Sentiment Analysis of Tweets using pre-trained BERT and ordinal regression

Text Mining: Final Assignment

Arsen Ignatosyan$^{[s4034538]}$ and Severin Holtmann$^{[s3624099]}$

Leiden University

1 Introduction

In 2017, Rosenthal et al. published their findings of the performance of a semantic evaluation of tweets collected from twitter as part of the 'SemEval' workshop series on semantic analysis tasks [5]. Task 4 of the 'SemEval-2017' research discusses the use of classification to identify sentiments within tweets in both English and Arabic. The SemEval dataset features $\sim 62,000$ english tweets collected between 2013 and 2016, and includes labels for the sentiment of the tweet on a 3-point scale (Positive, Negative, Neutral), as well as an identifier of the tweet's author. Based on this, we aim to identify the sentiment on a tweet-level, by training a support vector regression (SVR) and fine-tuning several pre-trained BERT LLM, tasked to predict the sentiment of tweets as a continuous measure, based on the 3-point ordinal scale (Negative: -1, Neutral: 0, Positive: 1) of the data. Our research evaluates the potential of a pretrained BERT model, fine-tuned for the ordinal regression task of measuring sentiment in the English tweets of the SemEval dataset. We developed a simple SVR as a baseline, as well as several fine-tuned BERT models. With this, we aim to discover the potential of pre-trained LLMs for novel tasks, and make a comparison to a much simpler SVR only trained on the data/task at hand, similar to the SVMs and other less advanced models used in the SemEval paper. Furthermore, our approach of using a continuous outcome measure aims to highlight the differences, benefits and challenges of using continuous or categorical outcomes.

2 Background

Our main source and influence for this undertaking is the research that Rosenthal et al. conducted themselves on Twitter sentiment analysis. In their paper 'SemEval-2017 Task 4: Sentiment Analysis in Twitter', the authors discuss sentiment classification of tweets into three categories under subtask A of the paper. The authors discuss the development of this task over the years, seeing as their research has already been conducted since 2013, and elaborate on it by also including Arabic tweets, thereby introducing an unprecedented complexity to the task. Rosenthal et al. largely focus on providing an overview of the culmination of this task, including the data collecting and labelling procedure, model evaluation, and model performance of the several teams which participated in the workshop.

The data was collected from twitter, with a large focus on popular events and trending topics on a global scale. All data from past years (2013 - 2016) was used as well. Ultimately, only relevant topics which featured more than 100 tweets were retained (Rosenthal et al., 2017). To obtain labels of the sentiment of these tweets such that supervised learning techniques are possible, the authors used human annotators, with each tweet being labelled by at least five people. Labels were provided on a five-point scale and featured both overall- and topic-level sentiment labels. However, we will only look at a three-point scale and overall sentiment.

To measure the performance of models, the authors measure the average recall across the three sentiments. This metric is particularly favourable as it is less sensitive to class imbalance than other common metrics, such as F1. Nevertheless, overall accuracy and an average F1 of the positive and negative classes is measured as well. Given that we are taking a regression approach, our primary performance metrics will be more focused on error-based indicators. However, we will also provide classifications based on a simple discretisation procedure, hence allowing us to adopt the categorical metrics used in the paper.

Further, the authors discuss the performance of the 38 teams which participated in the English subtask A. Popular approaches included CNNs and LSTMs, SVMs, as well as more statistical approaches such as Naive Bayes and Logistic Regression. However, no continuous regression was mentioned, which highlights the scientific value added by our enquiries. The best performing team was 'DataStories' and 'BB_twtr', with an average recall of 0.681 on the test set. 'BB_twtr' is likely the superior team, as their model performed better on the secondary metrics F1 and accuracy. These results were achieved using CNNs and LSTMs. Many of the high ranking teams also used SVM, which provides further relevance to our baseline approach of using SVR.

Another relevant resource is the 'TweetEval' paper by Barbieri et al., 2020 [1], which discusses the development of the RoBERTa model base for tweets that we decided to use for comparison. The authors discuss how they used the RoBERTa architecture developed by Liu et al., (2019) [3] to develop their own model or using the original RoBERTa as a baseline for fine tuning on a twitter corpus of 60m english tweets which were provided with automatic emotional labels by Twitter. In conclusion, the researchers find that fine tuning the original RoBERTa base on their Twitter corpus leads to the best performance, underlining the superior power of using a well-trained general model and using data specific to one's task to fine tune it. Based on this, it is of interest to us to see how we can benefit from a model trained on a task very similar to ours, and then specialise the model even further for our task.

