Statistical sentiment analysis of Twitter as a review platform

Ankitkumar Jain
Masters of Computer Science
NC State University
Raleigh, North Carolina
aajain2@ncsu.edu

Chirag Jain
Masters of Computer Science
NC State University
Raleigh, North Carolina
csjain@ncsu.edu

Nirav Jain
Masters of Computer Science
NC State University
Raleigh, North Carolina
najain@ncsu.edu

Pratik Kumar Jain
Masters of Computer Science
NC State University
Raleigh, North Carolina
pjain22@ncsu.edu

Rishabh Jain
Masters of Computer Science
NC State University
Raleigh, North Carolina
rjain10@ncsu.edu

ABSTRACT

Reviews being the most important thing in deciding whether to visit a place or not and having a genuine review is the most essential part. Generally, reviews available on websites and platform such as Yelp are more inclined to the positive aspects of a place. Moreover, the amount of usage of social media platforms has increased also, these platforms contain posts which expresses user's opinion in a better way. Twitter being an important contributor, understanding the reliability of such tweets of being reviews is also important. To examine this, we have devised various approaches using natural language processing to analyze the tweets and obtain a statistical analysis of the same based on metadata obtains from tweets.

Keywords

Twitter, Sentiment Analysis, Reviews, Data Mining, Web Scrapping, Tweet, Natural Language Processing, Abbreviation Expansion, Yelp

1. INTRODUCTION

Reviews are the first thing explored by people before traveling to any place. Due to the digitalization, this can be collected from multiple sources. Some of these sources are Yelp, Zomato, Google, etc. A common observation found in these reviews is that they are mainly positive which can mislead the person and also create confusion as all the options are positively rated. Also, the number of reviews from these sources are fewer which might not be desirable to obtain a better picture of the place.

On the other hand, the use of social media has skyrocketed in the past few years. This can be proven by the figure (as shown in Figure 1) below which shows the growth of the number of users on various social media platforms from 2008 to 2017 in US. [5]

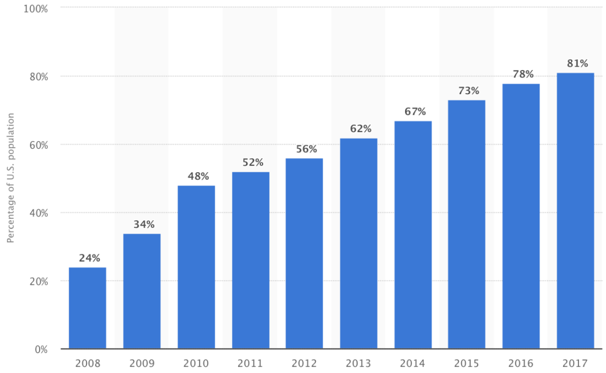


Figure 1: Percentage of U.S. population with a social media profile from 2008 to 2017

Also, the posts on such social media platforms align with the user's opinion. Therefore, these have a potential to be used as a reliable source for the reviews. As a majority of such posts are contributed by Twitter, we are doing a statistical sentiment analysis of tweets to verify whether tweets are a reliable source for reviews. We are using Natural Language Processing to do the same and obtain the polarity and subjectivity of such tweets.

2. CASE STUDY

In today's world there are multiple social media platforms such as Facebook, Twitter, Instagram etc. Out of all these, on Twitter the communication is more public so it is easier to connect to other people compared to other platforms. Also Twitter, being focused on providing news about various events happening in the world using the hashtag feature, so obtaining a tweet containing information about a place is easier.

Twitter is being used by almost all categories of users(as shown in Figure 2). [1] From the figure we can see, the distribution of Twitter usage among Age and Gender Demographics, Location Demographics, Education Demographics and Income Demographics. This supports our choice of Twitter for our analysis as the tweets are contributed from almost all categories of users.

