OPINION MINING OF TWITTER DATA ON IMPLEMENTATION OF GST USING MACHINE LEARNING ALGORITHM

Soudamini Hota
M.Tech
Computer Science & Engineering
Chandigarh University, Mohali, India

Abstract: GST stands for Goods and Services Tax and it is defined as the tax that is imposed on goods and services which are sold for domestic consumption. GST is an indirect tax that is added to the cost of goods and services. The business adds GST to the cost price of the goods and services to form the final retail price or service price. The consumer pays the cost price as well as the GST levied on the product or service. The amount that constitutes the GST is collected by the business and forwarded to the government. GST is a way to generate revenue for the government. GST merges all the central level taxes like value added tax, service tax, excise duty tax etc. and state level taxes like entertainment tax, luxury tax, entry tax etc. into one unified tax system. GST was implemented in India on 1 July 2017.

This work is based on studying the opinions of people in India towards the implementation of GST. The source data for opinion mining has been taken from Twitter as it is the most widely used microblogging site and can be considered as a valid indicator of people's opinions and sentiments. The source data for study and analysis has been pulled from Twitter using the Python library Tweepy. N-Gram modeling technique has been applied on the source data for feature extraction. The machine learning algorithm k-nearest neighbor is used for segregating opinions into positive and negative classes. The implementation of the classification methodology has been done in Python language.

***Index Terms* - GST, KNN, Python, Tweepy, Twitter.**

I. INTRODUCTION

Opinion mining, also known as sentiment analysis, has turned into an interesting and happening area of research in the last few years as it provides an insight into the opinions held by people towards any entity, ranging from a product, a brand, a person, an event, an organization and so on. People express their view points in the form of reviews, feedbacks and articles posted in newspapers, magazines, social networking sites, microblogging sites etc. With the world getting digitalized, social networking sites and microblogging sites have turned into places where people actively express their opinions. These opinions can be studied and mined and the results obtained can be used in a beneficial way.

Amongst all the microblogging platforms, Twitter has gained a huge growth in the last few years and is a potential source for studying the opinions and sentiments of people. Various organizations and companies are progressively looking for different ways to mine twitter information in order to figure out what individuals think and feel about their products and services. The microblogs posted on Twitter are known as tweets. Mentioned below are some key features of tweets:

- Message Length: Earlier the maximum length of tweets allowed was 140 characters, which was later doubled on November 7, 2017, for every language except Korean, Japanese and Chinese.
- Emoticons: Emoticons are pictorial representation of facial expressions that users include in their tweets to express their moods. Emoticons are formed by different combinations of letters, numbers and special characters.
- Target: In order to allude a particular user on twitter a "@" symbol is used. When usernames are mentioned this way, they automatically get alerted.
- Hash tags: Users make use of hash tags (represented using "#" symbol) in their tweets in order to categorize the tweet as belonging to a particular topic or theme. This feature facilitates the users to locate and follow the tweets which fall under a particular topic or theme.
- Trending topics: The topics, words or phrases that are mentioned higher number of times as compared to the others are considered as trending topics. Twitter web interface displays a list of trending topics on its home page.
- Related headlines: If a website includes a tweet in its article, the link to this article is mentioned with the tweet on Twitter. This feature enables users to get a better understanding of the tweet by visiting the article through the mentioned link.

II. LITERATURE REVIEW

Mihai Dascălu et al (2011), has presented an automated assessment system for evaluating each participant in an online discussion forum. The assessment system makes use of natural language processing methods, social network analysis methods as well as computed metrics for assessment and evaluation [14]. Genetic algorithm has been used to optimize the assessment results. This framework is capable of handling bigger corpuses in smaller amount of time; and it performs well under a wide assortment of conditions and loads.

Wen Hua et al, (2016) in his paper presents a prototype framework to understand short texts. There are numerous challenges that come up while analyzing short messages within various applications. Short messages do not conform to grammatical rules, which is why the traditional natural language processing tools cannot be applied on them. Besides short messages are usually ambiguous which renders them more difficult to handle using traditional methods [15]. The framework proposed in this paper makes use of semantic knowledge from an established knowledge base for better understanding of short texts. The performance of the framework is evaluated with the help of various simulation experiments. On the basis of the results achieved it is concluded that semantic knowledge is of utmost importance for the comprehension of short texts.

Ankur Goel et el, (2016) has proposed an improvement in the existing methods for sentiment analysis and classification of tweets into multiple classes [17]. The paper shows that the use of SentiWordNet along with Naïve Bayes classification algorithm can improve the accuracy of the classification results. The author has used dataset from Sentiment140 which contains 16 million tweets. When Naïve Bayes Classifier is used for classification, the accuracy obtained is 58%. However Naïve Bayes Classifier coupled with SentiWordNet for classification yielded higher accuracy.

