Sentiment Analysis and Topic Detection for Political Events Over an extended Course of Time

Akshay Singh
Northeastern University
Singh.aks@husky.neu.edu

Abstract

The goal of this project is to build a model that bridges the communication gap between the Indian government and population. With many social media platforms like Twitter gaining popularity, extremely large amounts of data are becoming available online. These contents are closely related to the preferences, attitudes and beliefs of the general public on a particular topic. To provide government with the approval or disapproval of the general population over their adopted policies and/or action (or inaction). I investigate the streaming tweets relevant to the government departments and ministers which are collected over different parts of the day (18th April, 2017). In the first part of the project Sentiment analysis is done on all tweets containing trending hashtags and in the second part of the project, the tweets containing these trending hashtags are used for topic modeling, thus linking each hashtag with a topic. Finally, the topics and sentiments are mapped. The results from this model can be used by governments as a reference for altering their unpopular policies and/or actions.

Keywords: Social Media; Twitter; Streaming tweets; hashtags; Sentiment Analysis; Topic Modeling

1. INTRODUCTION

The social media platforms like Facebook, Twitter, YouTube etc. in the recent times have become immensely popular, where millions of users register their response on various matters directly or indirectly affecting them. These Social networking services have very extended reach, transcending geographically boundaries and across various sections of the population as opposed to the traditional information gathering channels like surveys or traditional media etc. Communication and interactions in social media frequently reflect real-world events. Thus social media streams become perfect sensors of real-world events. It is difficult for the humans to read and summarize all the relevant content in terms of the expressed sentiment. Thus, there is a growing need for the automated analysis of this kind of data. This is a challenging task with foundations in natural language processing and text mining referred to as sentiment analysis. Sentiment analysis gives an effective and efficient means to expose public opinion timely which gives vital information for decision making in various domains. The focus of the model is to analyze the sentiments for recent events. Most of the research in sentiment analysis is done in analyzing product reviews on various E- commerce. But there is an increasing need to identify the real time events and analyze the related sentiments of population based solely on the textual data. For instance, identifying events like "Demonetization" (Indian government's decision to ban old currency notes), "Natural Disasters", "the riots during the Arab Spring" etc. from the tweets and then analyzing the population sentiment about it. There are Topic detection methods which can be used in identifying events by natural disasters, forming opinion about major political parties, urban monitoring, detecting abnormal activity which prevent possible misuses of online social platforms.

The steps implemented in this project are following:

- Collecting the tweets using twitter's Streaming API and filtering all the tweets in which Indian government departments or the minister heading the departments are tagged.
- This is done to make sure that all the tweets addressed to Government of India are captured.
- These tweets are collected on different parts of the day as the project also showcases the change in the sentiment of the population with time and potential stimulus that can change their preexisting opinion in a drastic manner.
- The change of the sentiment could be opposite i.e. either from being positive to negative or may reinforce the existing

sentiment as in, it might become more negative or more positive.

- After collecting the tweets, from different part of the days, a consolidated dataset is created and top ten most trending Hashtags are extracted from the generated dataset.
- Now, from the dataset set all tweets corresponding to each hashtag are segregated.
- After segregation of the tweets they are preprocessed to obtain the desired form.
- Now, to link the hashtags with a particular real time event or topic, a topic modelling (or detection) algorithm (LDA) is used on the segregated tweets for each hashtag and then hashtag is linked to the topic.
- After topic detection its turn to calculate sentiment associated with each hashtag.
- The sentiment of each tweet is calculated using the Supervised Learning Classifier called Naïve Bayes which is based on bag-of-words model.
- The Sentiment are categorized in to a 3 – class model, where the tweets will be categorized as either Positive, Negative or Neutral.
- Finally, the mapping between the topics and tweets is done using the hashtags is done for different part of the day.

2. RELATED WORK

Research in the area of sentiment mining started with product (Turney, 2002) and movie (Pang et al., 2002) reviews. Turney (2002) used Pointwise Mutual Information (PMI) to estimate the sentiment orientation of phrases. Pang et al. (2002) employed supervised learning with various set of n-gram features, achieving an accuracy of almost 83% with unigram presence features on the task of document level binary sentiment classification. Research on other domains and genres including blogs (Chesley, 2006) and news (Godbole et al., 2007) followed.

