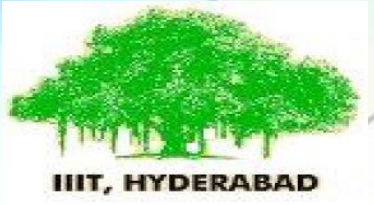


Sentiment Analysis in Twitter

Source Code: [http://github.com/mayank93/Twitter-Sentiment-Analysis]
Demo: [http://10.2.4.49:8000/TSAA/]

Contributed by:
Ayushi Dalmia (ayushi.Dalmia@research.iiit.ac.in)
Mayank Gupta(mayank.g@student.iiit.ac.in)
Arpit Kumar Jaiswal(arpitkumar.jaiswal@students.iiit.ac.in)
Chinthala Tharun Reddy(tharun.chinthala@students.iiit.ac.in)

Course: Information Retrieval and Extraction, IIIT Hyderabad
Instructor: Dr. Vasudeva Varma

What is Sentiment Analysis?

- It is classification of the polarity of a given text in the document, sentence or phrase
- The goal is to determine whether the expressed opinion in the text is positive, negative or neutral.

Negative

Praval Singh @Praval · 8m

Young techies leaving Infosys in droves | Attrition rate of 18.7% - bit.ly/1kwei68

Expand Reply Retweet Favorite Buffer Pocket More

David Pierce @piercedavid · Apr 14

The **Galaxy S5** is a very good (and very waterproof) smartphone that left me wanting more theverge.com/2014/4/14/5608… pic.twitter.com/x5SYQ1pcZe

View photo Reply Retweet Favorite Buffer Pocket More

Positive

Neutral

NDTV Gadgets @NDTVGadgets · 13h

Twitter buys social data provider Gnip ndtv.in/1hI5j1Y

View summary Reply Retweet Favorite Buffer Pocket More

Why is Sentiment Analysis Important?

- Microblogging has become popular communication tool
- Opinion of the mass is important
  - Political party may want to know whether people support their program or not.
  - Before investing into a company, one can leverage the sentiment of the people for the company to find out where it stands.
  - A company might want find out the reviews of its products

Using Twitter for Sentiment Analysis

- Popular microblogging site
- Short Text Messages of 140 characters
- 240+ million active users
- 500 million tweets are generated everyday
- Twitter audience varies from common man to celebrities
- Users often discuss current affairs and share personal views on various subjects
- Tweets are small in length and hence unambiguous

Problem Statement

The problem at hand consists of two subtasks:

- Phrase Level Sentiment Analysis in Twitter :

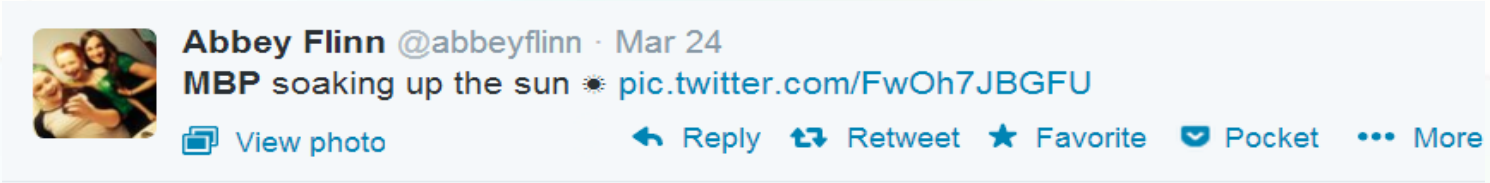
  Given a message containing a marked instance of a word or a phrase, determine whether that instance is positive, negative or neutral in that context.

