Exploring the opinions of people on GST tweets from different cities

Kumar P K*, Dr S Nandagopalanb

a Department of MCA, Kumar P K, VTU, PG Center Mysuru, India; b Department of IS&E, Dr S Nandagopalan, Vemana Institute of Technology, Bengaluru, India

Received: 2020-02-20
Revised: 2020-04-20
Accepted: 2020-04-30

Abstract: In the present world, people are extensively using the social networking applications such as twitter, face book, etc to share their opinions or thoughts on any particular product or service. Evaluation of the shared opinions through social networking applications is the trending research area. This work will help to analyze the opinions, views of the people about any product or service, and this study will help in decision-making. For exploring the opinions, Natural Language Processing and Machine Learning approaches being used to differentiate between the emotions such as positive, negative and neutral opinions. It is very tedious task to read the input text data given by the user in social media applications. Here proposed a novel bounded logistic regression with the GST data sets taken from different cities of India.

Keywords: Tweet cleansing, Sentiment Analysis, Machine Learning.

1. INTRODUCTION

With the extensive use of social applications for communications, people across the globe create tons of text data every day. An interaction usually happens through a public domain, and making all this data available to everyone. Analysis of emotions is the process to extract the opinions from social media applications. Location wise data is collected and opinions of users are analyzed to find out the location wise impact of any product or any scheme. Result analyses are helpful for taking optimal decisions in future. Here, Data is collected from different locations based on the GST scheme and used the bounded logistic regression method and categorize the opinions of the people. Twitter Data set on GST took from different locations of India and explored the sentiments from that data set. In this paper, only and extracting the information and process of identifying the emotions that the people express in the form textual reviews later categorized as positive, negative or neutral. Analysis of emotions is very complex and crucial in the growth of the company. This process will provide the useful information about the reviews. GST scheme was launched in the year of 2017 July 1 and decision took by the Indian government. The people to interact and express their feelings in the form of text majorly use twitter application. Share the post with someone is known as tweeting. Users are very curious to find out the tweets of their interest [1].

This manuscript is defined as follows: section II is about the earlier work and proposed method, section III is about the implementation and results from the proposed technique, section IV is conclusion and future work.

2. PROPOSED WORK

The work, which is proposed here, is having many phases such as downloading the tweets, tweets cleansing, labeling of the emotion, Analysis of data, developing machine leaning models, forecasting the emotion and rendering. The graphical representation of the proposed work is as shown in Figure 1.

*Department of MCA, VTU, PG Center Mysuru, India, Tel: +91 9945155090, E-mails: kumar.pk@vtu.ac.in
Fig. 1. Proposed work

<table>
<thead>
<tr>
<th>Downloading</th>
<th>Cleaning</th>
<th>Labeling</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Building ML Models</td>
<td></td>
<td>Data Vectorization</td>
</tr>
<tr>
<td>Interpreting the data</td>
<td></td>
<td>Prediction of Emo-</td>
</tr>
</tbody>
</table>

twitter application built a downloader engine, which consists of tweepy API (Application Programming Interface) and tweet cursor. Time duration and number of tweets is the input for tweet cursor. After downloading stored into the CSV file. It consists of different fields like tweet text, reply to, date of creation, status, retweet count, and is retweet. Here focus is on the tweet text field.

Data Cleansing:

Downloaded data from twitter is not in structured nature. It requires cleaning the data. Below steps are for data cleaning:

Step1: Removing all the html tags with the help of web scraping tool.

Step2: Removal of the special Notations such as non-letters.

Step3: Conversion of text to lower case

Step4: Removing the stop words using NL toolkit

Step5: Combining the words into one string after removing the stop words.

Labeling:

New algorithm proposed sticker calculation by making a nltk.sentiment.vader python library, in that API called SentimentIntensityAnalyzer have been utilized. SentimentIntensityAnalyzer will discover the extremity score on the each info tweet content given, from this we are computing compound extremity and subjectivity. Compound extremity gives the blend of positive, negative and unbiased extremity. In light of compound extremity extremes we are marking into positive, negative and nonpartisan feelings. Table 1 demonstrates the determined compound extremity and naming of the assumption. The proposed sticker calculation is as demonstrated as follows

New Algorithm:

If compoundPolarity > 0 Then

Label “+Ve”

Else if compoundPolarity == 0 Then

Label “neu”

Else if compoundPolarity < 0 Then

Label “-Ve”.

