PROJECT REPORT

ON

"SENTIMENT ANALYSIS OF TWEETS ABOUT UNION BUDGET OF 2019"

ABSTRACT

Social media gives a simple method of communication technology for people to share their opinion, and feelings.

The aim of this project is to analyze various sentiment behaviors of people about the UNION BUDGET OF INDIA, 2019 and thus, categorize sentiment of people as positive, negative or neutral.

Sentiment analysis of the tweets determine the polarity and inclination of the vast population towards a specific topic, item or entity. These days, the applications of such analysis can be easily observed during public elections, movie promotions, brand endorsements and many other fields.

ABOUT TWITTER

Twitter is an online micro-blogging and social-networking platform which allows users to write short status updates of maximum length 140 characters. It is a rapidly expanding service with over 320 million registered users - out of which 100 million are active users and half of them log on twitter on a daily basis - generating nearly 42 million tweets per day. Due to this large amount of usage we hope to achieve a reflection of public sentiment by analyzing the sentiments expressed in the tweets. Analyzing the public sentiment is important for many applications such as firms trying to find out the response of their products in the market, predicting political elections and predicting socioeconomic phenomena like stock exchange.

In order to directly address a tweet to someone we can add the target sign “@” or participate to a topic by adding an hashtag “#” to your tweet. Because of the usage of Twitter, it is a perfect source of data to determine the current overall opinion about anything.

ABOUT THE UNION BUDGET OF INDIA

The **Union Budget of India**, also referred to as the *Annual Financial Statement* in the Article 112 of the Constitution of India,is the annual budget of the Republic of India.

The Union Budget keeps the account of the government's finances for the fiscal year that runs from 1st April to 31st March. The Union Budget is classified into Revenue Budget and Capital Budget.

- Revenue budget includes the government's revenue receipts and expenditure. There are two kinds of revenue receipts - tax and non-tax revenue. Revenue expenditure is the expenditure incurred on day to day functioning of the government and on various services offered to citizens. If revenue expenditure exceeds revenue receipts, the government incurs a revenue deficit.

- The Capital Budget includes capital receipts and payments of the government. Loans from public, foreign governments and RBI form a major part of the government's capital receipts. Capital expenditure is the expenditure on development of machinery, equipment, building, health facilities, education etc. Fiscal deficit is incurred when the government's total expenditure exceeds its total revenue.

On 5th July 2019, Finance Minister Nirmala Sitharaman presented the union budget of 2019 in each of the following sectors:

Education

- This budget proposed a program named 'Study in India'. Through this program, the Indian government makes an effort to bring in foreign students to pursue higher education in India.
- Rs. 400 crores were allotted to make top Indian institutes as world-class institutes.
- The present government has promised in this budget that it will bring the new 'National Education policy', which is still in the draft stage.
- This budget proposed to set-up 'National Research Foundation' to fund and promote research.
- Experts always recommended spending 6% of GDP on education. Even though the money allotted for education has increased in this budget, the share of education in the total budget is just 3.4%. At present India needs reforms in education, so the funds allocation is disappointing.

Employment opportunities

- This budget reduced the corporate tax of companies with a turnover of up to Rs.400 crores per annum to 25%. Earlier only companies with turnover of up to Rs. 250 crores per annum had 25% corporate tax. For other companies, it was 35%. With the fresh move, 99.3% of companies will pay only 25% corporate tax. This is an encouragement to the MSME (Micro, Small & Medium Enterprises) sector. And this step has the potential to create new employment opportunities in this sector.

- In this budget, the government announced easing of Angel tax regulations, which is a boon for startups. Encouragement to startups will result in more employment opportunities.
- It was also announced that India will be transformed as the manufacturing hub of Electric Vehicles. This too has the potential to create so many jobs.
- It was also mentioned that 100 new clusters will be set up in this financial year under SFRUTI (Scheme of Fund for Regeneration of Traditional Industries) which will enable 50,000 artisans to join the economic value chain.
- As a part of the 'Housing for All' program, the government announced an additional tax deduction of 1.5 lakh on purchase of homes below Rs.45 lakh that are bought till March 2020. This is a boon to the real estate sector and has the potential to create lots of jobs in this sector.
- There was no mention of the unemployment crisis in the budget. Even though there are significant steps to create new jobs, there are no bold steps to solve the unemployment crisis in India.

