Research on Data Preprocessing and Categorization Technique for Smartphone Review Analysis

Vivek Agarwal
Department Of Computer Engineering,
Smt. Kashibai Navale
College of Engineering.
‘Ram Darshan’ bldg. 46/D,
Aundh Road,Pune -20

ABSTRACT

Given the current market scenario, competition for dominance is at its peak. The best way for any smartphone manufacturer to deliver a quality product, is to perform a review analysis based on the end user experience and their demands. This helps the manufacturers to make necessary product changes and launch more relevant features in their smartphones. This paper aims to highlight the data preprocessing steps required for review analysis of a newly launched smartphones in the market by collecting tweets from the Twitter data feed. It aims to improvise the preprocessing steps and increase accuracy by filtering out the less relevant tweets and consider the more appropriate and genuine tweets.

General Terms

Categorization, reviews, iPhone, twitter, preprocessing, text.

Keywords

analysis, categorization, cleaning, dataframe, gibberish, preprocessing, processed, raw, review, structured, transformation.

1. INTRODUCTION

Today’s market scenario is a lot different than what it used to be a decade ago. With the introduction of smartphones and inbuilt sensors, the competition for high profile processors, high resolution cameras, screen size and performance has sky rocketed. Against this backdrop, the manufacturers risk a lot when planning to launch a new device in the market. It is important to carry out a detailed survey regarding the response of the end users to the previously launched products, its shortcomings and suggested improvisation by experts. After incorporating such changes and feature improvements, the situation is a lot safer for manufacturers to launch the new device and probability of success increases exponentially. Twitter is one of the most commonly and widely used platforms for users who wish to express themselves, and there is often a sentiment associated with a user’s tweet that expresses their likes and dislikes regarding their device experience. In this paper, I aim to target these tweets, categorize them according to various features that it expresses and apply preprocessing technique. The objective of this paper, is to demonstrate how Bot generated gibberish tweets can be separated from genuine human tweets and different aspects of a smartphone, can be categorized into the respective features that it talks about. For instance, a tweet saying, “Soo excited about my new iPhone 6S! Great camera and awesome processing speed! #iPhone6S #Selfies #Happy” expresses reviews regarding camera quality and processor performance. We need to categorize this tweet and apply preprocessing to transform this data into useful information. The data (tweets) that I have obtained was in an unstructured form. I have transformed this data by implementing a proper structure using dataframe, which allows us to apply SQL like queries to extract information.

2. DATA PREPROCESSING

Data Preprocessing is the process of simply transforming raw data into understandable format. Real world data is sometimes incomplete, inconsistent, redundant and noisy. Data preprocessing involves various steps that help to convert raw data into processed and sensible format.

The diagram below is used to depict the various steps involved in data preprocessing.

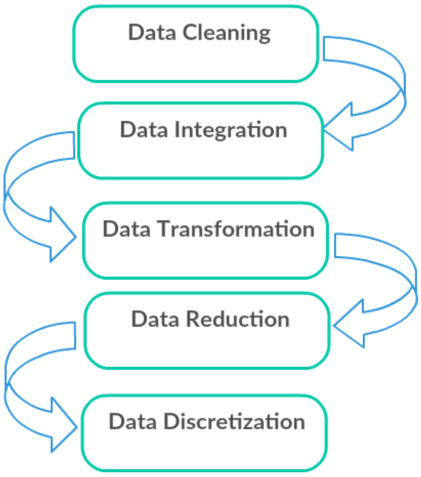

Fig 1. Data Preprocessing Steps.

A. Data Cleaning

Data cleaning is the process of detecting corrupt data and inaccurate records from a record set or database table. The main use of cleaning step is based on detecting incomplete, inaccurate, inconsistent and irrelevant data and applying techniques to modify or delete this useless data.

B. Data Integration

Data Integration focuses on unification of data residing in different sources and presenting a unified view of these data. Data with different representations are put together and any conflicts resulting from it are resolved. This process becomes vital in a number of scientific and commercial applications. With increasing volume and exponential growth of data, integrating it becomes even more significant.

