TweetEval:Emotion,Sentiment and offensive classification using pre-trained RoERTa

Usama Naveed
Reg: un21988
Department of Computer Science and Electronic Engineering
University of Essex
Colchester,UK
Usamabasra1122@gmail.com

***Abstract*—Pre-training of the language model has added to significant execution upgrades, yet it is hard to painstakingly think about different methodologies.Training of new model on different data sets and sizes is too expensive and required a lot of computation efforts.This paper present a RoBERTa pretraining replication that is specifically performed on three data sets emotion, sentiment and offensive respectively.In addition to the BERT MLM (masked-language modeling) training technique, the RoBERTa model incorporates some modifications related to training process. I examine that pretrained RoBERTa perform better score then training from scratch.Hugging face is providing RoBERTa based models for each of data sets. This RoBERTa based models are trained on 50M tweets and fine-tuned for tweet classification.In addition,the paper includes an overview and discussion of the processes and the results that is obtained.**

I. Introduction

Self-training models such as XLNET, GPT, ELMO, BERT have created performance gains, but evaluating which aspects of the techniques contribute the most can be difficult [1]. Varsity and the amount of data require a lot of computationally expense, limiting the amount of training data or doing training privately can also effects the performance of the actual model in advance.

Modern NLP(Natural language processing) systems perform inadequate when they are applied to noisy user generated textual data [2]. versatility and variety of conversational data requires the additional tackling challenges.In comparison, other seemingly simple tasks, such as sentiment analysis, have proved to be difficult on Twitter data due to less amount of cues in text and the lack of unified evaluation framework.This is especially important in the modern period of pre-training and Language Models (LMs), as these models exhibit a flexibility that can actually not be calculated comparably across Twitter datasets.In addition to SenEval and NLP performs only standerd task related to tweet data.

TEETEVAL which is considered as the benchmark for the tweet classification for English language is the inspiration for this work [2].TWEETEVAL is used for the seven tweet classification tasks but we choose three of them, These are Emotion,Sentiment and offensive tweet data set. For now we are just performing our task on pretrianed RoBERTa to visualize the data set. I compile and the check what type of dataset we have. we evaluate the RoBERTA pretrained model and try to improve the results.

As follows, the paper is structured. The datasets released to the participants for evaluating the model are listed in the next section then we have some visualization of data. [1]

II. Literature Review

TweetEval is a unified benchmark for Tweet classification which is based on seven datasets such as Emoji [3], Emotion [4], Hate speech [5], Irony [6], Offensive [7], Sentiment [8], Stance [9]. They use three different variations of RoBERTa which is pre-trained RoBERTa(ROB-Bs) and (ROB-Tw) which is retrained on their dataset and (ROB-Tw) which is trained from scratch [2].They trained the model on 60M English tweets.They use the same classification as used in this paper [10] which is multimedia based classification by adding one more dense layer [2].For Emotion Recognition task is

[1]TweetEval:Emotion,Sentiment and offensive classification using pre-trained RoERTa is available at: https://github.com/un21988/CE888-7-project-2.git

based on recognizing the emotions occur by tweet.For Emoji prediction they assign an emoji on the nature of tweet.According to them most of the emojis are labeled with three smiley , heart and smiling face with hearts.Irony detection consists of, that this tweet is irony or not and Hate speed detection consists of whether the tweet is hateful or not and offensive language detection consists of whether the tweet is offensive or not and sentiment analysis consists of that whether the tweet is negative, positive or neutral and stance detecting checks whether the tweet consists of favourable, neutral or negative.

RoBERTa which is abbreviated as robust optimized BERT pre-training approach [1] they BERT model which by adding some dynamically changes. RoBERTa achieve the state of the art results on GLUE, SQUAD and RACE. They also the CNN dataset and pretrain the model and release the code on github [1].We recommend modifications to the BERT pretraining protocol in the previous section that maximize the efficiency of the end mission.Then they aggregate changes and determine their cumulative effects. For the Robustly Configured BERT approach, they call this setup RoBERTa. RoBERTa is primarily trained with dynamic masking, large mini-batches and a wider byte-level BPE [1]. we are going to use pretrained RoBERTa for emotion, sentiment and offensive dataset. After that we compare the results and methodologies which is used in this paper.

