Twitter Sentiment Analysis

Contributed by:

Rucha Dubbewar

Mithil Amin

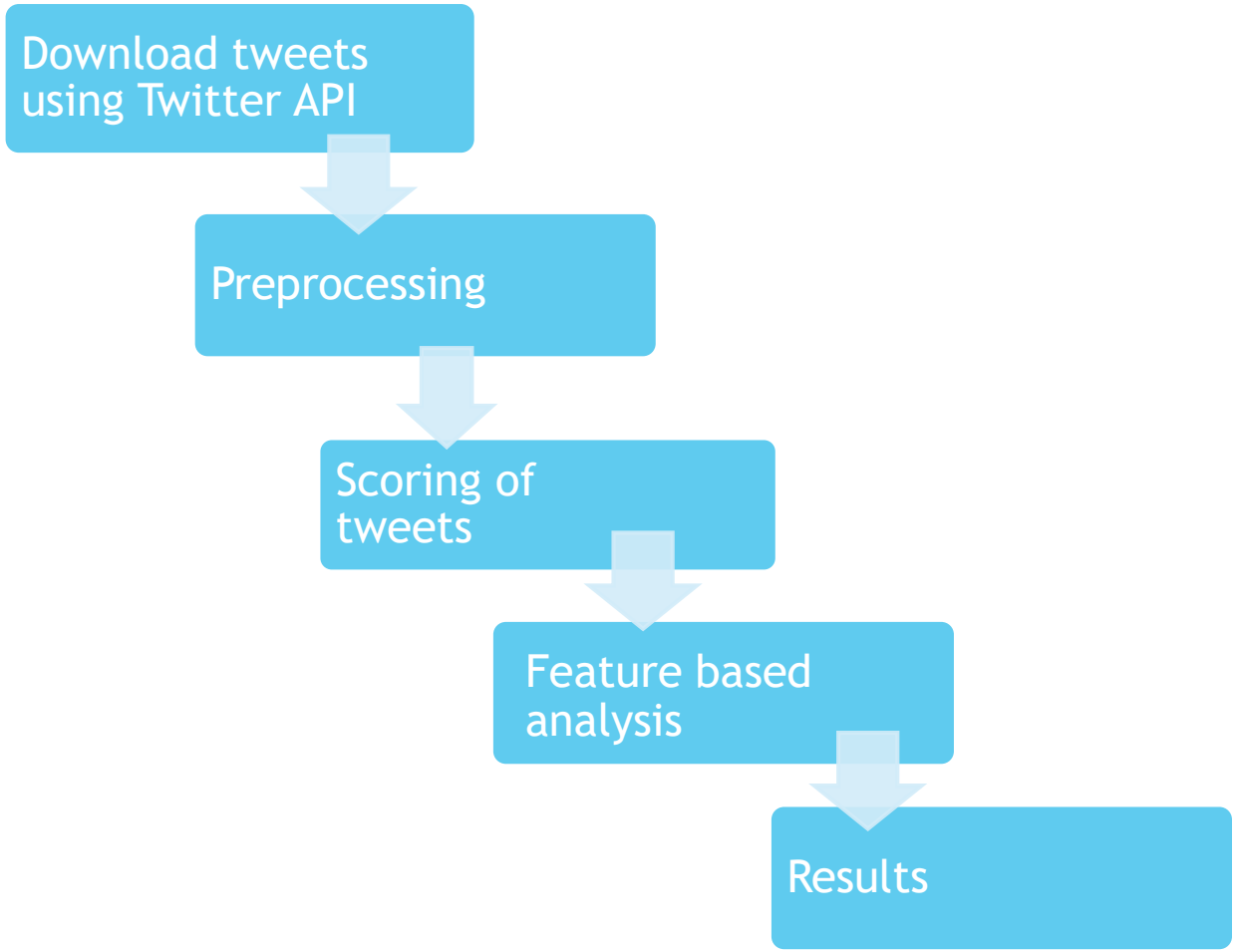
What is Sentiment Analysis and Why Twitter?

