TweetEval-based Tweet Classification Using a Transformer-based XLNet Model

Jacob Johnson
Computer Science and Electronic Engineering Department
University of Essex
Colchester, United Kingdom
jj17064@essex.ac.uk

***Abstract*—The TweetEval evaluation metric outlined in Tweeteval: Unified benchmark and comparative evaluation for tweet classification" proposes a standardised method of evaluation for performing classification tasks involving Twitter datasets. Along with the introduction of the TweetEval metric, the paper also details benchmarks of numerous models for 7 classification tasks using this devised metric. The goal of the experiment outlined and performed in this paper sought to contribute to the selection of models used to benchmark the classification tasks using TweetEval scores by testing a transformer-based XLNet model which had yet to be benchmarked. 3 classification tasks were completed using a transformer-based XLNet model (XLNet-base-cased). These tasks include sentiment analysis, hate speech detection and irony detection. The findings from this paper suggest that XLNet-base-cased on its own without any parameter tuning or pre-training on Twitter data has a rather noncompetitive performance in comparison to higher scoring models such as the BERTweet model or the RoBERTa based models for Twitter classification tasks according to the TweetEval metric.**

I. Introduction

Researchers at Cardiff university introduced the TweetEval evaluation metric in an attempt to standardise the evaluation procedure of classification tasks on Twitter datasets [1]. This is a welcomed attempt at evaluation standardisation, as it is often observed that there are far too many diverging evaluation methodologies when it comes to evaluating models in the natural language processing domain. This problem subsequently makes it more challenging for researchers to compare different models and solutions. TweetEval introduces a parsimonious way of evaluating models on twitter classification tasks by giving researchers and experimenters a script they can quickly run to evaluate their models' performance. TweetEval employs the use of a macro-average F1 score as the precise evaluation criteria of hate speech detection and the majority of other tasks. The exceptions relevant to this paper include sentiment analysis, in which the evaluation metric is macro-average recall and irony detection which the F1 value for the irony class [1]. Building on the foundations laid out in the original TweetEval paper, this experiment seeks to explore the performance of a baseline XLNet model (XLNet-base-cased), in regards to 3 out of the 7 classification tasks available for using the TweetEval dataset. The 3 tasks completed in this experiment include sentiment analysis [2], hate speech detection [3] and irony detection [4]. These tasks were selected to provide an idea of how XLNet-base-cased would perform classification on a Twitter dataset. The motivation for selecting the transformer-based XLNet model as the model of choice for this experiment is XLNet models are uniquely different from the other transformer-based models benchmarked so far on the Twitter datasets with TweetEval. XLNet uses an autoregressive formulation that enables learning bidirectional contexts by maximizing the expected likelihood over all permutations of the factorization order [5]. This is especially of note as it is a model that is designed to overcome the limitations of the BERT models, the derivatives of which are currently the highest-scoring TweetEval scores for the twitter classification tasks [6]. Furthermore, XLNet models have proven useful and have an excellent track record for performing well in numerous classification tasks [5]. 3 hypotheses have been devised to explore the effectiveness of the XLNet-base-cased model for classification on the 3 tasks. The experiment conducted in this paper will either support or refute the following hypotheses. Hypothesis I - The transformer-based XLNet-base-cased model will score higher than the BERTweet model in the sentiment analysis classification task according to the Tweeteval evaluation metric. Hypothesis II - The transformer-based XLNet-base-cased model will score higher than the BERTweet model in the irony detection classification task according to the TweetEval evaluation metric. Hypothesis III - The transformer-based XLNet-base-cased model will score higher than the BERTweet model in the hate speech detection classification task according to the TweetEval evaluation metric.