Agriculture

- For the first time, the importance of promoting organic farming is recognized. This budget announced that the government will promote 'Zero Budget Natural Farming'. This is a very progressive step and will promote self-sustainable agriculture. This can result in the end of reliance on loans in agriculture. But the government has not announced any funds for this, and without the government's support, it will be difficult for farmers to switch to this kind of farming. Moreover, on one side the government has announced zero budget farming, and on the other side, fertilizer subsidy allocation has jumped from Rs. 70,090 crore to Rs. 79,996 crore. This is contradictory and will make it more difficult for farmers to switch to natural farming.
- Agricultural growth is declining over the past few years. And the government did not take any significant steps to keep its promise of doubling the farmers' income by 2022. Mere support schemes like 'Kisan Samman Nidhi Yojana' cannot help farmers in the long run.
- It was also announced that new farmer producer organizations will be set up.
- This budget promised to set up 80 Livelihood business incubators and 20 technology business incubators under ASPIRE (A Scheme for Promotion of Innovation, Rural Industries and Entrepreneurship) to develop agro-rural

industries. This will greatly help startups that are based on agriculture. And thereby will help in lifting agriculture out of the crisis.

MSME sector

- It was proposed that the new e-commerce platform will be launched to sell products of Micro, Small and Medium enterprises. This is very helpful for MSME sector.
- It was also announced that there will be 2% interest subvention for GST registered MSMEs. This will encourage informal organizations to register and hence will help in formalization of the informal economy.

Startups

- Angel tax norms are eased, which is a boon for startups.
- A new TV program in Doordarshan was announced to promote startups and to help them in reaching to investors.
- Global investors meet will be conducted in India, which will be super helpful for startups of India.

Environment

- GST is reduced from 12% to 5% for Electric Vehicles (EV). And it was also announced that there will 1.5 lakhs worth tax exemption for interest paid on the purchase of e-vehicles. This is a very progressive step and will help in faster adoption of electric vehicles. If the production of renewable energy is increased to a great extent, electric vehicles will help in fighting against climate change.
- Along with this, the prices of petrol and diesel are increased, so more and more people may adopt electric vehicles.

Women Empowerment

- Every woman in SHG (Self Help Group) is eligible for an overdraft of Rs.5000 and one woman in every Self Help Group is eligible for loan up to 1 lakh rupees under Mudra Yojana. This will help in encouraging women to take part in economic development and also helps women in achieving financial independence.

- A new committee will be set up to suggest ways to encourage women to take part in economic development. This is a good step and helps in achieving gender equality.

National Security

- Defense budget hasn't changed. It is the same as the interim budget – 3.19 lakh crore rupees. This is disappointing because recently India has witnessed terror attacks, and terrorism is a growing threat. So national security should be given more importance.

Impact on common people

- Duties and cess on petrol and diesel were increased, so the effective price of petrol is increased by Rs. 2 per liter. This affects common people.
- Moreover, the rise in fuel prices will result in higher transport costs and hence the prices of essential commodities will go up. But on the other hand, the importance of connectivity is emphasized in this budget and also steps were taken to develop infrastructure, so the transport costs may come down. And as a result, this may cancel out the impact of the rise in fuel prices.
- As a part of the 'Housing for All' program, the government announced an additional tax deduction of 1.5 lakh on purchase of homes below Rs.45 lakh that are bought till March 2020. This will be very helpful for middle-class people.
- 'One Nation – One Grid' will make power affordable.
- 'Har Ghar Jal' was promised to make piped water supply accessible to all rural households by 2024 under the 'Jal Jeevan Mission' scheme.
- The health sector is neglected, which will impact common people negatively.

Conclusion

The Union Budget 2019-20 did not take effective steps to tackle the agricultural crisis and unemployment issue, which are the major concerns in the present times. But it is very practical and did not announce any short term beneficial schemes. It was made with a long-term vision and is focused on common people.

ABOUT SENTIMENTAL ANALYSIS

Sentiment analysis is the procedure of discovering and classifying opinions expressed in a piece of text(like comments\feedback text).The intended output of this analysis is to determine whether the writer's mindset towards that topic is positive, negative or neutral.

Sentiment analysis is employed on Twitter posts by means of following techniques

- Lexical based analysis
- Machine learning based analysis

Lexicon Based Approach

Lexicon based sentiment analyzer was proposed to determine the polarity and measures for tweet data of particular candidate

This technique is governed by the use of a dictionary consisting pre-tagged lexicons. The input text is converted to tokens by the Tokenizer. Every new token encountered is then matched for the lexicon in the dictionary. If there is a positive match, the score is added to the total pool of score for the input text. For instance if "dramatic" is a positive match in the dictionary then the total score of the text is incremented. Otherwise the score is decremented or the word is tagged as negative. Though this technique appears to be amateur in nature, its variants have proved to be worthy.