<table>
<thead>
<tr>
<th>Sanitized tweet text</th>
<th>Sentiment compound polarity</th>
<th>Sentiment neutral</th>
<th>Sentiment negative</th>
<th>Sentiment pos</th>
<th>Sentiment type</th>
</tr>
</thead>
<tbody>
<tr>
<td>fyi must tell clear doubts enhance general kno...</td>
<td>0.2960</td>
<td>0.548</td>
<td>0.151</td>
<td>0.301</td>
<td>POS</td>
</tr>
<tr>
<td>demonetization aap protests marching govt mov...</td>
<td>-0.5574</td>
<td>0.662</td>
<td>0.338</td>
<td>0.000</td>
<td>NEG</td>
</tr>
<tr>
<td>demonetization help combat terror pak able pri...</td>
<td>-0.4767</td>
<td>0.541</td>
<td>0.314</td>
<td>0.146</td>
<td>NEG</td>
</tr>
<tr>
<td>demonetization positive effect cigarette sales...</td>
<td>0.4019</td>
<td>0.593</td>
<td>0.141</td>
<td>0.267</td>
<td>POS</td>
</tr>
<tr>
<td>posts pic dead robber claims demoneti...</td>
<td>-0.9349</td>
<td>0.360</td>
<td>0.640</td>
<td>0.000</td>
<td>NEG</td>
</tr>
</tbody>
</table>

Table 1: Labeling of the sentiment for the sanitized tweet text.

Data Analysis

In the wake of naming of the conclusions, for example, constructive, pessimistic and unbiased on purified tweet information we can do exploratory information investigation to
comprehend the feeling of the general population GST. Table 2 demonstrates the investigation examination of the feeling on various plans. Graphical portrayal of EDA of feeling is appeared in

Table 2: Analysis of sentiment on GST

<table>
<thead>
<tr>
<th></th>
<th>GST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>2450</td>
</tr>
<tr>
<td>Negative</td>
<td>1712</td>
</tr>
<tr>
<td>Neutral</td>
<td>5657</td>
</tr>
<tr>
<td>Total</td>
<td>9819</td>
</tr>
</tbody>
</table>

Data Vectorization

Here need to find out how frequently the word occurred in the tweets by using Term frequency and Inverse docu freq. It is calculated using the below formula,

\[ W_{td} = TF_{td} \log \left( \frac{N}{DF_t} \right) \]

Where:
- \( TF_{td} \) — Occurrence of t in the entire document d
- \( DF_t \) — Gives the total of document which have the term t.
- N — number of words present the in the review

Here handling the number of words by specifying the rank from 1 to 3 counts. In tweets text it is crucial to calculate the frequency of words and positions of the word in the data.

The top 10 word found from the datasets is as given below:

| [gst, 1299], [tax relief, 973], [taxation, 1197], [tax law, 1259], [tax rate, 957], [tax sales tax, 900], [payroll taxes, 129], [tax, 998], [taxes, 239] |

3. BUILDING ML MODLES

Random Forest

Random Forests are prepared through the stowing strategy. Packing or Bootstrap Aggregating comprises of haphazardly examining subsets of the preparation information, fitting a model to these littler informational indexes, and totaling the expectations. This strategy enables a few occurrences to be utilized over and again for the preparation stage given that we are testing with substitution. Tree packing comprises of testing subsets of the preparation set, fitting a Decision Tree to each, and collecting their outcome.

The Random Forest strategy presents more arbitrariness and assorted variety by applying the sacking technique to the component space. That is, rather than looking covetously for the best indicators to make branches, it arbitrarily tests components of the indicator space, hence including greater assorted variety and lessening the difference of the trees at the expense of equivalent or higher inclination. This procedure is otherwise called "highlight stowing" and it is this groundbreaking technique what prompts an increasingly powerful model.

We saw now how to make forecasts with Random Forests. In the Random Forests calculation, each new information point experiences a similar procedure, however now it visits all the various trees in the gathering, which are were developed utilizing arbitrary examples of both preparing information and highlights. Contingent upon the job needing to be done, the capacities utilized for collection will contrast. For Classification issues, it utilizes the mode or most incessant class anticipated by the individual trees, while for Regression assignments, it utilizes the normal expectation of each tree.