C. Data Transformation

Data transformation plays a pivotal role in converting unprocessed data into understandable form. It consists of data normalization, aggregation and generalization. Data normalization helps to arrange the columns and tables of a database such that redundancy is minimum. This helps cut down on the processing time and complexity. Data aggregation helps in creating a brief summary for faster overview. The process of data generalization is also known as rolling-up data. It helps in generalizing data and creates successive layers of summary in evaluation database.

D. Data Reduction

Data reduction is the process of transforming digital info into ordered and simplified form. This data is generally derived through empirical and experimental means. It involves reducing large amounts of data into smaller and meaningful fragments.

E. Data Discretization

Data discretization is an important concept when you have a large amount of numeric data, but only want to classify it based on nominal values. In this scenario, the continuous data is split into discrete forms and the values of these discrete sets are said to be the nominal value. It is basically a process of converting continuous data attributes into a finite set of intervals with minimal loss of information.

3. PROPOSED APPROACH

In my approach, I have preprocessed the tweets that I received from twitter using tweepy API. The tweets that we receive are in unstructured JSON format. Our objective is to structure these tweets and preprocess it. I have mentioned the various steps involved in this process.

3.1 Overview

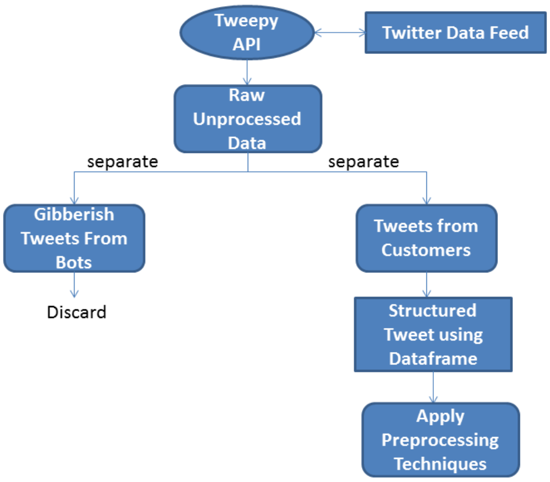

Fig 2. Overview of the entire procedure.

In my approach, I have used twitter data feed as my main source of dataset. Tweets obtained from twitter are extracted using the Tweepy API. This API is provided officially by twitter and is a RESTful API that requires special access token secret and consumer secret for authentication to access the live tweets. The tweets that we receive are obtained in a JSON format and have multiple attributes.

In order to search tweets relating to a particular device such as iPhone, I have set the filter as 'iPhone'. The Standard Listener then listens to all the tweets with 'iPhone' as the main keyword and hashtag. An example of sample JSON tweet is given as follows.

{"created_at":"Sat Oct 24 05:48:17 +0000 2015", "id":657795727796531200, "id_str":"657795727796531200", "text":"Love the HD resolution Camera for iPhone 6S. The body still looks a lil fragile to handle!", "source":"\u003ca href=\"http:\/\/twitter.com\" rel=\"nofollow\"\u003eTwitter Web Client\u003c\/a\u003e", "truncated":false, "in_reply_to_status_id":null, "in_reply_to_status_id_str":null, "in_reply_to_user_id":null, "in_reply_to_user_id_str":null, "in_reply_to_screen_name":null, "user":{"id":2644475012, "id_str":"2644475012", "name":"richard", "screen_name":"farmview_SZ", "location":"\u53f0\u7063", "url":null, "description":"RV & Trailer rear view system professional manufacture", "protected":false, "verified":false, "followers_count":2, "friends_count":18,"listed_count":0,"favourites_count":0,"statuses_count":4, "created_at":"Mon Jul 14 08:20:07 +0000 2014","utc_offset":null,"time_zone":null,"geo_enabled":false, "lang":"zh-cn", "favorite_count":0,"entities":{"hashtags":[],"urls":[{"url":"https:\/\/t.co\/yW7Eq2ZUCI","expanded_url":"http:\/\/www.amazon.com\/dp\/B015GWJK8Y\/ref=cm_sw_r_tw_dp_nzXkwb1982M94","display_url":"amazon.com\/dp\/B015GWJK8Y\/\u2026","indices":[81,104]}],"user_mentions":[{"screen_name":"amazon","name":"Amazon","id":20793816,"id_str":"20793816","indices":[108,115]}],"symbols":[]},"favorited":false,"retweeted":false,"possibly_sensitive":false,"filter_level":"low","lang":"en","timestamp_ms":"1445665697417"}

