

Report

Sentiment Analysis for Distance Learning Arabic Tweets Dataset Using Apache Spark

Summary

Social media such as Twitter, Blogs, and Facebook are become one of the top resources for dataset nowadays. In which twitter is one of them is most reliable due to privacy and best content policy. Tweets are using for sentiment analysis and to predict the meaningful information from the data. Sentiment analysis is the most famous and emerging field for data analysis. We are discussing about the distance learning domain in Saudi Arabia, to analyze the Arabic tweets in distance learning we used Arabic tweets for large dataset. We proposed a model to analyze the people's feedback by using twitter dataset in the distance learning domain, proposed model is based on the apache spark server to manage the large amount of dataset.

Our proposed model using Twitter API to get the tweets, and all the tweets are raw data which are stored on the apache spark server. We perform regex-based technique used for preprocessing to remove the retweet, links, hashtags, English and numbers, username, and emojis from the tweet dataset. After that a training model develop which is used to train the whole prepared data using regression model. To predict the sentiment inside from the tweets we used regression model from the machine learning domain.

At the end we used FLASK API for sentiment analysis of the given tweets in Arabic. Our proposed model gives best result as compared to various applied techniques, on the test data by passing the Arabic tweet and get 82% Accuracy, 81% F1 Score, 81% Precision and 81 Recall on the tweets dataset used.

Table of Contents

CHAPTER 1	8
1.1 Introduction	9
1.2 Distance Learning	1
1.3 E-Learning.....	1
1.4 Opinion Mining for the Distance Learning.....	1
CHAPTER 2	2
2.1 Literature Review.....	3
CHAPTER 3	5
3.1 Introduction	6
3.2 Twitter API Model and Dataset	6
3.3 Training Model	8
3.4 Experimental Setup	9
3.4.1 Research Methodology	9
3.5 Preprocessing Model.....	10
3.6 Machine Learning Model Implementation	11
3.6.1 Regression Model	11
3.6.2 Why we choose Regression Model.....	11
3.7 FLASK API for Results	12
3.8 Performance of Regression and SVM.....	12
3.9 Evaluation Measures	12
CHAPTER 4	14
4.1 Introduction	15
4.2 Results and Analysis	15
4.2.1 Graphical representation for Positive, Negative and Neutral Tweets.....	16
CHAPTER 5	17
5.1 Conclusion	18
5.2 Future Work	18
REFERENCES	19

List of Figures

<i>Figure 1- Create your Twitter Account</i>	<i>7</i>
<i>Figure 2- Developer account creation and Get Access to the twitter API</i>	<i>8</i>
<i>Figure 3- visual result for the given test tweet</i>	<i>15</i>
<i>Figure 4- Graphical representation of the whole results</i>	<i>16</i>

List of Tables

Table 1: Confusion matrix.....	12
--------------------------------	----

List of Flow Charts

Flow Chart 1: <i>Proposed methodology for the sentiment analysis of Arabic tweets</i>	10
---	----

LIST OF ABBREVIATIONS

BOF	Bag of Words
API	Application Programming Interface
SVM	Support Vector Machine

Chapter 1

Introduction

1.1 Introduction

Nowadays, the scheme of internet has changed, and the people are using social media as compare to the traditional media news. The blogs, twitter, face book and various other social media platform are using to get the instant information about anything. Mostly people are using the social media only for the latest news and any information about the latest trends in any field of life. Most of the social media channels are using by the people but the most useable media is twitter, which is very reliable due to its usage all over the world and better security on the people's data. It provides best information and always rely on the unique information. Many educational institutes are using the social media to share the latest news and information to the people. Mostly using twitter platform for this purpose, as other different educational systems using the twitter for their feedback analysis in the quality of education and to check the trust of their students on the educational system. The people are also discussing about the distance learning education techniques on social media including twitter [1]. In Saudi Arabia many institutes are facilitating for the distance learning education, to know about the people's behavior such as in the COVID-19 twitter dataset is playing important rule. To understand and know the actual thinking about distance learning can be checked.

Distance learning in any field of education plays some role to facilitate the students but there are various problems. In Saudi Arabia mostly people contacting with the distance learning educational methodology. In Saudi Arabia peoples are discussing about the distance learning educational concept. Our proposed methodology is concerning about the distance learning in Arabic tweets dataset [2] to analyze the people's behavior either they are thinking positive, negative or they have some neutral behavior.