The classification of a text depends on the total score it achieves. Considerable amount of work has been devoted for measuring which best lexical information works. An accuracy of about 80% on single phrases can be achieved by the use of hand tagged lexicons comprised of only adjectives, which are crucial for deciding the subjectivity of an evaluative text

Machine Learning Approach

Machine learning is one of the most prominent techniques gaining interest of researchers due to its adaptability and accuracy. In sentiment analysis, mostly the supervised learning variants of this technique are employed.

In the training data, a collection of tagged corpora is provided. The Classifier is presented as a series of feature vectors from the previous data. A model is created based on the training data set which is employed over the new/unseen text for classification purposes. In machine learning technique, the key to accuracy of a classifier is the selection of appropriate features. Generally, unigrams (single word phrases), bi-grams (two consecutive phrases), tri-grams (three consecutive phrases) are selected as feature vectors. There are a variety of proposed features namely number of positive words, number of negative words, length of the document, Support Vector Machines (SVM) and Naïve Bayes (NB) algorithm. Accuracy is reported to vary from 63% to 80% depending upon the combination of various features selected.

The machine learning technique faces challenges in: designing a classifier, availability of training data, correct interpretation of an unforeseen phrase. It overcomes the limitation of lexical approach of performance degradation and it works well even when the dictionary size grows exponentially.

NOTE: Naïve Bayes and Support Vector Machine (SVM) are the most used machine learning approaches to formulate the sentiment analysis classifying into positive, negative or neutral. Nowadays researchers are using unsupervised learning using natural language processes and other intelligent techniques such as the Neural Network approach to improve the accuracy but working better with a huge amount of data.

Conclusion

Depending upon the application, the success of any approach will vary. Lexical approach is a ready-to-go and doesn’t require any prior information or training. While on the other hand machine learning requires a well-designed classifier, huge amounts of training data sets and performance tuning prior to deployment.

DID YOU KNOW?

Sent meter, an application for analyzing opinion data was used during a campaign of Swachh Bharat Abhiyan in India in 2014 using unstructured data from Twitter, and it achieves 84.47% of accuracy using machine learning approach with unigram words.

PROBLEM STATEMENT

The objective of this project is to detect and analyze the tweets written by the general public about the UNION BUDGET OF INDIA, 2019 and thus, categorize sentiment of people as positive, negative or neutral. Also, we want to train the sample of tweets and labels generated and then predict the labels on the given test dataset. The evaluation metric for this practice problem is F1 score.

METHODOLOGY

1. DATA COLLECTION

Data in the form of raw tweets is retrieved by using the Python library called "tweepy" which is an easy to use python library for accessing the twitter API. The API(Application Program Interface) is a code that allows us to communicate with each other. The API requires us to register a developer account with Twitter and fill in parameters such as consumer Key, consumer Secret, Token access, and Token Secret. This API allows you to get all random tweets or filter data by using keywords.

I have extracted the tweets from twitter about the general public's opinion about the recent UNION BUDGET OF INDIA,2019.

1. DATA PREPROCESSING

Pre-processing the data is the process of cleaning and preparing the text for classification.

But, before this too, I checked for null texts in tweets extracted and removing the duplicate tweets.Online texts contain usually lots of noise and uninformative parts such as

a) **Hashtag**: A hashtag is any word or phrase immediately preceded by the # symbol. When you click on a hashtag, you'll see other Tweets containing the same keyword or topic.
b) **@username**: A username is how you're identified on Twitter, and is always preceded immediately by the @ symbol. For instance, Katy Perry is @katyperry.
c) **MT**: Similar to RT (Retweet), an abbreviation for "Modified Tweet." Placed before the Retweeted text when users manually retweet a message with modifications, for example shortening a Tweet.
d) **Retweet**: RT, A Tweet that you forward to your followers is known as a Retweet. Often used to pass along news or other valuable discoveries on Twitter, Retweets always retain original attribution.
e) **Emoticons**: Composed using punctuation and letters, they are used to express emotions concisely, ";) :) ...".

In addition, on word level, many words in the text do not have an impact on the general orientation of it. Keeping those words makes the dimensionality of the problem high and hence the classification more difficult since each word in the text is treated as one dimension. Here is the hypothesis of having the data properly pre-processed: to reduce the noise in the text should help improve the performance of the classifier and speed up the classification process, thus aiding in real time sentiment analysis.

And thus, in my dataset, I have removed the @username and urls. Also, RT was present in the tweets which is considered irrelevant in the analysis so these were removed.