The "text" attribute used in this file contains the actual tweet that we are interested in and every other attribute is the meta-data that we receive. In order to increase accuracy and decrease processing time, we can reduce the dataset by neglecting out the gibberish tweets. Many of the tweets that we obtain are generated by Bots which do not express any sentiment or review. These tweets are gibberish and not generated by humans. They need to be filtered out in order to reduce the size of the data. In order to identify the Bots, we make use of the "user" attribute of the JSON tweet. The Bots can be easily identified from the "name" attribute which consists of random alphanumeric characters as opposed to proper alphabetical strings. Also the source of the tweet gives us a good indication of the genuinity of the tweet. For example, a Bot tweet with a source attribute is given as follows -

"source":"\u003ca href=\"http:\/\/twittbot.net\/\".

Here, we can trace the origin of the tweet to a twitter bot URL indicating a non-human generated tweet. In the previous example of JSON tweet, note that the "source" attribute contains "http:\/\/twitter.com\" as the URL and "name" attribute is set to "richard" indicating a plausible human name.

3.2 Structured Form.

In order to define a structure for the tweets that we receive, I have defined it in a dataframe. To create a dataframe, I've made use of 'Pandas Data Analysis Library' in Python. Pandas is an open source, BSD-licensed library providing high performance, easy to use data structures and data analysis tool.

It can be used for reading and writing data between in-memory data structures and different formats: CSV and text files, Microsoft Excel and SQL databases.

Pandas Dataframe helps us structure all the tweets in various categories such as tweets relating to 'camera', 'screen', 'battery', 'app', 'processor', 'bend', 'connectivity' and 'performance' as the various categories. These categories also form the columns of our structured database table. Once we have categorized the tweets into their respective matching columns, we shall apply preprocessing techniques to each one.

This method of categorization is performed using synonym builder and bag of words clustering. Pandas also supports querying language for information extraction from dataframe. We can extract particular columns and/or rows using querying.

3.3 Preprocessing Technique

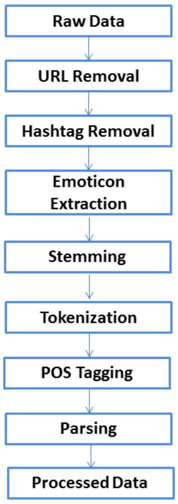

Fig 3. Preprocessing Steps

The data that we have obtained from twitter data feed is in a raw unprocessed JSON format. Our objective would be to process this data in a cleaner format such that we can extract information further. Fig 3 demonstrates the step by step process of the techniques that we plan to apply to the data for preprocessing.

The tweepy API is a restful web service that extracts live tweets from twitter database. This collection of tweets is saved into a JSON file in our computer that acts as our primary source of dataset. In order to achieve high precision and accuracy, I have saved a large amount of dataset.

In order to increase accuracy of the reviews that we plan to extract at a later stage, we need to filter this dataset. The tweets that we have obtained contain gibberish tweets that have been created by BOTs rather than humans, as demonstrated in Fig 2. It is vital to filter out these kinds of tweets as they can cause noise and irrelevancies in our results. Filtering out these tweets can help us get a closer estimation of human reviews.

3.3.1 URL Removal

The urls used in the tweet play no vital role in conveying the sentiment behind a user's review. It thus needs to be discarded. We make use of regular expression to find the occurrence of a hyperlink. Regular expressions library helps in pattern matching and url detection. Once a link has been detected, it is discarded from the tweet.

3.3.2 Hashtag Removal

The hashtags used in the tweets denote the subject that the tweet is about. In this analysis I have removed neutral hashtags and tokenized the ones demonstrating sentiments

Table I. Examples showing the tweets and their corresponding expression denoted using hashtags.

Tweet	Expression
"Got a new iPhone 6S Today!! ☺ ☺ #iPhone #awesome #happyyy"	Hashtags represent 2 positive and 1 neutral expressions.
"The back cover could be a lil darker! #iPhone6S #nevermind"	Hashtags represents 2 neutral expressions relating to back cover.
"RT @jason123: thanks for the new iPhone bday gift!! Yay! #iPhone #Happpyy"	Hashtags represents 1 neutral and 1 positive expression.