- Sentence Level Sentiment Analysis in Twitter:

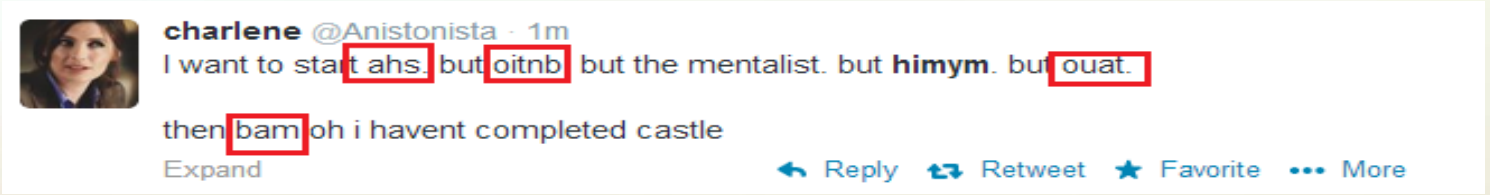
  Given a message, decide whether the message is of positive, negative, or neutral sentiment. For messages conveying both a positive and negative sentiment, whichever is the stronger sentiment should be chosen.

Challenges

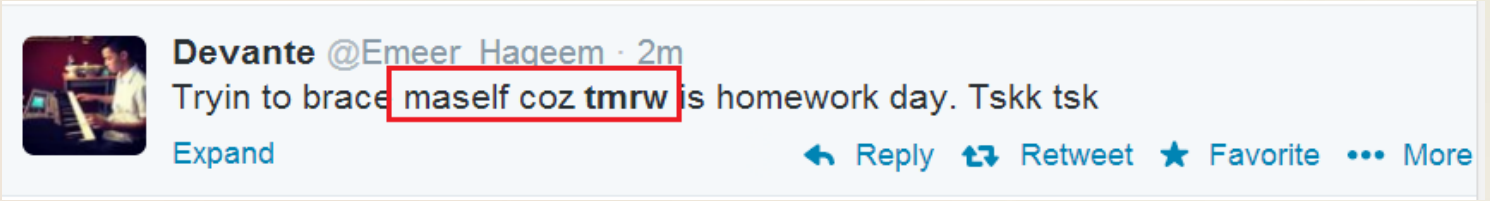
- Tweets are highly unstructured and also non-grammatical



