JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4767-4775
https://publishoa.com
ISSN: 1309-3452

A Sentiment Analysis to Forecast the Dimensions of Well Being during Pandemic Outbreak Using Machine Learning Algorithms

Dr. P. Tamije Selvy[1], Ms. G. Nivedhitha[2] and Ms. K. Saranya[3]
[1-3]Department of Computer Science and Engineering,Sri Krishna College of Technology,Coimbatore.

Received: 2022 March 15; **Revised:** 2022 April 20; **Accepted:** 2022 May 10

Abstract

In recent times, our world has been massively affected by a global pandemic. Due to the swift increase in the infection and the death frequency, it has been caused an extensive public health crisis globally and activated some issues such as economic catastrophe, mental and physical worries and so on. In the course of this period, the internet community involvement and dealings rise vigorously and people are able to share one's perspectives and state of wellbeing. This paper focuses mainly on the dimensions of well-being of every individual during the pandemic outbreak. Initially, from the user-generated content (UGC) on social platform, we can examine the public's thoughts and sentiments on different aspects such as health grade, concerns and awareness about the pandemic. Further, the analysis is done based on the supervised machine learning approach. The accuracy of the algorithm was around 93%.Through this research work, health organizations and volunteers can better assess and understand the public's needs in order to convey appropriate and effective information. It can also eventually assist in developing health interposition strategies and design operative drives based on public insights.

Keywords: Sentiment Analysis, Machine Learning, eXtreme Gradient Boosting.

1. Introduction

Sentiment analysis also referred as opinion mining, which integrates natural language processing (NLP) and machine learning techniques to assign weighted sentiment scores to the entities, topics, themes and categories within a sentence or phrase[1]. It has attracted a lot of attention from researchers over time and has seen radical changes in the way the analysis is done to gather and extract the emotion that has been depicted directly or indirectly by humans when stating themselves on social media boards [2]. It is used to comprehend the perception of the people about the pandemic and also to know what range people livelihood is affected [5]. This analysis scrutinizes the problem of studying texts, like posts and reviews, uploaded by users on microblogging platforms, forums, and electronic businesses[11], regarding the opinions they have about this worldwide pandemic outbreak. In this research paper, the main objective is to obtain a better understanding of the social opinions and perspectives about their wellbeing during the pandemic times and how it has changed people's thinking over the past few months. The approach mainly consists of the following steps: Gathering the datasets, from standard repository, which is the foremost step in sentiment analysis. Then, data preparation and analysis, vectorization, model training and prediction were done. As a result of this study we can determine how people are affected on basis of dimensions of wellbeing and can take necessary steps by the government and health departments. The remaining paper is discussed as follows: Background works, Methodology, Performance Analysis, Applications, Conclusion and References.

2. Background Works

This section deals with literature study about sentiment analysis .With the major increment in the amount of data generated online, the field has attracted a significant number of researchers to invest in the study of social media records and to make the most out of the information available. In recent times, sentiment analysis has involved a substantial amount of work and is continuing to grow at a rapid pace.

Mohammad Abu Kausar .et al, [3] proposed a study which aims to capture, process and evaluate people's emotional state within the certain timeframe on the tweets posted on twitter. The process is depicted as follows. Initially, Collection of the tweets were done through Twitter API using RTweet package in R programming was used. Several hashtags were used for collecting the tweets. Next stage is to preprocess the tweets by data cleaning. Then, calculate the sentiment using syuzhet package and analyze the result. In short, The collected tweets will be used, preprocessed and applied with text mining algorithms for performing the sentiment analysis. This study affords a good analysis on sentiments and mind sets of individuals on Covid-19 and enabled to recognize that almost the same level of thinking of individuals all over the world.

Mrityunjay Singh .et al, [6] performed sentiment analysis with the aid of the BERT model on the twitter data sets. The BERT model is used for emotion classification. The sense of a word in a given sentence depends on the other words surrounding it. It feeds all input at once to handle dependencies among words. In this research work, implementation of sentiment analysis were done on two data sets; At the initial stage, one kind of data set is collected by tweets done by people from all over the world, and the other kind of data set contains the tweets done by people of India. The accuracy of the emotion classification from the GitHub repository have been validated and the performance of the model have been increased.