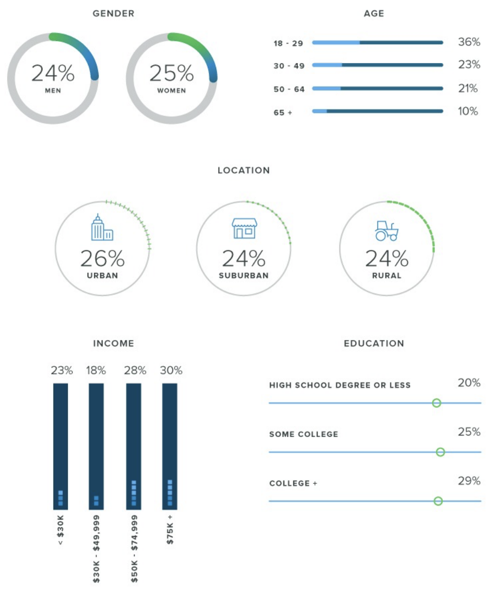


Figure 2: Twitter usage among key Demographics

3. CURRENT SYSTEM

Performing sentiment analysis on the reviews from Yelp is not much useful, since the rating for a particular location is already available on Yelp. Therefore, the current system is mining reviews from social media websites like Twitter and Google places to perform sentiment analysis.

Current system is making use of JavaScript, HTML, CSS for front-end development and Python for the back-end development. The system has also included few libraries such as textblob, python-twitter, python-google-places, spacy and standard libraries for python. Also, have used few APIs namely Google Places, Google Maps, Twitter API.

Some of the challenges in the current system are:

- Vocabulary Correction : Users do not use standard vocabulary therefore the system needs to find a way to translate user vocabulary into standard vocabulary.
- Different languages : Users use different languages therefore the system needs to understand the context in different languages.
- API check : There is a limited number of reviews that can be extracted from Twitter/Google places daily.
- Unrelated tweets : Some tweets mention the location of a place but are not reviews.
- Different names for a location : Users can refer to same location using different names or different location using same name.

4. USER SURVEY

In this section, we will describe all the different questions we put into our survey, the reason why we ask those questions and what we can conclude from the answers. We have used Google Forms to collect the data.

Question 1: Which source you generally use to obtain reviews before visiting a place?

We asked this question to divide the user group into categories based on the platforms they used to check the reviews before visiting any place. The user had option to choose between a platform containing reviews or a social media platform.

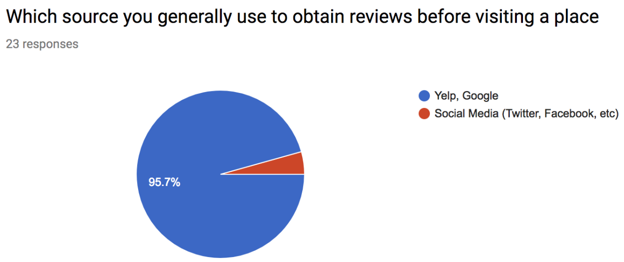


Figure 3: Responses for Question 1

Question 2: Do you write a tweet about a place while you are at that particular place?

We asked this question to see whether the users use Twitter more frequently while they are at that place or they tweet about it when they have time.

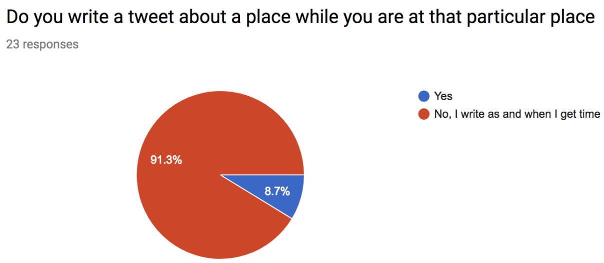
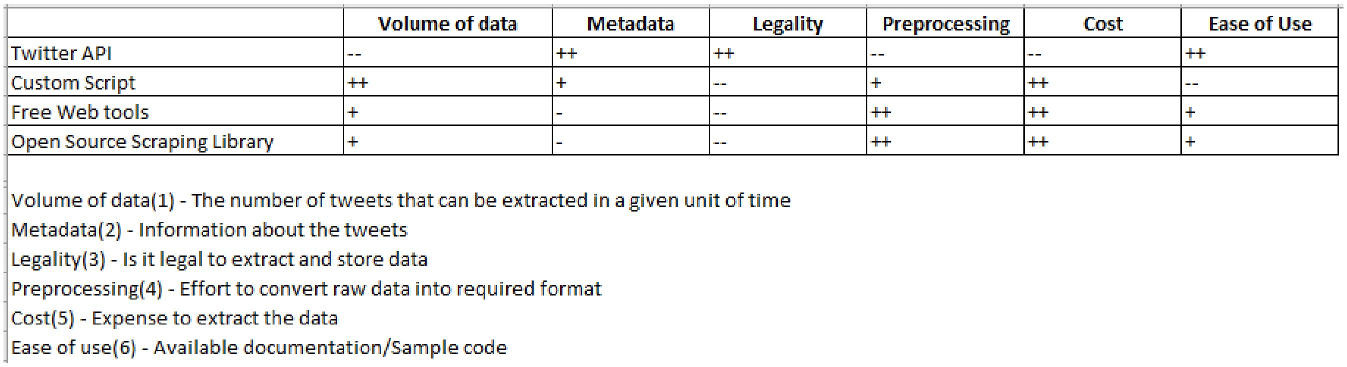


Figure 4: Responses for Question 2

Question 3: Would it matter if the review (tweet) about a location is from a celebrity?