Shweta Rana et el, (2016) has proposed methods for analyzing the reviews on movies posted by users, and has categorized the reviews into positive and negative classes [18]. Three different algorithms Naïve Bayes, Synthetic words and Linear SVM have been used and compared. The results generated by these algorithms indicate that Linear SVM algorithm provides the highest accuracy.

Wiraj Udara Wickramaarachchi et al, (2017) has proposed a methodology to identify the emotion conveyed through an image. Users express their views and thoughts by posting comments consisting of different elements like text, images, recordings or videos. This increases the effectiveness of the communication among users since they have no face-to-face cooperation [20]. This new methodology is an improvised version of the previous works. It uses the concept of Latent Semantic Analysis (LSA) as it is a light weight approach. The research work proved to be more efficient than the previous works because of the light weight nature of the approach. In addition, the author has designed a prototype of a Graphical User Interface to identify the emotion conveyed by the image uploaded in it.

III. RESEARCH METHODOLOGY

This research work aims at learning a classifier for sentiment analysis and classification of twitter data into positive and negative classes. The learned classifier has been used categorize tweets on GST into positive and negative classes for the purpose of analyzing the opinions held by people towards the implementation of GST in India. The classifier to distinguish the input tweets into positive and negative classes is modeled using KNN algorithm. The research methodology followed is depicted in Fig.1.

i. Data extraction

Twitter data has been used as the source data. Twitter offers several APIs for the delivery of tweets to its users and developers for training, testing and analysis. The REST API, the Search API, and the Streaming API are the three unique API variants that are made available by Twitter: In order to get tweets that match particular criteria, the filtering parameters can be set while extracting tweets from Twitter. Once the query has been constructed it can be kept running by the API and all the tweets that match the criteria will be delivered as the output of the program. The data set for this work was extracted from Twitter using ‘Tweepy’, a Python library that facilitates communication between Python and Twitter.

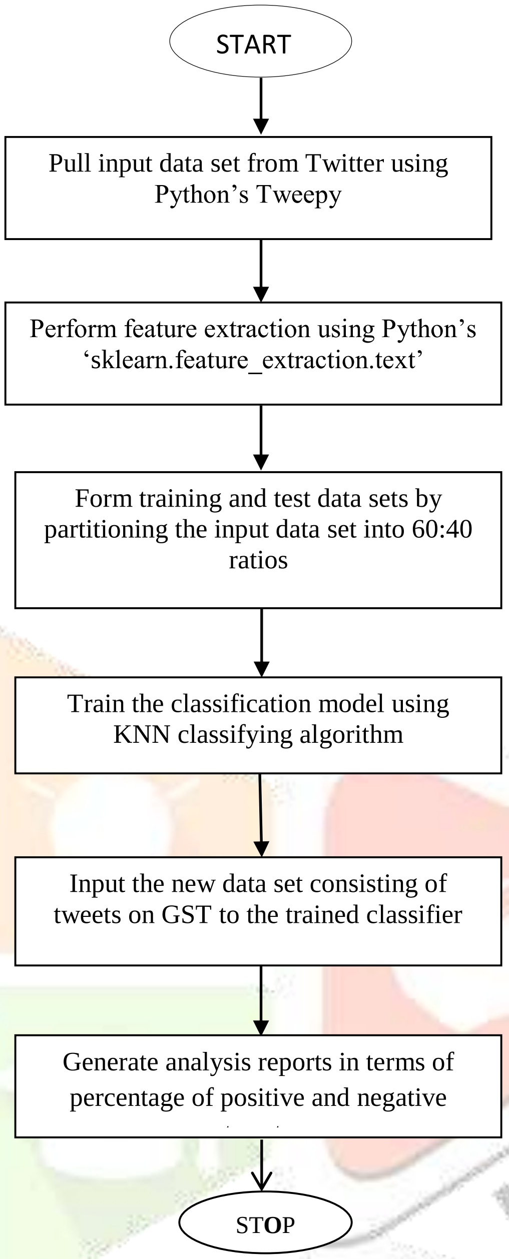
START
Pull input data set from Twitter using Python's Tweepy
Perform feature extraction using Python's 'sklearn.feature_extraction.text'
Form training and test data sets by partitioning the input data set into 60:40 ratios
Train the classification model using KNN classifying algorithm
Input the new data set consisting of tweets on GST to the trained classifier
Generate analysis reports in terms of percentage of positive and negative
STOP

Fig 1: Proposed Methodology

ii. Data preprocessing

Data preprocessing is an integral step in any data mining task. Data preprocessing is done to remove incomplete, inconsistent and noisy data from the dataset.