There are various techniques applied to analyze the distance learning on Arabic tweets. We are dealing with the problem to analyze the twitter dataset for distance learning using the machine learning techniques. To use the large amount of dataset needs to use some server to manage and compute the whole dataset on the system [3]. Apache Spark is using mostly for this problem we choose the apache spark for server usage and to compute the large amount of dataset. The complete flow of the proposed model is discussed in that way as, introduction section, literature review, proposed methodology including all the detailed steps. After that discussed about the results and conclusion. At the end we discuss all the references which we discuss in our work.

1.2 Distance Learning

Distance Learning is basically the approach in which the student not always appear physically in the classrooms to attend the classes. Normally they are communicated via some paperwork resources in the past. But nowadays it has been transferred to the online method where students attend the classes via online resources like skype, team etc. Due to the COVID-19 pandemic all almost all the institutes are using some online resources as mentioned above. Distance learning saves the time and are very good for the students to get the education.

1.3 E-Learning

The traditional methodology has been changed to the online method in which all the resources are transferred to online using various resources named skype, zoom, meat, etc. By using this technique students can get the content easily and if they are unable to attend the class physically can get all classes. There are various benefits using the e-learning systems as recently we face COVID-19 pandemic and all the institutes are transferred to the e-learning system.

E-Learning has some drawback as well which are discussed in the previous techniques like the quality of education disturbed. Because various students are from those places where internet issues and various problems due to non-technical staff. Our research is focusing on these points where are drawbacks and where the benefits for the e-learning education system.

1.4 Opinion Mining for the Distance Learning

Based on distance learning twitter dataset, it can be getting the solutions either using the e-learning benefit or not. By using the tweet, it can be checked either the peoples are thinking positive or negative also can check the neutral about the distance learning in Saudi Arabia. We can identify either the distance learning is better approach or not in the COVID-19 pandemic. Mean we can take decision based on tweet and analyze the data to make some rules on the given dataset.

Chapter 2

Literature Review

2.1 Literature Review

In this section we discuss about various techniques for the sentiment analysis using twitter dataset and distance learning domain. Ujjanta et.al. [4] proposed a technique for sentiment analysis on the twitter dataset, the dataset used to get from the twitter is from different domain like graduates, and officials as well. In this technique they get the tweets from twitter API after that apply some techniques to pre-process the tweets and then pass this tweets data to the VADER model to get the sentiment analysis on the given tweets.

Sentiment analysis for the distance learning tweets are discussed in the given technique, Vishal A. et.al. proposed a method to analyze the twitter dataset for the distance learning domain. The tweets are scrapped using the twitter API and bulk amount of data has scrapped to analyze it. They use the machine learning techniques to sentiment the tweets using classifier to classify the positive, negative tweets [5].

A novel approach implemented in this article; they used the twitter distance learning tweets used for the model. They scrapped the twitter dataset by using the twitter API. After that apply some preprocessing techniques to remove the outlier like link, username and special characters etc. They applied the machine learning techniques to sentiment he tweets for the different dataset including distance learning tweets [6].

The very latest technique applied to get the sentiment for the distance learning tweets. Deep learning techniques applied on this model and get the best results for the sentiment of positive, and negative for the overall dataset [7]. Lei et.al. discussed the deep learning methodology for the sentiment analysis of twitter dataset. Discussed about the opinion mining and decision mining techniques for the user review. A deep learning approach is best for the sentiment analysis which is implemented in this paper.

We proposed a novel approach for using distance learning Arabic tweets, in our proposed technique first designed the API model which scrap the tweets using twitter API. When we get bulk amount of dataset that data preprocessed to remove the unnecessary content like link, username, and special characters from all the scraped tweets. This data pass to the training model which train for the sentiment and after that used FLASK API for the three type of sentiments positive, negative, and neutral tweets.

Chapter 3

Implementation

3.1 Introduction

Implementation of the proposed model are discussed in this chapter. The twitter dataset used for the proposed model which are collected via twitter API. The experimental setup of the methodology implemented in the proposed model are discussed. The evaluation measures which are used to evaluate the proposed model elaborate in detail. Complete details about the proposed system is discussed in detail.

