LTL-UDE at SemEval-2019 Task 6: BERT and Two-Vote Classification for Categorizing Offensiveness

Piush Aggarwal, Tobias Horsmann, Michael Wojatzki and Torsten Zesch
Language Technology Lab
University of Duisburg-Essen
piush.aggarwal@stud.uni-due.de,
{tobias.horsmann, michael.wojatzki, torsten.zesch}@uni-due.de

Abstract

This paper describes LTL-UDE's systems for the *SemEval 2019 Shared Task 6*. We present results for Subtask A and C. In Subtask A, we experiment with an embedding representation of postings and use a Multi-Layer Perceptron and BERT to categorize postings. Our best result reaches the 10th place (out of 103) using BERT. In Subtask C, we applied a two-vote classification approach with minority fallback, which is placed on the 19th rank (out of 65).

1 Introduction

The Internet is frequently used for online debates and discussions, where individuals or groups are increasingly often verbally attacked. Online platform providers aim to remove such attacking posts or ideally, prevent them from being published. Manual verification of each posting by a human moderator is infeasible due to the high amount of postings created every day. Consequently, automated detection of such attacking postings is the only feasible way to counter this kind of hostility.

In this work, we present our results for the *SemEval 2019 Shared Task 6: Identifying and Categorizing Offensive Language in Social Media* (Zampieri et al., 2019b) on the OLID dataset (Zampieri et al., 2019a). Subtask A focuses on the binary distinction if a post is offensive or not, while Subtask C determines if the target is an individual, group, or other entity. Our submission for Subtask A ranks 10th, for Subtask C ranks 19th.

For Subtask A, we experiment with word list-based classification, using classifiers such as SVM or logistic regression based on sentence embeddings, and neural network-based models such as a Multi-layer Perceptron (MLP) and Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018). We find that the SVM performs best on our development set, but BERT reaches the best result on the test dataset. Moreover, a learning curve experiment suggests that more training data will lead only to minor improvements. In Subtask C, we choose a two-vote classification approach, where we let two systems compete with a fallback to the minority class in case the systems disagree. This fallback approach has a high robustness between our development and the official test dataset.

2 Related Work

Detection of offensive or potentially hurtful online postings is investigated under a variety of names. Waseem et al. (2017) focuses on *abusive language*, Kumar et al. (2018) tackles the problem as *aggression* while Macbeth et al. (2013) approaches this problem as *cyberbullying* to mention just a few. Furthermore, the field of hate speech detection is strongly related, which aims at detecting a similar kind of online statements (Waseem and Hovy, 2016; Wojatzki et al., 2018).

Common approaches to detecting such socially unacceptable statements utilize rich feature sets consisting of word ngrams, surface forms and syntactical features (Warner and Hirschberg, 2012; Nobata et al., 2016). Human-knowledge is provided by word lists containing offenses as key words or phrases (Bassignana et al., 2018; Wiegand et al., 2018b). Xiang et al. (2012) approaches the task as topic modelling problem using Latent Dirichlet Allocation (Blei et al., 2003).

These tasks are tackled with feature engineering-based approaches such as SVM or regression models but also with convolutional neural networks (Wiegand et al., 2018b).

3 Subtask A: Offensiveness

Subtask A is a binary classification task. A posting is either offensive or not offensive. For this task,

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 678–682
Minneapolis, Minnesota, USA, June 6–7, 2019.

we experiment with the following approaches:

Preprocessing We lowercase all postings and use the Ark Tokenizer (Gimpel et al., 2011) for word splitting. These preprocessing steps are used in all experiments.

Lexical Matching We use the following handcrafted word lists of abusive words: (i) Profane Word List[1] containing more than 1,300 English tokens, (ii) UdS Lexicon of Abusive Words[2] having 1,651 entries (Wiegand et al., 2018a), and (iii) *Multilingual Lexicon of Words to Hurt* from HurtLex (Bassignana et al., 2018) with 9,313 terms.[3] A posting is classified as offensive if it contains any words in the before mentioned lists.

