Open Journal of Physical Science (OJPS)
ISSN: 2734-2123
Article Details:
DOI: 10.52417/ojps.v2i2.276
Article Ref. No.: OJPS0202002-261
Volume: 2; Issue: 2, Pages: 10 - 17 (2021)
Accepted Date: 20th October, 2021

Open Journals Nigeria (OJN)
Open Access | Bi-annual | Peer-reviewed
www.openjournalsnigeria.org.ng
editorial@openjournalsnigeria.org.ng

RESEARCH ARTICLE **OJPS0202002-261**

A NATURAL LANGUAGE PROCESSING APPROACH TO DETERMINE THE POLARITY AND SUBJECTIVITY OF IPHONE 12 TWITTER FEEDS USING TEXTBLOB

[*1]Abubakar, B. U. & [2]Uppin, C.

*[*1] Department of Computer Science, Faculty of Computing and Applied Sciences, Baze University, Jabi Abuja, Nigeria*
[2]Department of Computer Science, Faculty of Computing and Applied Sciences, Baze University, Jabi Abuja, Nigeria
***Corresponding Author's E-mail:** usman.abu@bazeuniversity.edu.ng, cv.uppin@bazeuniversity.edu.ng*

ABSTRACT

Sentiment analysis and opinion mining is a branch of computer science that has gained considerable growth over the last decade. This branch of computer science deals with determining the emotions, opinions, feelings amongst others of a person on a particular topic. Social media has become an outlet for people to voice out their thoughts and opinions publicly about various topics of discussion making it a great domain to apply sentiment analysis and opinion mining. Sentiment analysis and opinion mining employ Natural Language Processing (NLP) in order to fairly obtain the mood of a person's opinion about any specific topic or product in the case of an ecommerce domain. It is a process involving automatic feature extractions by mode of notions of a person about service and it functions on a series of different expressions for a given topic based on some predefined features stored in a database of facts. In an ecommerce system, the process of analyzing the opinions of customers about products is vital for business growth and customer satisfaction. This proposed research will attempt to implement a model for sentiment analysis and opinion mining on Twitter feeds. In this paper, we address the issues of combining sentiment classification and the domain constraint analysis techniques for extracting opinions of the public from social media. The dataset that was employed in the paper was gotten from Twitter through the tweepy API. The TextBlob library was used for the analysis of the tweets to determine their sentiments. The result shows that more tweets were having a positive subjectivity and polarity on the subject matter.

Keywords: *Cybercrime, Deep-Learning, Digital Forensic, Denial of Service Attacks, Network-monitoring system, Network Forensics*

Acronyms:
NLP - Natural Language Processing
NLTK - Natural Language ToolKit

INTRODUCTION

A smart phone is a mobile phone that performs many of the functions of a computer, typically having a touchscreen interface, internet access, and an operating system. An iphone12 is a smart phone produced by Apple Tech company. With the explosion of the semantic web (Web 3.0) and the increasing number of social networks, forums, blogs amongst others, people express their ideas about various topics on internet platforms. In the business sector, opinions from users which are shared on blogs or forums are useful in business growth, customer satisfaction and product ratings. An ecommerce business can utilize feeds posted by users on the internet on products by polling social network sites, forums, blogs to conduct market research and determine a general perspective of customer sentiment and opinion about their product listing. There are obvious economic benefits that sentiment analysis brings forth with and thus, organizations are utilizing solutions to analyze the sentiment of customers on their product.

Sentiment analysis is a branch of computer science that is geared towards automatic processing of the information provided in a text in order to carry out polarity classification. Polarity classification is the process used to determine the neutrality, positivity or negativity from analysis of textual content (lexalytics, n.d.).