Another classification of polarity in tweets was (Go et al., 2009). The authors conducted a supervised classification study on tweets in English, using the emoticons (e.g. ":)", ":(", etc.) as markers of positive and negative tweets. (Read, 2005) employed this method to generate a corpus of positive tweets, with positive emoticons ":)", and negative tweets with negative emoticons ":(". Subsequently, they employ different supervised approaches (SVM, Naive Bayes and Maximum Entropy) and various sets of features and conclude that the simple use of unigrams leads to good results, but it can be slightly improved by the combination of unigrams and bigrams.

In (Pang, Lee and Vaithyanathan 2002), the authors apply supervised learning methods such as naïve Bayesian and support vector machines (SVM) to classify movie reviews into two classes.

In (Pak and Paroubek 2010), tweets containing emoticons are used as training corpus to avoid manual annotation. They split up the data so that happy emoticons form the positive annotated set and sad emoticons form the negative annotated set. Their results show that along SVM and CRF (conditional random fields), Naïve Bayes classifier performed best and bigrams outperformed unigrams as features.

In (Brody and Diakopoulos 2011), the authors present an automatic method which leverages word lengthening to adapt a sentiment lexicon specifically for Twitter and other social messaging networks. A different approach is made in (Davidov, Tsur and Rappoport 2010), where several Twitter tags and smileys have been used as sentiment labels to build a classification framework. The authors use different feature types (punctuation, words, n-grams and patterns) for classification and show that the framework successfully identifies sentiment types of the untagged tweets.

Wang et al. (2012) proposed a real-time sentiment analysis system for political tweets which was based on the U.S. presidential election of 2012. They collected over 36 million tweets and collected the sentiment annotations using Amazon Mechanical Turk. Using a Naive Bayes model with unigram features, their system achieved 59% accuracy on the four category classification.

O'Connor et al. (2010) investigated the extent to which public opinion polls were correlated with political sentiment expressed in tweets. Using the Subjectivity Lexicon (Wilson et al., 2005), they estimate the daily sentiment scores for each entity. A tweet is defined as positive if it contains a positive word and vice versa. A sentiment score for that day is

calculated as the ratio of the positive count over the negative count. They find that their sentiment scores were correlated with opinion polls on presidential job approval but less strongly with polls on electoral outcome.

3. DATASET

In this project, I used the dataset for training Naïve Bayes classifier, provided by sentiment140.com [1] which consisted of labelled dataset classifying tweets into Negative, Positive and Neutral.

The Dataset is a CSV file with emoticons removed. Data file format has 6 fields:

0 – The polarity of the tweet (0 = negative, 2 = neutral, 4 = positive)
1 – The id of the tweet
2 – The date of the tweet
3 – The query.
4 – The user that tweeted
5 – The text of the tweet.
Link in references: [1]

As the model is a three class model with positive, negative and neutral sentiments, this dataset is appropriate to train the naïve Bayes classifier.

4. METHODOLOGY

In this model the sentiment analysis and topic modelling of the tweets is done in parallel and then are mapped using the common hashtag.

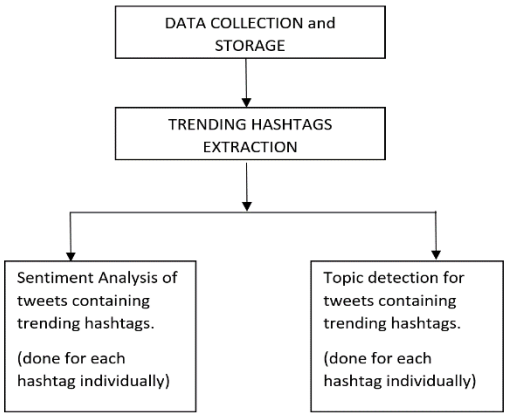

Figure 4 (a)

4.1 Data Collection and Storage

For the implementation of the model a total of 34,310 tweets were collected from twitter using its Streaming API's (via Python) on 18th April 2017. The tweets were filtered from the twitter stream only if they were tagged with the twitter handle of the various Government of India's ministries or the ministers who were in charge of those ministries. The name of each ministry were collected from the official Government of India website.

The tweet data downloaded were in JavaScript Object Notation (JSON) Format. The tweets were collected on four parts of the day i.e. 18th April 2017 as per the Indian Standard Time (IST). First batch of the tweets were collected in the morning, second batch of tweets were collected in the evening, third batch was collected in night and fourth batch was collected late after mid night.