Following from this, we also inspected the 'TimeLMs' paper by Loureiro et al., 2022 [4], which discusses an expansion of the RoBERTa model, which we will be using as a base for fine tuning. The primary addition of this model is that the corpus has since been expanded to 90m tweets. We will be making use of the sentiment classification variant of the model as a baseline to be fine tuned for our continuous sentiment analysis task.

3 Data

We decided to join all data provided for subtask A and then make our own training, validating and testing split. We made this decision to provide a more general picture of the way sentiment is expressed in tweets, seeing as the data were collected at different points in time and thus may express their views in different ways. This also allows the seamless incorporation of specialised data, such as the '2014sarcasm' dataset. Any duplicates were deleted from the joined sample, leaving us with 49,467 tweets, of which 22,000 positive, 19,500 neutral and 7,700 negative. We notice that far less tweets labelled as negative are present in the data. This may impact the performance (particularly of ambiguously negative tweets), but will not be taken into account explicitly during the modelling process. Lastly, we split the data into a training (64%), validation (16%), and testing set (20%), leaving us with 31,658, 7,915 and 9,894 observations respectively. Having re-coded the labels 'negative', 'neutral', 'positive' into -1, 0, 1 respectively, we may proceed with development of the models (We actually use 0, 1, 2 as encoding during the modelling process but deemed a zero centre to be more intuitive for explanation).

4 Methodology

In general, we aim to explore a fundamentally different approach to what is discussed by Rosenthal et al.; regression of a continuous parameter. While ordinal regression bears difficulties due to the inherent violation of regression assumptions (categorical data, bounded support, ambiguous scale), we believe that the indiscrete nature of the resulting predictions can provide more nuance to the notion of a tweet. A three-point scale may often be too simple to give an accurate representation of the sentiment. Our hope is that a continuous metric can reinstate more granular information about the sentiment of a tweet, thereby becoming a more informative predictor.

As a preliminary reformulation of the text data to a usable format for SVR, we applied a countvectorizer transformation. We omit stop words, make everything lower-case and limit the ngram range to (1, 3), based on a word-level analysis. The intention behind this is to simplify the model and provide a format in line with the uncased BERT base, which we will be using later on. Finally, we compute the tf-idf scores of the vectors. This metric reflects both the relevance of an n-gram to a given tweet, as well as the overall (inverse) relevance of the n-gram to our corpus of tweets. Such vectors of tf-idf scores will be used as input for our support vector regression.

For the BERT models, the raw data is simply passed through an AutoTokenizer, which refers to the tokenizer used by the respective BERT base. For the bases we used, this is typically the WordPiece tokenizer, which is a closed-source tokenization algorithm by Google.

As we decided to use a simple support vector regression as our baseline, we simply used the default model provided by scikit-learn. This model features a radial basis function kernel, which creates complex, non-linear decision spaces. The influence of individual observations 'gamma' is set to $\frac{1}{n_{features} \cdot Var(X)}$, where $Var(X)$ represents the overall variance of a collective set of all features. The broadness of the no-penalty area 'epsilon' is set to 0.1. As the SVR is supposed to merely provide a fundamental benchmark, we chose to not tune it any further and provide more attention to the various BERT models.

To prepare the BERT bases for an ordinal regression task, we implemented a few specifications, such that the model produces a single continuous outcome. This entails setting the number of outcomes of the pretrained model to 1, and designing a custom loss function, which extracts the output from the current hidden state as a float and passes it through an MSE loss function.

Using this fundamental structure, we tested several BERT bases and also experimented with different learning rates. The bases we used were the 'BERT-base-uncased', which is a standard of english language modelling. The model uses 110 million parameters and was trained in an unsupervised environment on a huge range of publicly available texts. As the model is such a standard, we deemed it a necessary foundation for our testing of the BERT architecture.