3.3.3 Emotion extraction

The emoticons play a vital role in conveying the sentiment of the end-user regarding his product experience. For example, ":-)" is labeled as positive whereas ":(" is labeled as negative. We assign each emoticon a label as follows: Extremely-positive, Extremely-negative, Positive, Negative, and Neutral.

3.3.4 Stemming

Stemming is the process of reducing the derived words to the word's stem. For example, words like thoughtful, doggy, catlike, exciting are reduced to thought, dog, cat, excite as their stem words. In order to match the synonyms of the stems, I have used a synonym builder. For example, synonyms of 'excite' are 'thrilled', 'exhilarated', 'animated', 'electrified', 'moved' etc. Stemming helps reduce a large variety of words into a single word denoting its stem or synonym.

In this, I have also expanded the commonly used acronyms into their English equivalent words.

Table 2. Examples denoting the acronym and corresponding expansion.

Acronym	English Expansion
gr8, gr8t	Great
lol	Laughing out loud
rofl	Rolling on the floor laughing.

I have also replaced sequence of repeated characters by three characters. For example, we convert 'awsoooome' to 'awesooome' and 'coooool' to 'coool'. We do not replace it by 2 characters as we wish to differentiate between regular usage and emphasized usage. I have made use of regular expressions library in python for reduction.

3.3.5 Tokenization

Tokenization is the process of creating tokens for the words obtained in the tweet. The entire sentence is split into fragmented words which can later be tagged for POS analysis and parse tree generation. For example, consider the tweet –

"RT @SkyNews: Facebook admits its app drains iPhone battery but says a fix is here https:\/\/t.co\/oHDFrSwaGa https:\/\/t.co\/bnWVUIVjy0"

The tokenized form that I obtained is as follows

[u'Facebook', u'admits', u'its', u'app', u'drains', u'iPhone', u'battery', u'but', u'says', u'a', u'fix', u'is', u'here']

This shows that the tweet has been successfully rid of 'RT @ SkyNews:' and the URL links. I have made use of Natural Language Toolkit library to perform tokenization.

3.3.6 POS Tagging

POS (Parts of Speech) Tagging, also known as word-category disambiguation, is the technique of marking up a word in corpus as corresponding to a part of speech. Using NLTK library, I have created the POS tags for the tokens earlier generated.

For the tweet mentioned in the previous step, here are the resultant POS tagged words.

[(u'Facebook', 'NNP'), (u'admits', 'VBZ'), (u'its', 'PRP$'), (u'app', 'NN'), (u'drains', 'VBZ'), (u'iPhone', 'JJ'), (u'battery', 'NN'), (u'but', 'CC'), (u'says', 'VBZ'), (u'a', 'DT'), (u'fix', 'NN'), (u'is', 'VBZ'), (u'here', 'RB')]

3.3.7 Parsing

After we are done generating token and tagging them, our final step would be to perform Semantic analysis. For this we need to define a CFG and generate a parse tree accordingly.

For this purpose I have used the Penn Treebank corpus. If the token is a stop-word, we simply add a subtree "(STOP ('stop-word'))" to the parse tree. These subtrees need not be traversed when looking for the review sentiment. Before constructing parse trees I have generated named entities for the processed tweets. One example of named entities for the above tweet is

(S (ORGANIZATION Facebook/NNP) admits/VBZ its/PRP$ app/NN drains/VBZ (ORGANIZATION iPhone/JJ) battery/NN but/CC says/VBZ a/DT fix/NN is/VBZ here/RB)

Notice that the entities Facebook and iPhone are named as ORGANIZATION.

4. RESULTS

The preprocessing has been applied on a variety of smartphones such as iPhone, OnePlus Two, Samsung Galaxy S6. Below are the results obtained after compiling the tweets about the above mentioned devices.

Table 3. Demonstrates the statistics involved in the analysis.