We propose a novel approach for the sentiment analysis on Arabic tweets dataset. Because it's first and most important of data collection for the e-learning domain. Sentiment analysis is the task where we get the feedback for anything and based on the dataset design some model to predict the best decision upon the dataset. There are various techniques for sentiment analysis, like used SVM, Decision Tree, MLP and random forest etc. also various datasets available for this work. We choose the very reliable social media twitter dataset because Twitter is best suited for sentiment analysis data. It contains a limit to text length and can easily be scrapped or collected using twitter API or third-party libraries. We proposed a solution for the distance learning feedback in Saudi Arabia. The below diagram shows all the preliminary steps performed to implement this novel approach. The details for this methodology are given below.

3.2 Twitter API Model and Dataset

This module is used to get Arabic Tweets, it uses Twitter developer credentials and saves the tweets in raw format. First, we create a twitter account for our general use. After completing this process, we need to create the developer account where we create an app which facilitate us to scrap the tweets. In this module first time we write a code which connect our model with twitter API.

For all this work we have used the twitter developer account to get the credentials of scrapping the tweets. There is a limit of queries to get tweets, we paid to get the API to extract a lot of tweets and then used it in our dataset. By using all this process, we get maximum tweets dataset in which the tweet size is generally 512 characters. We extracted around 7k tweets which are enough to build the sentiment dataset for narrow domain.

In this diagram shown where we need to create our twitter account in the very initial step. Here we create our twitter account which is the same account we used normally for our profile.

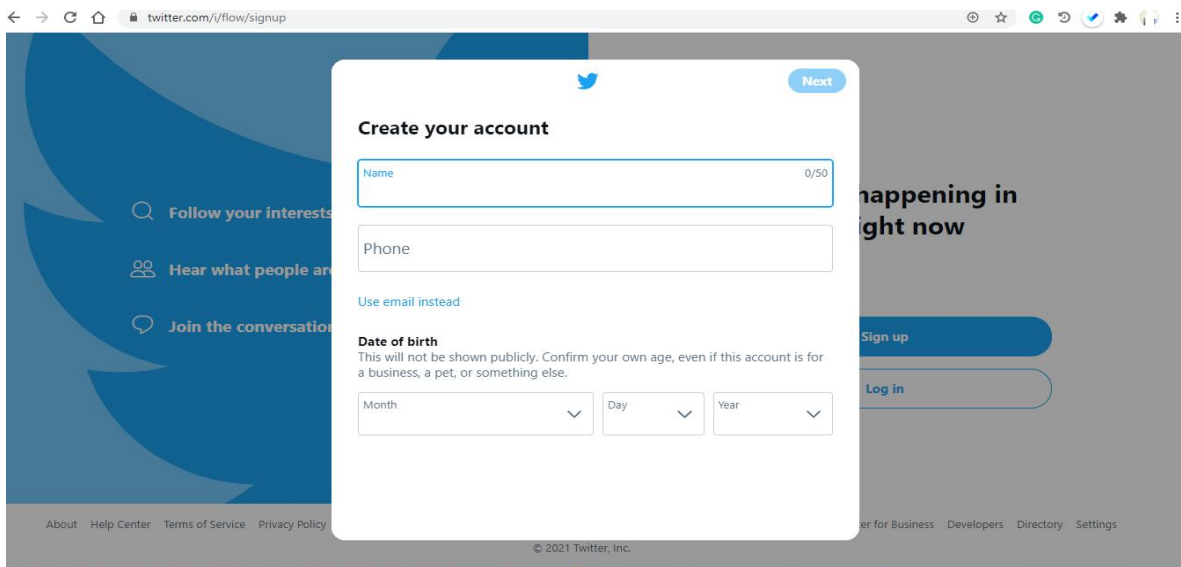
A screenshot of the Twitter 'Create your account' form. The form is titled 'Create your account' and has a 'Next' button in the top right corner. It contains several input fields: 'Name' (with a character count of 0/50), 'Phone', and 'Date of birth' (which is split into 'Month', 'Day', and 'Year' dropdown menus). Below the 'Phone' field, there is a link that says 'Use email instead'. The background of the page shows the Twitter logo and some navigation links like 'Follow your interests', 'Hear what people are saying', and 'Join the conversation'. At the bottom, there are links for 'About', 'Help Center', 'Terms of Service', and 'Privacy Policy', along with a copyright notice for 2021 Twitter, Inc.

Figure 1- Create your Twitter Account

In the second picture shown how we create the twitter developer account, because it is very important to get the twitter credentials. There are four important keys required so that's why we use this twitter developer account, the keys which are required are mentioned.