Despite the fact that this is an incredible and exact technique utilized in Machine Learning, you should constantly cross-approve your model as there might be Overfitting. Additionally, in spite of its power, the Random Forest calculation is moderate, as it needs to develop numerous trees during preparing stage and as we definitely know, this is a ravenous procedure.

Bounded Logistic Regression

It (LR), uses the logistic method called sigmoid function, it is used to determine the development in the population in ecology. It is an vector curve which have floating values ranking from 0 and 1, but its limits to as specified below:

\[ \frac{1}{1 + e^{-\text{value}}} \]

Where, \( e \) indicates the log of eulers in the exponential way and actual value need to be transform. LR uses the function as the report of the linear regression. The input data (x) combined together with unit of weight to find the value of the output. (y). The main difference from LR is the value modeling is zero or one rather that numbers. Example for LR equation is as below:

\[ y = \frac{e^{(b_0 + b_1 x)}}{1 + e^{(b_0 + b_1 x)}} \]

Where, \( y \) is the output prediction, \( b_0 \) indicates the bias value and \( b_1 \) is the coefficient value for the input. Every column in the data need to related to the coefficient values.

In the proposed technique, enhanced by LR by modifying the default value of the LR parameter such as Cs, state of the
randomness, solution and multiclass. It is an integer value lies from 1e-4 to 1e5.
To get the best result, divided the proposed approach into 2 parts namely train and predict phase. It will be discussed as below:

**Algorithm: Initial Phase**

**Input:** Dataset has GST tweet

**Output:** Classification of Tweets emotions into pos, neg and neutral

**Method:**
1. Load the csv file for train phase
2. Apply tweet cleansing for cleaning data
3. Labeling the emotions using proposed approach.
4. Data Vectorization and division of Vectorised data.
5. By using term frequency and its inverse document frequency, extract the properties.
6. Divide the preprocessed data for training (80:20)
7. Construct a ML models for analyzing data
8. Apply proposed BLR.
9. Evaluate the performance.

**Algorithm: Training Phase Ends**

**Algorithm: Prediction Phase**

**Input:** Tweet without labels i.e pos, neg, neutral

**Output:** Labels are predicted (pos, neg, neutral)

**Method:**
1. Load the new csv file for prediction phase
2. Apply cleansing technique to clean data
3. Save the cleaned data into data frame
4. From data frame extract the properties using term frequency and its inverse document frequency
5. Apply proposed BLR to predicting the label neg, pos and neutral emotions

**Algorithm: prediction phase ends**

4. IMPLEMENTATION AND RESULTS

To implement the approached ML model we used python. GST Data is fetched using twetpyAPI, which is in python. Here cleansing that is removal of the tags, is done by using web-scraping library, removing the retweets, special symbols and other. After cleaning the data, creating of bag of word. Then these words are giving as an input to the algorithm. Here used the proposed BLR approach. Finally, after train the algorithm, we get the number of pos, neutral and neg response by the people on GST.
CONCLUSION

Emotion analysis is the way to know the people views in the world. Here examined the Govt of India scheme GST to know the people emotion about that. For analyzing used tweets, they are not proper format. Hence initially transferred to proper format by using cleaning algorithms’ on GST dataset. Then Used new BLR algorithm to find out new data sets without having any labels and predicting the approached BLR.

REFERENCES


Author Biographies
P.K. Kumar, Department of MCA, Post Graduate Studies, VTU, Mysore Region, Mysuru, India. Currently pursuing PhD under Visvesvaraya Technological University, Belagavi and currently working as an Assistant Professor in Department of MCA, Post Graduate Studies, VTU, Mysore Region, Mysuru. His research interest is in image analysis, data mining. He has around 5 years of Experience in teaching and 7 years of experience in industrial field. Also, published 05 International Journal Articles, 3 International Conference paper and 2 National Conference papers.

Dr. S. Nandagopalan, An M. Tech from NIT, Surathkal and Ph. D from Amrita University is a highly qualified and experienced faculty member to impart quality education to the students. His overall teaching experience of 34 years. It includes teaching, research, and administration. He has presented more than 11 research papers in the international conferences organized by IEEE, Springer, WASET, ACATA, etc., in various countries such as USA, France, China, Singapore, Thailand, and Malaysia. To his credit, he has also published 8 research papers in the leading international journals.