Posting Embeddings We represent each posting by a dense embedding, which we create from word embeddings by summing up the vector values of the word representations. The resulting posting vector is re-scaled into the range zero to one. We use the pre-trained embeddings provided by Mikolov et al. (2018), which are trained on the common crawl corpus.

Classifiers We apply the following classifiers: SVM (Chang and Lin, 2011), Logistic Regression (Fan et al., 2008), Random Forest (Breiman, 2001) and a Decision Tree (Breiman et al., 1984). We use the implementation provided by scikit-learn (Pedregosa et al., 2011) using default parameters.

Multi-Layer-Perceptron (MLP) With the same pre-processing and feature extraction steps used as for shallow models described above, we train a MLP with 100 hidden units in Scikit-Learn with ReLu as activation function and Adam optimizer (Kingma and Ba, 2014). We initialize the neural network with the *fasttext* word embeddings provided by Mikolov et al. (2018).

BERT We use the provided pre-trained BERT-base model (Devlin et al., 2018) to create a vector representation of a posting. We fine-tune the model on the training data set using a sequence-length of 128 and batches of 32. We also investigate the impact of enriching the training dataset with additional data by using machine translation. We back and forth translate the training data to obtain paraphrases of the original training data, which we expect to improve model performance. We translated the data into Russian, Chinese, and Arabic and back to English using Google's translation service. We repeated the fine-tuning with this enriched dataset.

[1] https://www.cs.cmu.edu/~biglou/resources/

[2] https://github.com/uds-lsv/lexicon-of-abusive-words

[3] http://hatespeech.di.unito.it/resources.html

Set	Approach	F_1	Acc
dev	SVM	**.795**	**.814**
	BERT	.771	.799
	Ensemble	.767	.789
	BERT-trans	.732	.768
	Logistic Reg.	.704	.728
	MLP	.687	.705
	Random Forest	.641	.678
	Lexical Matching	.619	.680
	Decision Tree	.567	.585
	Baseline - all NOT	.400	.667
test	Ensemble	.748	.782
	BERT	**.798**	**.839**
	SVM	.729	.761
	Baseline - all NOT	.418	.720

Table 1: Subtask A: Results in term of macro F_1 on a held-back development dataset containing 1,048 offensive postings and 2,192 not offensive (NOT) ones.

Ensemble We combine the best three approaches (BERT, SVM, and Logistic Regression) in an ensemble, which was reported to often account for improvements in a similar shared task for German (Wiegand et al., 2018c). We use the majority vote of these classifiers as the prediction.

3.1 Results

Table 1 shows the results for Subtask A. We report results on a self-created development dataset (25% of the original training data, 3,240 postings of which 1,048 postings are labeled as offensive and 2,192 as not offensive). We use the majority class as a baseline. On our dev dataset, we find that a SVM with the posting vector-representation achieves the best F-Score, followed by BERT. Contrary to our expectation, BERT's performance decreased by adding the machine-translated data. On the test dataset, we find BERT to perform best followed by the ensemble, which seems to add some additional robustness to the classification.

Learning curve A central question for shared tasks such as this one is if the amount of provided training data is sufficient to train a reliable clas-

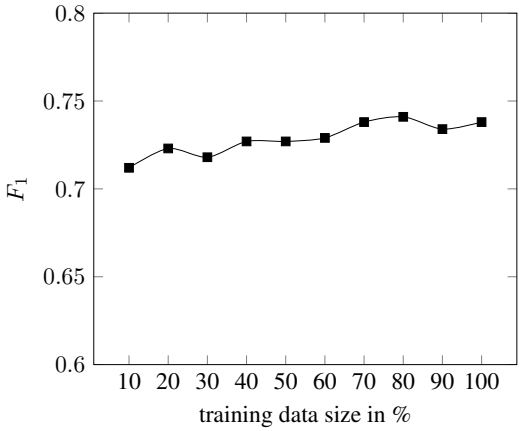

Figure 1: Learning curve on the training dataset. F-Score performance for adding an increasing amount of training data evaluated against the development set.