II. Literature Review

There were 6 models trained, tested and validated on all 7 classifications In the "Tweeteval: Unified benchmark and comparative evaluation for tweet classification" paper. The observations that can be gleaned from the results of this paper show a general superiority of the transformer-based RoBERTa family of models over the more traditional, non-transformer base of models (SVM, LSTM and FastText). The best overall performing model according to the average TweetEval score is the RoBERTa-retrained model, which scored the highest out of all the classification tasks except for irony detection (RoBERTa-twitter attained the highest score for that specific

task). Explanations for RoBERTa-retrained models' rather dominant performance over the other RoBERTa transformer-based models could be due to the retraining process of the model on Twitter data. Whereas the RoBERTa-base model was not pre-trained on any Twitter data and the RoBERTa twitter model was pre-trained solely on Twitter data from scratch [1]. Interestingly the RoBERTa twitter model pre-trained solely on Twitter data generally underperformed the RoBERTa base model, thus providing evidence that solely pretraining a model on Twitter data may not be optimal, but rather utilizing a mixture of context-specific training data, as well as more general training data, builds a superior classification model at least in the case of classifying Twitter data.

In 'BERTweet: A pre-trained language model for English tweets. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations', the BERTweet model devised in this paper sets the current benchmark for the highest overall model benchmarked for the 7 classification tasks on the TweetEval dataset. Within the paper that demonstrates the effectiveness of BERTweet, the model was compared and evaluated alongside XML-R and RoBERTa transformer models with the result showing the BERTweet model outperforming these models on two sentiment analysis tasks, even when the parameters of the XLM-R model and RoBERTa model were adjusted as well as different data normalization methods [7]. This paper interestingly undermines the point made in the prior literature analysis on whether it is more ideal to pre-train a transformer-based model on a mixture of datasets rather than just Twitter data. In this case, BERTweet was pre-trained on a huge amount of Twitter data (850 million tweets) and did not only manage to outperform the XLM-R and RoBERTa models in the paper but also outperformed all the other models on 6 of the TweetEval classification tasks. BERTweet sets the current top benchmark of the best classifier for most of the classification tasks, with the only exception being the offensive speech task (RoBERTa-retrained narrowly outperformed BERTweet).

The 'generalized autoregressive pre-training for language understanding' paper plays a large role in the inspiration of this experiment. This paper details and presents the XLNet model with its potential uses in classification tasks as well as auto generative features. This paper demonstrates that in tasks involving more general corpora than Twitter data, XLNet outperforms both BERT and RoBERTa transformer-based models in classification [5]. Whether the findings of this paper transfer to the domain of Twitter classification tasks is one of the main points of inquiry this experiment seeks to explore.

To conclude this literature review, it is important to note that TweetEval is a relatively recently devised evaluation and benchmark metric. Therefore, the literature is somewhat limited on this subject as only a relatively select amount of models have been tested for these select classification tasks. With this in mind, it presents many opportunities for researchers and experimenters to add and develop models with this evaluation system in mind, with the future potential of TweetEval having a wealth of models tested using its metrics. Therefore fully realising a standard to compare the quality of different models for 7 commonplace classification tasks on Twitter data. This premise is what inspired this experiment to be conducted as there is currently no benchmark for any XLNet based models, so the findings of this paper could contribute relevant knowledge as the XLNet model is a unique transformer-based model separate from the current BERT and RoBERTa based model benchmarked using TweetEval.

III. Methodology

The dataset used for the TweetEval benchmark classification tasks conducted in this experiment can be retrieved and downloaded from Cardiff NLP's GitHub https://github.com/cardiffnlp/tweeteval/tree/main/datasets. The 3 datasets that were downloaded correspond with the 3 tasks outlined in the introduction. The sentiment analysis dataset contains 45,389 individual tweets and labels for training. The testing data contains 11,906 tweets and labels. The second dataset, the irony detection text files contain training text and labels with 2,862 samples while the test text/labels have 784 samples. Hate speech detection data ha training text/labels with 9,000 samples and test text/labels that contain 2,970 samples.

The main tool used for setting up and running this experiment was Google Colabs. This environment provided an ideal setting to conduct this experiment as the utilizable GPU runtime allows a faster training process for training the XLNet-base-cased model without the requirement of expensive local hardware. The specific XLNet-base-cased model that will be utilised for this experiment is imported from HuggingFace using the Simple transformers module with the called model being labelled as the 'xlnet-base-casesd' model. This specific XLNet model's architecture consists of 12-layer, 768-hidden, 12-heads, 110M parameters. Each classification task conducted in this experiment is completed in their own respective notebook. The datasets for this experiment have already been split into training and a testing holdout set, therefore there was no requirement to further split the dataset. The steps outlined in the methodology were repeated for each of the 3 separate classification tasks.