The table depicting the tweets collected per hour during different parts of the day are following:

Time of the Day (18th April 2017)	# of Tweets per hour
Morning	5279
Evening	6598
Night	5759
Late-Night (After Mid-Night)	2159

Figure 4.1 (a)

The graphical representation of the above table is given below:

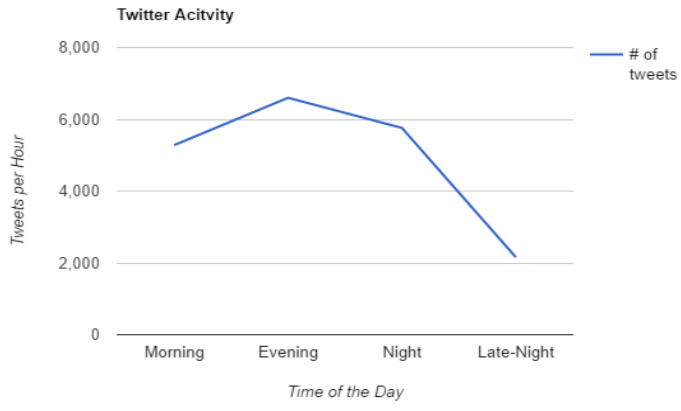

Figure 4.1 (b)

4.2 Trending Hashtags Extraction

For extracting the trending hashtags only text part of the tweets is needed, thus from tweet corpuses (JSON files) collected during various time of the day I calculated the cumulative count for each hashtag and selected the hashtags with top ten counts as the trending hashtags. The details of which are given in the table below:

Trending Hashtag	# of Tweets in which hashtag is present
#lahukalagaan	621
#bjymnec	276
#vijaymallya	265
#privateschoolfeehikeatrocity	129
#makeinindia	117

Figure 4.2 (a)

Brief background of these hashtags:

#lahukalagaan: A campaign which urges Indian Finance Minister Arun Jaitley to exempt the tax on women sanitary napkins by making it a non-luxury item under the new GST bill.

#bjymnec: National executive meet for young members of the current ruling party in India i.e. BJP and promotion of BHIM app aimed at encouraging digital payments in India. This app is seen crucial to track down money transaction and control black money.

#vijaymallya: An Indian businessman and former politician who absconded India to evade arrest, as he is unable to pay his debts. It became a major issue, with people asking that, law should be equal for everybody. Indians were unhappy with the current ruling party which allowed him to sneak to London during the trial. His arrest in London suddenly started trending on the evening of 18th April 2017, while the tweets for this model were being collected, more details on this are discussed in the results section.

#privateschoolfeehikeatrocity: Parents expressing their concern over the unregulated hike in fees by the private schools every year.

#makeinindia: A campaign started by Indian Prime Minister to attract foreign direct investment in India. It after a major company announced to set up its manufacturing plant in India.

4.3 A. Sentiment Analysis of Tweets containing Trending Hashtags

In this model sentiment analysis runs parallel to the topic modelling (or detection) which will be discussed later.

The flow diagram for the sentiment analysis is given below:

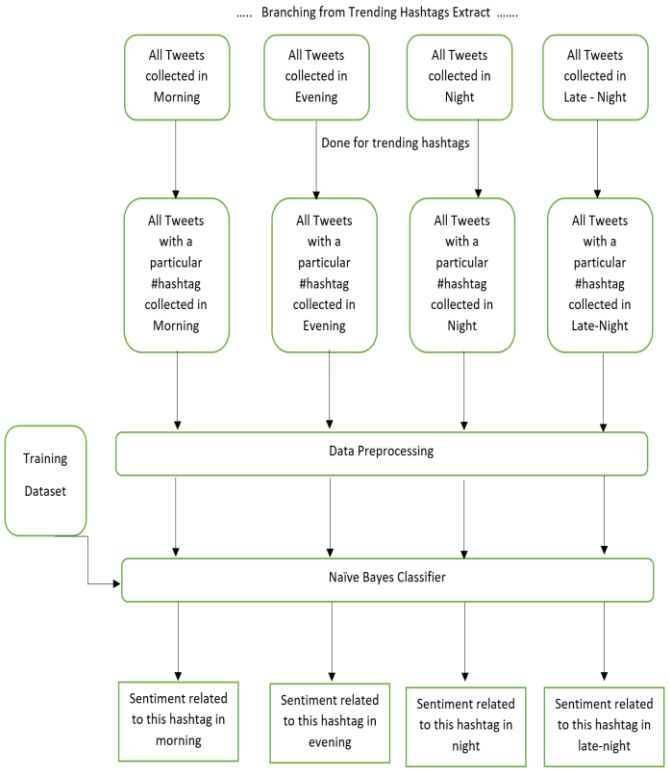

Figure 4.3 (a)

For each trending hashtags, the tweets containing them at different timestamps are extracted and then processed and finally sentiment analysis is done on them using the Naïve Bayes classifier.

