

# Twitter Sentiment Analysis

Shikha Jain, Rashi Shrishrimal, Itisha Rajat Dewan

---



## 1. Introduction

With the enormous increase in web technologies, number of people expressing their views and opinions via web are increasing. This information is very useful for businesses, governments and individuals. With over 500+ million Tweets (short text messages) per day, Twitter is becoming a major source of information. Twitter is a micro-blogging site, which is popular because of its short text messages popularly known as Tweets. Tweets have a limit of 140 characters. Twitter has a user base of 240+ million active users and

---

---

thus is a useful source of information. Users often discuss on current affairs and share their personal views on various subjects via tweets. Out of all the popular social medias like Facebook, Google+, Myspace and Twitter, we choose Twitter because of the following reasons:

- Twitter contains an enormous number of text posts and it grows every day. The collected corpus can be arbitrarily large.
- Twitter's audience varies from regular users to celebrities, company representatives, politicians, and even country presidents. Therefore, it is possible to collect text posts of users from different social and interests groups.
- Tweets are small in length, thus less ambiguous
- Tweets are unbiased in nature

We build models for two classification tasks: a 3-way classification of tweets into **positive**, **negative** and **neutral** classes. We experiment with the baseline model and feature based model. We do an incremental analysis of the features.

## 2. Timeline

The entire timeline was divided into two phases. The first phase comprised of creating a sentiment analysis tool based on unigrams only with various features. In the second phase, the leftover features were implemented and then the already unigram-feature was modified to check whether the performance upgrades or not. The entire project was successfully completed and the code for it can be found on GitHub. A report corresponding to the project was also hosted, and a video was also made.

## 3. Dataset

Twitter is a social networking that allows users to post real time messages, called tweets. Tweets are short messages, restricted to 140 characters in length. Due to the nature of this microblogging service, people use acronyms, make spelling mistakes, use emoticons

---

and other characters that express special meanings. Following is a brief terminology associated with tweets.

- Target: Users of Twitter use the @ symbol to refer to other users on the microblog. Referring to other users in this manner automatically alerts them.
- Hashtags: Users usually use hashtags to mark topics. This is primarily done to increase the visibility of their tweets.
- Emoticons: These are facial expressions: pictorially represented using punctuation and letters; they express the user's mood.

The dataset consists of 8222 manually annotated tweets and was tested by using cross validation techniques.

## 4. Resources and libraries

Language used : Python

In this work we use following external resources:

- emoticons.py - a library with function for assigning polarity to the emoticons from Christopher Potts' tokenizing script.
- Sentiment lexicon dictionary - mpqa lexicon, etc.
- AFINN data

Tools / libraries used

- nltk ( tokenize, sentiwordnet, pos\_tag, treebank, etc )
- sklearn ( svm and naive bayes )
- numpy
- Scikit , sklearn classification report

## 5. Feature Generation

Data was cleaned by removing those tweets which were not in english. It was then tokenized into words in order to create various features. Lemmatization ,stop words

---

removal from the tweets and Replacement of the acronyms with their full-forms by looking up at the acronyms dictionary was used in phase 1 but removed in phase 2 due to poor accuracy .

Each tweet was represented as a feature vector made up of the following groups of features:

1. word n-grams: presence or absence of contiguous sequences of 1, 2, 3, and 4 tokens;
2. non-contiguous ngrams (ngrams with one token replaced by \*);
3. character n-grams: presence or absence of contiguous sequences of 3, 4, and 5 characters;
4. all-caps: the number of words with all characters in upper-case;
5. POS: the number of occurrences of each part-of-speech tag;
6. hashtags: the number of hashtags
7. Punctuation: the number of contiguous sequences of exclamation marks, question marks, and both exclamation and question marks; whether the last token contains an exclamation or question mark;
8. emoticons: The polarity of an emoticon was determined with a regular expression adopted from Christopher Potts' tokenizing script : presence or absence of positive and negative emoticons at any position in the tweet;
9. whether the last token is a positive or negative emoticon;
10. elongated words: the number of words with one character repeated more than two times, for example, yaaaaaaaaay';
11. ends with one of the punctuation marks: ? , ! , ' , " , etc.
12. A negated context affects the n-gram and lexicon features: we add NEG' suffix to each word following the negation word ('perfect' becomes 'perfect NEG'). The ' NEG' suffix is also added to polarity and emotion features ('POLARITY positive' becomes 'POLARITY positive NEG'). The list of negation words was adopted from Christopher Potts' sentiment tutorial.

