Subjectivity Mining: Assignment 3

Group AI-11

Matthias Agema[1] (2707650), Noa van Mervennée[2] (2710633), and Matteo De Rizzo[3] (2749303)

[1] Business Analytics, Vrije Universiteit, 1081 HV Amsterdam
m.j.agema@student.vu.nl
[2] Artificial Intelligence, Vrije Universiteit, 1081 HV Amsterdam
n.van.mervennee@student.vu.nl
[3] Artificial Intelligence, Vrije Universiteit, 1081 HV Amsterdam
m.de.rizzo@student.vu.nl

1 Introduction

Due to the rapid growth of social media and subsequently offensive ellipsis on the internet, the need for automatic classification of hate speech related content is a must nowadays. In the previous two assignments we got acquainted with annotation schemes and used a pipeline for automatic classification by making use of conventional machine learning models. Transformer models have shown state-of-the-art solutions already in the field. Still a challenge is the degree of generalizability for such models. Therefore, in this paper, we performed experiments with a selection of models on both in- and cross domain. For this task, two datasets were used for training from Zampieri et al. [6] and Mandl et al. [4].

The paper is organised in the following way. First, the data is described in Section 2, shedding light on data that is used for in-domain and cross-domain experiments. Next, Section 3 presents the methods that we used. Based on this, the experimental setup will be discussed in Section 4, including preprocessing steps and the models' settings. This is followed by the results per model in Section 5. Subsequently, error analyzes will be provided of both in- and cross domain experiments in Section 6 and is provided with a quantitative analysis. At last, conclusions are drawn in Section 7, which also covers the future work.

2 Data

For this study, two datasets have been used, OLID and HASOC. In-domain experiments will be performed on OLID, whereby models will be trained on a subset of OLID's traindata. Therefore, the sizes of both labels in the traindata of both datasets are equal as can be seen in tables 1 and 2. With respect to the cross-domain experiments, classifiers will be trained on the HASOC traindata. At last, results in terms of performance are measured only on OLID's testdata. In this section we will elaborate on both datasets.

Table 1. Data distribution OLID

	Trainset	**%**	**Testset**	**%**
NOT	3591	61.4	620	72.1
OFF	2261	38.6	240	27.9
Total	5852		860	

Table 2. Data distribution HASOC

	Trainset	**%**
NOT	3591	61.4
HOF	2261	38.6
Total	5852	

2.1 OLID

The original Offensive Language Identification Dataset (OLID) contains over 14.000 English tweets. This dataset is annotated by Zampieri and colleagues for offensive content using a three layered annotation scheme. In the first layer A, the goal is to discriminate tweets between offensive and non-offensive. Where offensive

content is defined as: "Posts containing any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct. This includes insults, threats, and posts containing profane language or swear words." [6]. Layer B distinguishes targeted insults from untargeted ones. At last, if posts were targeted, in layer C they categorized the targets of insults/threats in three categories. In this paper, we focus on layer A, whether posts are offensive or not. For our in-domain experiments a subset of OLID's traindata was used. The data is preprocessed resulting in a binary label, whereby 0 indicates a non-offensive tweet and 1 corresponds to an offensive message.

2.2 HASOC

The Hate speech and Offensive Concent Identification (HASOC) in Indo-European languages track was introduced in 2019 by Mandl et al. [4]. Three datasets were developed for three different languages, namely: English, Hindi and German. For our experiments we only make use of the English tweets and Facebook messages. HASOC was inspired by OffensEval from Zampieri et al. [7] and comprises three sub-tasks, where the data was created for. The annotation architecture looks like the from Zampieri and colleagues, but differs at some points. In the first layer, they distinguish two labels for a binary classification task, namely: Hate and Offensive (HOF) and Non- Hate and offensive (NOT). They labeled posts as HOF if they contain any form of non-acceptable language such as hate speech, aggression and profanity, which is quite similar to OLID's rule for layer A, all others get NOT. Next, in layer B all HOF labeled cases are further categorized into three classes, (i) hate speech (ii) offensive and (iii) profane, which corresponds reasonable with the third layer from OLID's annotation scheme. At last, layer C considers the type of offense, targeted or untargeted. We will again focus on the first layer, for which a binary classification can be performed. Preprocessing steps are already done by labeling hate and offensive as 1, and a 0 corresponds to non-hate and non-offensive messages. Only traindata is used from HASOC, because of the cross-domain experiments that we are going to perform on this dataset.

