	79.84	79.36	77.63
Universal Encoder	78.57	76.75	58.52	84.90	69.67	56.43	82.41	81.28
ELMo	78.24	76.67	83.54	82.20	82.66	80.72	81.27	79.68

Table 2: *Dev* Set Accuracy and Macro-F1 scores (in percentage) for **Sub-Task B**

Model	LR		RF		SVM-RBF		XGB	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
InferSent	66.92	64.98	69.29	65.24	60.49	32.51	68.30	69.03
Concat-p	55.37	60.40	60.29	66.93	66.17	64.35	70.37	68.93
Lexical Vectors	62.80	61.80	64.48	63.44	41.30	29.29	71.87	66.83
Universal Encoder	64.57	60.68	58.52	69.55	61.67	63.47	62.14	69.55
ELMo	80.00	60.48	73.54	64.65	71.66	67.00	69.47	67.76

Table 3: *Dev* Set Accuracy and Macro-F1 scores (in percentage) for **Sub-Task C**

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.

4 Dataset

The data collection methods used to compile the dataset used in OffensEval is described in (Zampieri et al., 2019). Sub-task A (Offensive language Detection) deals with classifying posts as offensive (OFF) vs not (NOT). Sub-task B (Categorization of Offensive Language) deals with categorization of offense as: targeted (TIN) and untargeted (INT). Sub-task C (Offensive Language Target Identification) categorizes the targets of insults and threats as individual (IND), group (GRP), and other (OTH). The overall dataset across the three sub-tasks consists of 14100 posts. Fig. 1 (reproduced from (Zampieri et al., 2019)) shows dataset size details.

A	B	C	Training	Test	Total
OFF	TIN	IND	2,407	100	2,507
OFF	TIN	OTH	395	35	430
OFF	TIN	GRP	1,074	78	1,152
OFF	UNT	—	524	27	551
NOT	—	—	8,840	620	9,460
All			13,240	860	14,100

Figure 1: Distribution of label combinations in the data (taken from (Zampieri et al., 2019))

True Label	Predicted Label		
		NOT	OFF
	NOT	605	15
	OFF	172	68

Table 4: Sub-task A, ELMo sentence embeddings with SVM classifier using RBF kernel

True Label	Predicted Label		
		TIN	UNT
	TIN	198	15
	UNT	19	8

Table 5: Sub-task B, Concatenated p mean sentence embeddings with XGBoost classifier

5 Results and Analysis

Note that we have not used any external datasets to augment the data for training our models.

In Tables 1, 2, and 3, we provide the dev set macro-averaged F-1 and accuracy for each of the three sub-tasks A, B and C respectively.

We notice the best performance across tasks with ELMo embeddings with SVM (using the RBF kernel).

The confusion matrices for our test set classifications are also given in Tables 4, 5, 6 respectively for each of the sub-tasks A, B and C.

Similar trends are observed for the final classification results on the test set (scored on CodaLab) for the sub-tasks A, B and C in Tables 7, 8, 9 respectively. Our system performed the third best in sub-task C of the 2019 SemEval task.

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

True Label	Predicted Label			
		GRP	IND	OTH
	GRP	52	18	8
	IND	9	85	6
	OTH	11	12	12

Table 6: Sub-task C, Universal Encoder sentence embeddings with XGBoost classifier

System	F1 (macro)	Accuracy
All NOT baseline	0.4189	0.7209
All OFF baseline	0.2182	0.2790
Lexvec	0.4317	0.7233
Concat p-means	0.5572	0.7558
ELMo	**0.6436**	**0.7826**

Table 7: Results for Sub-task A using LexVec, Concatenated p-mean and ELMo sentence embeddings with SVM classifier using RBF kernel

System	F1 (macro)	Accuracy
All TIN baseline	0.4702	0.8875
All UNT baseline	0.1011	0.1125
Concat p-means	**0.6205**	**0.8583**
InferSent	0.5953	0.8792
Universal	0.5950	0.775

Table 8: Results for Sub-task B. using Concatenated p-mean, InferSent and Universal sentence embeddings with XGBoost classifier

System	F1 (macro)	Accuracy
All GRP baseline	0.1787	0.3662
All IND baseline	0.2130	0.4695
All OTH baseline	0.0941	0.1643
InferSent	0.4425	0.6009
Universal	**0.6258**	**0.6995**
ELMo	0.5176	0.6103

Table 9: Results for Sub-task C. using InferSent, Universal and ELMo embeddings with XGBoost classifier

6 Conclusions and Future Work

It is also important to note that the experiments are performed using the default parameters, so there is further scope for improvement with a lot of fine-tuning, which we plan on considering for future research purposes. Further, we observe that the class distribution is highly imbalanced due to which there might be a bias introduced by the training algorithms. We plan to explore SMOTE (Chawla et al., 2002) for making the class labels balanced and then train the classification which will prevent the bias towards the unbalanced classes.

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 *WWW*.

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 *Proceedings of ICWSM*.

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 Salakhutdinov, Richard S Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-Thought Vectors. *arXiv preprint arXiv:1506.06726*.

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.

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*.

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*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*.

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.