

**Twitter Sentiment Analysis**

**Application of Data Mining Tools & Techniques**

**DATA MINING - B9DA103**

**TERRI HOARE**

BHARATH MELAPPILLY RAJAGOPAL

STUDIENT ID: 10376807

DATE: 18-12-2018

Contents

Introduction ……..……………………………………………………………………… 3

Business Understanding ……...…………………………………………………… 4

Data Understanding …………………………………………………………………. 5

Data Preparation ………………………………..……………………………………. 7

Modelling ………………………………………………….……………………………… 9

Evaluation ……………………………………………………………………………….. 12

Deployment …………………………………………………………………………….. 12

Conclusion ………………………………………………………………………………. 16

References ………………………………………………………………………………. 17

**Introduction**

Data Mining is supposed to be the process of exploration through data in order to understand and predict new algorithms to understand and predict the future. It helps in taking a step ahead for analyzing the future. The process of data mining contains three important major steps, which is considered to be foundation. They are Statistics which is the numeric study of data relationships, Artificial Intelligence which is the human-like intelligence displayed by software and/or machines, and Machine Learning which is the algorithms for understanding the data for the purpose of prediction.

With the advance technologies, the processes are becoming much liberal and much easy and understandable to implement compared with the manual and time-consuming processes which we had. Business sectors like Banks, Manufactures, Retailers etc. are now using the advance technologies in data mining in order to understand the future with regards to the future price predictions, to understand the consumer trends etc.

Twitter is an American online social and news networking services on which the different people use different modes to communicate their emotions which are collectively called as tweets. These tweets can be seen by the public, anyhow the senders have the option to restrict them as well. Each tweets are restricted to 280 characters. The tweets are widely being utilized for many purposes like, for movie reviews, to understand the current trends, the ideologies etc. Our main aim of the assignment is to build a model to detect and summarize the sentiments using the tweets on random topics.

The whole idea of this analytical process was carried out with the CRISP-DM data modelling steps. By keeping in mind the advantages of this model, we started to build ahead by splitting the whole data mining process into six steps, and these are as follows,

* Business Understanding,
* Data Understanding,
* Data Preparation,
* Modelling,
* Evaluation,
* Deployment.

No matter which organization is implementing this model, the result can be accurately understood by providing a schema to follow while the preparation, planning and the final execution of data mining work.

**Business Understanding**

Twitter sentimental analysis is the model which has been created here is to mainly understand the sentiment from the given text. It is defined as a process in which the text mining techniques are extracted and further it is being studied ahead to understand the polarity of the given document or text. In this implemented principle, we are mainly focusing in understanding how accurate it is when we compare it with the human judgments and other conclusions to understand the similarity and focus. These understanding and comparisons are mainly calculated by different measures which is based on precision and recall oven three variables which are, Negative, Positive and Neutral.

While we consider from a business point of view, the final output from the prepared data can be helpful in many ways. Example it could be helpful in making a decision, it could be helpful in understanding end result of a political or a social media campaign etc. Here in this assignment, we are mainly focusing on implementing the sentimental analysis on three different topics which are,

* Avengers End Game, which is a Marvel Movie.
* #MeToo, which is campaign against the sexual harassment in the Film Industry.
* Demonetization in India.

The main benefits of Sentimental Analysis are as follows,

* **Adjust marketing strategy,**
* **Measure ROI of your market campaign,**
* **Develop product quality,**
* **Improve customer service,**
* **Crisis Management,**
* **Lead Generation,**
* **Sales Revenue.**

Social media has not only become an integral part of our daily lives but also has become a valuable source of resource for businesses. With low costs, businesses are now able to gather actionable insights easily. From customer service to marketing, sentiment analysis can improve everything. However, it should be kept in mind that no machine learning algorithm can provide an accuracy of 100% due to complexity and limitations of the technology. So, despite some of its limitations, sentiment analysis has proven to be a valuable opportunity for businesses to grow.

**Data Understanding**

We got the dataset from https://www.kaggle.com/kulgen/elon-musks-tweets which includes the tweets by Elon Musk from November 2012 to December 2017. This Excel file have 3219 rows and five columns which shows the row number, tweet, date, retweet from and user. From this, we took only the tweets for analyzing.

**![A close up of a newspaper

Description automatically generated]()**

We got the training dataset from https://old.datahub.io/dataset/twitter-sentiment-analysis/resource/486ee031-d3d9-4d50-82de-7aa8893a564f. This have 300 rows and 2 columns. The Columns are tweets and sentiment. In tweets column there are 3 different values as 0, 2 and 4, in which 0 shows negative, 2 shows neutral and 4 shows the tweet is positive. For making it user friendly we changed the values to Negative, Neutral and Positive as explained before.

**![A close up of text on a white background

Description automatically generated]()**

**Data Preparation**

The data which we are using need filtering and should get prepared for processing. The data may not be in a clean version and lot of filtering will be needed. So, we used an operator called ‘Process Document From Data’. In that operator, we used different operators for filtering data.

![A picture containing screenshot

Description automatically generated]()

The operators are

* **Transform cases**

This operator is used to transform all the characters to lowercase.

* **Tokenize**

This operator will split the sentences in to tokens. There are several options how to specify the splitting points. We used non-letters to split in to tokens. For example, the operator will create token when there is a full stop, coma, space etc.

* **Filter Stopwords (English)**

This operator will remove English stopwords. For example, am, is, are etc.

* **Filter Tokens (by length)**

This operator filters the tokens according to their length. We have given the range 4 to 25. So, the operator will filter every words except this range.

* **Stem (Porter)**

This will change all the words to its base form. For instance, gone, going will be changed to go.

So, after getting the result from these operators we give this to set role operator.