3 Methodology

Three models have been developed and executed with the aim of classifying hateful/offensive content. As we are dealing with a (binary) classification task, we searched for classification models that are commonly applied to text-classification tasks. The methodology of two pre-trained language models (BERT and BERTweet) and one machine learning model (SVM) will be explained in the following sections. For this study, both the BERT and BERTweet models have been implemented using the default hyperparameters.

3.1 BERT

The first chosen model is a pre-trained language model, called BERT. BERT stands for Bidirectional Encoder Representations from Transformers, which is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers [1]. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks [1].

Pre-training The model is pre-trained on the BooksCorpus (800 million words) and English Wikipedia (2,500 million words). Pre-training had two supervised tasks: (i) masked language modeling and (ii) next sentence prediction. In task (i), 15% of all tokens in each sentence are masked at random, then the model predicts the masked tokens. For task (ii), two masked sentences are concatenated, which may or may not have been consecutive in the corpus. It is now up to the model to predict whether those sentences followed each other or not. Both tasks (i) and (ii) are essential for BERT to find context in two directions and associate with relations between sentences.

3.2 BERTweet

As the OLID dataset consists only of texts originating from Twitter, we chose to implement BERTweet as our second pre-trained language model. Tweets are recognized by their short length, use of abbreviations, emojis and informal grammar, making it very difficult to use the content for a classification task. However, BERTweet is the first public large-scale pre-trained language model for English Tweets [5]. The BERTweet uses the same architecture as BERT, which is, as mentioned, trained with the masked language modeling. The pre-training procedure is based on RoBERTa [2] which optimizes the BERT pre-training approach for more robust performance [5]. RoBERTa is significantly improved compared to BERT by training the model longer, with bigger batches over more data; removing the next sentence prediction objective; training on longer sequences; and dynamically changing the masking pattern applied to the training data [2].

Pre-training As mentioned in the paper of Liu et al. [5], the model is trained using a 80GB corpus of 850 million English Tweets. The Tweets consist in total of 16 billion tokens where each Tweet has at least 10 and at most 64 tokens. First, Tweets from 01/2012 to 08/2019 have been downloaded. The Tweets are tokenized using the "Tweet Tokenizer" from the NLTK library and use the emoji package to translate emoticons into text. Normalization is performed by anonymizing user-data with @USER and webpage related data with HTTPURL. At last, retweeted Tweets are filtered out resulting in a corpus with 845 million Tweets. More Tweets originate from 01/2020 to 03/2020 and are related to the COVID-19 pandemic. The same procedures have been applied to these Tweets, resulting in the remaining trained corpus of 5 million English Tweets. Finally, all 850 million Tweets are segmented with subword units resulting in 25 subword tokens per Tweet on average.

3.3 SVM

The third model that has been used to classify the datasets is the best performing algorithm trained on the OLID dataset with a size of 13,240 from the previous assignment, a supervised machine learning model called support vector machine (SVM). This model employs classification techniques to solve two-group classification problems. A SVM can classify new data once it has been provided with a dataset containing labeled entries for both categories. SVM has gained popularity in the field throughout the past years on account of its text classification capabilities. When used for this purpose, it usually provides good results in terms of classification recall rate, accuracy rate and F1-score [3]. As for assignment two, the SVM was utilized in combination with Count Vectorizer. As the training-procedure is already optimized on a larger batch-size, we chose not to perform the hyperparameter-tuning procedure again on the small traindata. Therefore, the model is not validated on a development set, and is directly trained and evaluated on the train- and testset provided in table 1 and 2.

4 Experimental setup

The objective of this research is to investigate the generalizability of several methods for automatic classification of hate speech. To this end, the three models will first be trained on the in-domain setup (OLID dataset), and then with the same parameters they will be trained on the cross-domain setup (HASOC dataset). This will produce a total of six algorithms, all of which will then be evaluated on the OLID test dataset. In order to evaluate them we will be using macro-averaged precision, recall, and F1-score.

4.1 Hyperparameters transformer models

For the first pre-trained transformer model BERT, we implemented the default hyperparameters of $\text{BERT}_{\text{BASE}}$[4]. The model adopts the Adam optimizer with a learning rate of 0.0001. $\text{BERT}_{\text{BASE}}$ applies 12 layers, a hidden size of 768 and a number of self-attention heads of 12, as stated in the paper of Devlin et al. [1]. The

[4] `https://huggingface.co/bert-base-cased`

BERTweet model is initialized with the $\text{BERTweet}_{\text{BASE}}$ model [5] architecture that implements the same architecture as the $\text{BERT}_{\text{BASE}}$ and the same pre-training procedure as RoBERTa [5], resulting in 135 million parameters [6].