Data Preprocessing

For data processing, Natural Language Toolkit (NLTK) is used, which is an open source platform for building python programs to work with human language data.

This toolkit provides a variety of text processing techniques such as tokenization, stemming, tagging, word segmentation etc. which were used for data processing in this model. Preprocessing mainly included the following three steps:

- Tokenization [2]: In a raw post, terms can be combined with any sort of punctuation and hyphenation and can contain abbreviations, typos, or conventional word variations. I used the tokenizer provider by the NLTK toolkit to extract the bags of cleaner terms from the original messages by removing stop words and punctuation, compressing redundant character repetitions, and removing mentions, i.e., IDs or names of other users included in the text for messaging purposes.
- Stemming [3]: In information retrieval, stemming is the process of reducing inflected words to their root (or stem), so that related words map to the same stem. This process naturally reduces the number of words associated with each document, thus simplifying the feature space. In my model I used NLTK stemmer packages.
- Aggregation. Topic detection methods based on word or n-grams co-occurrences, or any other type of statistical inference, suffer in the absence of long documents. This is the case of social media, where user-generated content is typically in the form of short posts. In information retrieval, it is common practice to partially address this problem by concatenating different messages together to produce super-documents of larger size. We build super-documents based on two strategies. The first involves temporal aggregation that glues together N messages that are contiguous in time. The second involves similarity-based aggregation that attaches to a message all the near-duplicate messages posted in the same time slot, identified through an efficient document clustering method.

Naïve Bayes Classifier [4]

The Naïve Bayes method for classification is often used in text classification due to its speed and simplicity. A flow diagram of a Naïve Bayes classifier is given below:

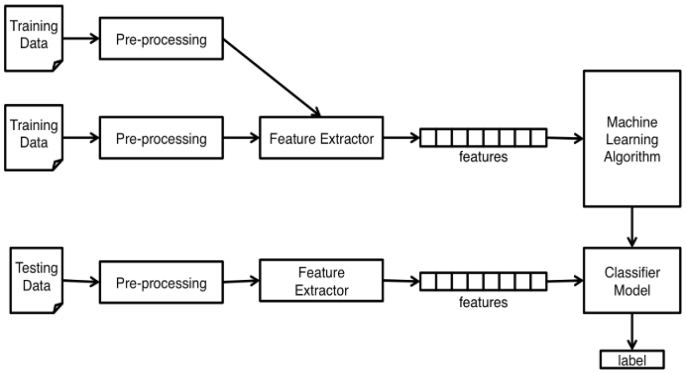

Figure 4.3 (b)

It makes the assumption that words (or k-grams) are generated independently of word position. For a given set of classes, it estimates the probability of a class, c, given a document, d, with terms, t, as:

$$P(c|d) = P(c) \prod_{1 \leq k \leq n_d} P(t_k|c)$$

The classifier then returns the class with the highest probability given the document. In practice, the log probability is estimated, given by:

$$\arg\max_c \log(\hat{P}(c)) + \sum_{1 \leq k \leq n_d} \log(\hat{P}(t_k|c))$$

i. Unigram Naïve Bayes

For unigram Naïve Bayes, the probability of a term given the class is given by the empirical counts of that term in messages with the same class. There are two methods to this, however, with one being the multinomial model and the other the Bernoulli model. In the multinomial model, the probability is given by

$$\hat{P}_{multi}(t_k|c) = \frac{T_{ct_k}}{|V_c|}$$

Which is the total number of times the term appeared associated with that class divided by the total number of terms seen for the class. This contrasts with the Bernoulli model where the count applies to the number of documents containing the term for that class over the total number of document for the class.

ii. Bigram Naïve Bayes

The bigram language model departs from the bag-of-words model and calculates the probability that a document belongs to a class by calculating the probability that each word comes after the word before it given the class based on the empirical counts of word pairs seen for the given class.

In this case, however, the training set appears to be too sparse for solely using bigrams for classification, so a linear interpolation model as well as a back off model is used. The linear interpolation model weights the unigram and bigram probabilities to determine the overall probability of a document belonging to a class:

$$P(c|d) = \alpha P_{unigram}(c|d) + (1-\alpha) P_{bigram}(c|d)$$

The back off model, on the other hand, uses the bigram probability for a term if the bigram has been seen with this class. Otherwise it backs off to the unigram probability.