Set	Approach	F_1	Acc
dev	MLP+MLP	**.565**	.708
	SVM+MLP	.550	.688
	MLP+SVM	.541	.699
	SVM+SVM	.535	.701
	MLP	.523	.699
	SVM	.480	**.733**
	BERT	.492	.730
	Random Forest	..461	.676
	Decision Tree	.462	.604
	Logistic Regression	.508	.707
	Baseline - all IND	.255	.621
test	MLP+MLP	**.556**	**.666**
	SVM+MLP	.498	.615
	Baseline - all IND	.213	.469

Table 2: Subtask C results

sifier. Figure 1 shows a learning curve computed over the provided training data with testing against the hold-out development set. We split the training data into equal-sized data blocks which are randomly distributed over labels and add an increasing number of data blocks to see the performance improvement by adding more data. The results shows that improving the machine learning model is a more promising strategy than providing even more data as the slope indicates only minor improvements if more data is added.

4 Subtask C: Offense Targets

The goal in this subtask is to identify the kind of target at which a tweet is directed at (i.e. at this point it is already known that the tweet is a targeted offense, just the target itself is not yet determined). A target is either an individual (IND), a group (GRP), or other (OTH), if none of the previously mentioned two categories apply. We apply the same approaches as already used in Subtask A.

Two-Vote Classification with Minority Fallback Furthermore, an analysis of the class distribution showed that the class for *other* has comparatively few instances. This makes it challenging for a classifier to reliably detect such an under-represented class. Therefore, we attempt to redefine the problem as a binary classification problem using two classifiers. If the two classifiers agree in their prediction, we take the predicted class (either *individual* or *group*). In case of an disagreement, we select the minority class, *other*, as prediction. Thus, we also alter the training data to contain only two classes. The labels of the under-represented *other* class are mapped for one classifier to *individual* and for the other one to *group*, which creates a kind of minority-class noise. Our intuition is, if both classifier overcome the uncertainty added by the (small) amount of noise, the prediction is considered reliable. Consequently, we consider a disagreement as evidence for assigning the minority class.

Results Table 2 shows the results. We find that our two vote classification approach, using two MLPs, reaches the highest F-Score on the development and test set. On the development set, we reach the best accuracy result with an SVM but the considerably lower F-Score shows a strong bias towards a single class. Moreover, MLP+MLP shows a high robustness when comparing the F-Score performance between development and test set.

5 Conclusion

In this paper, we presented our approach on identifying and categorizing offensive language in social media. We mostly rely on lexical and semantic features for all subtasks. Results shows that semantic features have a significant impact on system performance. In general, our system leaves much room for improvement. Detection of offensiveness could probably benefit from more semantically oriented features that go beyond the surface form of words. We make the source code of our experiments publicly available[4].

[4] https://github.com/aggarwalpiush/OffensEval2019

References

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A Multilingual Lexicon of Words to Hurt. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018.*

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Leo Breiman. 2001. Random Forests. *Mach. Learn.*, 45(1):5–32.

Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. 1984. *Classification and regression trees*. The Wadsworth statistics/probability series. Wadsworth and Brooks/Cole Advanced Books and Software, Monterey, CA.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *J. Mach. Learn. Res.*, 9:1871–1874.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 42–47. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6980.

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 Cyberbullying (TRAC-2018)*, pages 1–11. Association for Computational Linguistics.

Jamie Macbeth, Hanna Adeyema, Henry Lieberman, and Christopher Fry. 2013. Script-based story matching for cyberbullying prevention. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 901–906.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in Pre-Training Distributed Word Representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

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*, WWW '16, pages 145–153, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

William Warner and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

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 Language Online*, pages 78–84. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018a. Inducing a Lexicon of Abusive Words – a Feature-Based Approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056. Association for Computational Linguistics.

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

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

Michael Wojatzki, Tobias Horsmann, Darina Gold, and Torsten Zesch. 2018. Do Women Perceive Hate Differently: Examining the Relationship Between Hate Speech, Gender, and Agreement Judgments. In *Proceedings of the Conference on Natural Language Processing (KONVENS)*, pages 110–120, Vienna, Austria.

Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting Offensive Tweets

via Topical Feature Discovery over a Large Scale Twitter Corpus. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1980–1984, New York, NY, USA. ACM.

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

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval)*.