Sentiment analysis and opinion mining employs Natural Language Processing (NLP) in order to fairly obtain the mood of a person's opinion about any specific topic. It is a process involving automatic feature extractions by mode of notions of a person about a service and it functions on a series of different expressions for a given topic based on some predefined features stored in a database of facts (Yse, 2019). This proposed research will attempt to implement a sentiment analyzer on twitter feeds for iPhone12 using python TextBlob library. The reason iphone12 was chosen for the study is because of the current argument in the tech company as to the performance improvement of the iphone12 over the iphone11.

SENTIMENT ANALYSIS

Sentiment analysis is the process of determining whether a piece of writing is positive, negative or neutral (lexalytics, n.d.).

NATURAL LANGUAGE PROCESSING

This is a field of Artificial Intelligence that gives the machines the ability to read, understand and derive meaning from human texts (Yse, 2019).

ARTIFICIAL INTELLIGENCE

This is the simulation of human intelligence processes by machines (Margaret Rouse, 2020).

AIM

The aim of this research is to propose a scheme as a sentiment analyzer that will identify the sentiments in twitter feeds sentence-level on iPhone products.

OBJECTIVES

1. To invoke tweepy OAuthHandler function to retrieve dataset
2. To preprocess tweet emoji's using regular expressions
3. To classify sentiments of tweets as positive, negative or neutral.
4. To visualize the sentiments on a scatter plot.

RELATED WORKS

Hii (2019) focused on negation handling which is an aspect of sentiment analysis. This technique is employed in order to detect the sentiments of a particular input with the use of certain words. These words are associated with negativity such as never, not, no. The technique that was used by the researchers was the concept of negation resolution in which there is an assumption of working negation scope detector been in existence. Negation resolution is a technique in which a machine learning algorithm tends to decide and hence, performs some adjustments that would be needed to track the negation effect on input. A novel scheme was proposed in which meaningful components were incorporated into negation handling. A validation through mathematical calculations of information gain to determine specificity was first done and later on it was compared with previous negation resolution techniques in a sentiment analysis pipeline (Hii, 2019). The specificity approach offered higher information gain and higher performance over non-semantic approaches when tested on product reviews of selected topics of interest (Hii, 2019)."

Junyi Jessy Li (2015) proposed a prediction system to predict the specificity of sentences by exploiting only certain features with minimum processing requirements. These are then proceeded to be trained in a semi-supervised machine learning environment. The system proposed performed sentiment analysis without needing syntactic parsing nor part of speech tagging. The system, as opposed to state-of-the-art. (Junyi Jessy Li, 2015) developed a tool called SPECITELLER that pinpointed the need for specificity in opinion mining by depicting its usefulness for highlighting input sentences that are in need of simplification.

Anish Kumar Varudharajulu (2019) proposes a method aimed at consolidating customer opinions gotten from internet on websites like twitter, Instagram amongst others. This method tends to focus on the analysis of certain clustering algorithms and classification algorithms at more than one granularity levels. This aimed at monitoring and measuring customer satisfaction. The approach tends to be an automated sentiment analyzer whereby certain machine learning algorithms play a helpful role in identification of knowledge and its analytics (Anish Kumar Varudharajulu, 2019). The knowledge base was approved by machine learning experts and the dataset was fed to the system for effective analysis of sentiments. "The system identifies opinion expressions as phrases containing opinion words, opinionated features and also opinion modifiers. These expressions are categorized as positive, negative or neutral. Opinion expressions are identified and categorized using localized linguistic techniques (Anish Kumar Varudharajulu, 2019).

Opinions can be congregated at any desired level of specificity i.e. feature level or product level, user level or service level and so on. It has been found that J48 classification algorithm and simple k-means clustering algorithm are most suitable for restaurant customer reviews. This Indicates that J48 is the beat classifier with highest accuracy and Simple K- mean is used for a smaller number of cluster when compared with other techniques (Anish Kumar Varudharajulu, 2019).