We asked this question to obtain a user group who are influenced by tweets from a celebrity or a highly influential person.



	Volume of data	Metadata	Legality	Preprocessing	Cost	Ease of Use
Twitter API	--	++	++	--	--	++
Custom Script	++	+	--	+	++	--
Free Web tools	+	-	--	++	++	+
Open Source Scraping Library	+	-	--	++	++	+

Volume of data(1) - The number of tweets that can be extracted in a given unit of time

Metadata(2) - Information about the tweets

Legality(3) - Is it legal to extract and store data

Preprocessing(4) - Effort to convert raw data into required format

Cost(5) - Expense to extract the data

Ease of use(6) - Available documentation/Sample code

Figure 5: Assessment of various products/services

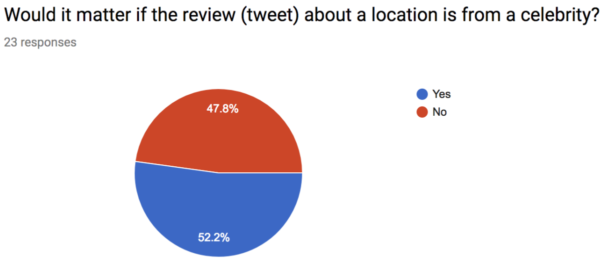


Figure 6: Responses for Question 3

4.1 Survey Conclusion

Based on the 23 evaluations we can observe that the majority of people prefer review-based platform as compared to social media platform. Also the survey shows that greater number of user writes tweet when they have time and not when they are at that particular place. Along with that the influence of a celebrity on the review of a place affects almost half of the users who took the survey.

5. PROPOSED IMPLEMENTATION

5.1 Data Collection Techniques

This is a very crucial step of the proposed system. The volume of data or tweets collected will determine the accuracy of our results and the generality of the evaluation. Also the amount of metadata collected in this step will help us define the proposed Indices [Section 5.4] better and with precise thresholds. The various available approaches for data collection are follows :

1. Custom python script
2. Twitter API
3. Open Source Scrapping Libraries
4. Free Web based tools

Each method has it's own advantage and disadvantage as shown in Figure 5. Method 1 will give us full control over the volume and amount of information captured for each tweet. This will also enable us to extract more data with parallel processing. However, one of the major disadvantage with this approach is the amount of metadata captured will be less compared to what the official Twitter API[9] provides.

Method 2, which is the official Twitter API, provides a lot of metadata about a tweet[6]. More than the amount, the structure of data is well formatted eliminating need for pre-processing. However, one of the major setbacks for this approach is the rate limitation[7] for 'Standard Subscription'.

Other methods mentioned in the list have all the disadvantages of Method 1 but with the incentive that they provide data scrapping with minimum or no configuration overhead. The amount of metadata and search query customization in these methods vary but are generally same. We are planning to use a hybrid approach to incorporate the advantages of all the above methods into our tool.

5.2 Abbreviation and Acronym Expansion

Abbreviation and Acronym Expansion is the process of determining the correct expansion of an abbreviation / acronym in a given context.

One approach is to use a general library that maps abbreviations and acronyms to their expanded full-forms. However, the challenge with using a general library is acronyms are almost always domain dependent. That is why it is not a good idea to have a "general" library. NLP, for example, could mean 'natural language processing' or 'neuro-linguistic programming', depending on the domain. Another approach is if we have a collection of textual documents where both the acronym and its expansion occur you can apply an algorithm to extract (acronym, extension) pairs. That will not be possible as the data in context is *tweets*

