Sentimental Analysis on Tweets

1 Introduction

The purpose of this project is to analyze a number of tweets and predict their sentiment. Sentiment can be positive, neutral, negative.
For example a tweet like this : ”I am so happy today!!!”, is, with high possibility, a positive tweet.
In order to predict the sentiment of each tweet we use data mining machine learning techniques such as preprocessing on data, vectorization, classification, etc.

2 Methodology

In this project we used some tweets to train our model. These tweets were taken from International Workshop on Semantic Evaluation (2017).
Using these train sets, we managed to train our model to guess random tweets that it has never ”seen” with 66% (2/3) accuracy.
To achieve this, we separated the project in 3 parts.

- **Preprocessing** : We removed symbols like hashtags, links, emoticons, emojis, stopwords etc. We tokenized the sentences in words in order to perform vectorization.

- **Vectorizing Preprocessed features** : Features in machine learning is basically numerical attributes from which anyone can perform some mathematical operation such as matrix factorisation, dot product etc.
  But these features are in the form of string so first we need to convert these string features into numerical features. To convert string data into numerical data one can use following methods :

  - **Bag Of Words**
  - **Tf-idf**
  - **Word embeddings** : By using word embeddings we take a matrix with 300 values (features). We managed to add more values to this matrix using some dictionaries that we were given.
    These dictionaries contain many words and each one of them has a real value that belongs to [-1, 1] which represents a sentimental analysis.
    In order to expand our word embedding matrix we increased the number of features by including the length of a tweet, maximum and minimum real value of each word in a tweet, using the dictionaries mentioned above.

- **Classfication** : Classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.
  We classificated our data using 3 different classifiers.

  - **SVM classifier**
  - **KNN classifier**
  - **Round Robin Classification** : Round Robin Classification is a technique that splits the problem in binary problems and then uses KNN classifier for each pair of classes.

3 Results

We used different kinds of graphs, wordclous, etc to present our results. All of them can be found in our jupyter notebook file.

4 Discussion and Conclusions

In this project we managed to grasp some really important techniques in data mining and machine learning.
Moreover, we familiriazed ourselves with new python libraries like :

- matplotlib
- pandas
- nltk
- pickle
- sklearn