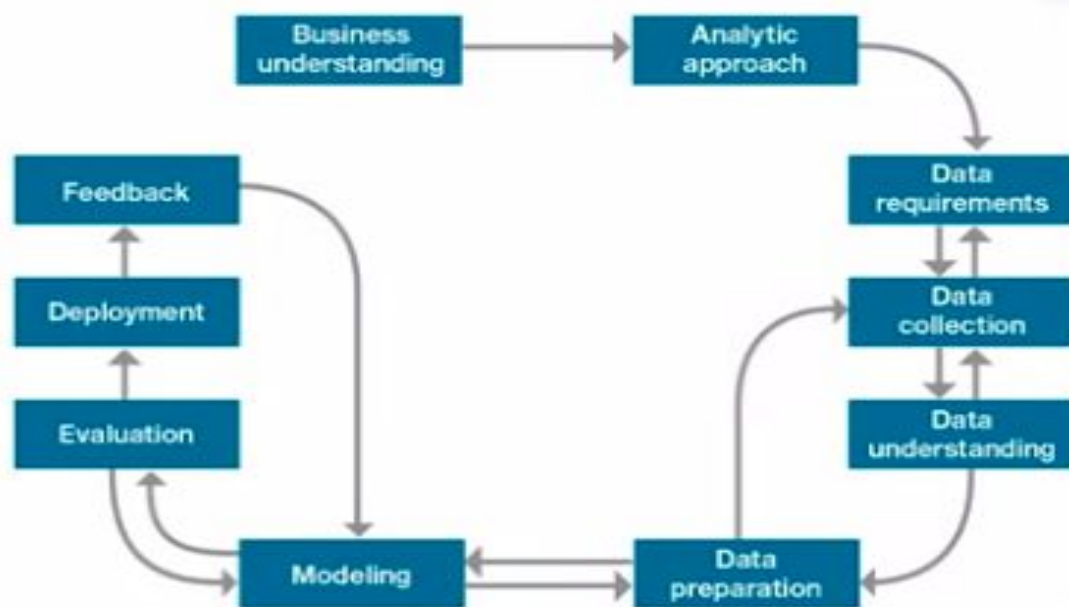


Twitter Sentiment Analysis by Bhuvana Chandra Atche and KartheekK Jampana

1 Problem Statement

Twitter is a popular social networking website where members create and interact with messages known as “tweets”. This serves as a mean for individuals to express their thoughts or feelings about different subjects. Various different parties such as consumers and marketers have done sentiment analysis on such tweets to gather insights into products or to conduct market analysis. Furthermore, with the recent advancements in machine learning algorithms, we are able improve the accuracy of our sentiment analysis predictions. In this report, we will attempt to conduct sentiment analysis on “tweets” using various different machine learning algorithms. We attempt to classify the polarity of the tweet where it is either positive or negative. If the tweet has both positive and negative elements, the more dominant sentiment should be picked as the final label.



We use the dataset from [Kaggle](https://www.kaggle.com/datasets/rohitkumar9001/tweets-sentiment-analysis) which was crawled and labeled positive/negative. The data provided comes with emoticons, usernames and hashtags which are required to be processed and converted into a standard form. We also need to extract useful features from the text such unigrams and bigrams which is a form of representation of the “tweet”. We use various machine learning algorithms to conduct sentiment analysis using the extracted features. However, just relying on individual models did not give a high accuracy so we pick the top few models to generate a model ensemble. Ensembling is a form of meta learning algorithm technique where

Twitter Sentiment Analysis by Bhuvana Chandra Atche and Kartheek Jampana

we combine different classifiers in order to improve the prediction accuracy. Finally, we report our experimental results and findings at the end.

Data Description

File descriptions


- `train.csv` - the training set
- `test.csv` - the test set
- `sampleSubmission.csv` - a sample submission file in the correct format

Data fields


- `id` - sample id
- `positive` - emotion, 1 for positive, 0 for negative
- `tweet` - text of the tweet


Data (37 MB)

API




 kaggle competitions download -c cs5228-project-2

?

 Download All






Data Sources

 sampleSubmission.c...	200k x 2
 test.csv	200k x 2
 train.csv	800k x 3

About this file

No description yet

Columns

 id
 positive
 tweet

2 Data Description

The data given is in the form of a comma-separated values files with tweets and their corresponding sentiments. The training dataset is a csv file of type `tweet_id,sentiment,tweet` where the `tweet_id` is a unique integer identifying the tweet, sentiment is either 1 (positive) or 0 (negative), and tweet is the tweet enclosed in `"`. Similarly, the test dataset is a csv file of type `tweet_id,tweet`.