An approach that we will explore is a method for predicting the most appropriate expansion of acronyms using Wikipedia.[13] It is based on tf-idf word frequencies. Given a document containing an unknown acronym, the algorithm makes its prediction in two general phases. First, training phase where Wikipedia documents are scraped and acronyms and their expansion is discovered. Second, tweets are fed into the system containing unknown abbreviation. The input is converted to a tf-idf frequency vector. For each possible expansion (class), it computes the dot product of the class's learned coefficient vector and the tweet's frequency vector (and add the class's bias term). This results in a set of scalar values, one for each possible class. The class with maximum value is predicted as the expansion of the current abbreviation/ acronym. The process is depicted in Figure 7[2]

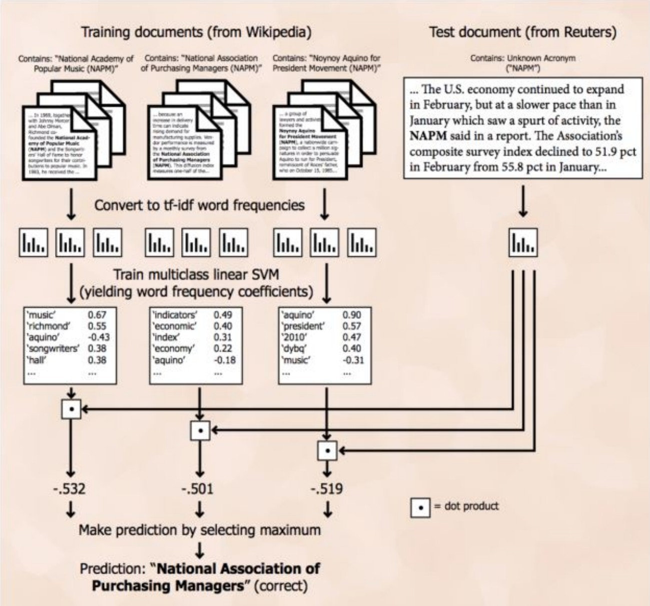

Training documents (from Wikipedia)
Test document (from Reuters)
Contains: "National Academy of Popular Music (NAPM)"
Contains: "National Association of Purchasing Managers (NAPM)"
Contains: "Noynoy Aquino for President Movement (NAPM)"
Contains: Unknown Acronym ("NAPM")
... The U.S. economy continued to expand in February, but at a slower pace than in January which saw a spurt of activity, the NAPM said in a report. The Association's composite survey index declined to 51.9 pct in February from 55.8 pct in January...
Convert to tf-idf word frequencies
Train multiclass linear SVM (yielding word frequency coefficients)
'music' 0.67
'richmond' 0.55
'aquino' -0.43
'songwriters' 0.38
'hall' 0.38
'indicators' 0.49
'economic' 0.40
'index' 0.31
'economy' 0.22
'aquino' -0.18
'aquino' 0.90
'president' 0.57
'2010' 0.47
'dybq' 0.40
'music' -0.31
-.532
-.501
-.519
= dot product
Make prediction by selecting maximum
Prediction: "National Association of Purchasing Managers" (correct)

Figure 7: Abbreviation/Acronym Expansion process

5.3 Vocabulary Correction

Vocabulary Correction is the process of rectifying words like "abt" that stand for about and "ordr" which stands for "order". These words might occur in tweets because of the informal nature of the medium, character limit or mere typing error.

We would use n-gram model to find the most probable word and then use substitution. Example, we substitute the word "ordr" by the word "order".[12] An n-gram model is a type of probabilistic language model for predicting the next item in such a sequence. A probabilistic model of word sequences could suggest the desired substitution in tweets.

5.4 Tweet Analysis Indices

Twitter corpora is a humongous database of tweets, however every tweet is not a review. Therefore, it is important to filter tweets that are reviews and perform sentiment analysis only on those tweets. Also, Twitter consists of various types of tweets, such as General Tweets, Retweets, Quote Tweets, Promoted Tweets and so on [11]. It also consists of various types of users, such as general users, bots and celebrities. The current implementation does not consider these factors while analyzing tweets for sentiments. However, we need to handle the bias of tweets introduced by these various factors instead of ignoring them, in order to generate correct sentiment about the reviews. Based on the meta-data of the tweets, we define the following indices to weigh a tweet for its review quality

5.4.1 Review Relevance Index

In order to identify whether a tweet can be considered a review or not, we first need to understand what is a review. There is no universal guideline for defining a "review", thus we will use machine learning for this task. We know for a fact that websites such as Yelp and Google provides dedicated platforms for reviews. Therefore, we will create a bag-of-words model on data extracted either from Yelp or Google or some other similar service. The extracted data will be pre-processed as follows:

- Consider reviews less than 280 characters (Same as Twitter word count)
- Ignoring case
- Ignoring punctuation
- Ignoring frequent words
- Fixing misspelled words
- Reducing words to their stem (e.g. "play" from "playing") using stemming algorithm

Once we're ready with bag-of-words model, we will use term presence to identify whether a tweet is a review or not. Note that we are not using term frequency because Twitter is a micro blogging website and therefore, checking frequency of words in the tweet will produce poor results compared to term presence. Review Relevance Index will simply be a binary index, with value 1 indicating that a tweet is a review and 0 indicating otherwise.

