HAD-Tübingen at SemEval-2019 Task 6: Deep Learning Analysis of Offensive Language on Twitter: Identification and Categorization

Himanshu Bansal
University of Tübingen
himanshu.bansal@student.uni-tuebingen.de

Daniel Nagel
University of Tübingen
daniel.nagel@student.uni-tuebingen.de

Anita Soloveva
University of Tübingen
Lomonosov MSU
anita.soloveva@student.uni-tuebingen.de

Abstract

This paper describes the submissions of our team, HAD-Tübingen, for the SemEval 2019 - Task 6: "OffensEval: Identifying and Categorizing Offensive Language in Social Media". We participated in all the three sub-tasks: Sub-task A - "Offensive language identification", sub-task B - "Automatic categorization of offense types" and sub-task C - "Offense target identification". As a baseline model we used a Long short-term memory recurrent neural network (LSTM) to identify and categorize offensive tweets. For all the tasks we experimented with external databases in a postprocessing step to enhance the results made by our model. The best macro-average F_1 scores obtained for the sub-tasks A, B and C are 0.73, 0.52, and 0.37, respectively.

1 Introduction

The use of offensive language is an ubiquitous problem one faces when using social networking services like Twitter. Users of such services often take advantage of the anonymity of the individual platforms for using the computer-mediated communication to engage in offensive behaviour against individuals, groups and/or organizations. Due to increasing problems with offensive language and a raising demand for offensive language detection on platforms like Twitter, tasks, similar to the current one have already become popular for several different languages: English (Waseem et al., 2017), German (Wiegand et al., 2018) and Spanish (Rosso et al., 2018). With increasing popularity of Twitter, over 1.48 billion users (June 2013) and still new accounts signing up every day, the need for improvement on tackling the well known problem of insults inside the platform has become more and more necessary.

The Twitter platform[1] describes itself as a connection to "what's happening in the world and what people are talking about right now". For this reason alone, its data attracts more and more NLP researchers all over the world. "Tweets", the messages one can send over this platform can be described as micro-texts, limited to 280 characters, over which users can interact with each other or simply post statements. Since the input is up to the user, one could include misspellings, emoticons, hashtags but also slang and abusive words, what makes those messages a valuable source for different analyses.

As was mentioned in the beginning, the goal of this paper is to consider our approach for the SemEval 2019 - Task 6: "OffensEval: Identifying and Categorizing Offensive Language in Social Media", for task information (see Zampieri et al. 2019b) and for dataset description (see Zampieri et al. 2019a). We took part in all of the three subtasks, using an LSTM based classifier. In the remainder of the paper, we describe our methods and discuss both our results and suggestions for further work.

2 System description

Neural network models have recently gained more and more popularity for text classification tasks, since they perform quite efficiently in modeling of sequences and offer advantages for computation. For this competition, we used unidirectional LSTM, where the recurrent component took a sequence of words as an input. We set the basic parameters in the model as follows: 30 as the number of epochs, a batch size of 43 for sub-task A, since it was the smallest batch size that the 860 tweets could be divided by, where our model

[1] https://twitter.com/

still performed well. For the other sub-tasks we went with 30 and 71 as batch sizes for the test sets of 240 and 213 tweets, accordingly. We used 4 hidden layers with 50 neurons per each, since our overall score declined by decreasing and increasing their number. Our dropout ratio was set to 0.95, the embedding size to 100, learning rates varied between submissions from 0.003 to 0.005.

The model was implemented in Python and makes use of Tensorflow (Abadi et al., 2015) and Scikit-learn (Pedregosa et al., 2011) libraries for training the classifier. We optimized our architecture parameters by predictions of support vector machine (SVM) model, described in (Rama and Çöltekin, 2017) and (Çöltekin and Rama, 2018). It used 'bag of n-grams' as features, and took not only word n-grams, as in our LSTM based model, but combined character and word n-grams, weighted by sublinear TF-IDF scaling. We picked the epoch with the best F_1-score for each parameter setting according to these SVM predictions. Our repository can be found on github https://github.com/cicl2018/semeval-2019-task-6-HAD.