A. Importing Necessary Modules and Scripts

Necessary Python modules that were imported include simple transformers, which allows a parsimonious implementation of transformer-based models. Other modules imported include Pandas, NumPy, Sci-Kit learn and PyTorch. The TweetEval script was also imported as a necessity to complete the evaluation for this experiment.

B. Preprocessing

To keep this experiment replicable, the preprocessing methodology followed closely to that in the original TweetEval paper and did not deviate in a significant way. The benefit of using the ready-made TweetEval datasets the preprocessing has mostly already been completed. The preprocessing steps the researchers at Cardiff took can be viewed in the original

TweetEval paper [1]. The only additional preprocessing steps taken in this experiment was removing the newline token ‘’ to reduce the unnecessary tokens in the training text, as well as converting the training labels from a text format into a Pandas DataFrame which is then merged with the training text. Observably the preprocessing stages taken were kept to a minimum in the creation of the TweetEval dataset for each classification task. This is likely due to the fact Twitter Corpora are a rather unconventional set of data due to it being a social media platform where you would obtain many unique tokens specific to Twitter and other social media platforms. Therefore over-zealously adopting rigorous preprocessing methodologies may accrue too much information loss to create an effective generalisable classification model.

C. Model Training

Once the training data and labels have been preprocessed using the method above the XLNet-base-cased model is imported from Simple Transformers, calling the ClassificationArgs function. Within this function, we set the number of training epochs to 1 as well as setting the number of classification labels relevant to the classification task (3 labels for sentiment analysis, 2 labels for irony detection and 2 labels for hate speech detection). Once the parameters were set for the XLNET-base-cased model was then trained on the model on the training DataFrame. Benefits of using the Simple transformer module over setting up the transformer model manually include that the straight forward implementations allow this experience to be replicable for researchers without having advanced programming skills. Another benefit to using Simple Transformers is it allows any transformer model uploaded to HuggingFace to be used, allowing researchers to easily expand the scope of this experiment to trial other transformer-based by adjusting one line code.

D. Model Evaluation

Once the training of the model on the respective dataset was completed the test text dataset was imported and the same preprocessing methodology for the training text was performed on the test data. After this stage, the predictions of the test labels were obtained using the trained XLNet-base-case model. The prediction values obtained were then converted from a DataFrame back into a text file format with the same structure as the test label set. The TweetEval script was then run comparing the predicted labels to the test labels and then returning the TweetEval score for the model.

E. Data Analysis

To produce the results for this experiment, the TweetEval evaluation scores of the 3 classification tasks were compared to the models in the Cardiff TweetEval leaderboard that can be found at https://github.com/cardiffnlp/tweeteval. This is a necessary comparison to provide evidence to either support or refutes the 3 hypotheses outlined in the introduction. To ascertain and provide an idea of the overall performance of the XLNet-base-cased model, the mean TweetEval score was calculated and computed for all 3 tasks as well as the other models TweetEval scores for the 3 classification tasks conducted in this experiment.

IV. Results

Results show that following the methodology outlined in this paper was valid for obtaining TweetEval scores that can be compared with the model scores benchmarked in the Cardiff NLP TweetEval leaderboard. The following results below show the TweetEval score of the XLNet-base-cased model obtained as well as direct comparisons to other models scores for that respective task.

A. Sentiment Analysis

Model	Sentiment Analysis TweetEval Score
BERTweet	73.4
RoBERTa-Retrained	72.6
RoBERTa-Base	71.3
XLNet-Base	70.2
RoBERTa-Twitter	69.1
FastText	62.9
SVM	62.9
LSTM	58.3

Fig. 1. TweetEval Scores Table for Sentiment analysis.