- Out of Vocabulary Words

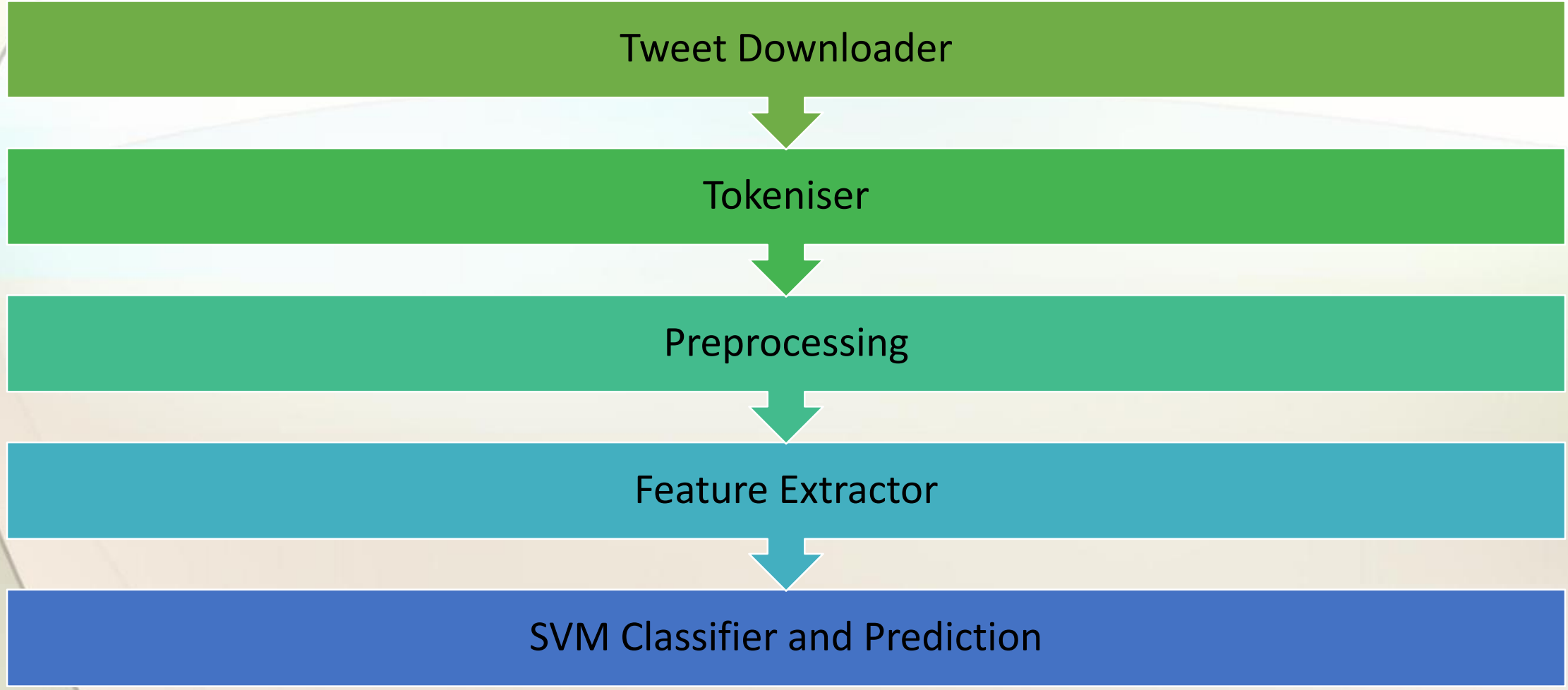


- Lexical Variation



- Extensive usage of acronyms like *asap, lol, afaik*

Approach


Tweet Downloader
Tokeniser
Preprocessing
Feature Extractor
SVM Classifier and Prediction

Approach

- Tweet Downloader
  - Download the tweets using Twitter API
- Tokenisation
  - Twitter specific POS Tagger developed by ARK Social Media Search
- Preprocessing
  - Removing non-English Tweets
  - Replacing Emoticons by their polarity
  - Remove URL, Target Mentions, Hashtags, Numbers.
  - Replace Negative Mentions
  - Replace Sequence of Repeated Characters eg. ‘cooooooooool’ by ‘coool’
  - Remove Nouns and Prepositions

Approach

- Feature Extractor
  - Polarity Score of the Tweet
  - Percentage of Capitalised Words
  - Number of Positive/Negative Capitalised Words
  - Number of Positive/Negative Hashtags
  - Number of Positive/Negative/Extremely Positive/Extremely Negative Emoticons
  - Number of Negation
  - Positive/Negative special POS Tags Polarity Score
  - Number of special characters : ?,!,*
  - Number of special POS
- Classifier and Prediction
  - The features extracted are next passed on to SVM classifier.
  - The model built is used to predict the sentiment of the new tweets.

Results

A baseline model by taking the unigrams, bigrams and trigrams and compare it with the feature based model for both the sub-tasks

Sub-Task	Baseline Model	Feature Based Model	Baseline + Feature Based Model
Phrase Based	62.24 %	77.33%	79.90%
Sentence Based	52.54%	57.57%	58.36%

Accuracy

Sub-Task	Baseline Model	Feature Based Model	Baseline + Feature Based Model
Phrase Based	76.27*	75.23	75.98
Sentence Based	55.70	59.86	60.55

F1 Score

*Classifies in positive classes only, hence high recall.

Conclusion

- We investigated two kinds of models: Baseline and Feature Based Models and demonstrate that combination of both these models perform the best.

- For our feature-based approach, feature analysis reveals that the most important features are those that combine the prior polarity of words and their parts-of-speech tags.