Amrita Mathur Purnima**,** [4] executed sentiment analysis based on lexical oriented method which aims to fetch tweets from dataset and after applying prepossessing, classification of tweets have been done into positive and negative and further classify into six emotions using lexical oriented method using R. There are many libraries, dictionaries and packages available in R to assess emotions prevalent in a text. Three of the general purpose lexicons used in the paper are Bing, AFINN and NRC (from the text data package). For handling the cleaned information the inbuilt sentiment analyzer is utilized in R, which utilizes the NRC emotion dictionary reference to figure out the proximity of six emotions. A complete score is determined for every supposition and plotted utilizing ggplot2 library.

3. Methodology

This section deals with the methodology that has been implemented in this research paper. Sentiment analysis has become an dynamic research area due to the accessibility of many opinionated data through amplified activity in all the social sites. Due to unforeseen occurrence of pandemic, it has been severely affecting people all over the world, and hence there is a compulsion to analyze the opinion of people on the pandemic. Here, the sentiment analysis plays a major role where the people opinion can be segregated into three different segments.Andthis work helps to understand the people perception about pandemic and its impact on the public. This research work aimed at analyzing the sentiments to predict the impact of pandemic on the users wellbeing.The preliminary stage is to collect the datasets. Once the datasets has been gathered, they are pre-processed,analysed,extracted and a model is developed which is effective for detecting the actual sentiment behind a data's related to pandemic.All the steps performed in this study are shown in Fig. 1.

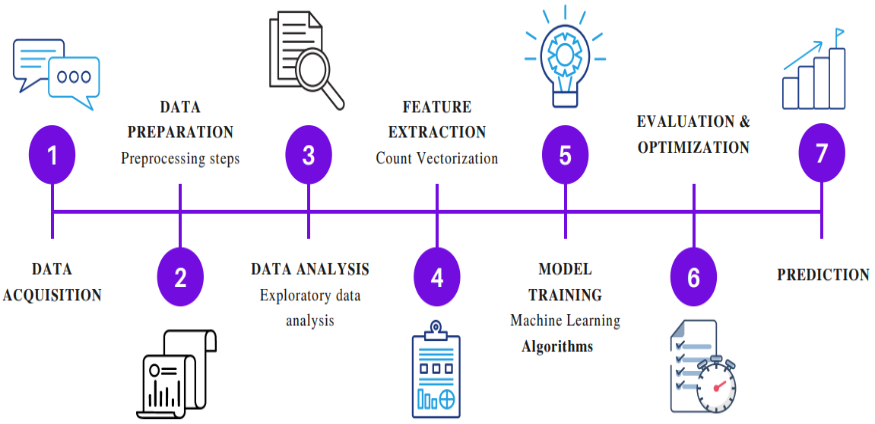


Fig.1. Schematic Representation of Proposed System

A. Data Acquisition

To create a machine learning model, the primary entity is a dataset which is required to execute the problem statement. Data Acquisition is a process of gathering the dataset. A dataset is a structured collection of data. There are enormous data available to run analytics platform. In this stage, public dataset correlated to pandemic are gathered. This data set contains of nearly 4000 records with various columns which is obtained from standard kaggle dataset repository for machine learning. The dataset is represented in CSV format with a file size of 660kb.

B. Data Preparation

The data, gathered was never in the proper format for model training. Here, Data preprocessing plays a major role in removing the redundant variables, duplicate values, missing values and so on. By eliminating such irregularities is very imperative because they may lead to wrongful calculations and predictions. Thus, this stage makes it suitable for a machine learning model which also raises the accuracy and efficiency of a respective model. Below diagram (Fig.2.) depicts the workflow of preprocessing steps.

