1. Project Title and Team Members

<https://github.com/dldowning/Fall2022-5222>

**Analysis of Impact on F1 Score of Permutations of**

**RegEx Expressions of Tweets for Sentiment Analysis**

David L Downing

Justin Roberts

5222 Feature Engineering

Dr Sayed Shah

Fall 2022

University of North Texas

2. Idea description

Using natural language processing skills, features will be extracted from tweets using regular expressions. The corpus is a collection of hundreds of tweets mined from Twitter usage. The training set has labels for positive or negative sentiment of the tweet that has been done by hand. Using regular expressions, we will extract a diverse array of quantitative features. Examples might be the mean word length or the ratio of alpha characters to punctuation characters. Using wrapper methods, we would generate different lists of independent variables to train a binary classifier through logistic regression. The F1 score of each combination of features will be compared to each other to review the impact of the features on the F1 score.

3. Motivation

Due to a limited experience with NLP, we seek to explore the space more fully. The first part of this course covered NLP functions and methods so we will apply them on a larger dataset to assess their efficacy. We got this dataset to practice referencing a pre-trained lexicon but it’s an interesting example of natural language processes. Regular expressions still show up in a variety of sources so this is an attempt to practice using regular expressions on the tweets to extract the information we’re looking for. What is interesting about tweets is that although they are of limited characters they are highly unstructured and varied. This will mean that data cleaning and checking for our regex will be critical to ensure we are extracting the information we believe we are extracting.

4. Significance

The size of the corpus and the amount of text data generated on Twitter is too large to be manually processed by humans. The ability to automate this analysis would be valuable to save time but also to improve accuracy. The interrater reliability of qualitative assessments is often suspect. We are relying on the experts who hand labeled our training set in order to have good labels and then evaluating our predictions for usability. Having well written regular expressions that are flexible enough to be used on any number of tweets that a user could submit allows future pipelining of data to continue to be input into the model. The outputs of pipelined tweets could be viewed as a time series to see how sentiment is changing over time. If each pipeline is restricted to certain hashtags, then you could review how the sentiment for different topics or products is changing with time. This would be a useful and usable datapoint for data based decisions for executives to be recommended by the data science team.

5. Goals and Objectives:

The goal is to take a combination of research and creativity in order to generate new features. The goal will be to create a high quantity of features as well as a diverse range of features. We want to have enough features to be able to have varying groups of them but have a diverse range of features of different types we engineered in order to analyze the quality of the features and usefulness. The objective will be to understand how the features we engineered will impact the F1 score and come to some understanding of why some features have greater power than others.

6. Features

Here are some preliminary features to be extracted:

1. Pre-trained values from existing lexicons as has been used in the literature
2. Length of the tweet
3. Length of the longest word in the tweet
4. Mean length of the words in the tweet
5. Ratio of longest word to the mean length of words
6. Ratio of quantity of punctuation to full length of tweet
7. Ratio of capitalization to non-capitalization
8. Presence of emojis
9. Ratio of number of words in the 500 most common words lexicon

vs number of words not

1. Ratio of words in the dictionary compared to misspelled words / non-dictionary words

The features will be considered with and without scaling or normalization.

The features will also be considered with stopwords vs without stopwords.

7. Related Work (Background) Literature Survey

The current literature has a wide variety of tools to use from BERT, deBERTA, Transformers, convolutional neural networks. The goal of our project is not to find the best model but to evaluate the features. For this, we will use a multiple logistic regression model to evaluate which features are better than which other features.

Many of the papers focus on transfer learning from other pre-trained lexicons to bring in scores that can be used as features. While we intend to investigate the usage of these features, we also want to compare the engineering of new features through regular expressions.

