Project 2: TweetEval-based Multi-class tweet classification

January 12, 2021

1 Description

The experimental landscape in natural language processing for social media is too fragmented. Each year, new shared tasks and datasets are proposed, ranging from classics such as sentiment analysis to irony detection or emoji prediction. Therefore, it is unclear what the current state of the art is, as there is no standardized evaluation protocol, neither a strong set of baselines trained on such domain specific data. To tackle this challenge, a few researchers from Cardiff University proposed a new evaluation framework (TweetEval) consisting of seven heterogeneous Twitter-specific classification tasks. They also provided a strong set of baselines as starting point, and compare different language modeling pre-training strategies.

2 Aim

Your job is to perform multi-class tweet classification on any 3 out of 7 datasets available at TweetEval (i.e. emotion, emoji, hate, irony, offensive, sentiment, and stance). Compare your findings with their leader-board available at https://github.com/cardiffnlp/tweeteval. We expect you to achieve better results than TweetEval leader-board on **test set**, if possible (i.e. best case). Please, go through their manuscript https://arxiv.org/pdf/2010.12421.pdf and have in-depth analysis for their GitHub repository [1]. Please use proper evaluation metrics as recommended in section 2.2 of the manuscript. You are free to use use their Google Colab codes with proper acknowledgment in your assignment report. We also recommend to read Transformer-based Language Models such as GPT [1], BERT [2] or XLNET [3] .

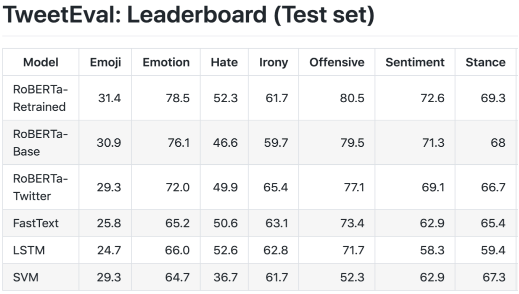
3 Material

A benchmark for tweet classification in English (i.e. TweetEval [4]) is avialable at https://github.com/cardiffnlp/tweeteval. There are the seven datasets of TweetEval, with its corresponding labels.

**Note: Pleas use any 3 out of 7 datasets available at TweetEval to report your performance and compare your performance.

4 Current Results

The current accuracy and other metrics are given as follows:



TweetEval: Leaderboard (Test set)

Model	Emoji	Emotion	Hate	Irony	Offensive	Sentiment	Stance
RoBERTa-Retrained	31.4	78.5	52.3	61.7	80.5	72.6	69.3
RoBERTa-Base	30.9	76.1	46.6	59.7	79.5	71.3	68
RoBERTa-Twitter	29.3	72.0	49.9	65.4	77.1	69.1	66.7
FastText	25.8	65.2	50.6	63.1	73.4	62.9	65.4
LSTM	24.7	66.0	52.6	62.8	71.7	58.3	59.4
SVM	29.3	64.7	36.7	61.7	52.3	62.9	67.3

Figure 1: Test set results for 7 datasets

[1] https://github.com/cardiffnlp/tweeteval

5 Help

- Please keep in mind before starting your project that you should have finished reading their manuscript `https://arxiv.org/pdf/2010.12421.pdf` and go through their GitHub repository `https://github.com/cardiffnlp/tweeteval`.
- To know how to use the pre-trained models, you can check their Google Colab Notebook, with sample code for masked language modeling, extracting embeddings from tweets and tweet classification `https://colab.research.google.com/drive/18cNn4cJ-bAi-Luiqi8V6c09Tj-iiX0oG`.
- Please check their Pre-trained models and code and you can download the best Twitter masked language model (RoBERTa-retrained in the paper) from HuggingFace.
- Discuss with CE888 academic staff and GLAs.
- Use are free to use any NLP tool.

References

[1] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding with unsupervised learning," *Technical report, OpenAI*, 2018.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[3] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and X. Le Q, "generalized autoregressive pretraining for language understanding. arxiv 2019; 1906.08237," 1906.

[4] 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.