Mai & Le (2021) made use of comments from social media platforms (such as YouTube) to examine public opinion toward products. They proposed a novel framework for automatically collecting, filtering, and analyzing comments from YouTube for a given product. First, they devise a classification scheme to select relevant and high-quality comments from retrieval results. These comments are then analyzed in a sentiment analysis, where they introduced a joint approach to perform a combined sentence and aspect level sentiment analysis. The study aimed to achieve the following:

1. Capture the mutual benefits between these two tasks, and
2. Leverage knowledge learned from solving one task to solve another. Experiment results on our dataset show that the joint model achieves a satisfactory performance and outperforms the separate one on both sentence and aspect levels.

The framework does not require feature engineering efforts or external linguistic resources; therefore, it can be adapted for many languages without difficulties (Mai & Le, 2021).

Among various neural architectures applied for sentiment analysis, long short-term memory (LSTM) models and its variants such as gated recurrent unit (GRU) have attracted increasing attention. Although these models are capable of processing sequences of arbitrary length, using them in the feature extraction layer of a DNN makes the feature space high dimensional. Another drawback of such models is that they consider different features equally important. To address these problems, (Basiri, Nemati, Abdar, Cambria, & Acharya, 2021) proposed an Attention-based Bidirectional CNN-RNN Deep Model (ABCDM). By utilizing two independent bidirectional LSTM and GRU layers, ABCDM extracts both past and future contexts by considering temporal information flow in both directions. Also, the attention mechanism is applied on the outputs of bidirectional layers of ABCDM to put more or less emphasis on different words. To reduce the dimensionality of features and extract position-invariant local features, ABCDM utilizes convolution and pooling mechanisms. The effectiveness of ABCDM is evaluated on sentiment polarity detection which is the most common and essential task of sentiment analysis. Experiments were conducted on five review and three Twitter datasets (Basiri, Nemati, Abdar, Cambria, & Acharya, 2021).

Hate speech detection on Twitter is critical for applications like controversial event extraction, building AI chatterbots, content recommendation, and sentiment analysis (Badjatiya, Gupta, Gupta, & Varma, 2017). (Badjatiya, Gupta, Gupta, & Varma, 2017) defined this task as being able to classify a tweet as racist, sexist or neither. The complexity of the natural language constructs makes this task very challenging. (Badjatiya, Gupta, Gupta, & Varma, 2017) performed extensive experiments with multiple deep learning architectures to learn semantic word embeddings to handle this complexity. Their experiments on a benchmark dataset of 16K annotated tweets show that such deep learning methods outperform state-of-the-art char/word n-gram methods (Badjatiya, Gupta, Gupta, & Varma, 2017).

In the paper, they experimented with multiple classifiers such as Logistic Regression, Random Forest, SVMs, Gradient Boosted Decision Trees (GBDTs) and Deep Neural Networks (DNNs). The feature spaces for these classifiers are in turn defined by task-specific embeddings learned using three deep learning architectures: FastText, Convolutional Neural Networks (CNNs), Long Short-Term Memory Networks (LSTMs). As baselines, we compare with feature spaces comprising of char n-grams."

Main contributions of the paper are as follows (Badjatiya, Gupta, Gupta, & Varma, 2017): “

1. Investigated the application of deep learning methods for the task of hate speech detection.
2. Explored various tweet semantic embeddings like char n-grams, word Term FrequencyInverse Document Frequency (TF-IDF) values, Bag of Words Vectors (BoWV) over Global Vectors for Word Representation (GloVe), and task-specific embeddings learned using FastText, CNNs and LSTMs.
3. The methods proposed from their conclusion indicates that it beats state of-the-art methods by a margin (∼18 F1 points better)”

METHODOLOGY

Data Source and Collection

CREATE TWITTER DEVELOPER APP

For the purpose of this study, a twitter developer app must be created in order to get access to the live tweets from twitter. To target the API to pull tweet data, we require Oauth credentials that must be passed with each request made from our machine learning application. The authentication credentials are in a variety of forms depending on the specific endpoint authentication architecture. In our instance, we require four credentials that are specific to our user profile which can be used for making the API request on our behalf

1. Consumer Key (API key)
2. Secret key for consumer
3. OAuth Token for Access (Public and Secret)

CONNECT TWITTER DEVELOPER APP TO GOOGLE COLAB

In order to interface our machine learning application with twitter, we need to invoke the tweepy OAuthHandler function and pass in our consumer key and consumer secret as arguments. The function returns an auth object. We further invoke a function called set_access_token on the returned auth object and pass our access token and access token secret as arguments. After properly configuring and authenticating our credentials with the OAuthHandler functions, we finally invoke the API method of the tweepy module so as to get an API object which can be used to make http calls to twitter endpoint with details of tweets needed.