III. Methodology

RoBERTa pretrained we is used for this dataset.we did not choose emoji prediction data because the low results in the task of emoji prediction compared to those obtained in the official SemEval task are due to the down-scaling of the training 1648 data. Because of data distribution in TweetEvaL was most 50k tweets per task, whereas in the original competition, by id sharing, the training data was one order of magnitude bigger [5].

Data that is selected for RoBERTa is Sentiment, Emotion and offensive dataset. In sentiment analysis we have mapping dataset which has three labels such as positive, negative and neutral. Test label and test text are used for testing the data set. In test dataset we have almost 8800 tweets [8].In emotion analysis we have mapping dataset which has four labels such as anger, sadness, joy and optimison. Test label and test text are used for testing the data set. In test dataset we have almost 1421 tweets [4].In offensize analysis we have mapping dataset which has three labels such as offensive and non-offensive. Test label and test text are used for testing the data set. In test dataset we have almost 860 tweets [7].BERT is trained on 16 GB English language based data which is WIKIPEDIA's combination [1].

TABLE I
Data Description

Task	*Lab*	*Val*	*Test*
Emotion Det.	4	374	1421
Offensive lg.	2	11916	860
Sent analysis lg.	3	2000	8800

Table 1: Number of labels, val and test set for each datasets

Pre-training approaches have been developed for various training targets, including machine translation of language modeling and masked language modeling. Several recent papers have used a simple finetuning process recipe for each end task and pre-training for a variation of the target of a masked language model.We based on a fixed area for this initial benchmark and for the sake of reproducibility and usability like classification. We agree, however, that other substantial activities can need to be judged accordingly. Thus, we would like to do more activities in the context of other four dataset analysis for future work.

IV. Result

RoBERTa pretained perform very well in every tasks. The fact is that twitter data is not only noisy but the format of the data is also matter. For that we perform pre-processing a function which takes the string which is tweet here, if it founds the username like @anyusername then is it will change it to @user for the sake of fairness in the testing data of if any link is found then it will also remove it and return the string which is ready to pass into the model. For now prepossessing on data is continued. we put a tweet into the model for emotion detection, the text of tweet is:"/Depression is real. Partners w/ depressed people truly don't understand the depth in which they affect us. Add in /anxiety amp; makes it worse" and the output for the tweet is giving in fig1.

Fig. 1. Model output on one instance

```
[['1) sadness 0.966', '2) anger 0.0184', '3) optimism 0.011', '4) joy 0.0046']]
```

Figure is showing the output from the Model after testing on single tweet

V. Discussion

As discussed Tweet classification is become very difficult not only because the size and verity of data but it's fomate which is not specific and less caus in English language.Althoogh the taining on tweet data is also resouce and time consuing.BERT trained from scratch are trained on the validity break with an early stop 1.0e-5 instruction and learning rate. After around 8/9 days on 8 NVIDIA V100 GPUs Twitter Corpus, both versions converged. On 60M tweets obtained through the extraction of a large corpus of English tweets [1].The working of these tweet classification language models is very simple. First, on a large unlabeled corpus, they are trained. Then, if a sufficient training set exists, they are fine-tuned to the mission.However, one can doubt whether the current pretrained models trained on standards that are suitable for social media text. Therefore we studies different models and we found that the pretrained RoBERTa is feasible to use. we extract our data using colab file open() function and then write the file on google drive for later use.Then we analysis the data and try to run the model.

VI. Conclusion

we present RoBERTa pretrained model used on emotion, sentiment and offensive classification.we also evaluate three types of BERT models which is pre-trained RoBERTa(ROB-Bs) ,(ROB-TW) and retrained on their dataset and (ROB-RT) which is trained from scratch [1].After that found that pretrained models are very handy and requir less comutational resources and if we want to train the model on our dataset then it will cost alot and not possible for us.we run RoBERTa for our every dataset and it is giving result very accurately.we additionally use the model and release the code publicly at:https://github.com/un21988/CE888-7-project-2.git

References

[1] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[2] F. Barbieri, J. Camacho-Collados, L. Neves, and L. Espinosa-Anke, "Tweeteval: Unified benchmark and comparative evaluation for tweet classification," *arXiv preprint arXiv:2010.12421*, 2020.