4.2 Hyperparameters SVM

For this paper, we attempted to tune of the model on these two specific datasets. A wide variety of hyperparameters have been manipulated so to discover the best performing model on this task. Nonetheless, while certain combinations of hyperparameters would just slightly improve the macro average F1 score, it would also create a higher discrepancy in both precision and recall among the two categories. In fact, the models ability to categories non-offensive messages slightly improved, on the countrary to offensive ones, which showed a reduction on both scores. Therefore we came to the logical conclusion to utilize the best performing hyperparameters from the previous assignment, as they have been previously finetuned on a much larger batch. Hence, we decided to use a combination of unigrams and bigrams, that is, n-gram = (1,2). In addition to the Count Vectorizer, the parameter of the SVM are a min_df parameter value of two and a max_df value of 0.5. The choice for the regularization parameter C was 0.5 as it showed the best results in previous work.

5 Results

The models' performance in terms of precision, recall and F1-scores for the in- and cross domain is shown in respectively table 3 and 4. All metrics are divided over the two different classes non-offensive (NOT) and offensive (OFF) and are depicted in bold if that particular score is best in that row. As mentioned in the experimental setup, section 4, all models are tested on the OLID testdata.

Table 3. Metrics in-domain experiments for BERT, BERTweet and SVM (trained on OLID data)

In-domain		**BERT**	**BERTweet**	**SVM**
Precision	NOT	**0.89**	0.88	0.83
	OFF	0.68	**0.74**	0.60
	macro avg.	0.78	**0.81**	0.72
Recall	NOT	0.87	**0.91**	0.86
	OFF	**0.72**	0.68	0.55
	macro avg.	**0.80**	0.79	0.70
F1-score	NOT	0.88	**0.89**	0.84
	OFF	0.70	**0.71**	0.57
	macro avg.	0.79	**0.80**	0.71

We can observe that all models on both domains yield to higher scores for the non-offensive class, with a slight improvement of the BERT and BERTweet compared to the SVM model. The scores for the non-offensive class are approximately similar for both the BERT and BERTweet model. The recall score for the OFF class is tremendously smaller compared to the recall score of the NON class for all models on both domains, which implies that it is hard for these models to correctly predict offensive messages. In addition, for both domains, it is noteworthy that the BERTweet model outperforms the BERT model exceptionally in terms of precision for the offensive messages. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative [7], i.e. how good is the model at predicting a certain category. Additionally, SVM can hardly compete with the transformers models. In-domain results are descent, but SVM explicitly falls short in the cross-domain experiment. The in-domain F1-scores for the non-offensive class are fairly close to each other, especially for BERT and BERTweet with a negligible difference.

[5] `https://huggingface.co/vinai/bertweet-base`
[6] `https://datquocnguyen.github.io/resources/NVIDIA_GTC_Talk.pdf`
[7] `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html`

Table 4. Metrics cross-domain experiments for BERT, BERTweet and SVM (trained on HASOC data)

Cross-domain		**BERT**	**BERTweet**	**SVM**
Precision	NOT	**0.83**	0.84	0.76
	OFF	0.54	**0.76**	0.38
	macro avg.	0.69	**0.80**	0.56
Recall	NOT	0.80	**0.94**	0.77
	OFF	**0.59**	0.53	0.35
	macro avg.	0.70	**0.73**	0.56
F1-score	NOT	0.82	**0.88**	0.76
	OFF	0.56	**0.63**	0.36
	macro avg.	0.69	**0.76**	0.56

Overall, cross-domain results are comparable with the in-domain results, but with a drop in performance for most metrics. One can see that the precision for the offensive class is the only metric that have higher performance in the cross-domain experiment. This is offset by a relatively large drop in recall scores for the offensive class. What means that models in cross-domain have more difficulty in finding a large part of the offensive cases.

6 Analysis

For this section, a certain amount of errors regarding the in-domain and cross-domain experiments are analysed on a qualitative- and quantitative-level. In the first two sections, predicted messages are qualitativetly analyses. After that, a certain amount of statements are quantitatively substantiated.