Twitter Sentiment Analysis by Bhuvana Chandra Atche and Kartheek Jampana

3 Methodology and Implementation

3.1 Pre-processing

Raw tweets scraped from twitter generally result in a noisy dataset. This is due to the casual nature of people's usage of social media. Tweets have certain special characteristics such as retweets,

emoticons, user mentions, etc. which have to be suitably extracted. Therefore, raw twitter data has to be normalized to create a dataset which can be easily learned by various classifiers.

We have applied an extensive number of pre-processing steps to standardize the dataset and reduce its size. We first do some general pre-processing on tweets which is as follows.

Convert the tweet to lower case.

Replace 2 or more dots (.) with space.

Strip spaces and quotes (" and ') from the ends of tweet.

Replace 2 or more spaces with a single space.

We handle special twitter features as follows.

3.1.1 URL

Users often share hyperlinks to other webpages in their tweets. Any particular URL is not important for text classification as it would lead to very sparse features. Therefore, we replace all the URLs in tweets with the word URL. The regular expression used to match URLs is `((www\.[\S]+)|(https?://[\S]+))`.

3.1.2 User Mention

Every twitter user has a handle associated with them. Users often mention other users in their tweets by `@handle`. We replace all user mentions with the word `USER_MENTION`. The regular expression used to match user mention is `@[\S]+`.

Emoticon(s) Type Regex Replacement

:), :), :-), (:, (:, (-:, :') Smile (:s?)|:-)|\s?:|(-:|'\')) EMO_POS

:D, : D, :-D, xD, x-D, XD, X-D Laugh (:s?D|:-D|x-?D|X-?D) EMO_POS

;-), ;), ;-D, ;D, (;, (-; Wink (:s?\(|:-\(|\)\s?:|)\-;) EMO_POS

<3, :* Love (<3|:*) EMO_POS

:-), : (, :(,), -): Sad (:s?\(|:-\(|\)\s?:|)\-;) EMO_NEG

:(, :'(, :(Cry (:,\(|:'\(|\)\s?:|)\-;) EMO_NEG

Twitter Sentiment Analysis by Bhuvana Chandra Atche and KartheekK Jampana

Table 3: List of emoticons matched by our method

3.1.3 Emoticon

Users often use a number of different emoticons in their tweet to convey different emotions. It is impossible to exhaustively match all the different emoticons used on social media as the number is ever increasing. However, we match some common emoticons which are used very frequently. We replace the matched emoticons with either EMO_POS or EMO_NEG depending on whether it is conveying a positive or a negative emotion. A list of all emoticons matched by our method is given in table 3.

3.1.4 Hashtag

Hashtags are unspaced phrases prefixed by the hash symbol (#) which is frequently used by users to mention a trending topic on twitter. We replace all the hashtags with the words with the hash symbol. For example, #hello is replaced by hello. The regular expression used to match hashtags is #(\S+).

3.1.5 Retweet

Retweets are tweets which have already been sent by someone else and are shared by other users. Retweets begin with the letters RT. We remove RT from the tweets as it is not an important feature for text classification. The regular expression used to match retweets is \brt\b.

After applying tweet level pre-processing, we processed individual words of tweets as follows. Strip any punctuation ['"?!,.():;] from the word.

Convert 2 or more letter repetitions to 2 letters. Some people send tweets like I am sooooo happpppy adding multiple characters to emphasize on certain words. This is done to handle such tweets by converting them to I am soo happy.

Remove - and '. This is done to handle words like t-shirt and their's by converting them to the more general form tshirt and theirs.

Check if the word is valid and accept it only if it is. We define a valid word as a word which begins with an alphabet with successive characters being alphabets, numbers or one of dot (.) and underscore(_).

Some example tweets from the training dataset and their normalized versions are shown in table 4.

3.2 Feature Extraction

We extract two types of features from our dataset, namely unigrams and bigrams. We create a frequency distribution of the unigrams and bigrams present in the dataset and choose top N unigrams and bigrams for our analysis.

3.2.1 Unigrams

Probably the simplest and the most commonly used features for text classification is the presence of single words or tokens in the the text. We extract single words from the training dataset and

Twitter Sentiment Analysis by Bhuvana Chandra Atche and Kartheek Jampana

create a frequency distribution of these words. A total of 181232 unique words are extracted from

3

Normalized misses swimming class URL

Raw @98PXYRochester HEYYYYYYYYYY!! its Fer from Chile again

Normalized USER_MENTION heyy its fer from chile again

Raw Sometimes, You gotta hate #Windows updates.

Normalized sometimes you gotta hate windows updates

Raw @Santiago_Steph hii come talk to me i got candy :)

Normalized USER_MENTION hii come talk to me i got candy EMO_POS

Raw @bolly47 oh no :(r.i.p. your bella

Normalized USER_MENTION oh no EMO_NEG r.i.p your bella

Table 4: Example tweets from the dataset and their normalized versions.