Another pretrained model we used was the 'Distilbert-base-uncased', which is trained on the BERT base, but uses a distillation process, which aims to learn to produce the same output as the original BERT and also maintain similar hidden states, while using a simpler and more efficient architecture [6]. This model was specifically intended for downstream fine-tuning, making it a suitable candidate for our research. We will be testing the default model which uses a learning rate of 5e-5, as well as a much higher rate of 1e-3.

Lastly, we opted to explore 'Twitter-RoBERTa-base for Sentiment Analysis (2022)', which is a sentiment analysis BERT model specifically trained on tweets using the RoBERTa base, which was trained on 124M tweets [2]. Seeing as this model has been trained on the data we are handling (tweets) for the task we are interested in (sentiment analysis), we deemed this a most ideal base to use. Again, we test the default model, as well as a variant with an extended maximum tokenizer length of 512 instead of the default 256. The increased toke length may improve the amount of contextual information considered by the model.

In order to allow for a more direct comparison to the models discussed in the SemEval paper, we also chose to discretise our continuous estimates. This means that we rounded our predictions to the nearest whole number and set cutoffs at -1, 1, such that values could only be -1, 0, 1. Based on this, we can make a direct qualitative comparison to the metrics originally used, allowing us to provide accuracy, average recall, and an F1 score of the positive and negative predictions.

5 Results

Overall, we observe that the RoBERTa base, which was already fine tuned for sentiment analysis of tweets, performs best. Given the circumstances, these results are not a surprise. Hence, it is almost more interesting to reflect on the often times negligible improvement when compared to the more generic BERT models tested, as well as the difference between the BERT models and the SVR baseline, or the best performances in each category measured in the SemEval-2017 paper.

Table 1. Model Performances

Model	**MSE**	**MAE**	**R2**	**Avg. Recall**	**F1 (P/N)**	**Accuracy**
SVR Baseline	0.357	0.481	0.278	0.305	0.326	0.606
Bert-base-uncased	0.246	0.367	0.501	0.730	0.737	0.761
Distilbert-base-uncased	0.246	0.359	0.502	0.675	0.676	0.717
Distilbert-base-uncased, lr: 0.0001	0.250	0.370	0.493	0.669	0.672	0.707
Twitter-RoBERTa-base	**0.199**	**0.299**	**0.596**	**0.756**	**0.767**	**0.767**
Twitter-RoBERTa-base, 512	0.204	0.321	0.587	0.736	0.745	0.758
SemEval 2017 - Best performances	-	-	-	0.681	0.685	0.664

We notice an improvement in MSE of 30%-40% between the SVR and BERT models. The tendency is even more extreme in the discrete measures. Nevertheless, this shows that the use of a much more complex model which has been pretrained on similar tasks can provide substantial improvements.

Between the standard BERT and the distilled variant, we notice no difference in the continuous metrics. This is very positive, as it implies that the distillation procedure is highly efficient at capturing the BERT base, using ∼40% less parameters. Interestingly, for the classification metrics, the difference becomes more apparent with the distilled BERT performing ∼5-10% worse than the standard BERT model.

As mentioned before, the RoBERTa model, which has been pretrained for sentiment analysis of tweets performed best. Our extended input vector appears to perform slightly worse than the default parametrisation.

When comparing the discrete predictions of our models to the models discussed in the original 'SemEval' paper, we observe that our models perform significantly better, which is particularly impressive considering our predictions are not originally categorical. Nevertheless it should be noted that comparison to a categorical BERT would be more indicative of this. This emphasises the major improvement that transformer LLMs have provided to the world of text mining and the far-reaching potential of fine tuning.

6 Discussion

Our research highlights the difference in approach between a multi-classification, and a standard (ordinal) regression task. A major limitation of our approach is the lack of labels that are present in the format that our model has been trained to predict. On the other hand, our model is able to give a point prediction, which can provide more nuance about the exact sentiment. A classification approach, such as the original RoBERTa model used may only describe the probability mass distribution across the three categories. Based on this, one could compute a weighted average to arrive at a metric similar to ours, vice versa to how we discretised our predictions by rounding them to the nearest class.