**Set Role Operator**

It is used to change the role of one or more attributes and here we make sentiment as the target attribute so that the system can identify the sentiment attribute as the target attribute. Output from this operator is given to another operator called Select Attribute.

**Select Attribute.**

This is used to filter out the attributes which have any missing values. So, we have given attribute filter type as no\_missing\_values. The output of this operator is given to the Validation Operator. We used split validation here.

**Modelling**

3 different models are used in this process for finding out which model gives the greater accuracy. We used different data and found out the performance of each model to determine which one is better. After trying 3 different models the most accurate model is chosen for creating the final model.

So, here we used a Split Validation operator. Here we are splitting the entire data set to training and test data set. We have given the split ratio as .7. Now in the validation operator, we used 3 different models as explained before. They are SVM, Decision Tree and Naïve Bayes. The training data is given to the operators and they create models and ‘Apply Model’ operator is used to apply the model on the testing dataset. Then we used ‘Performance’ operator to find out the performance by each model.

![A screenshot of a cell phone

Description automatically generated]()

The performance we obtained on each model is as follows

1. **SVM**

![A screenshot of a cell phone

Description automatically generated]()

1. **Decision Tree**

![A screenshot of a cell phone

Description automatically generated]()

1. **Naïve Bayes**

![A screenshot of a cell phone

Description automatically generated]()

From all these models Naïve Bayes Gave more accuracy with 65.52%.

**Evaluation**

Now we have a model, we have data and its time to implement. But before implementation, we must look through everything again. So here in this step, we check whether we are going to achieve the objective by reviewing the model. Also, we check whether all issues are considered or not. The steps for evaluation are

* Evaluate the Results.
* Process review.
* Determination of next steps

So, we have the performance of every model and we decided to use Naïve Bayes. Naive Bayes is a family of simple "probabilistic classifiers" based on applying Bayes theorem with strong (naive) independence assumptions between the features. With an accuracy of 65.52, this model created better prediction of tweets as neutral, negative and positive. We also used other datasets to make sure the prediction is right.

**Deployment**

This is the final stage of the process. We have already created a model for implementation, but we have to implement it. This stage is for that. The implementation can be in any form. For example, visualization, making we application, web page etc. We selected to visualize the result so that it will be easy to understand by any people who have good knowledge in technology and also for those don’t have any experience in technology.

We used three popular topics in social media to get the prediction of the tweet on them whether they are positively tweeted, negatively tweeted or neutrally tweeted. For that, we used ‘Search Twitter’ operator to collect the tweets from twitter and the ‘Process Document’ operator as used in training data, for cleaning the data.

The topics we selected are

1. **Avenger End Game**

This is the most awaited movie of 2019 so that we think most of the people will be tweeting positively about the topic and recently the trailer of the movie was also released. So, the hype of the movie will be really high.

1. **#MeToo**

It is a hash tag campaign against the sexual harassment against women in film industry. So, we think the most tweets will have a negative character.

1. **Demonetization in India.**

This is a process happened in India introduced by the current Prime minister Mr Modi which withdrawn all the 500 and 1000 rupee notes in the country. From medias of India and the footages they shown which reveals views by common people, this event badly affected the people of India. So, we think the tweets will have negative character.

Following are the results of the different topics

1. **Avenger End Game**

![A picture containing text

Description automatically generated]()

From the ring graph, we can see that most of the tweets are neutral. So, our thinking was not correct. People tweeted about this topic with more neutrality.

1. **#MeToo**

![A screenshot of a cell phone

Description automatically generated]()

The results in this is as expected. The more tweets are negative compared to others. So, in this our prediction was correct.

1. **Demonetization in India**

![A close up of a mans face

Description automatically generated]()

In this, our prediction was more negative comments but as per the output the tweets are of more neutral than negative. Here also our thinking was wrong ant output by rapidminer was interesting.

The results shows that the prediction is accurate but the prediction by a common man like us is not the correct one. Different people have different thinking and a person can’t predict only by media or his own opinion but the actual prediction using datamining shows the correct prediction and it may be same and it can be different.

**Conclusion**

While concluding we can say that the project had successfully being driven by the CRISP-DM methodology in order to attain the final goal.

In the first phase, we had analyzed the twitter sentiment analysis of the upcoming movie “Avengers End Game, using the CRISP-DM methodology and we were able to predict that the tweets about the movie were mostly neutral which was quite shocking as the movie was tagged as the most awaiting movie of 2019.

In the second phase, the analysis clearly shows the emotions of people about the #MeToo campaign. Without mush surprises, the result was favoring. The negative tweets were more comparing with the other two.

In the final phase, we had picked up the most important topic which is the Demonetization in India. As per the press and the several social media reports most of the people were against the government with regards to the implementation of demonetization. But as per our twitter sentimental analysis, the response was neutral which was again quit shocking.

Overall the predications were accurate with regards to the three selected topics and was quite shocking to. And we were able to understand the fact that the thoughts of a common man would be different when we compare it with the thoughts of various other sectors in the economy.

**References**

Wikipedia.org. (2018). *Wikipedia*. [online] Available at: https://www.wikipedia.org/ [Accessed 05 Dec. 2018]

RapidMiner. (2018). *Visual Workflow Designer for Data Scientists | RapidMiner Studio*. [online] Available at: https://rapidminer.com/products/studio/ [Accessed 17 Jul. 2018].

Kaggle.com. (2018). *Kaggle: Your Home for Data Science*. [online] Available at: https://www.kaggle.com/ [Accessed 05 Dec. 2018].

aylien.com. (2018). AYLIEN: Making Sense of the Incomprehensible. [online] Available at: https://aylien.com/ [Accessed 05 Dec. 2018]

old.datahub.io. (2018). DATAHUB: The easy way to get, use and share data. [online] Available at: https://old.datahub.io/ [Accessed 05 Dec. 2018]