# Sentiment Analysis of Twitter Data

Yagyansh S. Kumar, Daivat Vaishanani, Prabhav Pandya, Jiten Sidhwani
The LNMIIT

## I. The Problem Statement

Microblogging websites have evolved to become a source of varied kind of information. This is due to nature of microblogs on which people post real time messages about their opinions on a variety of topics, discuss current issues, complain, and express positive sentiment for products they use in daily life. Infact, companies manufacturing such products have started to poll these microblogs to get a sense of general sentiment for their product. Many times these companies study user reactions and reply to users on microblogs. One challenge is to build technology to detect and summarize an overall sentiment.

Using this social media we built models for classifying ”tweets” into positive, negative and neutral classes. We build models for two classification tasks:

1. **A 3-way classification of already demarcated phrases** in a tweet into positive, negative and neutral classes.
2. **A 3-way classification of entire message(i.e the whole tweet)** into positive, negative and neutral classes.

We experiment with the baseline model and feature based model.
We use manually annotated Twitter data for our experiments. In this work we use three external resources:
1) A hand annotated dictionary for emoticons that maps emoticons to their polarity.
2) An acronym dictionary collected from the web with English translations of over 5000 frequently used acronyms.
3) A lexicon which provides a prior score between -5 and +5 for commonly used English words and phrases.

Table I. EMOTICON DICTIONARY

| Emoticon | Polarity |
|---|---|
| $:-):):o):]:3$ | Positive |
| $:-D:D8DxDXD$ | Extremely Positive |
| $:-/:/=/=</3$ | Negative |
| $D:D8D=DXv.vDx$ | Extremely Negative |
| $>:)B)B-):):-)>$ | Neutral |

Table III. AFINN-111 DICTIONARY

| Word | Score |
|---|---|
| adore | 3 |
| aggressive | -2 |
| bitch | -5 |
| breathtaking | 5 |
| celebrate | 3 |

Table II. ACRONYM DICTIONARY

| Acronym | Expansion |
|---|---|
| admin | administrator |
| afaik | as far as I know |
| omg | oh my god |
| rol | rolling over laughing |
| wip | work in progress |

## II. The Dataset

In this project we have used the dataset provided in **SemEval(Task 9)**. The dataset consists of tweet IDs which are annotated with positive , negative and neutral labels. For sentiment analysis at the phrase level, the dataset contains 15022 phrases from different tweets. Since, some of the tweets were not available while downloading, we are left with 11880 phrase.

For the second sub task, which is analysing the sentiment of the entire tweet we have 13497 tweets IDs. As mentioned before some of the tweets were not available while downloading. This leaves us with 12640 tweets.

## III. Challenges to Begin with

1. Tweets are highly unstructured and also non-grammatical.
2. Out of Vocabulary Words.
3. Lexical Variation.
4. Excessive usage of acronyms like asap, lol etc.

## IV. Preprocessing & Approach

### A. Tokenization

After downloading the tweets using the tweet IDs provided in the dataset, we first tokenize the tweets. This is done using the Tweet-NLP developed by ARK Social Media Search. This tool tokenises the tweet and returns the POS tags of the tweet along with the confidence score.

### B. Remove Non-English Tweets

Twitter allows more than 60 languages. However, this work currently focuses on English tweets only. We remove the tweets which are non-English in nature.

### C. Replacing Emoticons

Emoticons play an important role in determining the sentiment of the tweet. Hence we

replace the emoticons by their sentiment polarity by looking up in the Emoticon Dictionary.

**D. Remove Url**

The URLs which are present in the tweet did not carry much information regarding the sentiment of the tweet. Thus these are removed.

**E. Remove Target**

The target mentions in a tweet done using '@' are usually the twitter handle of people or organisation. This information is also not needed to determine the sentiment of the tweet.
Hence they are removed.

**F. Replace Negative Mentions**

Tweets consists of various notions of negation. In general, words ending with 'nt' are appended with a not. Before we remove the stopwords 'not' is replaced by the word 'negation'. Negation play a very important role in determining the sentiment of the tweet which is is discussed later in detail.

**G. Hashtags**

Hashtags are basically summariser of the tweet and hence are very critical. In order to capture the relevant information from hashtags, all special characters and punctuations are removed before using it as a feature.

**H. Sequence of Repeated Characters**

Twitter provides a platform for users to express their opinion in an informal way. Tweets are written in random form, without any focus given to correct structure and spelling. Spell correction is an important part in sentiment analysis of user-generated content. People use words like 'coooool' and 'hunnnnngry' in order to emphasise the emotion. In order to capture such expressions, we replace the sequence of more than three similar characters by three characters. For example, wooooow is replaced by wooow. We replace by three characters so as to distinguish words like 'cool' and 'cooooool'.

**I. Numbers**

Numbers are of no use when measuring sentiment. Thus, numbers which are obtained as tokenized unit from the tokenizer are removed in order to refine the tweet content.

**J. Nouns and Prepositions**

Given a tweet token, we identify the word as a Noun word by looking at its part of speech tag given by the tokenizer. Noun words don't carry sentiment and thus are of no use in our experiments. The same goes for prepositions too.

**K. Stop-word Removal**

Stop words don't carry any sentiment information and thus are of no use to us. We create a list of stop words like he, she, at, on, a, the, etc. and ignore them while scoring the sentiment.

## Models used -

**A. Baseline Model**

In the baseline approach, we first clean the tweets. We perform the preprocessing steps and learn the positive, negative and neutral frequencies of unigrams, bigrams and trigrams in training.