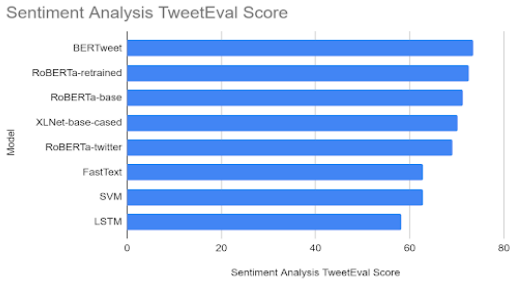

Fig. 2. TweetEval Scores Graph for Sentiment analysis.

The results from Figures 1 and 2 show that the TweetEval score of the pre-tuned XLNet-base places this specific transformer based model in the middle of the pack with a TweetEval score of 70.2. While outperforming RoBERTa Twitter by a small margin of 1.1 points, the XLNet model underperforms the RoBERTa base model, the RoBERTa retrained model and the BERTweet model, falling short of outperforming the top model by 3.2 points. Still, the XLNet-Base model attained a higher TweetEval score than any of the non-transformer based models (SVM, LSTM and FastText), for this specific task.

Model	Irony Detection TweetEval Score
BERTweet	82.1
RoBERTa-twitter	65.7
FastText	63.1
LSTM	62.8
RoBERTa-retrained	61.7
SVM	61.7
XLNet-base-cased	60.3
RoBERTa-base	59.7

Fig. 3. TweetEval Scores Table for Irony Detection.

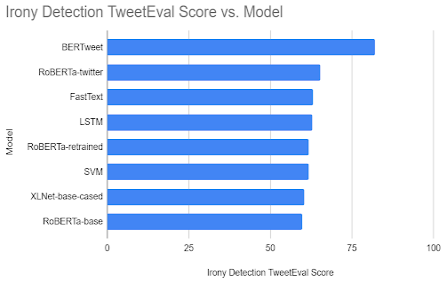

Fig. 4. TweetEval Scores Graph for Irony Detection.

B. Irony Detection

For the irony detection classification task results shown in Figures 3 and 4, the XLNet-Base model underperforms coming second last with a TweetEval score of 60.3. The model narrowly outperformed the RoBERTA-Base model by 0.6. The non-transformer based models all performed superiorly to the XLNET-Base model as well. The BERTweet model outperformed the XLNet-Base model by a large margin of 21.8 points showing a clear outperformance according to the TweetEval metric on this particular dataset and task.

C. Hate Speech Detection

Model	Hate Speech Detection TweetEval Score
BERTweet	56.4
LSTM	52.6
RoBERTa-retrained	52.3
FastText	50.6
RoBERTa-twitter	49.9
RoBERTa-base	46.6
XLNet-base-cased	39.1
SVM	36.7

Fig. 5. TweetEval Scores Table for Hate Speech Detection.

The results for the hate speech detection task also provide comparatively low scores for the XLNET-Base model (Figures 5 and 6), with the model coming second last in comparison to other models according to the TweetEval metric with a score of 39.1. The XLNet-Base model performs better than the SVM model outperforming it by 2.4 points, though falls short of all the other models with the best model benchmark set by BERTweet at a score that is 17.3 points higher than that of the XLNET-Base model.

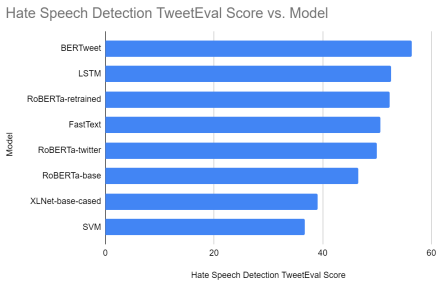

Fig. 6. TweetEval Scores Graph for Hate Speech Detection.

D. Mean TweetEval Score

Model	Mean TweetEval Score
BERTweet	70.6
RoBERTa-retrained	62.2
RoBERTa-twitter	61.5
LSTM	59.4
RoBERTa-base	59.2
FastText	58.9
XLNet-base-cased	56.5
SVM	53.8

Fig. 7. Mean TweetEval Scores Table.