To compare these approaches, we will fabricate a few tweets, gradually developing from negative to positive, to show how our best performing model, the fine-tuned RoBERTa, as well as the original RoBERTa by Loureiro et al., 2022 predict the outcome both as a continuous and categorical measure. As a continuous measure, we will provide the original output of our model, as well as a weighted sum of the probability predictions of the three sentiments provided by the classification RoBERTa model:

$$Sentiment = \begin{bmatrix} P_{Neg} \; P_{Neut} \; P_{Pos} \end{bmatrix} \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

To provide a categorical metric, we will simply apply the discretisation transform described above to our point prediction and use the most probable class for the original RoBERTa. Predictions using the original RoBERTa were simply made using the API on huggingface.

As both our corpus, as well as the corpus used by Loureiro et al. likely featured a lot of political commentary on Donald Trump, we decided to opt for the following tweets to highlight the gradual differences in sentiment. We chose to do this on a five-point scale, rather than a three-point scale used by our training data and the original RoBERTa model in order to highlight the potential benefits of using a continuous outcome. It should be noted that the actual sentiment of these sentences is still highly subjective, so interest predominantly lies in how much the two models differ, rather than the prediction they provided in general.

Table 2. Sentences

Aspired Sentiment	**Sentence**
Strongly Negative	'Donald Trump is the worst president!'
Weakly Negative	'Donald Trump is not the best president ever.'
Neutral	'Donald Trump is a tolerable president.'
Weakly Positive	'Donald Trump is a decent president.'
Strongly Positive	'Donald Trump is the best president ever!'

Table 3. Prediction Comparison

Aspired Sentiment	**Base Cont.**	**Tuned Cont.**	**Base Class.**	**Tuned Class.**
Strongly Negative	-0.923	-0.944	Negative	Negative
Weakly Negative	-0.876	-0.911	Negative	Negative
Neutral	0.713	0.820	Positive	Positive
Weakly Positive	0.704	0.827	Positive	Positive
Strongly Positive	0.978	0.988	Positive	Positive

We observe that both models struggle to produce gradual predictions. Surprisingly, the Base model, for which we took the weighted sum of the categorical probabilities, produced more gradual results than our tuned model. An explanation for this may be that the model never observed values besides -1, 0, 1, paired with the fact that our calculation of the continuous measures are inherently different. Classification results are identical between the models. The results remain inconclusive.

7 Conclusion

In summation, our research aimed to identify the potential of using a variety of BERT bases to finetune a model for continuous sentiment prediction using a corpus of twitter posts. In doing so, we discovered that this strategy proved to be very effective and led to our best performing model, far better than our SVR baseline. Indeed, this model was not only trained on the English language, as with 'BERT-base-uncased', but also trained on a much larger corpus of tweets than our own, hence familiarising the model with the tweet structure. The biggest conclusion we may draw from this is that a hierarchical approach where a rather general base model is gradually fine tuned for increasingly niche tasks can lead to the most robust and well-rounded results.

In addition, our choice of applying an ordinal regression-style continuous outcome has proven to struggle expressing sentiment in finer detail than a mere classification model when trained on categorical data. Nonetheless, it can be argued that such a format is also more in line with the actual subjectivity of sentiments expressed on twitter. Seeing as discretised performance does not suffer too much (our discrete performance is still better than the SemEval-2017 references), we may regard continuous outcomes as an enriching approach to sentiment analysis tasks.

References

1. Barbieri, F., Camacho-Collados, J., Neves, L., Espinosa-Anke, L.: Tweeteval: Unified benchmark and comparative evaluation for tweet classification. Snap Inc., Santa Monica, CA 90405, USA; School of Computer Science and Informatics, Cardiff University, United Kingdom (2020)
2. Camacho-collados, J., Rezaee, K., Riahi, T., Ushio, A., Loureiro, D., Antypas, D., Boisson, J., Espinosa Anke, L., Liu, F., Martínez Cámara, E.: TweetNLP: Cutting-edge natural language processing for social media
3. 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)
4. Loureiro, D., Barbieri, F., Neves, L., Espinosa Anke, L., Camacho-collados, J.: TimeLMs: Diachronic language models from Twitter. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 251–260. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-demo.25, https://aclanthology.org/2022.acl-demo.25
5. Rosenthal, S., Farra, N., Nakov, P.: SemEval-2017 task 4: Sentiment analysis in Twitter. In: Proceedings of the 11th International Workshop on Semantic Evaluation. SemEval '17, Association for Computational Linguistics, Vancouver, Canada (August 2017)
6. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. ArXiv **abs/1910.01108** (2019)