5.4.2 Tweet Support Index

Meta-data about a particular tweet, such as number of retweets, and number of likes are important factors to consider while mining reviews from tweet. If the number of likes of a negative tweet, such as "I hate the traffic near Empire State Building", is high then it means that many people feel this way about traffic around Empire State Building. Therefore, it is crucial to not disregard such information, but rather include it while analyzing a tweet. For this reason, we define a tweet support index associated with every tweet, as follows: Tweet Support Index = a1 x Likes + a2 x Retweets, where a1 and a2 are configurable parameters. For the purpose of our study, we have a1 = 1.2 and a2 = 1. Thus, Tweet Support Index = 1.2 x Likes + Retweets.

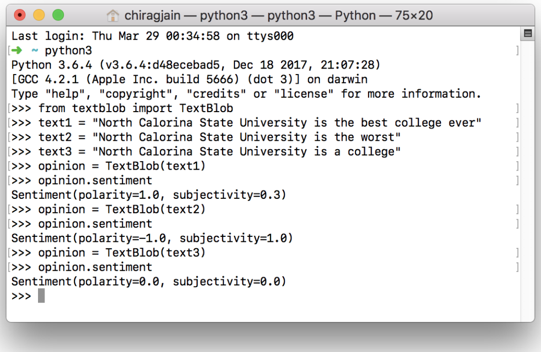
5.4.3 Influential Index

Influencers on Twitter are people who inspire people in their industry. A good review of such an influencer will introduce influencial bias in tweets of their followers. Consider the following tweet by Elon Musk: "Dinner in an old Belgian ironmongery. Best menu art ever." [10] Since Elon Musk loves the menu of this restaurant, chances are that Elon Musk's followers would also love this menu. It is important to factor in this information while analyzing the reviews of this particular restaurant, because influencers may be bought for marketing purposes and thus their tweet might not reveal the real underlying sentiment. For this purpose, we will associate an influential index with every tweet. A straightforward approach to identify an influencer is to simply consider the number of followers of the user who is tweeting.

5.5 Sentiment Analysis

Sentiment analysis or emotion AI refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine [4].

For our project we plan to perform sentiment analysis on the Tweets' text. The process will provide us information about the subjectivity and polarity of the text. This will be done using the existing python library TextBlob [8]. Once we have the value indicating the subjectivity of a particular tweet, we can determine whether the tweet is a general factual statement or a polar opinion or comment. The polarity value will give a sense of negative or the positive tone of the tweet. Figure 8 shows the example of possible types of opinion which can be tweets and their corresponding sentiment's polarity and subjectivity.



```
Last login: Thu Mar 29 00:34:58 on ttys000
➜  ~ python3
Python 3.6.4 (v3.6.4:d48ecebad5, Dec 18 2017, 21:07:28)
[GCC 4.2.1 (Apple Inc. build 5666) (dot 3)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> from textblob import TextBlob
>>> text1 = "North Calorina State University is the best college ever"
>>> text2 = "North Calorina State University is the worst"
>>> text3 = "North Calorina State University is a college"
>>> opinion = TextBlob(text1)
>>> opinion.sentiment
Sentiment(polarity=1.0, subjectivity=0.3)
>>> opinion = TextBlob(text2)
>>> opinion.sentiment
Sentiment(polarity=-1.0, subjectivity=1.0)
>>> opinion = TextBlob(text3)
>>> opinion.sentiment
Sentiment(polarity=0.0, subjectivity=0.0)
>>>
```

Figure 8: Sentiment Analysis on various opinions

Alternatively, instead of using the standard sentiment analysis module provided by TextBlob we can also train our own learning model to provide better control over the accuracy of the sentiment measure. This is because the TextBlob sentiment analyzer is trained on a dataset containing movie reviews. These are similar to reviews for a particular location but have subtle difference in the range and variety of words that are used. A custom sentiment analyzer can be built and trained using the python NLTK library[3]. There is a great guide[14] about implementing the same.

