Emotion Detection, Twitter sentiment analysis

Thanawat Thanaponpaiboon
220038055
Data Science (Master's Degree)
Guy.Thanaponpaiboon@city.ac.uk

1 Introduction

The nature of humans consists of various emotions such as happiness, sadness, joy, anger etc. Emotional state can change throughout the day. People's emotions are predictable through their voice or their face. However, without that sort of information detecting emotion from only the text seems to be a very challenging task to accomplish.

In this study, we shall apply sophisticated machine-learning approaches to tackle this problem and evaluate whether this cutting-edge approach can predict human emotions precisely or not.

2 Background

Language is not only used to convey information to the receiver but also to express the emotion of the speaker. The growth of sentiment analysis experiments in natural language processing drive people to construct creative study and research. Some of these examples are related and beneficial to our research, Wang and Zhou (2018) determine the intensity of sentiment as a continuous value lie between 0 to 1 which represent the most negative and most positive comment respectively. In addition, showing evidence that increasing the performance of the learning model by ensembling several models together leads to outperforming the baseline model as a result. Correspond to Ge et al. (2019) research. Using ensemble learning techniques to deep learning models such as CNN and LSTM to capture the emotions through the conversational corpus. Another research using twitter dataset, also shows that hashtag is the tag that comes after the tweet in the sentence *significantly increases the performance of the classification model to detect human emotion.* (Mohammad, 2012)

Hopefully, this previous research from expertise merged with knowledge that we gained throughout this course will lead us to meaningful results and prevent us from getting blind or misleading analysis at the end of this research.

3 Proposed methodology

Labelling tag emotions by manually using human look like a tedious task and an endless cycle to do because it is required people's effort to read through all the word in the sentence and decide which sentence expresses which emotion. Applying machine learning approaches to assist humans in the process of analysing emotions is required.

The purpose of this research is to create a model that can detect the emotions given the text data from twitter platform. Includes evaluation and identifies advantages and disadvantages of using technology to leverage a corpus. Along with examples of previous research that we mentioned which provide inspiration and guidelines to achieve the main goal. Below is an initial plan for this research

- **Data selection and pre-processing**: Collect the raw text data including transforming data into the proper format and make it ready to use for the analysis.
- **Construct the machine learning algorithm**: Create machine learning model baseline and improve the model using various techniques such as feature engineering or using different algorithms.
- **Model evaluation and comparison model**: Choose the method to compare the result between the baseline model and the comparison model. This process included writing the report and making a conclusion.

Finally, the main objective of our project is to compare the performance of the model to predict human emotions using various methods in order to increase the model's accuracy. With a reasonable difficulty of the problem and a plentiful approach, this project should be completed around a month

after we start. As a consequence, we expect that our model will be useful for analysing and extracting the emotions of humans based on the given information

3.1 Data

The dataset we will use in this research is available from Kaggle website https://www.kaggle.com/datasets/pashupatigupta/emotion-detection-from-text The data consist of a collection of tweets that have been already annotated with the emotions behind them.

Originally, this dataset is having 13 different emotions with around 40000 records in total. To build a reliable classification model we have to control the label of emotion into 5 main different states. Finally, the dataset we will use in this study has only three columns that are unique tweet identification, raw text messages and label of emotion. The original public data can be obtained from data.world platform under Public License.

3.2 Baselines model in this study

A baseline algorithm that we will apply in this project is a simple linear model (Logistic Regression) for the classification in detect emotion task. Provide consistent results with less effort to set up and the concept behind this model is easy to comprehend which makes it reasonable to use as a baseline for this research.

3.3 Proposed timeline

- Retrieve Data and pre-processing (1 week)
- Modelling and experiment (2 weeks)
- Evaluation performance and write the report (1 week)

This project should not take longer than 1 month to accomplish and submit the research to the lecturer.

4 Experimental setup and tools

We will launch our code in google colab platform. With the additional benefits such as cloud storage capabilities making it much more efficient for our study to train the machine learning model.

Important Library that we aim to use

- Natural Language Toolkit (NTLK) : Useful library for text pre-processing. It contains crucial text processing libraries for tokenization, parsing, classification, stemming etc.
- scikit-learn : A good starting point to train baseline model and apply metrics for evaluation. Provide various classification, regression and clustering algorithms include logistic regression.
- PyTorch : Provides a significant deep-learning tool in python which use for building deep learning models. It is widely used in the field like image recognition and language processing.

Data pre-processing state is required once we obtained the raw text from Kaggle. We will use the tokenisation technique to break down the sentence to the word and extract the stop word from the data. Lemmatisation is also necessary to convert words into base form. After this, we will try to apply hashing and TF-IDF to observe the performance of our model.

We will apply logistic regression as a baseline model. Another approach we consider using in order to compare the performance to our baseline is the deep learning model called Long short-term memory (LSTM). It is commonly used in deep learning applications such as speech recognition, and natural language processing. We expect that our deep learning model will outperform our baseline model at the end of the experiment.

References

Suyu Ge, Tao Qi, Chuhan Wu, and Yongfeng Huang. 2019. THU_NGN at SemEval-2019 task 3: Dialog emotion classification using attentional LSTM-CNN. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 340–344, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Saif Mohammad. 2012. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.

Min Wang and Xiaobing Zhou. 2018. Yuan at SemEval-2018 task 1: Tweets emotion intensity prediction using ensemble recurrent neural network. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 205–209, New Orleans, Louisiana. Association for Computational Linguistics.