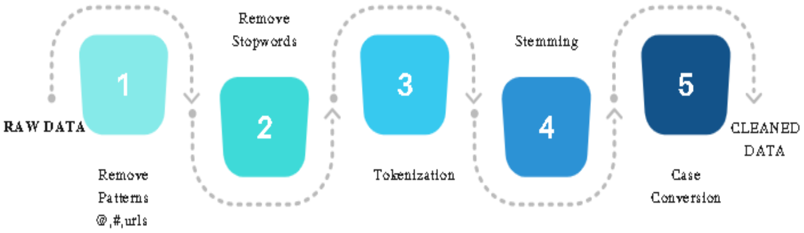
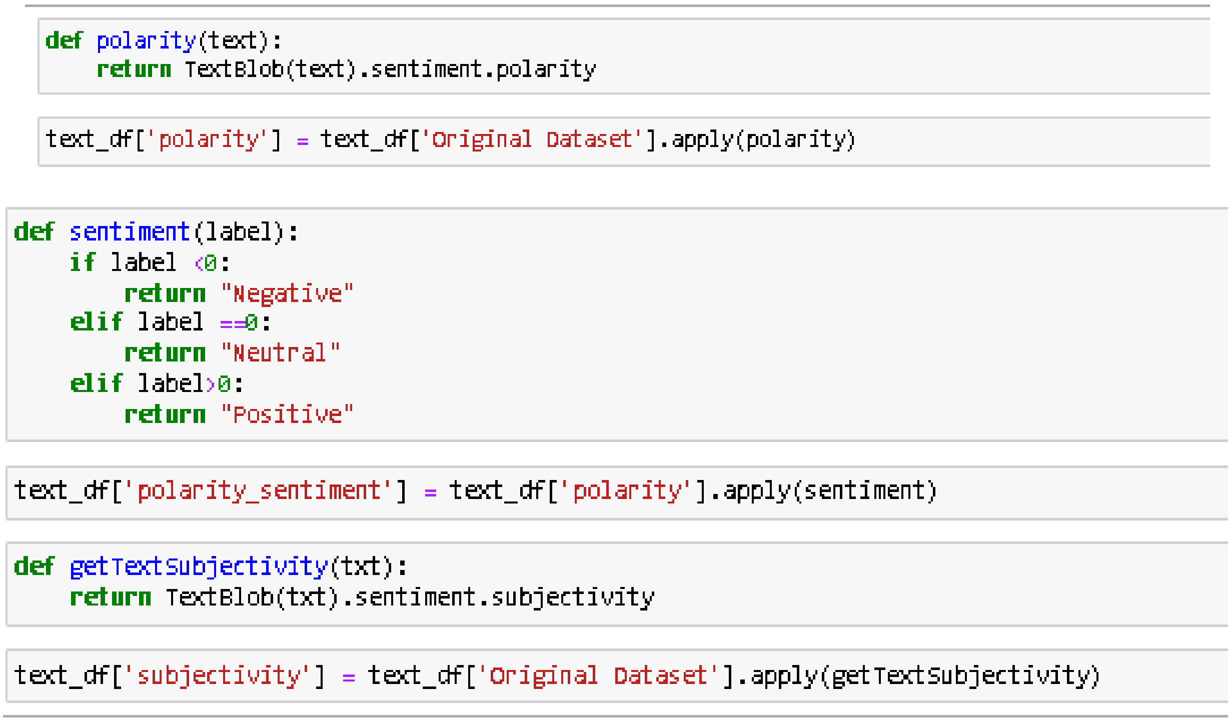


Fig.2. Workflow of Preprocessing

C. Data Analysis

This stage is the conceptualized stage of machine learning. The main purpose of EDA (Exploratory Data Analysis) [12] is to comprehend the patterns and trends in the data. At this point, all useful insights are drawn and correlations between the variables are better understood. It helps to determine how best to manipulate data sources to get the answers we need, making it easier to discover patterns, spot anomalies, test a hypothesis, or check assumptions. The batches of pre-processed tweets were passed to Text Blob[1,8,9] for sentiment score calculation, and then, polarity, and subjectivity were found depending upon the sentence and stored separately for further analysis.Fig.3. shows the snippet for the polarity and subjectivity calculation.



```
def polarity(text):
    return TextBlob(text).sentiment.polarity
```

```
text_df['polarity'] = text_df['Original Dataset'].apply(polarity)
```

```
def sentiment(label):
    if label <0:
        return "Negative"
    elif label ==0:
        return "Neutral"
    elif label>0:
        return "Positive"
```

```
text_df['polarity_sentiment'] = text_df['polarity'].apply(sentiment)
```

```
def getTextSubjectivity(txt):
    return TextBlob(txt).sentiment.subjectivity
```

```
text_df['subjectivity'] = text_df['Original Dataset'].apply(getTextSubjectivity)
```

Fig.3. Calculation of polarity and subjectivity score

Fig.4. represents the distribution of sentiments in the dataset (i.e.) positive, negative and neutral with respective percentage.