We will choose one implementation depending on the difference in the accuracy of the custom sentiment analyzer or the TextBlob library.

6. PROPOSED EVALUATION

6.1 Cross-Index Comparison

For each of the indices defined in section 5.4 above, we will perform a sentiment analysis with and without considering those indices. This will help us understand whether how the bias introduced by those indices affect the sentiment in the reviews. We plan to provide a side by side comparison for the same, on the left hand side the sentiment analysis considering the index, and on the right hand side sentiment analysis without considering the index.

6.2 Statistical Analysis

One of the key goals of the project is to perform Statistical Analysis tweets to find out parameters like what proportion of tweets can actually be used as review for a particular place / product, are the tweets highly polarized just to seek attention on social-media. Finally we would summarize and comment on whether social-media platforms like Twitter be used to replace / augment traditional review platforms like Yelp / Zomato. The results would be based on the insights we gain by applying techniques mentioned in section 5 on a significantly large number of tweets from varied range of users.

6.3 Cross-Product Review Comparison

There are tons of platforms available were a user can get reviews about a particular location and other related information. Some popular ones, in no particular order are as follows.

- Yelp
- Google Local Guide Reviews
- TripAdvisor
- Facebook
- Official website for the location
- USNews

The evaluation would involve multiple users to rate the quality, genuineness, length of reviews on few of the selected platforms mentioned above with the reviews mined from Twitter. The location will be some place that the user has not previously visited so that we can eliminate any prejudice. This result will supplement the result of the statistical analysis to infer the viability of Twitter as a platform for reviews.

7. ACKNOWLEDGEMENTS

Our special thanks to our Prof. Timothy Menzies and Teaching Assistant, Amritanshu Agrawal for helping and guiding us throughout the project development.

8. CONCLUSION

By the end of the project we shall be able to determine if Twitter can be used as a platform for reviews. Also, we would be able to identify which factors/parameters are relevant in analyzing the reviews. We should also be able to determine how good/bad are the twitter reviews as compared to reviews provided by existing platforms.

9. REFERENCES

[1] Social Media Demographic . `https://sproutsocial.com/insights/new-social-media-demographics/#twitter`, 2017. [Online; accessed 28-March-2018].

[2] Acronym Expansion Disambiguation. https://pdfs.semanticscholar.org/52d5/6c8de14b3d6640d048eb0fbd06ef95bd945d.pdf, 2018. [Online; accessed 26-March-2018].

[3] Natural Language Toolkit. https://www.nltk.org/, 2018. [Online; accessed 25-March-2018].

[4] Sentiment Analysis. https://en.wikipedia.org/wiki/Sentiment_analysis, 2018. [Online; accessed 25-March-2018].

[5] Social Media Platform User's Increase Graph . https://www.statista.com/statistics/273476/percentage-of-us-population-with-a-social-network-profile, 2018. [Online; accessed 28-March-2018].

[6] Standard search API. https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html#example-response, 2018. [Online; accessed 25-March-2018].

[7] Standard search API | Resource Information. https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html#resource-information, 2018. [Online; accessed 25-March-2018].

[8] TextBlob: Simplified Text Processing. http://textblob.readthedocs.io/en/dev/, 2018. [Online; accessed 25-March-2018].

[9] Twitter Developer Platform. https://developer.twitter.com/, 2018. [Online; accessed 25-March-2018].

[10] Twitter Status | Elon Musk. https://twitter.com/elonmusk/status/647209943515332608?lang=en, 2018. [Online; accessed 26-March-2018].

[11] Types of Tweets. https://help.twitter.com/en/using-twitter/types-of-tweets, 2018. [Online; accessed 26-March-2018].

[12] Vocabulary Correction. https://lagunita.stanford.edu/c4x/Engineering/CS-224N/asset/slp4.pdf, 2018. [Online; accessed 26-March-2018].

[13] Wikipedia Abbreviations Expansion. https://en.wikipedia.org/wiki/Wikipedia:Abbreviation_expansion, 2018. [Online; accessed 26-March-2018].

[14] L. Luce. Twitter sentiment analysis using Python and NLTK. http://www.laurentluce.com/posts/twitter-sentiment-analysis-using-python-and-nltk/, 2012. [Online; accessed 25-March-2018].