LOAD TWEETS

To load tweets, we made an API call to the twitter endpoint to get tweets. The tweets gotten were not live-streamed. They were tweets between a specific range of date as specified in the API call to twitter endpoint. The data that was returned from twitter endpoint was a normal python dataframe in form of rows and columns. Furthermore, to avoid unnecessary data, we limited the returned data frame to 500 instances of tweets. A filter is also applied on the search parameter to exclude any kind of retweets and to return tweets since a certain date.

EXPERIMENTAL TOOLS AND RESULTS

CLEAN TWEETS

The dataset requires preprocessing as it is not clean enough to be fed into a machine learning model. To do this preprocessing, we applied a regular expression python model and provide a series of patterns to be matched and removed if found in the text column of our dataset. The values to be removed are values like #, @, numbers, URL's and so on. These are words that have no value on the polarity of our sentiments and thus, they simply make our dataset unclean.

After cleaning the dataset, we need to associate each text/tweet with a polarity and subjectivity. A pre-trained sentiment analyzer library called TextBlob was used to determine the sentiments of tweets. The TextBlob module was used to iteratively calculate the sentiment polarity and sentiment subjectivity of each tweet and append the result set gotten into our original data frame.

The textblob returns the polarity as a floating-point number which lies in the range of [-1,1] where 1 means positive statement and -1 means a negative statement. Subjective sentences generally refer to personal opinion, emotion or judgment whereas objective refers to factual information. Subjectivity is also a float which lies in the range of [0,1].

ANALYZING TWEETS

To analyze the tweets and get a little insight on the occurrence of words and their relevance, wordcloud module was used to get an insight on the words that occurred more. Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. For generating word cloud in Python, several modules were used: matplotlib, pandas and wordcloud.

Figure 1 shows the codes and an intuitive visual representation of the data which was done using matplotlib to create a scatter plot of our data points.

CONCLUSION

On the completion of the research, the aims and objectives stated previously have been achieved. A machine learning approach was designed and implemented for sentiment analysis on tweets from twitter. The model was trained on iphone12 reviews. The sentiments of the tweets were analyzed based on the polarity and subjectivity of the sentences using a Natural Language ToolKit (NLTK). The sentiment analyzer was written using python programming language and google colabs as the IDE. The main development library that was used in developing the model was the TextBlob and the twitter developer API for making http request to retrieve tweets from twitter.

The final end goal of this study, as seen from the scatter plot, indicated most of the data points in red are more to the right of the midpoint. This shows that most of the tweets have a positive sentiment which concludes that iPhone 12 products have a more positive review.

SUGGESTION FOR FURTHER STUDIES

As seen from the implementation, the model has no Graphical User Interface (GUI). Furthermore, the model could be built and embedded into a web or mobile GUI which will enable for dynamically retrieving and analyzing tweets with different filter parameter.

```
plt.figure(figsize = (8, 8), facecolor = None)
for i in range(0,tweet_text.shape[0]):
  plt.scatter(tweet_text['polarity'][i], tweet_text['subjectivity'][i], color='red')
plt.title("Sentiment Analysis Scatter Plot of iphone12")
plt.xlabel("Polarity")
plt.ylabel("Subjectivity")
plt.show()
```