The Twitter Streaming API can be accessed in any programming language. The “twitter4j” is an open source, unofficial Java library, which provides a Java based module to easily access the “Twitter Streaming API”. The “twitter4j” provides a listener-based framework to access the tweets. To access the “Twitter Streaming API”, we need to sign in for Twitter developer account and should get the following OAuth authentication details.

- Customerkey
- CustomerSecret
- AccessToken
- AccessTookenSecret

Once the developer account is created, download the “twitter4j” jar files, and place it in the java class path. The sample diagram is shown for the process. Here we also need to create app to get all the dataset which is compulsory step. This module is used once to get the raw tweets in bulk, for this purpose we use twitter API and there was a problem to get the tweets was limited here we pay for more tweets. Then get about 7k tweets which are used in the proposed model which are enough for our model to train and testing the dataset.

Is everything correct?	
Primary use	Build customized solutions in-house
Account type	Organization
Twitter username	[Redacted]
Email	[Redacted]

Figure 2- Developer account creation and Get Access to the twitter API

3.3 Training Model

The scrapped dataset is copied to apache spark and there it is used for training spark model using conventions of parallel computing.

Apache Spark:

Apache Spark is basically open source general purpose distributed environment where we can compute our process on various parallel computing devices. We need the apache spark for analyzing the large amount of dataset such millions of tweets. Because locally it is very difficult to analyze and manage the big data. There various cloud-based server available like IBM, Teradata, AWS etc. Which are facilitating for the big data analysis. We choose apache spark which is open source and easily to understand. There various API's in apache spark which have different functionalities.

The apache spark model using the distributed environment for running the experiments. We use parallel execution for running the job on multiple Apache spark nodes. Since we are using single node with 2 threads data is distributed to these two threads.

Here we train all the large amount dataset by using the apache spark conventions in which maximum amount dataset can be easily compute. There are used some different packages like PySpark package, and MLlib which are required to compute the large amount of data are used, by using python. This is very important step which are required to manage the scrapped dataset from twitter API.

3.4 Experimental Setup

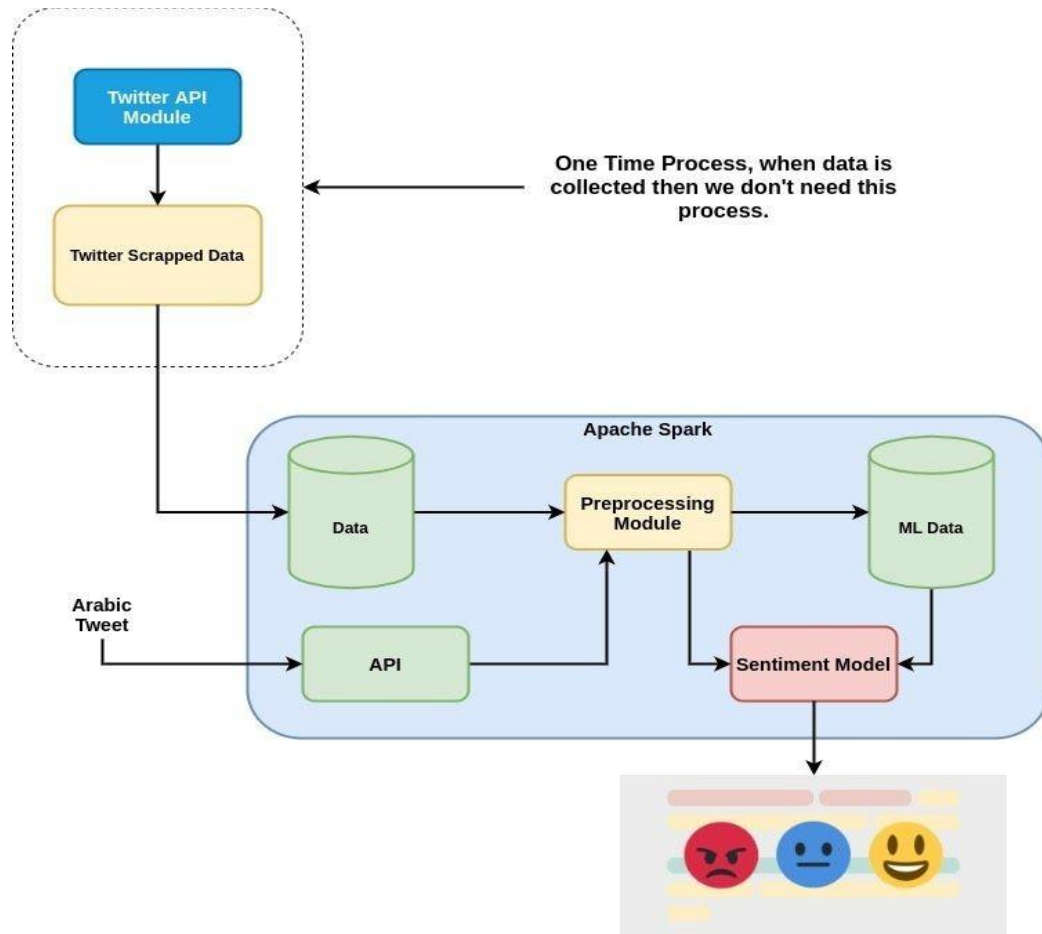
This section explains about methodology proposed in the pictorial form and explained in details. All the steps are discussed in detail for the proposed model.