<https://aclanthology.org/2020.findings-emnlp.148/>

<https://dl.acm.org/doi/abs/10.1145/2938640>

<https://ojs.aaai.org/index.php/ICWSM/article/view/14185>

<https://www.hpl.hp.com/techreports/2011/HPL-2011-89.pdf>

8. Dataset

This is a public tweet dataset. It has 30000+ tweets with associated labels of positive, negative, or neutral. Graphical user interface, text, application, email

Description automatically generated

9. Detail Design of Features

We have engineered almost 20 features so far, and we will have more in increment 2. The first ten features are comparing the tokens we pulled from the tweet to lexicons we pulled from subreddits where each word has a positive or negative score. We formed features by sampling 9 times from the tweet and summing the score. Other features come from regular expressions or comparing to lists. For example, we have the logarithm of the tweet length. We have the logarithm of the longest word length. We have the logarithm of the number of words of 5+ length. We have the ratio of unique words to total words. We have the ratio of stop words to total words. We have the number of times it says ‘no’. We have a number of features of this style that we will evaluate for how well they perform.

Text, application

Description automatically generated

10. Analysis

Graphical user interface, application

Description automatically generated

We built upon a logistic regression implementation from scratch from prior work and used those features to predict the label.

We achieved an F1 score that was considered high compared to previous existing work. This is using all of the features currently available. In increment 2, we want to test using different permutations of the features to see if removing poor features can increase the F1 score. We want to build a table that will tell us how much the F1 score will change if we Leave\_One\_Feature\_Out.

1. Implementation

We implemented a logistic regression. We’re using functions to do feature extraction so that once they are written we can apply the function to our data to quickly extract.

Graphical user interface, text, application

Description automatically generated

1. Preliminary Results

A screenshot of a computer

Description automatically generated with medium confidence

We are mainly looking at F1 score for our metric of choice. So when we evaluate future features, we will be looking at their change in the F1 score. F1 score change will be our ranking procedure for the quality of a feature.

13.. Objectives & Expected Outcome

The objective will be to have a cleaned dataset, tested functions that will generate a wide array of features, lists of features to be used as independent variables, and a visualization of which lists performed higher or lower on the F1 score analysis. With that information, it will give us the details we need to compare/contrast the features that were created and extrapolate practices to be used for future projects based on the types of features that were most successful.

14. Project management:

- Found datasets and a bassline comparison metrics for our model

- Created unique features relevant to the outcome of our prediction

- Implemented logistic regression by hand and through SkLearn

- Gathered F1, precision and recall scores to verify the accuracy of our findings

- Organized results

Regarding project management, our team has been focused making consistent progress towards our end goal. We began by thoroughly analyzing and our sourced datasets and verifying that they would be suitable for the analysis we want to conduct. It was also important that we had comparable metrics from other analysis that have a similar desired outcome. We created features that we believed relevant. After deciding on what data to use, we created relevant features for our predictions. Following that, we gathered accuracy metrics and organized all results.

Work Completed:

|  |  |  |
| --- | --- | --- |
| Testing Design | Outline of code blocks | David |
| Cleaning and File imports | Getting everything into dataframes | Justin |
| Feature Extraction | Writing functions to get features | 40% David / 60% Justin |
| LogReg & Metrics | Finding F1 | David |
| Writing Proposal |  | David |
| Writing Increment 1 | updates | Justin |

Work To Be Completed:

|  |  |  |
| --- | --- | --- |
| More features | Increasing our feature diversity | Justin |
| Best\_Feature\_Maker | Forward selection to determine best set of features | David |
| Try\_add\_one | Function to determine an F1 score if we use all features except one, iterate through all | David |
| Analysis Table | Generate a table to evaluate features by how much their F1 score changes | Justin |

15. References

1. [TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification](https://aclanthology.org/2020.findings-emnlp.148) (Barbieri et al., Findings 2020)
2. Anastasia Giachanou and Fabio Crestani. 2016. Like It or Not: A Survey of Twitter Sentiment Analysis Methods. ACM Comput. Surv. 49, 2, Article 28 (June 2017), 41 pages. <https://doi.org/10.1145/2938640>
3. Kouloumpis, E., Wilson, T., & Moore, J. (2021). Twitter Sentiment Analysis: The Good the Bad and the OMG!. Proceedings of the International AAAI Conference on Web and Social Media, 5(1), 538-541. Retrieved from <https://ojs.aaai.org/index.php/ICWSM/article/view/14185>
4. Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2011). Combining lexicon-based and learning-based methods for Twitter sentiment analysis. *HP Laboratories, Technical Report HPL-2011*, *89*, 1-8.