- Classification of polarity of given text, sentence or phrase in a document
- Goal : to determine the opinion of text is positive, negative or neutral
- Importance: Opinion of mass, marketing and business strategy, develop and improve product quality, improve customer service
- Popular
- Short messages
- Million active users
- Variety of mass
- Tweets are small in length

Problem Statement and Baseline

- Using feature extraction, we use machine learning algorithms to conduct sentiment analysis and estimate the model with best accuracy
- For a baseline, we use a simple positive and negative word counting method to assign sentiment to a given tweet. We use dataset of positive and negative words to classify tweets. In cases when the number of positive and negative words are equal, we assign positive sentiment.

Approach


Download tweets using Twitter API
Preprocessing
Scoring of tweets
Feature based analysis
Results

Data Description, Preprocessing and feature extraction

- Data extracted is divided into train and test
- This data is tweets with its sentiment and since it is before processing , it's a combination of words, emojis, symbols, hashtags and other user names.
- Analysis is done on Chicago cubs
- Extensive number of preprocessing steps are applied
- Tweets are converted to lower case
- dots and spaces are being replaced
- URL and '@' are replaced
- emojis are replaced with EMO_POS and EMO_NEG
- EMO_POS : ['(:\s?\)|:-\)|\(\s?:|\(-:|:\'\))',' (:\s?D|:-D|x-?D|X-?D)',' (<3|:*)', (;-?\)|;-?D|\(-?;)]
- EMO_NEG : ['(:\s?\(|:-\(|\)\s?:|\)-:)',' (:,\(|:\'\(|:"\()']
- Two features: Unigrams and Bigrams