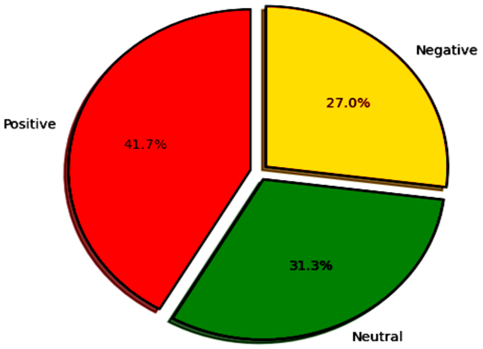


Fig.4. Distribution of Sentiments

Fig.5. depicts the categories of wellbeing with respect to their sentiments.

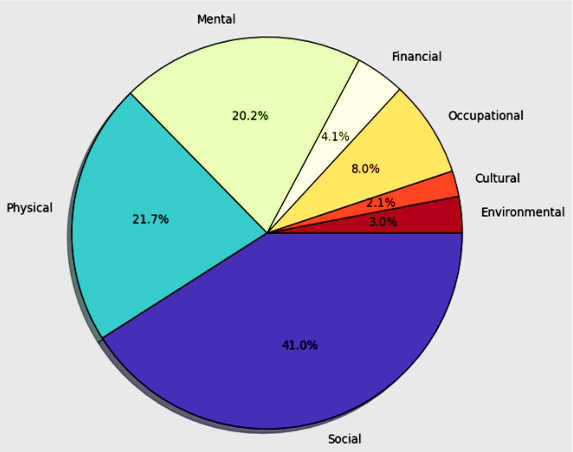


Fig.5. Distribution of categories

D. Feature Extraction

During this phase, various features are extracted from the dataset using vectorization techniques to construct a good classifier. In this stage, n-gram count vectorization[2,10] have been used to convert a given text into a vector on the source of the frequency (count) of each word that occurs in the entire text, while preserving the information in the original data set. This n-gram vectorizer uses parameter range

ngram-range = (a,z)

where, a is the lowest and b is the highest size of ngrams which we need to include in our features.The main goal is to get some distinct features out of the text for the model to train on, by converting text to numerical vectors. The features which are extracted should be in a well-defined setup that can be used as input directly to the classification algorithms. It produces better results than applying machine learning directly to the raw data.

E. Model Training

The entire insights and patterns obtained during data exploration are used to build the machine learning model. This stage always commences by splitting the data set into two different parts, train data, and test data. Training data will be utilized to build and analyze the model. The main logic of the model is based on the machine learning algorithm that is being implemented. In this paper, XGBoost machine learning algorithm have been used, which refers to eXtreme Gradient Boosting. XGBoost is a form of gradient boosted decision trees projected for speed and performance. It has an enormously high predictive power which makes it the optimal solution for accuracy compared to other algorithm .It is also referred as regularized boosting technique. Then,comparison of other supervised algorithms such as logistic regression[7] and Support Vector Machine were done with XGBoost algorithm.

F. Model Evaluation and Optimization

Once the model training have been done, it is time to evaluate the model using the test data which is used to check the efficiency of the model and how accurately it can predict the outcome. TABLE I provides the basic structure of confusion matrix for 3x3 matrix.

TABLE I. Basic Structure of Confusion Matrix

True Label	Positive(A)	Neutral(B)	Negative(C)
	Predicted Label		
Positive(A)	TA	FA1	FA2
Neutral(B)	FB1	TB	FB2
Negative(C)	FC1	FC2	TC

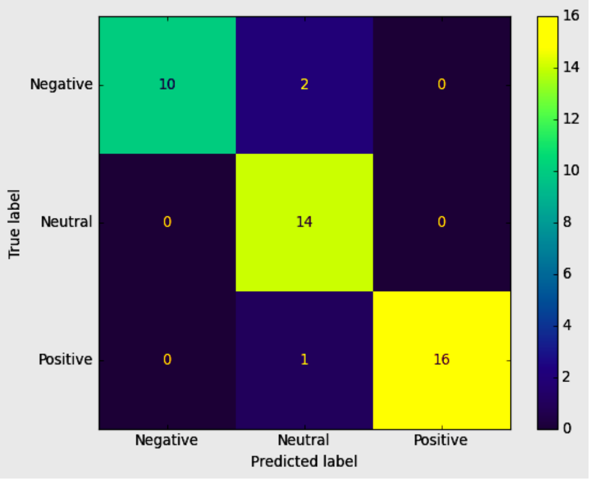


Fig.6. 3x3 Confusion Matrix of the proposed model

In order to calculate the accuracy and F1 Score for the above confusion matrix (Fig.6.) following equations can be utilized.

