Text Mining Project

Isha Pandya (m20210920), David Santos (r20181082), Foazul Islam (m20200750) , Mohamed Shamsudeen (m20210707)

1.Introduction:

Emotions are the feelings that people have in response to certain events or circumstances. Emotion may be demonstrated in a number of different ways, such as facial expression and body language, speech, and by written text [1]. Text document emotion detection is primarily a content-based classification challenge integrating principles from Natural Language Processing (NLP) and Machine Learning.

There are many ways of expressing emotion but the 8 basic emotions of a human are Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise and Trust. These are the emotions that mix and morph into something more profound.

2. Experimental Setup:

The main goal of our project is to read the text in a sentence and recognize the emotion with classifiers and compare their performances and the accuracy of the classifier. Initially, we imported all the necessary python libraries for our analysis, and then we imported the XED Multilingual dataset for sentiment analysis and emotion detection. The dataset consists of English and Finnish movie subtitles from the OPUS corpus.

Following this, we did the preprocessing and cleaning of our data to get more accurate results while applying the models. We applied some machine learning techniques and did the prediction and evaluation of the model after the preprocessing.

3. Preprocessing:

In preprocessing, the first step we do is to change the emotion values from numerical to categorical such as 1: 'Anger', 2: 'Anticipation', 3: 'Disgust', 4: 'Fear', 5: 'Joy', 6: 'Sadness', 7: 'Surprise', and 8: 'Trust' for the easy understanding of the sentence with the emotion. Following that, we simply checked the null values, emotion values, word counts, describing, and frequency counts. Also it is observed that the training dataset is quite balanced.

After observing sentences in the database cleaning process was done that consist of:

- removing all symbols that are not letters.
- make everything lowercase
- remove tags such as; &gt and stop words
- replace words with the corresponding lemma

4. Modeling:

As our dataset is not huge and sentences which are proved are not too long, we decided to start with basic machine learning models.

First step to extract the feature and convert the document into the token using the CountVectorizer. Convert a collection of text documents to a matrix of token counts.

While creating words using CV we did consider the "TF-IDF" method but apparently on the given database it was not giving the good results so we did not include its results into the report but our .py file has results.

model_name	Mean Accuracy	Standard deviation
DecisionTreeClassifier	0.284643	0.001263
KNeighborsClassifier	0.275429	0.007781
LinearSVC	0.332929	0.007351
LogisticRegression	0.343929	0.006967
MLPClassifier	0.295929	0.001966
MultinomialNB	0.329786	0.006501
RandomForestClassifier	0.214357	0.000160
SVC	0.333071	0.010204

Figure-1:Comparison between different model performance.

From this comparison we have considered Logistic regression (LR) is giving a more competitive model against the base model SVC. Below are the results for LR together with its confusion matrix.

	precision	recall	f1-score	support
Anger	0.34	0.56	0.42	211
Anticipation	0.39	0.41	0.40	170
Disgust	0.24	0.14	0.18	77
Fear	0.36	0.25	0.29	104
Joy	0.39	0.39	0.39	97
Sadness	0.26	0.20	0.22	87
Surprise	0.32	0.23	0.27	96
Trust	0.38	0.32	0.35	158
accuracy			0.35	1000
macro avg	0.34	0.31	0.32	1000
weighted avg	0.35	0.35	0.34	1000

Figure-2: Classification report for LR model with Dev set.

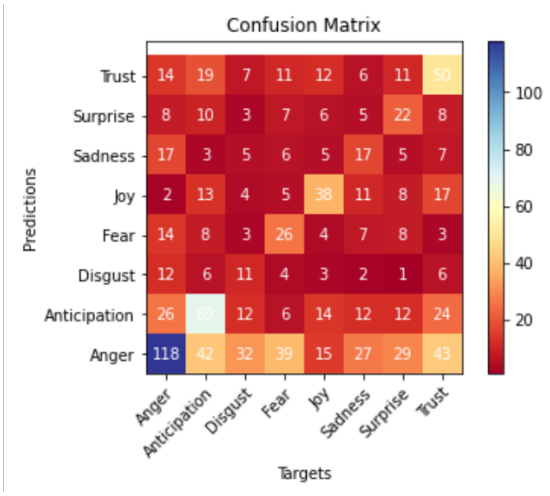

Figure 3: Confusion Matrix for LR in the dev set

5. Conclusion:

Emotion Detection, also known as Sentiment Analysis, is a critical area of study in the science of human-computer interaction. A considerable quantity of effort has been put into this project to detect emotions. As a result, the TFIDF of LR is the best performance model, and the second best is the SVC model. Against expectations, SVC is worse than LR. It is because there are many meanings in one word or sentence, not just one meaning.

Machine learning is based on training a model and then using that model to predict. The amount of data has a strong influence on the accuracy. The methods used to detect emotions from text are presented along with their limitations, which would make them perform efficiently.

6. Future Work:

For future work, more focus on preprocessing should be considered in order to avoid mistakes that result in the inclusion of incomplete words or sentences. Also, maybe some deeper exploration of the models' parameters can produce better results, as well as a bigger study and understanding of machine learning algorithms.

References

[1] https://airccj.org/CSCP/vol2/csit2237.pdf
[2]