3.4.1 Research Methodology

There are some important steps discussed here with details:

- i** In the very first step designs a model for the dataset scrapping. Which is used once to get the large dataset.
- ii** Second step is to develop a training model which are based on the apache spark.
- iii** In the third step apply preprocessing
- iv** Fourth steps discuss about prepared dataset or model implementation
- v** In the fifth step apply FLASK API for setting the sentiment analysis
- vi** In the last step discussed about the results against the evaluation measures

In the flowchart no 1 which shows a detailed description about the methodology proposed for the proposed model. The proposed model is basically used to analyze the tweets dataset for the distance learning domain to check the opinion of the peoples against the system. All the details are given below step by step.



Flow Chart 1: *Proposed methodology for the sentiment analysis of Arabic tweets*

3.5 Preprocessing Model

Preprocessing is applied to clean the tweet such as removing links, English and number, users, emojis, and labels are applied to each tweet for training purposes. It is very important to get the cleaned dataset (tweets) to get very accurate results after training and testing the model. There are some techniques which are applied which remove the unnecessary content from the dataset. The data we collected for applying the technique are bag of words where all the tweets without unnecessary information. The bag of words which are created using the five most commonly occurring words in all the tweets.

Bag of Words:

The bag-of-words model is a simplifying representation used in natural language processing and information retrieval. In this model, a text is represented as the bag of its words, disregarding

grammar and even word order but keeping multiplicity. In the proposed model used the most commonly occurring five words chosen to create the bag of words.

Based on these bags of words all the computation performed, and it saves the time and computation because when we get minimize the data by choosing this technique. It gives better results rather than we did not make bag of words in any technique.

3.6 Machine Learning Model Implementation

In this step we required the prepared dataset which need to pass the model by using some machine learning model. Let discuss about the machine learning model which we used in our proposed model named regression model for sentiment analysis.

3.6.1 Regression Model

Logistic regression is a popular method to predict a categorical response. It is a special case of Generalized Linear models that predicts the probability of the outcomes. In spark.ml logistic regression can be used to predict a binary outcome by using binomial logistic regression, or it can be used to predict a multiclass outcome by using multinomial logistic regression. Use the family parameter to select between these two algorithms, or leave it unset and Spark will infer the correct variant.

Data is now in a format that can be used to train a sentiment model. We have used the Regression model to predict the sentiment of tweets. There is need to set the dataset in the best form which model required for sentiment analysis. We pass this data to our model for the sentiment analysis. Various algorithms are available which can be used to predict the sentiment but reviewing various algorithms we used regression. It gives us best results.

3.6.2 Why we choose Regression Model

There are various other models named Support Vector Machine, Naïve Bayes, and many more using for the sentiment analysis. We choose the best model for sentiment analysis regression model which perform better and give the best results as wee compared in the result section.

3.7 FLASK API for Results

FLASK API is that which are used to get the sentiment results and it is drop-in-replacement for FLASK which is providing the browsable API implementation like the other Django type APIs. Flask API is provided for getting the results of tweets. It can be integrated into Websites and Smartphone apps for getting the sentiment. At the end we used flask API to get the results for our tweets, here we pass a tweet and it gives us result either it is positive, negative, and neutral tweet. Here we pass an Arabic tweet and our model give the result in the given three type of tweet.