Some of the steps involved in data preprocessing are:

a) **Removing twitter handles (@username) and also urls:**
Also, RT was present in the tweets which is considered irrelevant in the analysis so these were removed. Regular expressions are used to match alphabetical characters only and the rest are ignored. This helps to reduce the clutter from the twitter stream

b) **Convert to lowercase and tokenize:**
Tokenization which is the process of splitting the tweets into individual words called tokens. Tokens can be split using whitespace or punctuation characters. It can be unigram or bigram depending on the classification model used.

c) **Removing punctuations, numbers, and special characters:**
The punctuations, numbers, and special characters don't help much. it is better to remove them from the twitter texts and replace them with whitespaces.

d) **Removing stopwords:**
Tweets contain stop words which are extremely common words like "is", "am", "are" and hold no additional information. These words serve no purpose and this feature is implemented using a list stored in nltk.corpus library. We then compare each word in a tweet with this list and delete the words matching the stopwords list.

2. DATA FILTERING

a) **Stemming :**
It is the process of reducing derived words to their roots. It uses the algorithm that removes common word endings from English words such as "ly","es","ed","s".
Example: considering "care","caring","carefully","cared","cares" as "care" instead of separate words. Another, Example includes words like "fish" which have the same roots as "fishing" and "fishes" Most popular stemmers are Porter Stemmer, Lancaster Stemmer and Snowball Stemmer.
In our case, we have not employed any stemming algorithm due to time constraints.

b) **Lemmatization:**

It is the process of transforming to the dictionary base form. We have used WordNet which is a large lexical database for English words that are linked together by their semantic relationships. Words as a thesaurus, i.e., group words together based on their meanings. In this we also used POS-Tagging which is the process of assigning a tag to each word in the sentence as to which grammatical part of speech that word belongs to, i.e. noun, verb, adjective, adverb, coordinating conjunction etc.

NOTE: Stemming and lemmatization both generate the root form of the inflected words. The difference is that stem might not be an actual word whereas lemma is an actual language word. Another difference is that the stemming technique follows an algorithm with the steps on the words which makes it faster whereas lemmatization uses the wordnet corpus and corpus for stopwords as well to produce lemma which makes it slower than stemming. You also had to define the parts of speech to obtain the correct lemma.

3. SENTIMENT ANALYSIS

Now that we have discussed some of the text formatting techniques employed by us, we will move to the list of features that we have explored. As we will see below a feature is any variable which can help our classifier in differentiating between the different classes.

There are two kinds of classification in our system: the subjectivity classification and the polarity classification. As the name suggests the former is for differentiating between objective and subjective classes while the latter is for differentiating between positive, neutral and negative classes.
Sentiment analysis is basically the process of determining the attitude or emotion of the writer ie, whether it is positive or negative or neutral.

The sentiment function of textblob returns 2 properties i.e., polarity and subjectivity.

- Polarity is a float which lies in the range of [-1,1] where 1 means the most positive statement and -1 is the most negative statement. 0 meaning the neutral statement.
- Subjective sentences is a float value in the range of [0,1] which generally refer to the personal opinion, emotion or judgment whereas objective refers to a factual information.

  We then plot the polarity vs objectivity scores of the sentiment of people in a scatter diagram.

4. DATA EXPLORATION

a) **Pie chart for polarity among the tweets**: The pie chart is to see distribution of tweets in terms of their sentiment polarity and it turns out that 49% - neutral, 43% - positive and 8% are negative.

b) **Boxplot for length of tweets**: Distribution of the length of the reviews under each sentiment class is shown as a boxplot where there are outliers.

c) **Histogram for length of tweets**: histogram is to see the max length of tweet in data collected.
Also when split over the positive, negative and neutral reviews, it was found that the tweets which were of neutral sentiment were of largest length.

d) **Word clouds for positive and negative words**: Word cloud is a visual representation of the words used in a particular piece of text with the size of each word indicating its relative frequency.

5. FEATURE EXPLORATION

The texts have to be represented as numbers to be able to apply any algorithm. Bag of words is a method where you count the occurrence of words in a document without giving importance to the grammar and order of words. This is achieved by creating a term document frequency(TDM). It is the matrix with terms in rows and document names in columns and count of frequency of words as cells of the matrix

Sklearn provides a function under feature selection called countvectorizer which counts the number of times the word occurred.

While counting words is helpful, longer documents will have higher average count values than shorter documents, even though they might talk about the same topics.
To avoid this we use the concept of TF-IDF. TF-IDF stands for *term frequency-inverse document frequency*, and the tf-idf weight is a weight which is used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus.