6.1 In-domain errors

Figure 1, 2 and 3 visualizes the confusion matrices with respect to the models trained/tested on the OLID traindata. Both BERT and SVM obtained a larger amount of predicted offensive messages than BERTweet, whilst the true label is non-offensive (83 and 87). The following message is incorrectly classified offensive by BERT and SVM, and correctly classified non-offensive by BERTweet:

1. #SundayFunday EMOJI Go check out my girl the canncierge because she takes bomb photos and drops a hell of a lot of cannabis knowledge! She is the canncierge after all EMOJI I'm just a pot stock... URL

Hereby, the EMOJI is a replacement of three real emojis: two smiling faces and a thumbs up. It is expected that the BERTweet model classifies the message as non-offensive, as it translates emojis into text due to the implimitation of the emoji package during the pre-training procedure. Furthermore, an example of a message classified non-offensive by both BERT and SVM, and correctly classified offensive by BERTweet, is the following:

1. @USER #DavidHogg, you're nobody.

Another logical explanation for this result is the recognition of a name inside the hashtag. As the BERTweet model is specialized on tweets, this model might be better at detecting texts hidden in hashtags, and could therefore correctly predict that 'nobody' is a direct insult to the person. At last, the following message is correctly predicted as offensive by BERT, but non-offensive by SVM and BERTweet:

1. #LiberalHypocrisy #TacoBell When Liberals ask why your against illegal immigration? Taco Bell Employee: No Habla Ingles!" URL

A reason for this could be that the BERT model is trained on both WikiPedia texts and the BooksCorpus. Therefore it might be better at predicting racial comments due to the recognition of sentence structures.

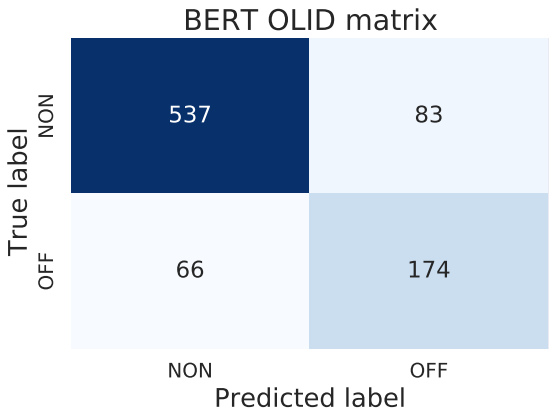

Fig. 1. Confusion matrix BERT OLID data

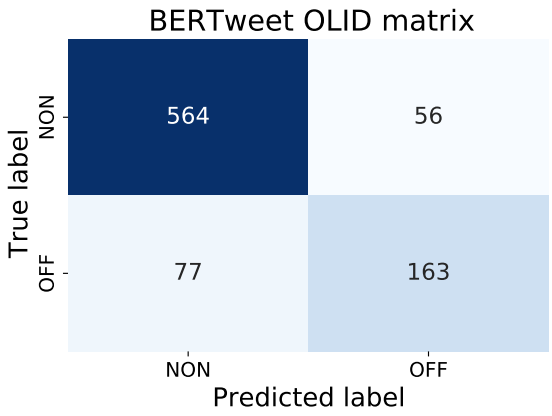

Fig. 2. Confusion matrix BERTweet OLID data

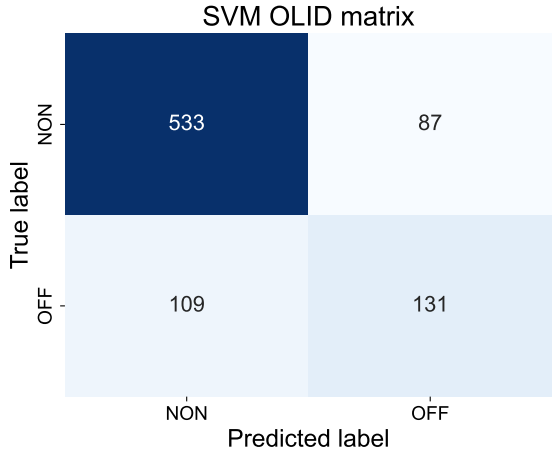

Fig. 3. Confusion matrix SVM OLID data

6.2 Cross-domain errors

As mentioned, in the cross-domain set-up, experiments have been performed that train the models on the HASOC traindata, and subsequently test the models on the OLID testdata. Figure 4, 5 and 6 visualizes the confusion matrices. It can be seen that BERT classifies much more content as offensive. An example of a message correctly predicted offensive by BERT and SVM, and incorrectly classified as non-offensive by BERTweet is shown below:

1. @USER @USER I'll use that one the next time im in a gun control debate or in a debate about free speech or taxes. Yes you can choose to be irresponsible or choose not to be. I argue responsible. Whats wrong with that? Don't justify murder by saying it was never alive or its my right.