After the process of Sentiment analysis using the Naïve Bayes classifier, we get the sentiment of each hashtag (or in other word all the tweets containing that particular trending hashtag) during different parts of the day i.e. in the morning, evening, night and late – night of 18th April 2017 since it is the day model was implemented.

4.3 B. Topic Modelling (Detection) of Tweets containing Trending Hashtags

In this model, topic modeling is done by using Latent Dirichlet Allocation (LDA) which is a generative probabilistic model of a corpus.

According to LDA, every document is considered as a bag of terms, which are the only observed variables in the model. The topic distribution per document and the term distribution per topic are instead hidden and have to be estimated through Bayesian inference

The flow diagram for topic modelling is given below:

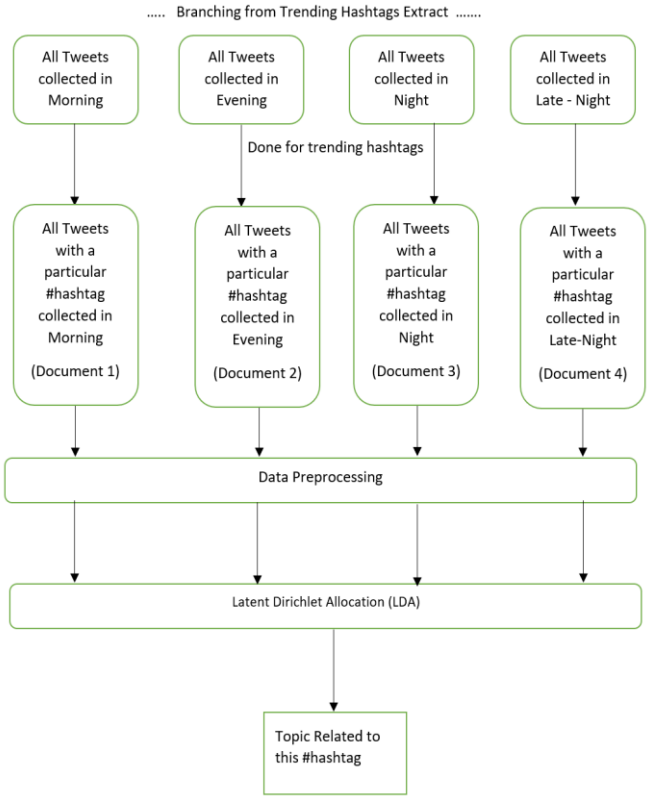

Figure 4.3 (c)

For each trending hashtags, the tweets containing them at different timestamps are extracted and Data Preprocessing similar to section 4.3 A. is performed on the tweets and then these processed tweets are passed as documents to the LDA algorithm and the detected topic for each trending hashtag is given as output.

Latent Dirichlet Allocation (LDA) [5]

LDA assumes the following generative process for each document w in a corpus D:

1. Choose N ~ Poisson (ξ).
2. Choose Θ ~ Dir(α).

3. For each of the N words w_n:
   (a) Choose a topic z_n ~ Multinomial(Θ).
   (b) Choose a word w_n from p (w_n| z_n,), a multinomial probability conditioned on the topic z_n.

First, the dimensionality k of the Dirichlet distribution (and thus the dimensionality of the topic variable z) is assumed known and fixed. Second, the word probabilities are parameterized by a k X V matrix β where $\beta_{ij} = p(w^j = 1 | z^i = 1)$. A k-dimensional Dirichlet random variable Θ lies in the (k - 1)- simplex, and has the following probability density on this simplex.

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1}$$

Where the parameter α is a k-vector with components $\alpha_i > 0$, and where Γ(x) is the Gamma function. The Dirichlet is a convenient distribution on the simplex – it is in the exponential family, has finite dimensional sufficient statistics and is conjugate to multinomial distribution.

Given the parameters α and β, the joint distribution of a topic mixture Θ, a set of N topics z and a set of N words w is given by:

$$p(\theta,\mathbf{z},\mathbf{w}|\alpha,\beta) = p(\theta|\alpha)\prod_{n=1}^{N}p(z_n|\theta)p(w_n|z_n,\beta)$$

Where $p(z_n|\Theta)$ is simply Θ_i for the unique I such that $z_n^i = 1$, Integrating over Θ and summing over z , marginal distribution of a document is obtained.