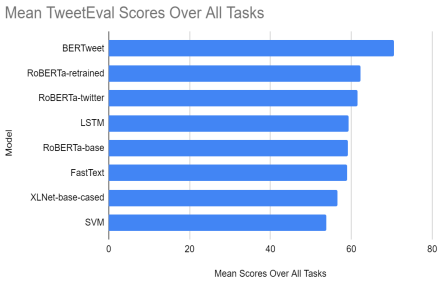

Fig. 8. Mean TweetEval Scores Graph.

The results are shown in Figures 7 and 8, where an overall score can be seen to estimate the model's comparative performance. The results from finding the mean TweetEval scores shows the XLNet-base-cased model once again coming second to last in terms of performance, only outperforming the SVM model and being the overall lowest-scoring transformer-based model present in the table. There is still a clear notable disparity between the mean TweetEval score for BERTweet and the rest of the models. BERTweet has a higher mean TweetEval score averaged over all tasks conducted in this

experiment by a significant 14.1 points over XLNet-base-cased.

V. Discussion

According to the results, hypotheses 1, 2 and 3 can be rejected as the XLNet-base-cased model failed to achieve a higher score in any of the 3 Twitter classification tasks outlined and conducted in this experiment. Overall, the XLNet-base-cased model had the most success on the sentiment analysis data coming at just 3.4 points under the best performing model. Although the outcome in the other two classification tasks show the XLNet-base-cased model substantially underperforming compared to most of the other models with the model being the second-worst model overall according to the mean TweetEval scores. An explanation for why the XLNet-base-cased model performed the best at the sentiment analysis task is likely due to the model being transformer-based, as comparing the results in Figures 1 and 2 show all the transformer bases models outperform the non-transformer-based models which suggest transformer-based models according to this experiment may indeed be superior in sentiment analysis classification tasks. This trend, however, was not found in the two other classification tasks in which the results were more mixed with the non-transformer based models outperforming the transformer-based models. This trend was most pronounced in the irony detection task where both the base models for XLNet and RoBERTA accrued the two lowest TweetEval scores for the task. A potential explanation for this is the irony detection task is one in which detecting irony is context-specific to Twitter data, which explains why transformer-based models that were pre-trained on Twitter data all performed better than the base models in which no pretraining on any Twitter data. While hate speech detection is the most 'difficult' classification task involved with Twitter data as even the highest-scoring BERTweet model only obtained a TweetEval score of 56.4, the XLNet-base-cased model still scores rather low and underperforms with a substantially lower TweetEval score. All in all the current results suggest and demonstrate the limitations of using the XLNet-base-cased model for Tweet classification tasks.

Although this experiment was insightful on how well an XLNet-base-cased model performs on Twitter classification tasks, due to the limited scope of this experiment it would not be appropriate to make too general conclusions on the classification capabilities of an XLNet transformer model on Twitter datasets. The main reason for this is this model has not been used on all the classification tasks in which the other models were trialled in the original TweetEval paper. Tasks such as offensive speech detection, emoji detection, stance detection and emotion detection were not present in this experiment, therefore the conclusions of this appear are not suited to be generalisable. Though the results of this paper provide a rough benchmark and premise on how this model will likely perform overall compared to the other models for each Twitter classification. No complete conclusion can be made until the experiment is repeated on all the remaining Twitter classification tasks.

While the XLNet-base-cased was used, the transformer model lacks any pre-training or hyperparameter adjustments that could potentially improve the performance of the model. The reasoning behind using an untuned model not specifically pre-trained on any Twitter data is that the primary goal of this experiment is to provide a baseline benchmark value for an XLNet model. With the information gained from this research paper, experimenters may deem whether it is worth pursuing building an XLNet model for Twitter classification tasks or whether it is more worthwhile using another transformer-based model to use in a similar experiment. Another reason for the lack of specific tuning is if a similar experiment is repeated, a researcher could build off the methodology outlined in this experiment to adapt a transformer-based model or tune a specific parameter of the XLNet-base-cased model to observe and have a baseline score of how the model would otherwise perform.