A reasonable explanation for this is the length of the message. As stated earlier, the BERTweet model is pre-trained only on tweets, which contain a small amount of characters. Therefore the model might be facing difficulties in predicting longer messages. SVM has only been trained on the HASOC data, which might conclude that the model is not biased when predicting longer messages. Moreover, it is again remarkable that many messages have been incorrectly classified as offensive by BERT and SVM, although the true label is non-offensive. Another example that caught our attention is the following:

1. An American Tail really is one of the most underrated animations ever ever ever. Fuck I cried in this scene

It would be expected that all models will predict this tweet as offensive because of the curse word *Fuck*. Remarkably, only BERTweet is able to correctly distinguish the Tweet as non-offensive. While BERTweet is not explicitly trained to detect hate speech, the model may be remarkable good at detecting it due to the common use of swear words on Twitter.

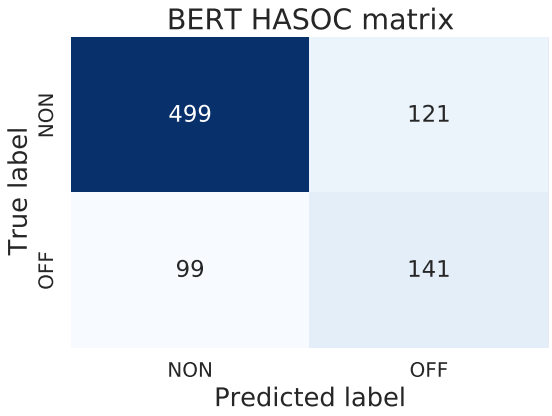

Fig. 4. Confusion matrix BERT HASOC data

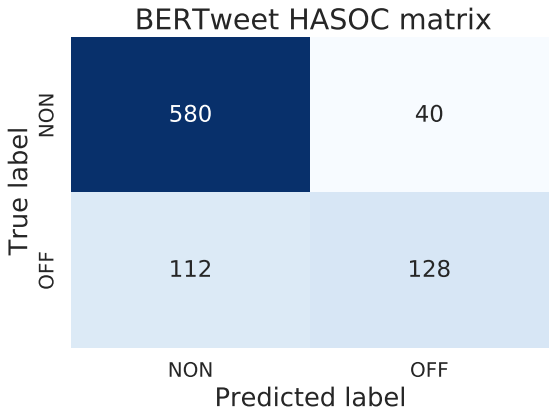

Fig. 5. Confusion matrix BERTweet HASOC data

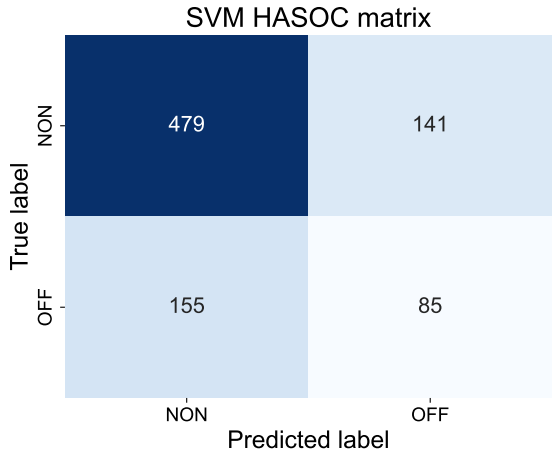

Fig. 6. Confusion matrix SVM HASOC data

6.3 Quantitative analysis

On table 3 we can notice how the three models performed on the in-domain setup. While SVM did not manage to reach the same level of neither transformer-based model, BERT and BERTweet performed similarly on the in-domain experiment. BERTweet made significantly less wrong predictions for the offensive class (See figures 1 and 2) and although it presents a lower Recall score on the offensive class than BERT ($0.68<0.72$) and a minimal higher score for precision in not offensive class ($0.88<0.89$) , it marginally outperforms it on all the other aspects. A high precision score is commonly preferable when conducting a classification task with imbalanced classes, as the precision does not account for the false negatives. The data distributions shown in table 1 and 2 confirms the class imbalance for this task with a ratio of 61.4% (Not Offensive) versus 38.6% (Offensive). In addition, BERTweet's better performance can be explained by its specialization to text derived from twitter. As cited earlier, the main difference between BERT and BERTweet it the data they were trained on, giving BERTweet an advantage when it comes to making predictions on the OLID