Total Tweets Analyzed	Tweets from BOTS (Gibberish)	Meaningful tweets
56,436	13,248	43,188

After the tweets have been extracted categorization is applied. The tweets are categorized according to the top features that they review.

Table 4. Total numbers of tweets in each category obtained.

Applications	Camera	Screen	Battery
5,966	6,649	5,506	6,131

Table 5. Total number of tweets in each category obtained.

Performance	iOS	Android	Processing and Memory
4,125	6,489	3,965	4,369

Table 6 . Top Tweets before Preprocessing

Applications	Camera	Screen	Battery
"Do you guys know how to make app shortcuts on iphone? Got some trouble figuring it out"	OnePlus 2 Review: It's Powerful, It's Cheap, and great camera resolution!! http://dlvr.it/BxynDl #oneplustwo #oneplus #mobile #oneplustworeview #lifehacker #tech	iPhone 6S nano review: lightning fast, slightly heavier, 3D Touch Screen is great.	One plus two battery life problems are hard to deal with :(... #onePlusTwo #battery
"Happy with the #OnePlusTwo, Sad that it has Bugs.. Hoping it's resolved soon"	I traded my Galaxy S5 for a Galaxy Note 5. Better camera, faster processor, faster charging. It only cost 2 years!	#LG LG V10 vs Samsung Galaxy Note 5: LG's dual-screen handset has better selfie camera.	Facebook addresses severe battery drain issue with updated iOS app
"RT @SkyNews: Facebook admits its app drains iPhone battery but says a fix is here https://t.co/oHDFrSwaGa https://t.co/bnWVUIVjy0"	#LG LG V10 vs Samsung Galaxy Note 5: LG's dual-screen handset has better selfie camera.	sometimes the screen of my #OnePlusTwo turns black and I have to restart the phone, Any recomendations?	#oneplustwo a device that doesn't let u down except for the Battery,barely 6:30 hrs backup.Truly Beast on performance

Table 7 . Top Tweets before Preprocessing (Cont.)

Performance	iOS	Android	Processing and Memory
The bend actually scares me, I know everybody's iPhone 6 does bend eventually but it's so worrying !!	Looking forward to the iOS update. #iOS #iPhone	The cynogen replaced Oxygen OS for android is quite impressive. #OnePlusTwo #Oxygen	The #OnePlusTwo finally arrived for my dad. The packaging is super sleek! Processor seems fast. This one is a #winner
Does anyone have any thoughts on #OnePlusTwo . Especially regarding the heating performance issues?	is my phone broken or is iOS 9 possibly this buggy? why???	One thing I like about #oneplustwo is it handles notifications beautifully! @OnePlus_IN I've used android phns be4 but this phone is unique	LG V10 vs Samsung Galaxy S6 Edge Plus: Samsung's device has faster processor
Happy with my one plus two @OnePlus_IN butter smooth performance	The "Moments" tab in the Twitter iOS update is trash. #iPhone #iOS	@iamkaran24 @dumbb786 The best Mobile for Dual Sim android I recommend Is One Plus Two.	Galaxy Note 5 - Stunning premium design, gorgeous display, power packed processor, Multitasking and stylus.