<b>Number of tokens</b>	<b>16469</b>
-------------------------	--------------

---

<b>Number of stop words</b>	<b>3280</b>
<b>Number of English words</b>	<b>7867</b>
<b>Number of punctuation marks</b>	<b>2110</b>
<b>Number of capitalized words</b>	<b>941</b>
<b>Number of twitter tags</b>	<b>3547</b>
<b>Number of exclamation marks</b>	<b>2307</b>
<b>Number of negations</b>	<b>1817</b>
<b>Number of other tokens</b>	<b>2108</b>

## 6. Classification

### Phase 1

- 1. Gaussian Naive Bayes**
- 2. Support Vector Machine with Polynomial kernel function with degree 2**
- 3. Support Vector Machine with linear kernel**

Classifier was modelled on the dataset. Tweets were preprocessed and passed onto a NLTK POS tagger. Features mentioned above from features 4 - 12 were calculated based on AFINN resources and sentiwordnet model of nltk. The features were then passed on to Gaussian Naive Bayes and Linear SVM classifier using 5 - fold cross validation.

### Phase 2

- 4. Support Vector Machine with linear kernel and increased features**

Features mentioned in the above section were used. External emoticons' library was used in order to identify the polarity of the sentiment. Also, external sentiment lexicons were used in order to identify the polarity of the words. The linear kernel and

---

the value for the parameter  $C = 0.005$  were chosen by 5-fold cross-validation on the training data.

## 7. Results

### Accuracy for Phase 1 for 5-fold cross validation

Linear Support Vector Machine Classification

[58.02421    55.67212    60.123121    57.12312    59.8901]

Support Vector Machine with Polynomial kernel function with degree 2

[68.1289    69.2267    67.54231    68.3941    69.4294]

**Maximum accuracy : 69.227**

### Phase 2 for 9-fold cross validation

Confusion matrix

```
[[2679  380    6]
 [  21 3926    1]
 [ 128  376  705]]
```

Classification report

	precision	recall	f1-score	support
positive	0.99	0.58	0.73	1209
neutral	0.84	0.99	0.91	3948
negative	0.95	0.87	0.91	3065
avg / total	0.90	0.89	0.88	8222

**Maximum accuracy : 90**

9 - Cross fold Validation

---

```

Cross Validation fold - 2
Confusion Matrix
/usr/local/lib/python2.7/dist-packages/numpy/core/fromnumerical
tion instead. To find the rank of a matrix see `numpy.linalg
VisibleDeprecationWarning)
[[324 46  0]
 [  1 451  0]
 [ 24 46 108]]
Classification Report
      precision    recall  f1-score   support

   positive      1.00      0.61      0.76       178
   neutral      0.83      1.00      0.91       452
   negative      0.93      0.88      0.90       370

avg / total      0.90      0.88      0.88      1000

Cross Validation fold - 3
Confusion Matrix
[[314 43  1]
 [  2 486  0]
 [ 17 53 84]]
Classification Report
      precision    recall  f1-score   support

   positive      0.99      0.55      0.70       154
   neutral      0.84      1.00      0.91       488
   negative      0.94      0.88      0.91       358

avg / total      0.90      0.88      0.88      1000

Cross Validation fold - 4
Confusion Matrix
[[260 49  0]
 [  4 545  0]
 [ 14 41 87]]
Classification Report
      precision    recall  f1-score   support

   positive      1.00      0.61      0.76       142
   neutral      0.86      0.99      0.92       549
   negative      0.94      0.84      0.89       309

avg / total      0.90      0.89      0.89      1000