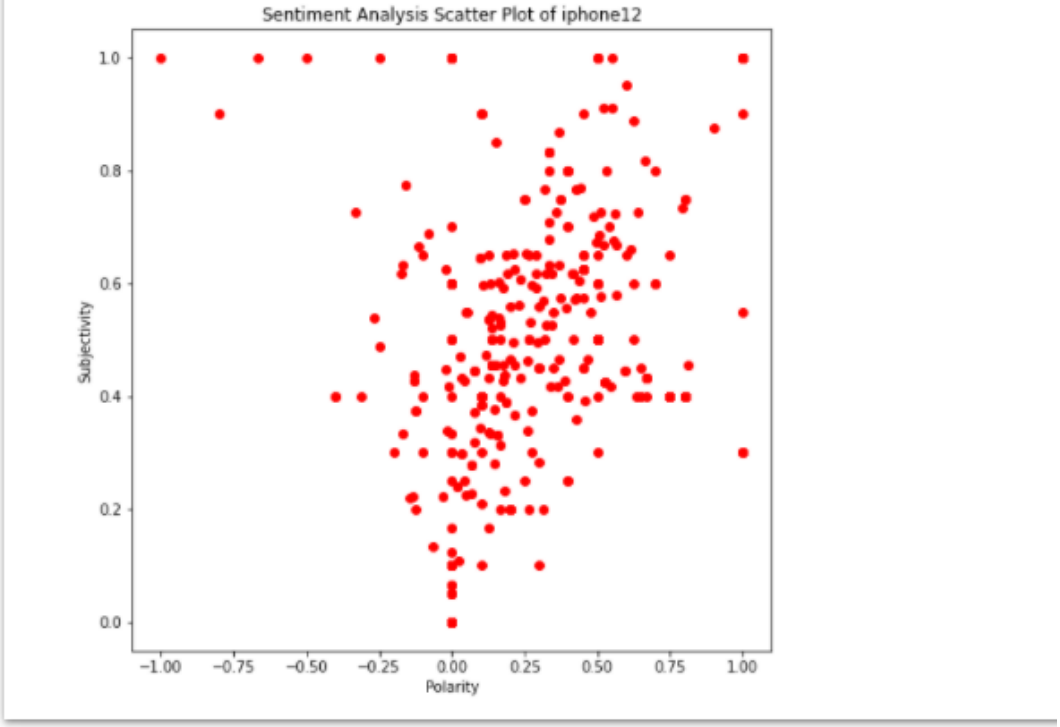

Figure 1: Sentiment Analysis Scatter Plot of iphone12

From figure 1, we notice that on the 2-D plane, the data points in red are more to the right of the midpoint which shows that most of the tweets have a positive sentiment.

REFERENCES

Anish Kumar Varudharajulu, Y. M. (2019). Data Mining Algorithms for a Feature-Based Customer Review Process Model with Engineering Informatics Approach. *Journal of Physics Conference Series*.

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017, June 1). Deep Learning for Hate Speech Detection in Tweets. *International World Wide Web Conference Committee (IW3C2).* Australia.

Basiri, M. E., Nemati, S., Abdar, M., Cambria, E., & Acharya, R. (2021, February). ABCDM: An Attention-based Bidirectional CNN-RNN Deep Model for sentiment analysis. *Future Generation Computer Systems*, pp. 279-294.

Hii, D. (2019). *Using Meaning specificity to aid negation handling in Sentiment analysis.* Retrieved November 22, 2020, from https://www.socsci.uci.edu/~lpearl/CoLaLab/papers/Hii2019_MeaningSpecNegSent.pdf

Junyi Jessy Li, A. N. (2015). Fast and Accurate Prediction of Sentence Specificity. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence.* Philadelphia.

Mai, L., & Le, B. (2021, May). Joint sentence and aspect-level sentiment analysis of product comments. *Annals of Operations Research*, pp. 493-513.

Vajjala, S., Majumder, B., Gupta, A., & Surana, H. (2020, January, 15). *Practical Natural Language Processing.* O'Reilly Media, Inc. Retrieved November 2020, 21, from https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1