Table 8. Top Tweets after Preprocessing

Applications	Camera	Screen	Battery
[('Do', 'VB'), ('you', 'PRP'), ('guys', 'VB'), ('know', 'VB'), ('how', 'WRB'), ('to', 'TO'), ('make', 'VB'), ('app', 'JJ'), ('shortcuts', 'NNS'), ('on', 'IN'), ('iphone', 'NN'), ('Got', 'NNP'), ('some', 'DT'), ('trouble', 'NN'), ('figuring', 'VBG'), ('it', 'PRP'), ('out', 'RP')]	[('OnePlus', 'CC'), ('2', 'CD'), ('Review', 'NN'), ('Its', 'PRP'), ('Powerful', 'NNP'), ('It', 'PRP'), ("'s", 'VBZ'), ('Cheap', 'NNP'), ('and', 'CC'), ('great', 'JJ'), ('camera', 'NN'), ('resolution', 'NN'),]	[('iPhone', 'NN'), ('6S', 'CD'), ('nano', 'JJ'), ('review', 'NN'), ('lightning', 'NN'), ('fast', 'RB'), ('slightly', 'RB'), ('heavier', 'JJR'), ('3D', 'CD'), ('Touch', 'NNP'), ('Screen', 'NNP'), ('is', 'VBZ'), ('great', 'JJ')]	[('One', 'CD'), ('plus', 'CC'), ('two', 'CD'), ('battery', 'NN'), ('life', 'NN'), ('problems', 'NNS'), ('are', 'VBP'), ('hard', 'JJ'), ('to', 'TO'), ('deal', 'VB'), ('with', 'IN')]
[('Happy', 'JJ'), ('with', 'IN'), ('the', 'DT'),('Sad', 'NNP'), ('that', 'IN'), ('it', 'PRP'), ('has', 'VBZ'), ('Bugs..', 'VBN'), ('Hoping', 'VBG'), ('it', 'PRP'), ("'s", 'VBZ'), ('resolved', 'VBN'), ('soon', 'RB')]	[('I', 'PRP'), ('traded', 'VBD'), ('my', 'PRP$'), ('Galaxy', 'NNP'), ('S5', 'NNP'), ('for', 'IN'), ('a', 'DT'), ('Galaxy', 'NNP'), ('Note', 'NNP'), ('5', 'CD'), ('Better', 'JJR'), ('camera', 'NN'), ('faster', 'JJR'), ('processor', 'NN'), ('faster', 'RBR'), ('charging', 'VBG'), ('It', 'PRP'), ('only', 'RB'), ('cost', 'VBD'), ('2', 'CD'), ('years', 'NNS')]	[('LG', 'NNP'), ('V10', 'NNP'), ('vs', 'NN'), ('Samsung', 'NNP'), ('Galaxy', 'NNP'), ('Note', 'NNP'), ('5', 'CD'),('LG', 'NNP'), ('dual-screen', 'JJ'), ('handset', 'NN'), ('has', 'VBZ'), ('better', 'JJR'), ('selfie', 'VBN'), ('camera', 'NN')]	[('Facebook', 'NNP'), ('addresses', 'VBZ'), ('severe', 'JJ'), ('battery', 'NN'), ('drain', 'NN'), ('issue', 'NN'), ('with', 'IN'), ('updated', 'JJ'), ('iOS', 'NN'), ('app', 'NN')]

[('Facebook', 'NNP'), ('admits', 'VBZ'), ('its', 'PRP$'), ('app', 'NN'), ('drains', 'VBZ'), ('iPhone', 'JJ'), ('battery', 'NN'), ('but', 'CC'), ('says', 'VBZ'), ('a', 'DT'), ('fix', 'NN'), ('is', 'VBZ'), ('here', 'RB')]	[('LG', 'NNP'), ('V10', 'NNP'), ('vs', 'NN'), ('Samsung', 'NNP'), ('Galaxy', 'NNP'), ('Note', 'NNP'), ('5', 'CD'), ('LG', 'NNP'), ('dual-screen', 'JJ'), ('handset', 'NN'), ('has', 'VBZ'), ('better', 'JJR'), ('selfie', 'VBN'), ('camera', 'NN')]	[('sometimes', 'RB'), ('the', 'DT'), ('screen', 'NN'), ('of', 'IN'), ('my', 'PRP$'), ('OnePlusTwo', 'NNP'), ('turns', 'VBZ'), ('black', 'JJ'), ('and', 'CC'), ('I', 'PRP'), ('have', 'VBP'), ('to', 'TO'), ('restart', 'VB'), ('the', 'DT'), ('phone', 'NN'), ('Any', 'DT'), ('recomendations', 'NNS')]	[('a', 'DT'), ('device', 'NN'), ('that', 'WDT'), ('does', 'VBZ'), ("n't", 'RB'), ('let', 'VB'), ('u', 'VB'), ('down', 'RP'), ('except', 'IN'), ('for', 'IN'), ('the', 'DT'), ('Battery', 'NNP'), ('barely', 'RB'), ('6:30', 'CD'), ('hrs', 'NN'), ('backup', 'NN'), ('Truly', 'NNP'), ('Beast', 'NNP'), ('on', 'IN'), ('performance', 'NN')]