$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n}p(z_n|\theta)p(w_n|z_n,\beta)\right)d\theta$$

Finally, taking the product of the marginal probabilities of single documents, probability of the corpus is obtained:

$$p(D|\alpha,\beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}}p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right)d\theta_d$$

The probabilistic graphical representation of the LDA model is given below:

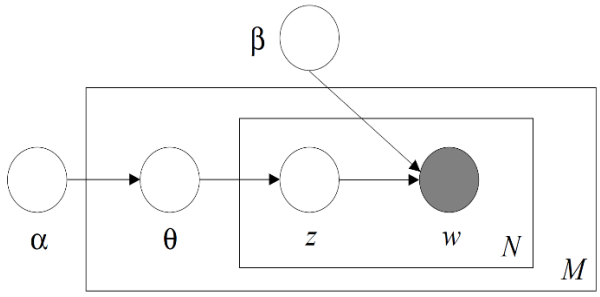

Figure 4.3 (d)

There are three levels to the LDA representation. The parameters α and β are corpus level parameters, assumed to be sampled once in the process of generating a corpus. The variables Θ_d are document level variables sampled once per document. Finally, the variables z_{dn} and w_{dn} are word – level variables and are sampled once for each word in each document.

4.4 Mapping the Sentiments obtained during different time of the day with the detected topics for the collected tweets using Hashtags.

The obtained sentiments for the particular trending hashtags during different part of the day (18th April 2017) are mapped with the topics detected for the same trending hashtags.

After mapping it will be clear that which topics had positive or negative or neutral sentiment during different part of the day (on which model was executed)

The flow diagram for this process is given below:

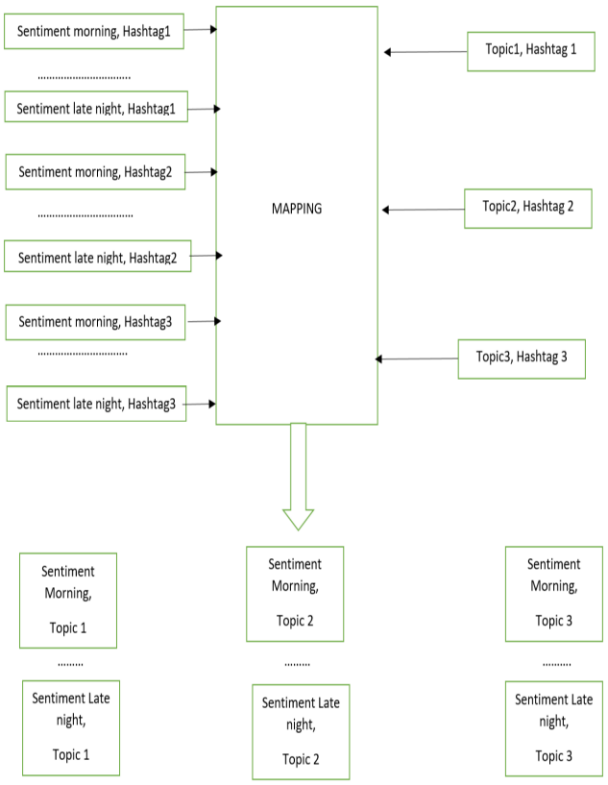

Figure 4.4 (a)

5. RESULTS

This project aims at not only identifying the current trending topics but also tries to analyze the sentiments of the population towards the overall governance of the Government. The project also tracks how the sentiments of the population mature with time. In the result section I will put forward the topic extracted for each trending hashtag and sentiments for the hashtag on the day of the model execution i.e. 18th April 2017. I will also point out some interesting observations which acted as a catalyst in triggering abrupt change in the sentiments of the population altered the sentiments of the people. The results for the Hashtags are following below:

5.1 Hashtag: #lahukalagaan

This hashtag symbolizes a campaign to change category of women sanitary napkins to non – luxury items, thus resulting in exemption of tax over it under the new GST Bill.

a) *Sentiment Analysis of #lahukalagaan:*

This hashtag showed a continued Negative sentiment (as the ration was less than 1) throughout the day (except morning) and population is, in general unhappy about the taxation of sanitary napkins.