3.8 Performance of Regression and SVM

The performance of Regression and SVM discussed as:

- The performance of Regression and SVM is tested based on the different evaluation measures like Accuracy.
- There are three types of tweets which labeled with Positive, Negative and Neutral
- Apache Spark used to analyze the tweets dataset
- Using the Logistic Regression Model our proposed model gets 82% Accuracy whereas using SVL only get 69% Accuracy value
- It means Logistic regression perform better than the SVM and finally we choose Logistic Regression

3.9 Evaluation Measures

The proposed system evaluates using the state-of-the-art measures including Precision, Recall, F1 Measure and Accuracy. And these based on the confusion matrix shown in table 1.

Table 1: Confusion matrix

	Negative	Positive
True Classify	N _{TP}	N _{FP}
False classify	N _{FN}	N _{TN}

- N_{TP}: represent the truly identify the negative tweet.
- N_{FP}: represent the falsely identify the negative tweet.

- N_{TN} : represent the truly positive tweets.
- N_{FN} : represent the falsely identify as positive tweets.

The accuracy can be defined as the percentage of correctly classified instances. The formula used in the model is shown in equation 1.

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \times 100\% \quad (1)$$

The recall is the fraction of relevant that are retrieved. The formula is shown in given equation (2)

$$\text{Recall} = \frac{N_{TP}}{N_{TP}+N_{FN}} \times 100\% \quad (2)$$

Precision is the most common evaluation measure used, precision is the ratio of the true positive to the sum of true positive and false positive. The formula to compute precision is shown in equation (3)

$$\text{Precision} = \frac{N_{TP}}{N_{TP}+N_{FP}} \times 100\% \quad (3)$$

Chapter 4

Results and Analysis

4.1 Introduction

Concluded after the implementation of the proposed model, our model performed best for the twitter sentiment analysis dataset on the distance learning domain. We used a large amount of dataset by using apache spark. The results we get are 82% with accuracy, which are best in related to the various other techniques. At the end we develop an app which receive any tweet from the user and give the feedback like positive, negative, or neutral tweet.

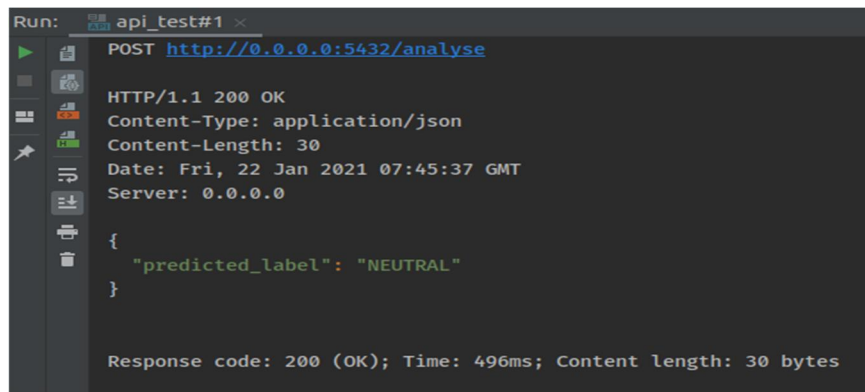


Figure 3- visual result for the given test tweet

4.2 Results and Analysis

The dataset used for this proposed model is about 7k tweets and all the dataset collected real time using twitter API. After allying al the steps we train and get the best results on the training dataset, The test tweets are unknow for the model which provide to test the performance of our proposed model, there are values which we get during the model evaluation. The value we get using the Accuracy measure is 82%, Precision 81%, Recall 81 and using F1 measure get 82% value against the twitter dataset.

4.2.1 Graphical representation for Positive, Negative and Neutral Tweets

Applying the different approaches to get better results for the sentiment analysis on the overall dataset we plot a graph which shows the results against all these types of tweets like positive, negative, and neutral tweets. In the graph clearly shown how the effect of distance learning on the people's behavior, we can easily identify now by seeing the graph.