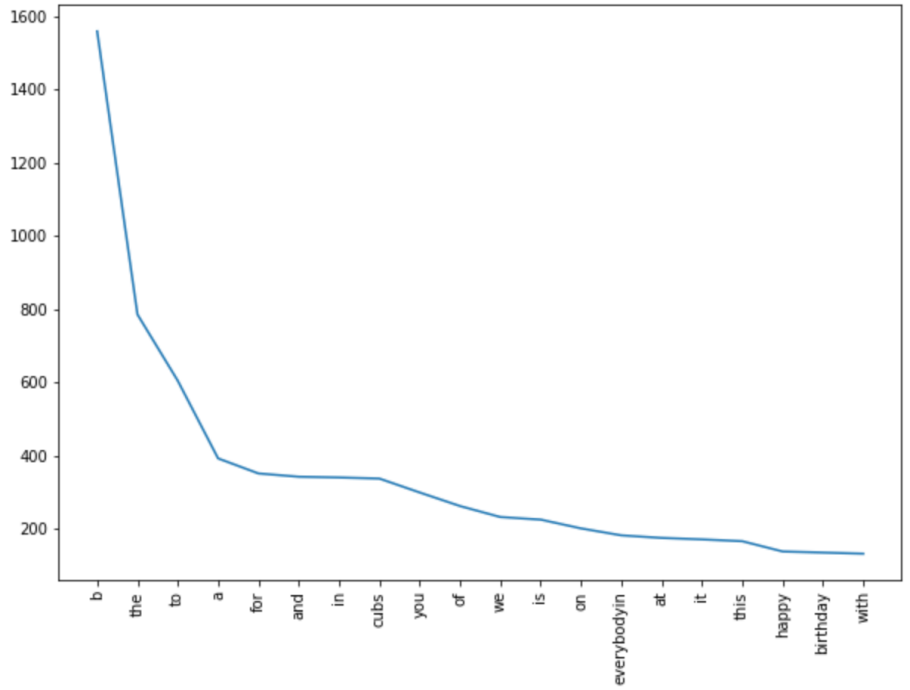
Statistics of Data

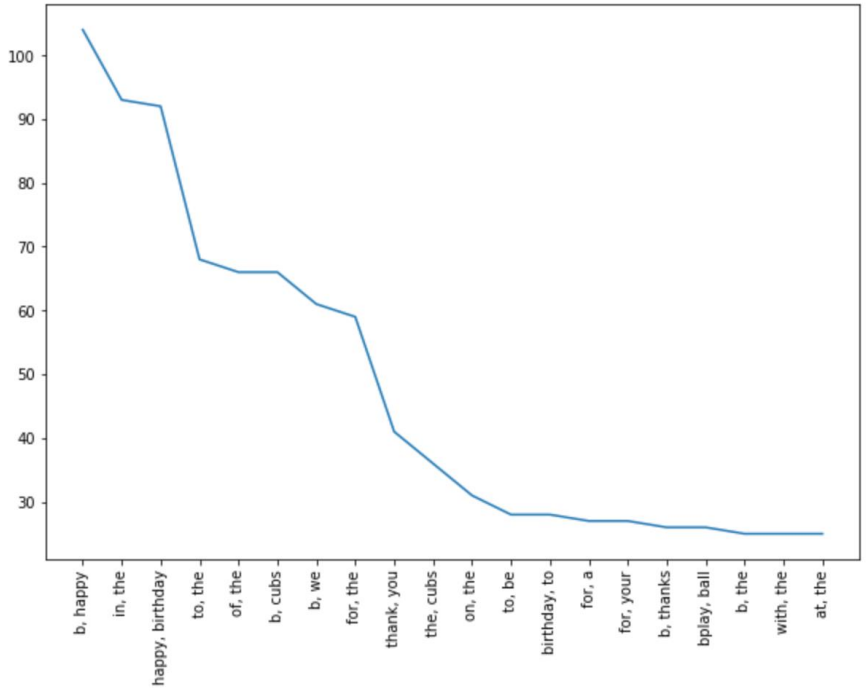
Train data	
Total	2590
Positive	1147
Negative	1443
@	2511
URL	1906
Emojis	4
Unigrams	20908
Bigrams	18340

Test data	
Total	647
Positive	278
Negative	369
@	507
URL	555
Emojis	12
Unigrams	5048
Bigrams	4406

Statistics of Data

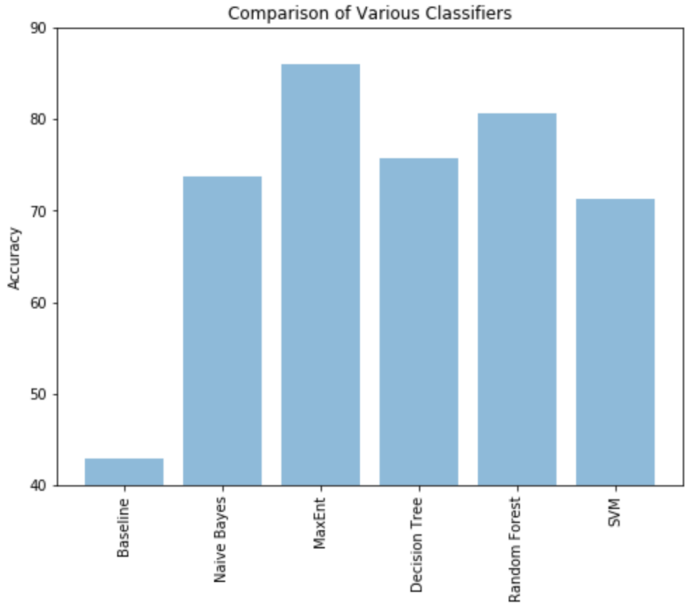
Top unigrams : b , the , to , a , for
Top bigrams : (b,happy),(in,the),(happy,birthday),(to,the),(of,the)





Results/Conclusion/Future Directions

Baseline	43
Naïve Bayes	73.74
Max Entropy	86
Random Forest	80.67
SVM	71.33
Decision Tree	75.67



- Maximum Entropy performed better among all classifiers with accuracy of 86%
- Handling emojis
- Symbols

References

- Sentiment analysis: capturing favorability using natural language processing - https://dl.acm.org/citation.cfm?id=945658
- Sentiment analysis of Twitter data - https://dl.acm.org/citation.cfm?id=2021114
- Opinion mining and sentiment analysis on a Twitter data stream- http://ieeexplore.ieee.org/abstract/document/6423033/
- A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle - https://dl.acm.org/citation.cfm?id=2390490
- Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis - https://search.proquest.com/openview/3bb4518ced02bbd25b28594d28b35094/1?pq-origsite=gscholar&cbl=2029261