Users include a number of elements in their tweets like acronyms, slangs, emoticons, URLs, misspellings, special symbols like @, *, #, ( ), { } , [ ] etc. The presence of such elements in tweets makes it difficult to extract the keywords. Therefore the tweets have to be preprocessed to prepare it for feature extraction.

Preprocessing of input data is done by making the following transformations in the tweets:

- Emoticons are identified as positive, negative and neutral depending on the sentiment state expressed by them. These emoticons are then replaced with labels representing their polarity.
- Acronyms are substituted with their expanded forms with the help of online dictionary for abbreviations.
- Slang words are replaced with their associated meanings with the help of slang word dictionary. Domain information contributes much to the interpretation of slang words.
- Stop word are removed from the tweets because they are less important to be considered for the feature space.
- URLs are removed from the tweets as they do not contribute in sentiment analysis and are hence considered redundant.
- Special characters are removed from the tweets as they carry no significance in the comprehension of the tweets.
- Target words used to mention a user by including @ symbol are replaced with "user".

- Tweets with low word counts are discarded as they are less informative.

The 'sklearn.feature_extraction.text' module in Python performs preprocessing and normalization of raw input data.

iii. Sentiment Classification

The research methodology is based on learning a classifier based on K Nearest Neighbor algorithm. KNN is a supervised machine learning classifier which learns by analogy, i.e. it compares a given test data point with the training data points that are similar to it and assigns the most common class amongst all the classes of the similar training data points. Each training data point is described by n attributes and is given by an n-dimensional vector, $X=\{x_1, x_2, x_3, \ldots.. x_n\}$. Thus all the training data points are represented in a pattern space of n dimensions. When the class of an unknown data point has to be predicted, KNN classifier looks for k training data points in the pattern space that are most similar to the given unknown data point. These k data points are considered as the k nearest neighbors of the unknown data point. Having found out the k nearest neighbors, the most common class of the neighbors is assigned as the class for the unknown data point.

Similarity of data points is measured in terms of Euclidean distance. The Euclidean distance between two data points $X_1=\{x_{11}, x_{12}, x_{13},\ldots, x_{1n}\}$ and $X_2=\{x_{21}, x_{22}, x_{23},\ldots.., x_{2n}\}$ is given by

$Dist(X_1,X_2) = \sqrt{\sum^{n}_{i=1}} (x_{1i} - x_{2i})^2$

Normalization of attributes is done so that the attributes with higher range do not outweigh the attributes with smaller range. Measuring of similarity using Euclidean distance applies in numerical attributes. In case of categorical attributes differential grading is used, wherein numerical scores are used to denote the differences between different attribute values.

IV. RESULT AND ANALYSIS

The classification results indicate that 60 percent of commenters hold negative opinions regarding the implementation of GST while 40 percent of the commenters have expressed their support towards GST. A survey of the tweets shows that majority of the people are unhappy with the format of GST that includes four different tax slabs. Individuals who are opposed to GST do not find any justification in the way in which 1211 items have been distributed across different tax slabs. Certain items which fall under the umbrella of basic necessities have been placed in higher tax slab; while certain items which fall within the bracket of luxury items have been placed in tax slab with lower tax rate. Also GST system includes a slab with 28% tax rate, which happens to be the highest among the 140 countries that have implemented GST. Multiple tax rates and a large list of exemptions have turned GST system into a complicated one. Commenters are of the opinion that a simplified GST system comprising of a flat tax rate would have been more effective in reducing the cascading effect of taxes along the chain of raw material suppliers, manufacturers, distributers and retailers. On the other hand, those who support GST have faith in the strength of this unified tax system that has subsumed all other forms of taxes at both central and state levels. It has increased the number of indirect tax payers massively. The commenters who favor GST opined that the problems that are occurring post implementations of GST are merely teething problems; and the system will stabilize in due course of time and will show its positive outcome in the long run.

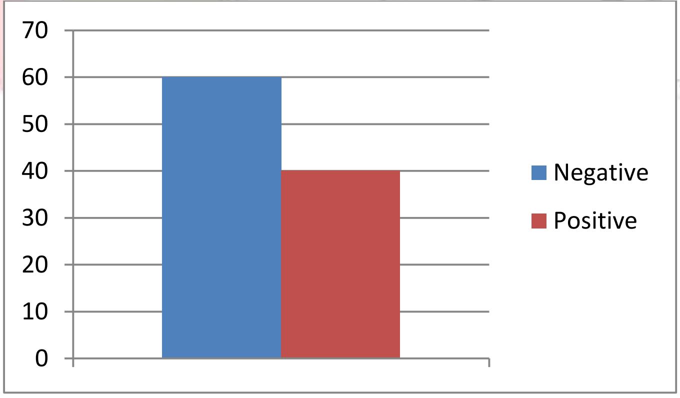

Fig.2 Analysis Results

V. CONCLUSION

The research done in this paper is qualitative in nature as Twitter users constitute only a part of the total population. A subjective research has been done to study the viewpoints held by people towards the implementation of GST and to analyze the underlying reasons that support their opinions. The analysis report generated by performing opinion mining on a sample of the population is an approximation of the opinions held by the people in general.