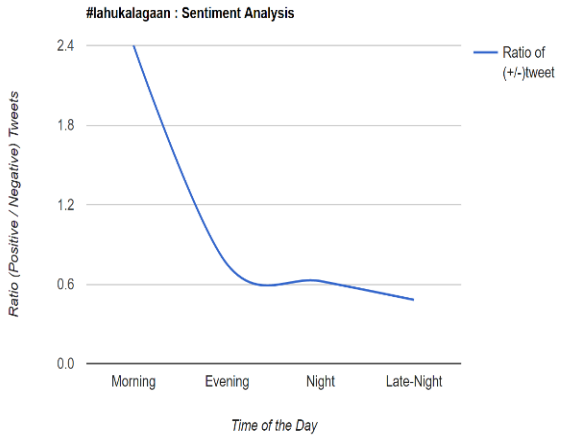

Figure 5 (a)

Observation:

1. The graph shows the steady negative sentiment after the evening time, after the initial positive sentiment during morning.
2. Although the graph shows the positive sentiments initially, upon analyzing the tweets I observed that the tweets were not supporting the taxation but the people were more optimistic about tax removal. This highlights the importance of the context in which the positivity or negativity of the tweets are analyzed for a particular hashtag.

b) *Topic Modelling (or Detection) of #lahukalagaan:*

The LDA algorithm was fed with four documents each corresponding to the part of the day under analysis and containing all the tweets with hashtag #lahukalagaan.

The output given out by the LDA algorithm is:

0.076*"lahukalagaan" + 0.056*"sanitari" + 0.044*"napkin" + 0.042*"tax" + 0.035*"free" + 0.024*"women"

Which is the accurate description of hashtag.

Each word is ordered with the probability of their occurrence and "+" shows that these words exists most frequently in the given cluster, thus hinting towards a topic.

c) *Mapping the results:*

The hashtag now serves as the linking entity between the sentiment and topic. The consolidated result for this topic is given below:

Time of the Day	**Sentiment for the topic** (Based on Ratio of +/- tweets): "lahukalagaan" +"sanitari" + "napkin" + "tax" + "free" +
Morning	Positive (ratio > 1)
Evening	Negative (ratio < 1)
Night	Negative (ratio < 1)
Mid - Night	Negative (ratio < 1)

Figure 5 (b)

5.2 Hashtag: #vijaymallya

This hashtag records the arrest of Vijay Mallya which happened in London on 18th April 2017. More details on how his arrest is related to government is explained in Section 4.2.

a) *Sentiment Analysis of #vijaymallya:*

This hashtag has the most interesting observations, as the event of the arrest was occurring in real time simultaneously as the tweets about his arrest were being collected. The population has a varying response to his arrest and then his bail after about 2-3 hours.

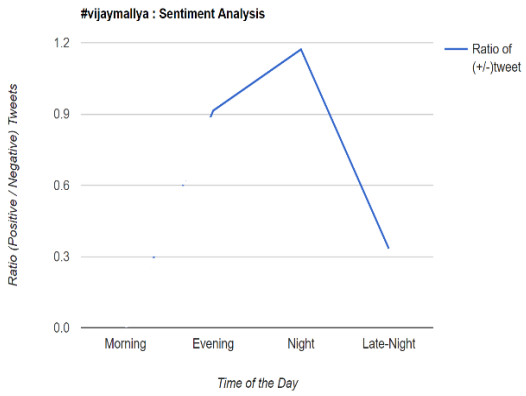

Figure 5 (c)

Observation:

1. In the Morning, there were no tweet about him as he was not arrested till then.
2. During the evening time, news of his arrest was received and people had negative sentiment about him, which gradually became little positive towards the night. The change was registered, because people started praising the Indian government after his arrest.
3. The period from Night to Late Night (i.e. after Mid Night) saw sentiment going negative abruptly with increased magnitude, this happened as the news about his bail just after 2-3 hours after his arrest started emerging. People were very upset, with some even suggesting that the arrest was staged by the government.

b) *Topic Modelling (or Detection) of #vijaymallya:*

The LDA algorithm gave the below topic for the hashtag #vijaymallya:

0.088*"vijaymallya" + 0.038*"arrest" + 0.030*"congratul" + 0.027*"promis" + 0.027*"fullfil" + 0.022*"govt" + 0.020*"bail" + 0.018*"ensur" + 0.018*"thevijaymallya" + 0.018*"london"

Topic is accurate and correctly describes the hashtag.