$$\text{Precision} = \frac{Correctly\ \ predicted}{Total\ \ Predicted}$$

$$\text{Recall} = \frac{Correctly\ \ Classified}{Actual}$$

$$\text{F1 Score} = \frac{2x\ precision\ \ x\ recall}{precision\ \ +recall}$$

$$\text{Accuracy} = \frac{Total\ \ Correctl\ y\ Classified}{Actual}$$

After the calculation of accuracy, further the model efficiency can be improved by tuning the model using hyper parameter tuning for each models to get the best fit.

G. Prediction

In this stage, once the model has been evaluated and improved, we can conclude which model is workingefficiently with good level of accuracy. Thus, we can use it for future predictions also.

4. Performance Analysis

This section displays the performance analysis of the algorithm which is used in this research paper. Here,TABLE II shows the comparison of various algorithm and using this table following graph (Fig.7.) has been projected which illustrates the contrast of accuracy and tuned accuracy.

TABLE II. Comparison of algorithms

Algorithm	Accuracy	Tuned Accuracy
Logistic Regression	73.58	77.36
Support Vector Classifier	77.71	78.34
XGBoost	87.50	93.02

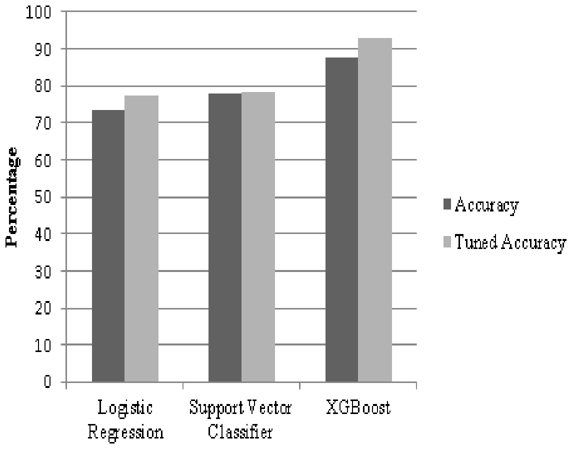


Fig.7. Graph showing the efficiency of algorithms

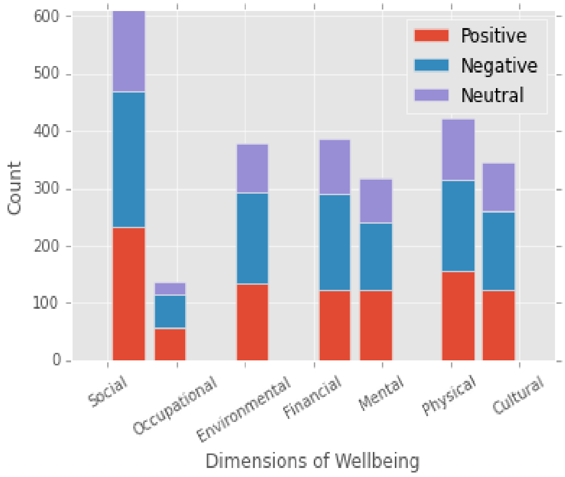


Fig.8. Graph analysis of wellbeing

Fig.8. visualize the percent of people affected during the pandemic under various categories of wellbeing with respect to their sentiments (i.e.) Positive, Negative and Neutral.

5. Applications of Sentiment Analysis

Sentiment analysis can be used in the following cases like product reviews, film reviews, social media posts and so on [14]. In Various Domains [13]: Recent researches in sociology and other fields like medical, sports have also been profited by sentiment analysis that illustrates developments in human emotions especially on social media. It is used in Business Intellect: It has been observed that people currently tend to look upon reviews of products which exist online before they buy them. And for many businesses, the online opinion decides the success or failure of the particular product. Thus, sentiment analysis plays a vital role in business. Businesses also demand to extract sentiment from the online reviews to improve their products and in turn their reputation and help in customer satisfaction.In Market and competitor research: This research is used to find out who is trending among the competitors and how the marketing efforts compare. We can analyse our competitor's content to find out what works with the public that we may not have noticed. It also helps the people to understand the strengths and weaknesses and how they relate to that of the competitors.