Table 9. Top Tweets after Preprocessing

Performance	iOS	Android	Processing and Memory
[('The', 'DT'), ('bend', 'NN'), ('actually', 'RB'), ('scares', 'VBZ'), ('me', 'PRP'), ('I', 'PRP'), ('know', 'VBP'), ('everybody', 'NN'), ("'s", 'POS'), ('iPhone', 'NN'), ('6', 'CD'), ('does', 'VBZ'), ('bend', 'VB'), ('eventually', 'RB'), ('but', 'CC'), ('it', 'PRP'), ("'s", 'VBZ'), ('so', 'RB'), ('worrying', 'JJ')]	[('Looking', 'VBG'), ('forward', 'RB'), ('to', 'TO'), ('the', 'DT'), ('iOS', 'NN'), ('update', 'NN')]	[('The', 'DT'), ('cynogen', 'NN'), ('replaced', 'VBD'), ('Oxygen', 'NNP'), ('OS', 'NNP'), ('for', 'IN'), ('android', 'NN'), ('is', 'VBZ'), ('quite', 'RB'), ('impressive', 'JJ')]	[('The', 'DT'), ('finally', 'RB'), ('arrived', 'VBD'), ('for', 'IN'), ('my', 'PRP$'), ('dad', 'NN'), ('The', 'DT'), ('packaging', 'NN'), ('is', 'VBZ'), ('super', 'JJ'), ('sleek', 'JJ'), ('Processor', 'NNP'), ('seems', 'VBZ'), ('fast', 'RB'), ('This', 'DT'), ('one', 'NN'), ('is', 'VBZ'), ('a', 'DT'), ('winner', 'NN')]
[('Does', 'NNP'), ('anyone', 'NN'), ('have', 'VB'), ('any', 'DT'), ('thoughts', 'NNS'), ('on', 'IN'), ('Especially', 'RB'), ('regarding', 'VBG'), ('the', 'DT'), ('heating', 'NN'), ('performance', 'NN'), ('issues', 'NNS')]	[('is', 'VBZ'), ('my', 'PRP$'), ('phone', 'NN'), ('broken', 'NN'), ('or', 'CC'), ('is', 'VBZ'), ('iOS', 'JJ'), ('9', 'CD'), ('possibly', 'RB'), ('this', 'DT'), ('buggy', 'NN'), ('why', 'WRB')]	[('One', 'CD'), ('thing', 'NN'), ('I', 'PRP'), ('like', 'VBP'), ('about', 'RB'), ('is', 'VBZ'), ('it', 'PRP'), ('handles', 'VBZ'), ('notifications', 'NNS'), ('beautifully', 'RB'), ('I', 'PRP'), ("'ve", 'VBP'), ('used', 'VBN'), ('android', 'JJ'), ('phns', 'NN'), ('be4', 'NN'), ('but', 'CC'), ('this', 'DT'), ('phone', 'NN'), ('is', 'VBZ'), ('unique', 'JJ')]	[('LG', 'NNP'), ('V10', 'NNP'), ('vs', 'NN'), ('Samsung', 'NNP'), ('Galaxy', 'NNP'), ('S6', 'NNP'), ('Edge', 'NNP'), ('Plus', 'NNP'), ('Samsung', 'NNP'), ("'s", 'POS'), ('device', 'NN'), ('has', 'VBZ'), ('faster', 'JJR'), ('processor', 'NN')]
[('Happy', 'JJ'), ('with', 'IN'), ('my', 'PRP$'), ('one', 'CD'), ('plus', 'CC'), ('two', 'CD'), ('butter', 'NN'), ('smooth', 'VBD'), ('performance', 'NN')]	[('The', 'DT'), ('Moments', 'NNPS'), ('tab', 'NN'), ('in', 'IN'), ('the', 'DT'), ('Twitter', 'NNP'), ('iOS', 'NN'), ('update', 'NN'), ('is', 'VBZ'), ('trash', 'JJ')]	[('The', 'DT'), ('best', 'JJS'), ('Mobile', 'NN'), ('for', 'IN'), ('Dual', 'NNP'), ('Sim', 'NNP'), ('android', 'NN'), ('I', 'PRP'), ('recommend', 'VBP'), ('Is', 'VBZ'), ('One', 'CD'), ('Plus', 'CC'), ('Two', 'CD')]	[('Galaxy', 'NNP'), ('Note', 'NNP'), ('5', 'CD'), ('-Stunning', 'VBG'), ('premium', 'NN'), ('design', 'NN'), ('gorgeous', 'JJ'), ('display', 'NN'), ('power', 'NN'), ('packed', 'VBD'), ('processor', 'NN'), ('Multitasking', 'NNP'), ('and', 'CC'), ('stylus', 'NN')]

