**Sentiment Analysis of Microblog Streams**

# Executive Summary

Sentiment analysis of Twitter microblog streams was explored using the VADER (for Valence Aware Dictionary for sEntiment Reasoning) model. The basic classifier based on valance score is enhanced using emoji and social features such as retweets and favourites. An early fusion approach was taken for the emoji handling, where emoji were converted to the descriptions and subject to the same handling as the text to get a compound score based on the VADER lexicon. A late fusion approach was taken for the social features, where the compound score of text is modified based on the product of an optimised weight and the number of retweets/favourites. Both features improve the f1 score of the classification. The optimal model using both emoji and social features has an f1 score of 0.734648.

# Introduction

Sentiment analysis the process of computationally identifying and categorizing opinions to determine if the user's attitude towards certain organizations, events or products is positive, negative, or neutral. The analysis of microblog data can help organisations understand the market and consumer sentiments better so as to make better business decisions.

# Methodology

## Program Structure

The programme consists of the following:

1. Pre-processing of text data
2. Handling of cases including emoji
3. Classification to get a Valance score
4. Enhancement with other features such as social features
5. Testing and optimisation based on precision, recall, and f1 scores

These will be further elaborated in the sections below.

## Data set

The data consist of a set of about 5000 tweets, crawled from Twitter and manually annotated with 3 labels, positive, negative, and neutral (ground truth). The data is split into split into training, development, and testing set for further processing.

## Pre-processing

The pre-processing section consists of the following steps:

1. Tweets in json format were read.
2. Text was extracted from tweet object. Text is the actual text of the status update. This provides the content of the microblog.
3. Tweets were treated with a pre-processing function, which does the following in the order given below:
   1. Convert to lowercase
   2. Remove URLs
   3. Remove Mentions
   4. Remove repeat characters
   5. Remove hashtag symbol
   6. Remove time
   7. Convert emoji into words
4. Treated data for each extracted attribute are saved into another text (.txt) file for further processing.

In the data cleaning and pre-processing, emoji are converted to the description of the emoji. The description is then treated with the rest of the text as part of the VADER analysis. An example of conversion is  which is converted to “grinning face”. As grinning is a positive word in the sentiment lexicon, this impact the classification of the tweet. The treatment of emoji in the pre-processing step is a case of early fusion, where the additional feature is added before the classifier.

## Handling of cases

From the treated data from pre-processing, different cases which affect the polarity and intensity of the sentiments are handled, including:

* negations
* punctuation emphasis & punctuation flooding
* word-shape as emphasis (capitalization difference)
* degree modifiers (intensifiers such as 'very' and dampeners such as
* 'kind of')
* slang words as modifiers such as 'uber' or 'friggin' or 'kinda'
* contrastive conjunction 'but' indicating a shift in sentiment;
* sentiment of later text is dominant
* use of contractions as negations
* sentiment laden emoticons such as :) and :D
* sentiment laden slang words (e.g., 'sux')
* sentiment laden initialisms and acronyms (for example: 'lol')

## Classification to obtain Valance score

The processed text are given valence score based on the VADER lexicon, where the words with stronger emotions are given a score with larger magnitude and positive sentiments are positive while negative sentiments are negative. VADER produces 4 sentiment metric s from the rating of the words, (1) the percentage of the text that is positive, (2) neutral, (3) negative and (4) the compound score where all of the lexicon ratings are summed.

Each tweet is then sorted based on the compound score, where tweets with compound score greater than 0.5 in positive, between -0.5 to 0.5 in neutral, and less than -0.5 in negative.

## Enhancement with other features such as social features

Other social features like retweet count and favourite count are extracted from the json file. Basing on the assumption that a more favorited and retweeted post is likely to have a stronger sentiment, the social feature was added as a modifier to the compound score, before sorting. The weight of retweets and favourites modifier to the compound score is optimised based on 10-fold cross validation. The formula for the modifier is given below:

## Testing procedure

The sentiment analysis model is tested based on Precision, Recall, and f1 score, where

As f1 score takes into account both precision and recall, it is the value that is optimised in the various models.

# Results and Discussion

The addition of emoji handling as a feature into the basic classifier leads to an improvement in precision, recall and f1 score. Refer to table 1 below for details. Overall, the addition of the emoji handling improves the f1 score of the classification to **0.726069** from the baseline of 0.719318.

The addition of social features like favourites and retweets improves the precision, recall and f1 score as well, with f1 score improving to **0.734648**, making the model with Text + Emoji + Social features the best model. The optimal weight of favourites and retweets are 0.45 and 0.1 respectively.

Table 1: Comparison of emoji handling and social features on precision, recall and f1 score

|  |  |  |  |
| --- | --- | --- | --- |
| Features Used | Precision | Recall | F1 score |
| Text (baseline) | 0.711225 | 0.746108 | 0.719318 |
| Text + Emoji | 0.717193 | 0.752228 | 0.726069 |
| Text + Emoji + Social features | **0.724998** | **0.760339** | **0.734648** |

# Discussion

## Emoji

Emoji are pictograms that function like emoticons, and have high sentiment value in microblog sentiment analysis. The emoji  at the end of a negative message is a form of negation and an enhancement at the end of a positive sentence. Emoji are handled by conversion into the description of the emoji, with examples listed in table 1 below. This method of handling emoji takes into account both the polarity and intensity of the sentiment. As the emoji is more positive, the description of the emoji also reflects accordingly, for example, beaming face with smiling eyes , has 2 positive words in the lexicon (beaming and smiling), and can be interpreted as more positive compared to the grinning face with only 1 positive word (grinning). Likewise, a negative emoji like crying will add a negative word in the lexicon.

Table 2: Examples of emoji and their description

|  |  |
| --- | --- |
| **Emoji** | **Description** |
|  | grinning face |
|  | beaming face with smiling eyes |
|  | rolling on the floor laughing |
|  | grinning face with smiling eyes |
|  | winking face |

## Social features

Social features such as retweets and favourites can indicate a stronger emotional inclination to the existing polarity, i.e. a highly favorited neutral post with a slight negative slant may be classified as a neutral post, when it is a negative on the ground truth. The addition of the social feature optimised by 10-fold cross validation shows an improvement to the f1 score, indicating that favourites and retweet counts do contain some value in sentiment analysis.

Other methods making use of social features like training of machine learning model and late fusion to the percentage polarity of the text shows poorer performance than just the text itself, indicating that using social features for prediction introduces too much noise into the classification. One possible reason is that retweet and favourite data is sparse, with most of the observations having 0 for both retweet and favourites, leading to high probability of overfitting. In addition, a high number of favourites and retweet can indicate strong sentiment for both positive and negative and in that sense, the numbers alone does not indicate whether the tweet is positive or negative.

# Conclusion

In conclusion, the VADER sentiment analysis can be improved by adding emoji handling and social features such as retweet and favourites as a modifier to the compound score resulting from the VADER analysis. The optimal model using both emoji and social features has an f1 score of 0.734648.