dataset (also consisting only of text taken from Twitter). In our case, BERTweet also performed better than BERT in the cross-domain experiment. This could be explained by the fact that the pre-training of BERTweet on the tweet format provided an advantage when testing on a congruent test set, even after being trained on a set of messages with different characteristics and different format. Indeed, as the HASOC set is not exclusively originating from twitter but it includes facebook messages, the overall characteristics of the data change. The messages tend to be longer, as well as containing more formal language (less abbreviations, emojis, etc.). It is plausible that these two aspects might have contributed to the drastic drop in performance observed in the cross-domain setup. In the case of BERTweet, this drop can be mostly observed in the recall score (0.68 to 0.53) and the F1-score (0.71 to 0.63) for the Offensive class. Moreover, the same effect can be noticed on all models used in this paper (see table 4).

7 Conclusion and future work

In this paper, two pre-trained transformer models and one conventional machine learning model have been developed and executed to gain further insights into the generalizability issues in the area of hate speech detection. Experiments show that all models perform approximately equally well with classifying messages as non-offensive. Difficulties arise when classifying hateful/offensive content. A practical reason for this could be the lack of availability of offensive data, resulting in class imbalance. As mentioned, the precision metric is a valuable metric to consider when class imbalances occur, where BERTweet has proven to excel. In contrast, the relatively low recall scores for both the in-domain and cross-domain setups regarding the Offensive class raise doubts. A trade-off must be accounted for. Is it more beneficial to trust your model on the quality, e.g. the model yields very few results, but with most of its predicted labels correct (precision). Or should you return many results, but included with many incorrect predictions (recall)? In the context of hate-speech it is difficult yet important to detect and act upon the detection of the content. To incorporate both the recall and precision metric, the F1-metric is evaluated. Here, it was again obvious that BERTweet outperformed the other models for both the in-domain and cross-domain setups. This advantage is most likely caused by the test environment of the experiments. The models have all been tested on content originating from Twitter. As the BERTweet model is pre-trained on 850 million tweets, it was expected that this model would outperform the other models. Concerns about the generalizability of the model may arise when this model has to classify offensive content coming from areas other than Twitter. In addition, it is recommended for future work to test the degree of bias towards the BERTweet model by testing the model on content other than tweets. Moreover, for future work, it might be interesting to test the performance of other transformer-based models compared to the ones described in this paper. For example, would HateBERT outperform BERTweet? As mentioned in section 6.2, BERTweet is already able to classify a tweet as non-offensive even though there is a curse word in the sentence, which could be the cause of the common use of swear words on Twitter. Would HateBERT not only identify hateful/offensive content better, but also account for the generalizability in case the BERTweet model fails on that part? In a follow-up paper we will aim to develop a more sophisticated method of classifying hateful content with the use of ensembles methods, that may answers some of the concerning questions.

Appendix

GitHub link

`https://github.com/teoderizzo/GrAI-11_assignment3_SubjectivityMining`

Contributions

Noa - 3.0, 3.2, 4.1, 6.0, 6.1, 6.2, 7
Matthias - 1, 2, 3.1, 5
Matteo - 3.3, 4.0, 4.3 6.3

References

1. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423
2. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach (2019), http://arxiv.org/abs/1907.11692, cite arxiv:1907.11692
3. Liu, Z., Lv, X., Liu, K., Shi, S.: Study on svm compared with the other text classification methods. In: 2010 Second International Workshop on Education Technology and Computer Science. vol. 1, pp. 219–222 (2010). https://doi.org/10.1109/ETCS.2010.248
4. Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., Patel, A.: Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In: Proceedings of the 11th Forum for Information Retrieval Evaluation. p. 14–17. FIRE '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3368567.3368584, https://doi.org/10.1145/3368567.3368584
5. Nguyen, D.Q., Vu, T., Tuan Nguyen, A.: BERTweet: A pre-trained language model for English tweets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 9–14. Association for Computational Linguistics, Online (Oct 2020). https://doi.org/10.18653/v1/2020.emnlp-demos.2, https://aclanthology.org/2020.emnlp-demos.2
6. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: Predicting the type and target of offensive posts in social media. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 1415–1420. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1144, https://aclanthology.org/N19-1144
7. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 75–86. Association for Computational Linguistics, Minneapolis, Minnesota, USA (Jun 2019). https://doi.org/10.18653/v1/S19-2010, https://aclanthology.org/S19-2010