6. Conclusion

This research work aimed at analyzing the sentiments and emotions of the people during the pandemic and which have been successfully conducted by supervised machine learning approach (XGBoost). To sustain the reliability of data and also to make ease of extracting information of users, the standard kaggle platform has been chosen for the study and analyzed. After the adequate analysis of data, Machine learning algorithms have been implemented and accuracy of the model is calculated which is around 93%. By this research work, we conclude that most of the people were affected socially and this work allows the health organizations and volunteers to evaluate and understand the public's opinion and take necessary measures during the pandemic outbreak. Hence, sentiment analysis is extremely beneficial as it allows us to gain an outline of the broader public view behind certain areas. Furthermore,this work helps to understand the people insights about pandemic and its impact on the general public.

References

[1] Nida Afroz.et al,“Sentiment Analysis of COVID-19 Nationwide Lockdown effect in India”- Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAIS-2021) IEEE Xplore Part Number: CFP21OAB-ART; ISBN: 978-1-7281-9537-7

[2] Manoj Sethi , Sarthak Pandey , Prashant Trar ,“Sentiment Identification in COVID-19 Specific Tweets” - Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020) IEEE Xplore Part Number: CFP20V66-ART; ISBN: 978-1-7281-4108-4

[3] Mohammad Abu Kausar .et al, “Public Sentiment Analysis on Twitter Data during COVID-19 Outbreak” - (IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 12, No. 2, 2021

[4] Amrita Mathur Purnima,“Emotional Analysis using Twitter Data during Pandemic Situation: COVID-19” - Proceedings of the Fifth International Conference on Communication and Electronics Systems (ICCES 2020) IEEE Conference Record # 48766; IEEE Xplore ISBN: 978-1-7281-5371-1

[5] Dr K B Priya Iyer, Dr Sakthi Kumaresh, “Twitter Sentiment Analysis On Coronavirus Outbreak Using Machine Learning Algorithms” - European Journal of Molecular & Clinical Medicine ISSN 2515-8260 Volume 07, Issue 03, 2020

[6] Mrityunjay Singh .et al, “Sentiment analysis on the impact of coronavirus in social life using the BERT model” - Social Network Analysis and Mining (2021)

[7] Nalini Chintalapudi .et al,“Sentimental Analysis of COVID-19 Tweets Using Deep Learning Models” - Infect. Dis. Rep. 2021, 13, 329–339.

[8] Maha A. Alanezi , Nabil M. Hewahi ,“Tweets Sentiment Analysis During COVID-19 Pandemic” - 2020 International Conference on Data

[9] Carol Shofiya and Samina Abidi,“Sentiment Analysis on COVID-19-Related Social Distancing in Canada Using Twitter Data” – International

Journal of Environment Research & Public Health

[10] Imamah .et al,"Twitter Sentiment Analysis of Covid-19 Using Term Weighting TF-IDF And Logistic Regresion - 2020 Information Technology International Seminar (ITIS) Surabaya, Indonesia, October 14-16, 202

[11] Md Shoaib Ahmed,"Detecting sentiment dynamics and clusters of Twitter users for trending topics in COVID-19 pandemic" - PLOS ONEhttps://doi.org/10.1371/journal.pone.0253300 August 9, 2021

[12] Rohitash Chandra,"COVID-19 sentiment analysis via deep learning during the rise of novel cases" - PLOS ONE | https://doi.org/10.1371/journal.pone.0255615 August 19, 2021

[13] Vishal A. Kharde, S.S. Sonawane ,"Sentiment Analysis of Twitter Data: A Survey of Techniques" - International Journal of Computer Applications (0975 – 8887) Volume 139 – No.11, April 2016

[14] Raktim Kumar Dey et al.,"A Literature Survey On Sentiment Analysis Techniques Involving Social Media And Online Platforms" - International Journal Of Scientific & Technology Research Volume 9, Issue 05, May 2020