2.1 Preprocessing

For neural network classification, data preprocessing has a great impact on the system's performance. Thus, at least one step from the following procedure was applied for all the submissions:

- lowercasing, since uppercased words can be both offensive and not
- hashtag parsing (e.g. #retrogaming → # retro gaming) (see, Baziotis et al. 2017)
  This tool is trained on 2 big corpora:
  - English Wikipedia
  - a collection of 330 million English Twitter messages
- removing tokens, containing "@USER"
  The user names are not given, thus this information is irrelevant for the classification task.
- character normalization
  We removed all the following charachters " : . , — ~ ", digits and single quotation marks except for abbreviations and possessors (e.g. *u're* → *u're*, but *about'* → *about*)
- using '=', '!', '?' and '/' as token splitters (e.g. *something!important* → *something important*)

2.2 Sub-task A - Offensive language identification

Sub-task A was a binary classification task. The goal was to identify whether the post is offensive (OFF) or not (NOT). The provided tweets were labeled as OFF if they contained any form of non-acceptable language or a targeted offense, and labeled as NOT in any other case.

2.2.1 System pipeline for sub-task A

Figure 1 describes the system architecture for sub-task A. For each of the three submissions we tried different approaches.

1. All the preprocessing steps (Section 2.1) + LSTM classifier with the use of SVM predictions, (see Section 2.2.2 and green arrows in Figure 1).
2. All the preprocessing steps + LSTM classifier with SVM predictions + additional manually created offensive word list, (see Section 2.2.3 and black arrows in Figure 1).
3. Hashtag parsing as a single preprocessing step + LSTM classifier with SVM predictions, (see Section 2.2.4 and red arrows in Figure 1).

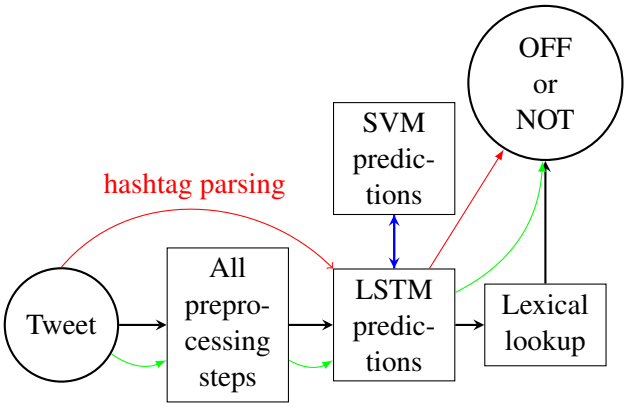

Figure 1: Pipeline of sub-task A

2.2.2 Submission 1, Sub-task A

In our first submission, we fed the preprocessed data into our LSTM model, setting the configurations (e.g. a learning rate of 0.003), according to the outcome of SVM predictions (Figure 1: green arrows).

2.2.3 Submission 2, Sub-task A

For the second submission we used a manually created additional offensive word list. After all the preprocessing steps, we ran the model with the same configurations as in the first submission except for the learning rate of 0.005, picking the epoch with the best F_1-score regarding SVM predictions. Then we postprocessed the results by using external manually collected offensive vocabulary, reannotating the tweets as offensive, if they contained abusive words from this list, but were labeled as not offensive by our model (Figure 1: black arrows).

2.2.4 Submission 3, Sub-task A

As a third submission, we preprocessed raw tweets only by hashtag parsing and let an LSTM model with a learning rate of 0.005 classify the data, choosing the epoch with the best F_1-score, according to the SVM predictions (Figure 1: red arrows).

2.3 Sub-task B - Automatic categorization of offense types

Sub-task B was a classification task of targeted (TIN) vs. untargeted (UNT) tweets. The test set contained only offensive (OFF) posts from the first sub-task. Tweets were considered as targeted, if they were insults/ threats to an individual or group, untargeted in any other case. For this sub-task we reduced the initial training set of 13.240 tweets to 4300, removing the tweets labelled with NOT, since non-offensive tweets would not add any improvement to the learning model and might even distort the learning process.

2.3.1 System pipeline for sub-task B

The system architecture for this sub-task is illustrated in Figure 2. Since the number of the representative tweets in the training data differed between categories a lot (i.e. 524 and 3876 for UNT and TIN, respectively), we used a weighted cross entropy to balance the data. Like in sub-task A, our approaches varied between submissions but this time we handed in 2 submissions.