REFERENCES

[1] G. Salton and C. Buckley, *Term-weighting approaches in automatic text retrieval. Information Processing and Management*, 24:513–523, 1988.

[2] O. Sandu, *Domain Adaptation for Summarizing Conversations,* PhD thesis, Department of Computer Science, The University Of British Columbia, Vancouver, Canada, 2011.

[3] S. Teufel and M. Moens, *Summarizing scientific articles: Experiments with relevance and rhetorical status*, Computational Linguistics, 28:409–445, 2002.

[4] D. C. Uthus and D. W. Aha, *Plans toward automated chat summarization*, In Meeting of the Association for Computational Linguistics, pp. 1–7, 2011.

[5] L. Zhou and E. H. Hovy, *Digesting virtual geek culture: The summarization of technical internet relay chats*, In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 298–305, 2005

[6] Tharindu Weerasooriya, Nandula Perera, S.R. Liyanage, *A method to extract essential keywords from tweet using NLP*, 2016 16th International Conference on Advances in ICT for Emerging Regions(ICTer).

[7] M. G. Ozsoy, F. N. Alpaslan, and Cicekli, *Text summarization using latent semantic analysis*, Journal of Information Science, vol. 37, no. 4, pp. 405-417, 2011.

[8] Adyan Marendra Ramadhani, Hong Soon Goo. *Twitter Sentiment Analysis using Deep Learning Methods*, 7th International Annual Engineering Seminar (InAES), Yogyakarta, Indonesia, 2017.

[9] K. Kaviya, C. Roshini, V. Vaidhehi, J. Dhalia Sweetlin. *Sentiment for Restaurant Rating*, 2017 IEEE International Conference on Smart Technologies and Management for Computing, Controls, Energy and Material (ICSTM).

[10] Rasim Alguliyev, Ramiz Aliguliyev, Nijat Isazade, *A Sentence Selection Model and HLO Algorithm for Extractive Text Summarization*, 2016, IEEE.

[11] Narendra Andhale, L.A. Bewoor, *An Overview of Text Summarization Techniques*, 2016, IEEE.

[12] Rupal Bhargava , Yashvardhan Sharma, *MSATS: Multilingual Sentiment Analysis via Text Summarization*, 2017, IEEE.

[13] Akshi Kumar, Aditi Sharma, Sidhant Sharma, Shashwat Kashyap, *Performance Analysis of Keyword Extraction Algorithms Assessing Extractive Text Summarization*, International Conference on Computer, Communication, and Electronics (Comptelix), 2017.

[14] Mihai Dascălu, Ciprian Dobre, Ştefan Trăuşan-Matu, Valentin Cristea, *Beyond Traditional NLP: A Distributed Solution for Optimizing Chat Processing*, 2011 10th International Symposium on Parallel and Distributed Computing.

[15] Wen Hua, Zhongyuan Wang, Haixun Wang, Kai Zheng, and Xiaofang Zhou, "*Understand Short Texts by Harvesting and Analyzing Semantic Knowledge*, 2016, IEEE.

[16] Pierre Ficamos, Yan Liu, Weiyi Chen, *Naive Bayes and Maximum Entropy approach to sentiment analysis: Capturing domain-specific data in Weibo,* 2017 IEEE International Conference on Big Data and Smart Computing (BigComp).

[17] Ankur Goel, Jyoti Gautam, Sitesh Kumar, *Real time sentiment analysis of tweets using Naive Bayes*, 2016 2nd International Conference on Next Generation Computing Technologies (NGCT).

[18] Shweta Rana, Archana Singh, *Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques*, 2016 2nd International Conference on Next Generation Computing Technologies (NGCT).

[19] Huma Parveen, Shikha Pandey, *Sentiment analysis on Twitter Data-set using Naive Bayes algorithm*, 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT).

[20] Wiraj Udara Wickramaarachchi, R. K. A. R. Kariapper, *An Approach to Get Overall Emotion from Comment Text towards a Certain Image Uploaded to Social Network Using Latent Semantic Analysis*, 2017 2nd International Conference on Image, Vision and Computing.