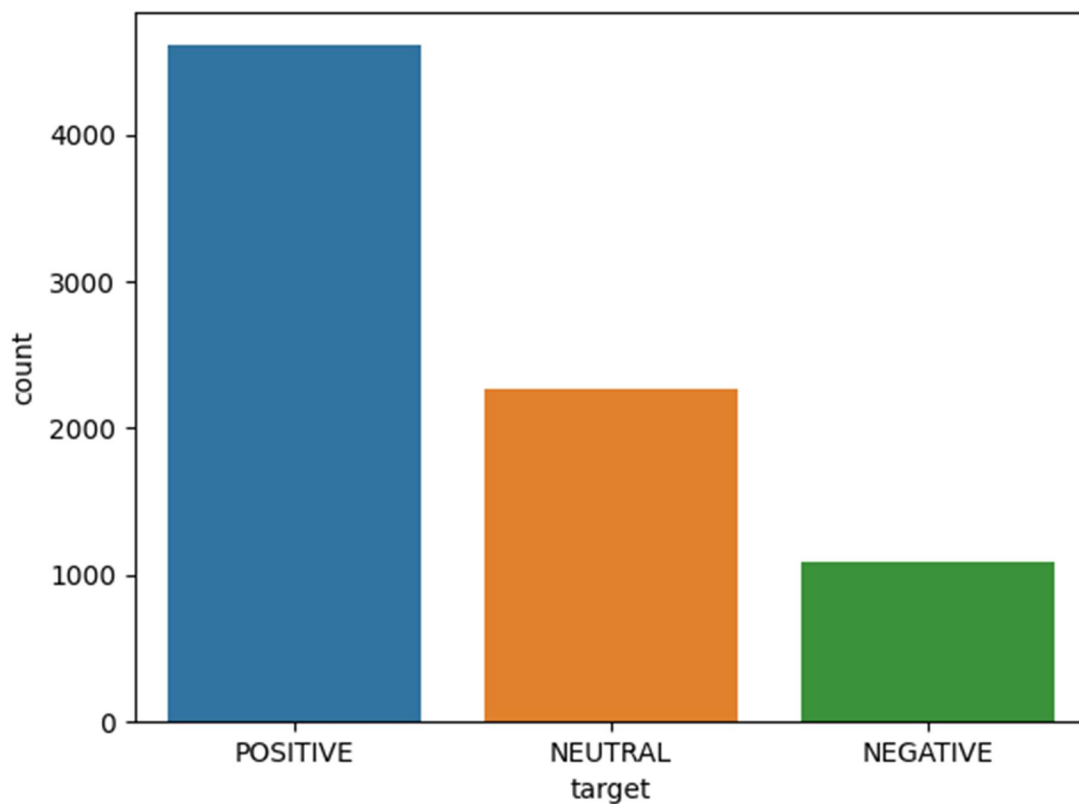


Figure 4- Graphical representation of the whole results

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Concluded after the implementation of the proposed model, outperform and get the maximum positive results as compare to the negative or neutral results. It is clearly shown here most people likes positively shown on fig 4 and discussed positive about the distance learning approaches at Saudi Arabia. Most of the people are rely on the system and few of them are wants to change and discuss negative about the distance learning. And very few peoples which are neutral on these all things.

.

5.2 Future Work

In the future there will be more work to be completed by applying some deep learning techniques to compute very large amount of the dataset for distance learning Arabic tweets and it can be predicting the various future years prediction. The curious problem we faced is by using Arabic tweets using apache spark on postman, it can be solved by creating the API which can resolve some different language syntax problems. As there are no problem for the English tweets because it has some built in API's.

References

- [1] R. M. Duwairi and R. Marji, "Sentiment Analysis in Arabic tweets," in *2014 5th International Conference on Information and Communication Systems (ICICS)*, 2014.
- [2] N. Arambepola, "Analysing the Tweets about Distance Learning during COVID-19 Pandemic using Sentiment Analysis," in *International Conference on Advances in Computing and Technology (ICACT-2020)At: University of Kelaniya, Sri Lanka*, Kelaniya, 2020.
- [3] M. Ahmed, "Arabic Sentiment Analysis using Apache Spark," in *reseacr h gate* , 2020.
- [4] U. Bhaumik, "Sentiment Analysis Using Twitter," in *Advances in Intelligent Systems and Computing*, vol 1276, Singapore, 2020.
- [5] V. . A. Kharde and S. S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of," *International Journal of Computer Applications*, vol. 139, no. April 2016, p. 0975 – 8887, 2016.
- [6] A. Krouska and C. Troussas, "The effect of preprocessing techniques on Twitter sentiment analysis," in *2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA)*, Chalkidiki, Greece, 2016.
- [7] L. Zhang, S. Wang and B. Liu, "Deep learning for sentiment analysis: A survey," in *Huawei Technologies Co. Ltd; National Science Foundation (NSF)*., 2018.