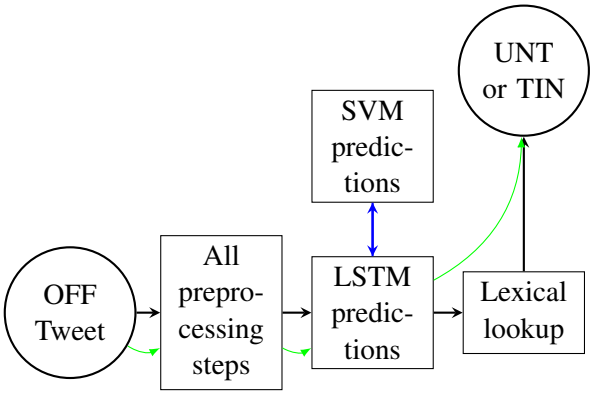

Figure 2: Pipeline of sub-task B

2.3.2 Submission 1, Sub-task B

The architecture of the first submission in this subtask is very much similar to the first submission in sub-task A with the only difference being that a learning rate was changed to 0.005 (Figure 2: green arrows).

2.3.3 Submission 2, Sub-task B

For the second submission we added a postprocessing step, where we reannotated the tweets that comprised particular word forms from a manually created list of potential insult victims as targets (Figure 2: black arrows). This database included following four parts:

- Names of representatives of top twitter profiles from the USA, the UK, Saudi Arabia, Brazil, India and Spain, since these countries have the most Twitter users[2] and Iran, Iraq, Turkey, Russia and Germany, because we predicted a possible aggression towards the users from these countries. The data was obtained from https://www.socialbakers.com/statistics/twitter/profiles/.
- A list of ethnic slurs, mostly extracted from https://en.wikipedia.org/wiki/List_of_ethnic_slurs
- A list of name-callings, primarily collected from https://www.urbandictionary.com/
- A list of 2nd and 3rd personal pronouns and abbreviations with them (e.g. *you*, *they've* etc.)

[2]This statistic can be found on https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-sele\protect\discretionary{\char\hyphenchar\font}{}{}cted-countries/

2.4 Sub-task C - Offense target identification

The third sub-task addressed offense target identification. This time we had three categories to choose from: Individual (IND), group (GRP), or other (OTH). The tweets were labeled as individually targeted, if a potential victim was a famous person, a named IND or an unnamed person interacting in the conversation. It was labeled as GRP, if the tweet was offensive with respect to a group of people considered as a unity due to the same ethnicity, gender or sexual orientation, political affiliation, religious belief, or similar, and was labelled as OTH, if the tweet intended to abuse an organization, a situation, an event, or an issue. The test data contained 213 offensive targeted tweets from sub-task B. The training set of 4300 offensive tweets was reduced to 3909 targeted ones for this sub-task.

2.4.1 System pipeline for sub-task C

The system architecture for this sub-task is illustrated in Figure 3.