5. FUTURE SCOPE

In this paper, I have taken into consideration, English as the primary source of language. However, NLTK supports multiple languages and thus, we can expand this work into a multi-lingual frame. The review accuracy can be increased by considering tweets from multiple languages such as German, French or Spanish.The algorithm used in this process does not take into consideration the concept of sarcastic tweets. For example, a tweet such as “The battery dries up in the blink of an eye. #great“ expresses sarcasm. However, this algorithm may fail to take it into consideration. We can further improvise on this system by modifying the algorithm to understand sarcasm.My next task would be based on using these categorized and processed tweets to classify and extract sentiment.

6. CONCLUSION

In this paper, I have focused on tweet categorization and preprocessing. I have been able to successfully categorize the tweets depending upon the various features of the smartphones they are reviewing about.

This paper demonstrates that the dataset can be reduced by discarding the gibberish tweets from bots and focusing on the human tweets. This reduces processing complexity and gives a cleaner dataset.

The analysis has been carried out for a wide variety of smartphone manufacturers and a combined result has been demonstrated. Using the concept of categorized review analysis, the manufacturers can be given a segmented and detailed review of the various features pertaining to their device. This gives them more accurate analysis regarding the strengths and weaknesses of many features in particular.

7. AKNOWLEDGEMENT

I would like to whole heartedly thank, our professor, Mr. Pramod Mali, and our mentor Mr. Shital Patil in helping me find the right directions and important resources required to complete this project. A special thanks to Mrs. Aaradhna Deshmukh, for being my mentor and motivating me. I would like to thank my institiution and Head Of Computer Department, Mr. P.N. Mahalle for giving me access to the required resources and backing me all the while. Without the kindness of my institute and blessing of my professors, I would not have been able to complete this project.

8. REFERENCES

[1] Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O'Reilly Media Inc.

[2] Sumian Peng, "Research on Data Preprocessing Process in the Web Log Mining," IEEE Conference. Information Science and Engineering (ICISE), 2009 1st International Conference, pp. 942 - 945, 26-28 Dec. 2009

[3] Sun Binli,"Research on data-preprocessing for construction of university information systems", IEEE Conference. Computer Application and System Modeling (ICCASM), 2010 International Conference Volume:1. 22-24 Oct. 2010

[4] Sudheer Reddy, K," An effective data preprocessing method for Web Usage Mining", IEEE Conference Information Communication and Embedded Systems (ICICES), 2013 International Conference. 7 - 10 Feb. 2013.

[5] Wei Jianping, "Research on Data Preprocessing in Supermarket Customers Data Mining", IEEE Conference Information Engineering and Computer Science (ICIECS), 2010 2nd International Conference, pp. 1 – 4, 25-26 Dec. 2010.

[6] Qing Ang, "Explored research on data preprocessing and mining technology for clinical data applications", IEEE Conference. Information Management and Engineering (ICIME), 2010 The 2nd IEEE International Conference, pp 327 – 330.16-18 April 2010.

[7] García, Salvador, "Data Preprocessing in Data Mining", Intelligent Systems Reference Library.

[8] Wei Zhang, "Data Preprocessing for Web Data Mining", Advances in Electronic Commerce, Web Application and Communication, pp 303-307.

[9] Sharon Christa, "Data Preprocessing Using Intelligent Agents", Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA), pp 197-204.

[10] Gaston L'Huillier, "Web Usage Data Pre-processing", Advanced Techniques in Web Intelligence-2, pp 11-34

[11] Amir R. Razavi, "A Data Pre-processing Method to Increase Efficiency and Accuracy in Data Mining", 10th Conference on Artificial Intelligence in Medicine, pp 434-443.