Next we create a feature vector of tokens which can distinguish the sentiment of the tweet with high confidence. For example, presence of tokens like am happy!, love love , bullsh*t ! helps in determining that the tweet carries positive, negative or neutral sentiment with high confidence. Also, before calculating the probability values, we filter out those tokens which are infrequent (appear in less than 10 tweets).It can be observed that the presence of such tokens guarantees the sentiment of the tweet with a high confidence.

Table IV. UNIGRAM BIGRAM AND TRIGRAMS

| Unigram | Bigram | Trigram |
|---|---|---|
| bad | negation wait | can negation wait |
| really | laugh loud | anywhere anytime great |
| win | looking forward | negation wait ! |
| shit | goodnite luck | wait eyes stye |
| laugh | negation miss | |
| fun | cant wait | |
| loud | love ! | |
| thanks | goodnite morning | |
| fuck | | |
| looking | | |
| forward | | |
| amazing | | |

**B. Feature Based Model**

As in the previous model, in this model too we perform the same set of preprocessing techniques.

**1) Prior Polarity Scoring**: A number of our features are based on prior polarity of words. For obtaining the prior polarity of words, we use AFINN dictionary and extend it

using senti-Wordnet. We first look up tokens in the tweet in the AFINN lexicon. This dictionary of abo
ut 2490 English language words assigns every word a pleasantness score between -5 (Negative) and +5 (Positive). We first normalize the scores by diving each score by the scale (which is equal to 5). If a word is not directly found in the dictionary we retrieve all synonyms from Wordnet. We then look for each of the synonyms in AFINN. If any synonym is found in AFINN, we assign the original word the same pleasantness score as its synonym. If none of the synonyms is present in AFINN, we perform a second level look up in the senti-Wordnet dictionary. If the word is present in senti-Wordnet, we assign the score retrieved from senti-Wordnet (between -1 and +1).

**2) Features**: We calculate these features for the whole tweet in case of message based sentiment analysis and for the extended phrase (obtained by taking 2 tokens on either sides of the demarcated phrase) in case of phrase based sentiment analysis. We refer to these features as Emotion-features throughout the paper.

| Feature Description |
|---|
| Polarity Score of the Tweet |
| Percentage of Capitalised Words |
| # of Positive Capitalised Words |
| # of Negative Capitalised Words |
| Presence of Capitalised Words |
| # of Positive Hashtags |
| # of Negative Hashtags |
| # of Positive Emoticons |
| # of Extremely Positive Emoticons |
| # of Negative Emoticons |
| # of Extremely Negative Emoticons |
| # of Negation |
| Positive POS Tags Score |
| Negative POS Tags Score |
| Total POS Tags Score |
| # of special characters like ? ! and * |
| # of POS |

## V. Issues Faced

For phrase level sentiment analysis the major challenge was to identify the sentiment of the tweet pertaining to the context of the tweet. We know that tokens can represent different aspects in different contexts. In order to capture this sentiment, we extend the phrase on either side by size two. That is given a phrase and the tweet in which it belongs, we extract the phrase which includes tokens on either side of the phrase. We believe that this helps in taking into consideration the context of the tweet. But after experimentation it was found that the accuracy of the system dropped for both the

models. It was less than what we achieve by taking the phrase only. Therefore we only use the phrases as demarcated to predict the sentiment.

For message level sentiment analysis the most difficult part was to resolve ambiguity. A message can contain both positive and negative sentiments and hence it is difficult to determine the stronger sentiment in the tweet. As a result the highest accuracy achieved is also not at par with the phrase based sentiment analysis.

# VI. Results

## A. Phrase Level Analysis

**1) Baseline Model**: For the baseline model, we consider the unigrams and bigrams. Taking the trigrams leads to drop in accuracy. The accuracy achieved for the baseline model is 62.24%.

**2) Feature Based Model**: For the feature based model we used the features as listed above. The model is trained using the features. We create feature vectors for the test samples and feed it to the model. The accuracy achieved using all the features is 77.86%.

## B. Sentence Level Analysis

**1) Baseline Model**: For the baseline model, we consider the unigrams and bigrams and trigrams. The accuracy achieved for the baseline model is 51.81%.This is 18% more than the chance probability.

**2) Feature Based Model**: For the feature based model we used the features as listed above. The model is trained using the features. We create feature vectors for the test samples and feed it to the model. The accuracy achieved using all the features is 57.43%.

# VII. Error Corrections

There are some incorrect predictions due to quite ambiguous nature of the phrases and messages. Also, since the model is not 100% accurate there will be some incorrect predictions. To improve the efficiency and accuracy of the model we manually or the user manually gives the correct label which is taken into consideration the next time the model is trained. Hence, the accuracy increases.

# VIII. Future Work

For future work we will integrate the whole set up in a Web Application. Also, we will explore even richer linguistic analysis, for example, semantic analysis and topic

modelling.

## IX. References & Resources Used


- Recognizing contextual polarity in phrase-level sentiment analysis.
  (Link to Paper)

- Sentiment Analysis of Twitter Data.
  (Link to Paper)

- Techniques and Applications for Sentiment Analysis.
  (Link to Paper)

- Emoticon Dictionary: http://en.wikipedia.org/wiki/List of emoticons

- Acronym Dictionary: http://www.noslang.com/dictionary/

- AFINN 111: http://www2.imm.dtu.dk/pubdb/views/publication details.php?id=6010

- Dataset : http://alt.qcri.org/semeval2014/task9/index.php?id=data-and-tools