Further investigating the impacts of hyperparameter tuning on the XLNet-base-cased model may provide a better classification model. Hyperparameter tuning was left out of this experiment due to the reasons described in the last paragraph but it is worth noting that they can play a large role in determining the effectiveness of the model. Future experiments should make use of hyperparameters tuning to specifically fit Twitter datasets, or by using GridSearchCV to initialise random hyperparameters to find the optimal values to enhance the XLNet model trialled in this experiment. Another way in which the experimental design could be improved is by incorporating the validation dataset. The validation holdout dataset while available for each classification task was not used in this experiment due to the lack of hyperparameter tuning and specified scope of the investigation. The ideal way to incorporate the validation data would be to use it as a tuning set to tune the hyperparameters of an XLNet model while leaving the testing set as the final holdout to test the model with the most optimal hyperparameters found.

To maximize the potential of the XLNet model's capabilities, it would be beneficial to pre-train the model on large swathes of Twitter data in a similar method on which the BERTweet and relevant RoBERTa models were pre-trained/retrained on. This would set the scene for a very interesting experiment as researchers could determine whether the current bottleneck on the XLNet models classification capabilities on Twitter data is constrained to the lack of context-specific training data for the architecture of the model to work with. It may very well be possible an XLNet model pre-trained on Twitter data can outperform the current top-performing models in the 7 classification tasks, despite the current underwhelming performance of the default XLNet-base-cased model.

VI. Conclusion

To conclude, from the results shown in Twitter classification tasks conducted in this experiment, the XLNet-base-cased

model shows little promise of supplanting the BERTweet model as the most useful model for Twitter classification according to the TweetEval metric. The results lead to the rejection of all 3 of the directional hypotheses laid out for this experiment and while the full scope of the investigation of the XLNet-base-cased model is limited due to only 3 out of the 7 Twitter classification tasks being completed, the current results show the XLNet-base-cased model is insufficient compared to the other available models on the TweetEval leaderboard. future investigations should aim to complete the model benchmark on all remaining classification tasks not attempted in this experiment. Furthermore, it could be prudent to adjust the hyperparameters of the XLNet-base-cased to obtain a more optimal classification model. Another avenue for future experiments is to investigate the effect of pre-training an XLNet model specifically on large amounts of Twitter data, to observe whether that would improve the performance of an XLNet transformer model to match or even potentially surpass the current BERTweet model benchmark for TweetEval Twitter classification tasks.

REFERENCES

[1] Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., amp; Neves, L. (2020). TweetEval: Unified benchmark and comparative evaluation for tweet classification. Findings of the Association for Computational Linguistics: EMNLP 2020. doi:10.18653/v1/2020.findings-emnlp.148

[2] Rosenthal, S., Farra, N., amp; Nakov, P. (2017). Semeval-2017 task 4: Sentiment analysis in twitter. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). doi:10.18653/v1/s17-2088

[3] Bauwelinck, N., Jacobs, G., Hoste, V., amp; Lefever, E. (2019). LT3 at SemEval-2019 TASK 5: Multilingual detection of hate speech against immigrants and women in Twitter (hatEval). Proceedings of the 13th International Workshop on Semantic Evaluation. doi:10.18653/v1/s19-2077

[4] Van Hee, C., Lefever, E., amp; Hoste, V. (2018). SemEval-2018 task 3: Irony detection in English tweets. Proceedings of The 12th International Workshop on Semantic Evaluation. doi:10.18653/v1/s18-1005

[5] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., amp; Le, Q. (2020, January 02). Xlnet: Generalized autoregressive pretraining for language understanding. Retrieved February 25, 2021, from https://arxiv.org/abs/1906.08237

[6] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., . . . Stoyanov, V. (2019, July 26). RoBERTa: A ROBUSTLY OPTIMIZED BERT Pretraining Approach. Retrieved February 25, 2021, from https://arxiv.org/abs/1907.11692v1

[7] Nguyen, D. Q., Vu, T., amp; Tuan Nguyen, A. (2020). BERTweet: A pre-trained language model for English tweets. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. doi:10.18653/v1/2020.emnlp-demos.2