Typically, the tf-idf weight is composed by two terms

a) **TF: Term Frequency**, which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization:
TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).

b) **IDF: Inverse Document Frequency**, which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following:
IDF(t) = log_e(Total number of documents / Number of documents with term t in it).

Thus,

Instead of the Countvectorizer, I have used the TfidfVectorizer which also creates a document term matrix as the CountVectorizer but instead of filling the token counts, it calculates the tfidf of each word.

6. MODELING

The features obtained after applying the feature extractions techniques on the text sentences are trained and tested using the classifiers like logistic regression,support vector machines,K-nearest neighbor, decision tree, and Bernoulli naïve bayes.

In Machine Learning, classification is a directed learning approach in which the computer program gains from the information input given to it and after that utilizes this figuring out how to group new perception.

a) Decision Tree

The Decision tree is the classifier model in which every node of the tree appears as a test on the feature of the dataset, and its progeny symbolize the endings. The leaf node represents the last classes of the information. It is a supervised classifier model which utilizes information with realized names to shape the decision tree and after that, the model is connected to the test data.

b) Random Forest:

Random forest is an ensemble learning algorithm for classification and regression. Random forest generates a multitude of decision trees classified based on the aggregated decision of those trees. It is the most popular ensemble technique of classification because of the presence of best features . For a set of rows and their respective sentiment labels, bagging repeatedly selects a random sample (with replacement. Each classification tree is trained using a different random sample. Finally, a majority vote is taken of predictions of these trees.

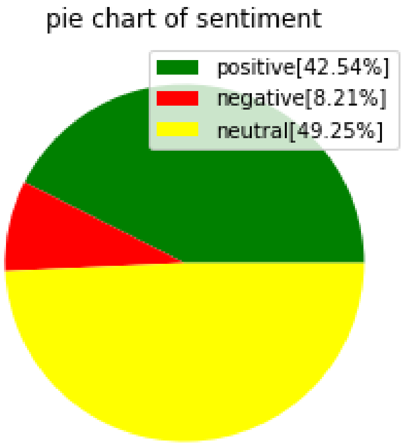
c) Support Vector Machine (SVM):

SVM (support vector machines) is a non-probabilistic double straight classifier. For a training set of points where is the feature vector, and is the class, with a need to find the maximum margin hyperplane that divides the points with and . The equation of the hyperp

d) Naïve bayes algorithm:

In ML, naïve bayes classifiers are a family of simple "probabilistic classifiers" based on applying the Bayes Algorithm with strong(naive) independence assumptions between the features. It is particularly useful for large datasets. Along with its simplicity naivebayes is known to outperform the highly sophisticated classification methods.

RESULTS OF THE STUDY



LIMITATIONS OF THE STUDY

Sentiment analysis on the raw data is always challenging due to the following reasons:

1. Positive and negative sentiment in n be each data.
2. Sarcasm using positive words in a negative way
3. There can be personal bias.

LIBRARIES USED

1. **TWEEPY**: is an open sourced python library which enables us to communicate with twitter platform and use its API.
2. **TEXTBLOB**: is more of a natural language processing library, but it comes with a rule based sentiment analysis library that we can use.
3. **VADER**(VALENCE AWARE DICTIONARY AND SENTIMENT REASONER): is a lexicon and rule based sentiment analysis tool that is specifically attuned to the sentiments expressed in social media.
4. **NLTK**(NATURAL LANGUAGE TOOLKIT): is a python library to make programs at work with natural language. It provides a user-friendly interface that are over 50 corpora and lexical resources such as WordNet Word repository. The library can perform operations such as tokenization, stemming, classification, parsing, tagging and semantic reasoning.
5. **MATPLOTLIB**: is a python library for 2d plots of arrays.it is a multi platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack.
6. **SEABORN**: python visualization library built on matplotlib. It allows you to make your charts prettier, and facilitates some of the common data visualization needs.
7. **PANDAS**: is a high level data manipulation tool built on top of the NumPy package and its key data structure is called the DataFrame. DataFrames allow you to store and manipulate the tabular data in rows of observations and columns of variables.
8. **SKLEARN**:is an open source python library that implements a range of machine learning, preprocessing, cross validation and visualization algorithms using a unified interface
9. **PIPELINE**: chains the several steps together once the initial exploration is done. For ex, some codes are meant to transform the features- normalize numericals, or turn the text into vectors to fill up the missing data, they are transformers; other codes are meant to predict the variables by fitting the algorithm, sucha s random forest or support vector machines, they are estimators. Pipeline chains all these together which can then be applied to the training data.