[3] F. Barbieri, J. Camacho-Collados, F. Ronzano, L. Espinosa-Anke, M. Ballesteros, V. Basile, V. Patti, and H. Saggion, "Semeval 2018 task 2: Multilingual emoji prediction," in *Proceedings of The 12th International Workshop on Semantic Evaluation*, 2018, pp. 24–33.

[4] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "Semeval-2018 task 1: Affect in tweets," in *Proceedings of the 12th international workshop on semantic evaluation*, 2018, pp. 1–17.

[5] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, and M. Sanguinetti, "SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter," in *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019, pp. 54–63. [Online]. Available: https://www.aclweb.org/anthology/S19-2007

[6] C. Van Hee, E. Lefever, and V. Hoste, "Semeval-2018 task 3: Irony detection in english tweets," in *Proceedings of The 12th International Workshop on Semantic Evaluation*, 2018, pp. 39–50.

[7] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval)," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019, pp. 75–86.

[8] S. Rosenthal, N. Farra, and P. Nakov, "Semeval-2017 task 4: Sentiment analysis in twitter," in *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, 2017, pp. 502–518.

[9] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, "Semeval-2016 task 6: Detecting stance in tweets," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016, pp. 31–41.

[10] D. Lu, L. Neves, V. Carvalho, N. Zhang, and H. Ji, "Visual attention model for name tagging in multimodal social media," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1990–1999.

Project Plan

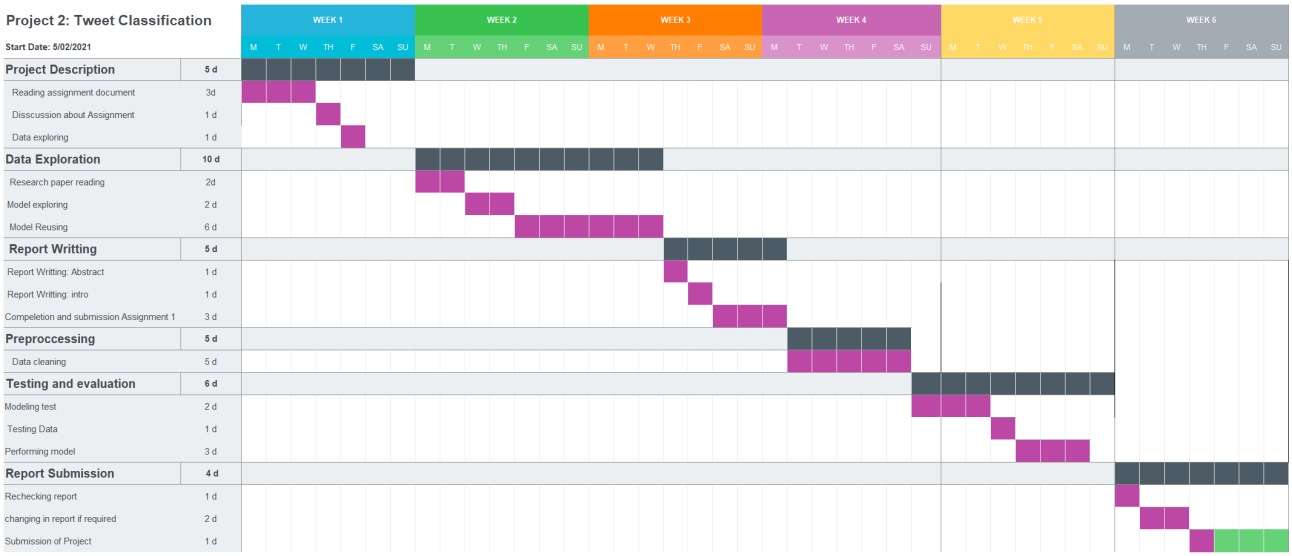

Fig. 2. Gantt Chat of the project