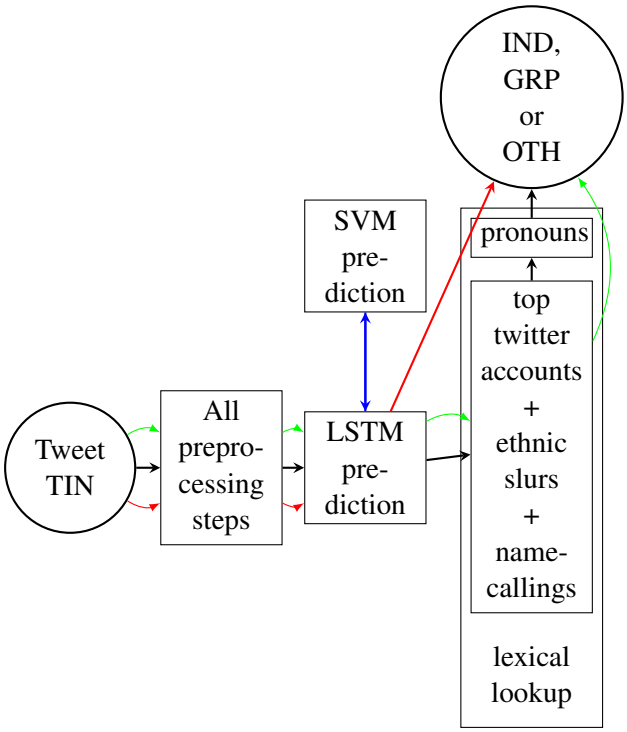

Figure 3: Pipeline of sub-task C

2.4.2 Submission 1, Sub-task C

This submission is reminiscent of the two previous first submissions, but the batch size was set to 71 and the learning rate to 0.003 (Figure 3: red arrows).

2.4.3 Submission 2, Sub-task C

In the second submission, we postprocessed the classified data, using the following datasets:

- Names of representatives of top twitter profiles from the USA, the UK, Saudi Arabia, Brazil, India, Spain, Iran, Iraq, Turkey, Russia and Germany. The data was obtained from https://www.socialbakers.com/statistics/twitter/profiles/:
  - celebrities and society/politics industries for identifying individual targets
  - community/political and community/religion industries for recognizing group targets
  - places, brands and entertainment/event industry for other targets
- A list of ethnic slurs, (see Section 2.3.3), for identifying group targets
- A dataset of name-callings, (see Section 2.3.3), for recognizing individual victims

These datasets helped to classify the categories IND, GRP or OTH by looking them up in our lists. (Figure 3: green arrow).

2.4.4 Submission 3, Sub-task C

The third submission differed from the previous one only in adding a list of 2nd and 3rd personal pronouns including their contractions to the existing database for the postprocessing step. We decided to try an approach with personal pronouns despite the fact, that they can both target individuals (e.g. "Take it out, you fucking wanker, or I'll take you out".) and groups (e.g. "All you democrats suck, and your momma's fat!").

3 Results

The results presented below were obtained using the macro-averaged F_1-score, provided by the organisers of OffensEval 2019. They included accuracy as well for comparison. Random baseline generated results by assigning the same labels for all instances were also added to the result Table 1. For example, "All OFF" in sub-task A represented the performance of a system that labels everything as offensive.

System	F1 (macro)	Accuracy
All NOT baseline	0.4189	0.7209
All OFF baseline	0.2182	0.2790
LSTM + Prepr.	0.6652	0.7337
LSTM + Prepr. + Lex. lookup	0.6487	0.7349
LSTM + Hashtag parsing	**0.7327**	**0.7977**

Table 1: Results for Sub-task A.

The best results for the first sub-task were produced by the simplest approach, which included only hashtag parsing as a preprocessing step and an LSTM based classifier with configurations, set according to SVM predictions. A plausible explanation to the bad performance of the second submission with a lexical lookup is that a task-specific lexicon should better be used as an input feature, which can only influence data classification, rather than as a decisive postprocessing step.

For sub-task B one can see the scores of our two submissions in Table 2. As before, the organizers have also included random baseline generated results by assigning the same labels for all instances.

System (Sumbission)	F1 (macro)	Accuracy
All TIN baseline	0.4702	0.8875
All UNT baseline	0.1011	0.1125
LSTM + Prepr.	**0.5246**	**0.8417**
LSTM + Prepr. + Lex. lookup	0.5022	0.8833

Table 2: Results for Sub-task B.

The best results for this sub-task were also achieved only by applying preprocessing steps to an LSTM model. Most likely, the problem was that our external dataset largely aimed to recognize names of top twitter accounts, which most frequently occur as usernames in tweets. However, in our case they were anonymized in both training and test sets (@USER). Last table shows the scores of our submissions for sub-task C:

System (Submission)	F1 (macro)	Accuracy
All GRP baseline	0.1787	0.3662
All IND baseline	0.2130	0.4695
All OTH baseline	0.0941	0.1643
LSTM + Prepr.	0.2027	0.3099
LSTM + Prepr. + Lex. lookup without Pronouns	0.3582	0.3709
LSTM + Prepr. + Lex. lookup with Pronouns	**0.3769**	**0.4883**

Table 3: Results for Sub-task C.