```

---



Cross Validation fold - 8

Confusion Matrix

```
[[366 42  1]
 [  3 426  0]
 [ 18  51 93]]
```

Classification Report

	precision	recall	f1-score	support
positive	0.99	0.57	0.73	162
neutral	0.82	0.99	0.90	429
negative	0.95	0.89	0.92	409
avg / total	0.90	0.89	0.88	1000

Cross Validation fold - 9

Confusion Matrix

```
[[377 41  1]
 [  2 428  0]
 [ 12  51 88]]
```

Classification Report

	precision	recall	f1-score	support
positive	0.99	0.58	0.73	151
neutral	0.82	1.00	0.90	430
negative	0.96	0.90	0.93	419
avg / total	0.91	0.89	0.89	1000

positive	Gas by my house hit \$3	positive	Gas by my house hit \$
negative	Theo Walcott is still	negative	Theo Walcott is still
negative	its not that I'm a GSP	positive	its not that I'm a GS
negative	Iranian general says I	negative	Iranian general says
positive	with J Davlar 11th. Ma	positive	with J Davlar 11th. M
negative	Talking about ACT's &a	positive	Talking about ACT's &
neutral	Why is "Happy Valentines Day"	positive	Why is "Happy Valenti
negative	They may have a SuperB	negative	They may have a Super
neutral	Im bringing the monster load o	neutral	Im bringing the monster load
neutral	Apple software, retail chiefs	neutral	Apple software, retail chiefs
positive	@oluoch @victor_otti @	positive	@oluoch @victor_otti
neutral	#Livewire Nadal confirmed for	neutral	#Livewire Nadal confirmed for
positive	@MsSheLahY I didnt wan	positive	@MsSheLahY I didnt wa
neutral	@Alyoup005 @addicted2haley hmm	neutral	@Alyoup005 @addicted2haley hm
neutral	#Iran US delisting MKO from gl	neutral	#Iran US delisting MKO from g
positive	Good Morning Becky ! T	positive	Good Morning Becky !
neutral	Expect light-moderate rains ov	neutral	Expect light-moderate rains o
positive	One ticket left for th	positive	One ticket left for t
negative	AFC away fans on Satur	negative	AFC away fans on Satu
negative	Why is it so hard to f	positive	Why is it so hard to
neutral	Game 1 of the NLCS and a remat	neutral	Game 1 of the NLCS and a rema
neutral	@TrevorJavier the heat game ma	neutral	@TrevorJavier the heat game m
true data	1,1	Top out/train0	1,1
		Top	



---

## 8. Conclusion

We implemented a variety of features based on surface form and lexical categories. The sentiment lexicon features (both manually created and externally generated) along with ngram features (both word and character ngrams) led to the most gain in performance.

We manually investigated the phrases or messages which were wrongly labelled by the system. We saw that the label and the phrase/message are quite ambiguous in nature. For example, big enough maybe does not give any positive sense. Similarly the message Desperation Day (February 13th) the most well known day in all mens life. is sarcastic in nature. Thus annotation error and sarcasm present in tweets leads to error propagation. Also, the training set is small. We feel that improving the size of the training set and incorporating sarcasm detection will push the accuracy.

Some examples of wrongly labelled tweets :

Tweets	Label
I'm bringing the monster load of candy tomorrow,I just hope it doesn't get all squished	Positive
Never start working on your dreams and goals tomorrow.....tomorrow never comes....if it means anything to U, ACT NOW!	Positive
My teachers call themselves givng us candy....wasn't even the GOOD stuff. I might go to Walmart or CVS tomorrow	Negative