For the last sub-task, which was devoted to categorizing targets of offense, a considerable increase in F_1-score can be observed by using the external datasets for postprocessing. Hence, the results showed that using a lexical lookup could be much more efficient in categorizing the possible victims than in identifying the presence of aggression per se. Below one can also find the confusion matrices of our best runs:

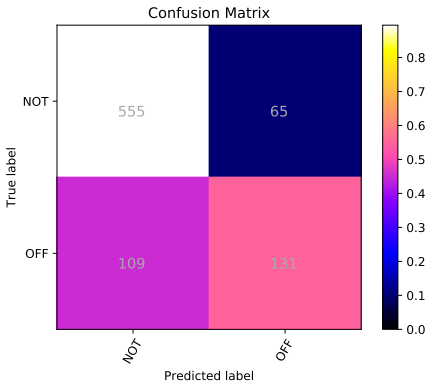

Figure 4: Sub-task A, HAD-Tübingen LSTM + Hashtag parsing.

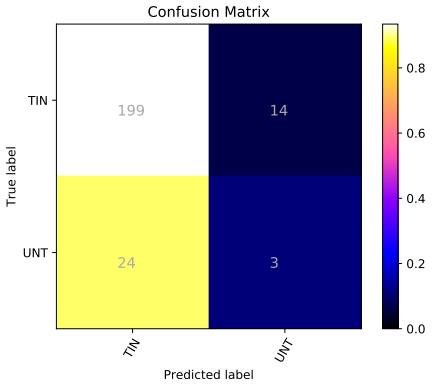

Figure 5: Sub-task B, HAD-Tübingen LSTM + Preprocessing.

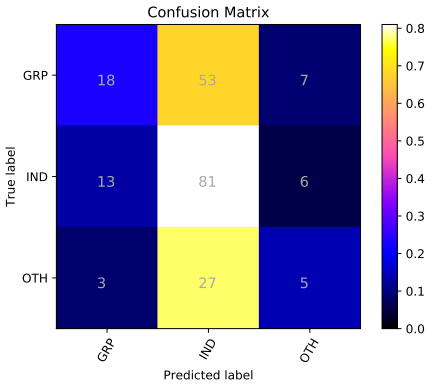

Figure 6: Sub-task C, HAD-Tübingen LSTM + Preprocessing + Lexical lookup with Pronouns.

It is also worth mentioning that a model choice and its settings should be made according to the training set size. In our case, the volume differed significantly for all the sub-tasks. However, a significantly lower performance of all the submissions can be observed on the last sub-task with the smallest training set.

4 Conclusion and future work

In our paper we presented the contribution of HAD-Tübingen to the OffensEval 2019 (SemEval 2019 - Task 6). Our approach combines sentence simplification as a preprocessing step and a lexical lookup as a postprocessing step with an unidirectional LSTM with 4 hidden layers. We picked the epochs according to the best F_1-score for our model configurations, according to SVM predictions. We found out that simple LSTM models are not likely to outperform SVM in such classification tasks. However, as a next possible step in working with an LSTM based classifier, could be using an external task-specific lexicon as an input feature to our model, but not as a postprocessing step. We would also like to make use of the pretrained vectors from Fastext library that are based on sub-word character n-grams for improving our model.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. Tensorflow: Large-scale machine learning on heterogeneous distributed systems.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada. Association for Computational Linguistics.

Çağrı Çöltekin and Taraka Rama. 2018. Tübingen-oslo at Semeval-2018 task 2: Svms perform better than rnns in emoji prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 34–38. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Taraka Rama and Çağrı Çöltekin. 2017. Fewer features perform well at native language identification task. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 255–260. Association for Computational Linguistics.

Paolo Rosso, Julio Gonzalo, Raquel Martínez, Soto Montalvo, and Jorge Carrillo de Albornoz, editors. 2018. *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages*. Sevilla, Spain.

Zeerak Waseem, Wendy Hui Kyong Chun, Dirk Hovy, and Joel Tetreault, editors. 2017. *The First Workshop on Abusive Language Online: Proceedings of the Workshop*. Association for Computational Linguistics (ACL), Vancouver, Canada.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of the GermEval 2018 Workshop*, pages 1–10, Vienna, Austria. Austrian Academy of Sciences.

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)*.