c) *Mapping the Results:*

Time of the Day	**Sentiment for the topic** (Based on Ratio of +/- tweets): "vijaymallya “,"arrest “,"congratul “,"promis “,"fullfil “,"govt “,"bail “,"ensur “,"thevijaymallya “,"london"'
Morning	Not Applicable as No tweets were collected.
Evening	Negative (ratio < 1)
Night	Positive (ratio > 1)
Mid - Night	Extremely Negative (ratio < 0.5)

Figure 5 (d)

5.3 Hashtag: #makeinindia

This hashtag symbolizes the Government on India's initiative to attract the foreign direct investment.

a) Sentiment Analysis of #makeinindia:

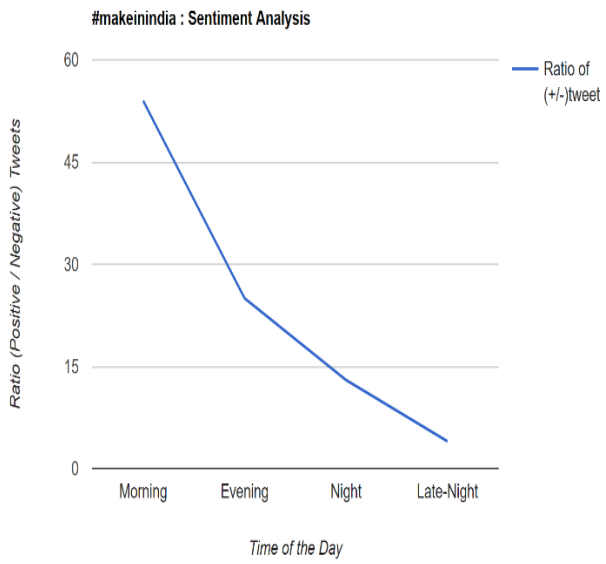

Figure 5 (e)

Observation:

1. This hashtag remains positive throughout the day and changes only in the degree by which it is positive. (ratio is > 1)
2. The hashtag started trending when KIA motors decided to invest 10 crore rupees (USD 1,552,000.00) in India which will boost the foreign direct investment.

b) Topic Modelling (or Detection) of #makeinindia:

The LDA algorithm gave the below topic for the hashtag #makeinindia –

"makeinindia" + 0.061*"boost" + 0.058*"plan" + 0.057*"get" + 0.057*"invest" + 0.056*"kia" + 0.055*"motor" + 0.051*"govt" + 0.050*"thank" + 0.050*"10k"

The calculated topic is fairly accurate and describes the hashtag properly.

c) Mapping the results:

Time of the Day	Sentiment for the topic (Based on Ratio of +/- tweets): "makeinindia","boost","plan","get","invest","kia","motor","govt","thank","10k"
Morning	Extremely Positive (ratio > 50)
Evening	Extremely Positive (ratio > 25)
Night	Highly Positive (ratio > 13)
Mid - Night	Positive (ratio > 4)

Figure 5 (f)

5.4 Hashtag: #privateschoolfeehikeatrocity

This hashtag shows the concern of the working class parents regarding the fees of the private schools which affects their monthly budget severely.

a) Sentiment Analysis of #privateschoolfeehikeatrocity:

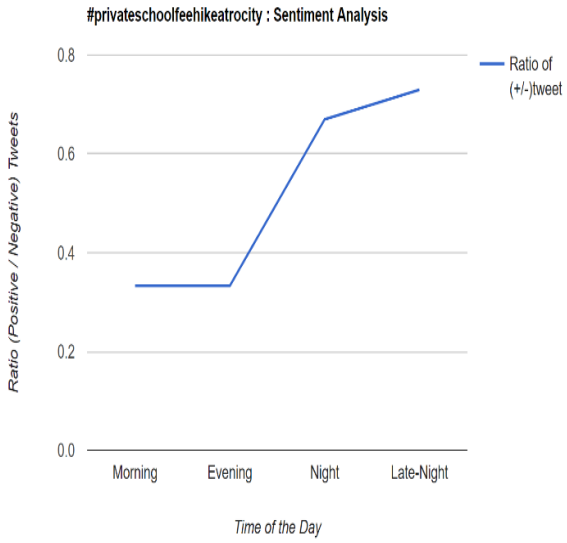

Figure 5 (g)

Observation:

1. The sentiment regarding the hike in fees remains negative throughout the day, just changes in magnitude from extremely negative to negative.

2. The other important thing to note is that the twitter activity with this tag increases after night indicating the certain section of working class, which find time to go online on social media after returning home after day of work, since the tweets were collected on a week day.

b) Topic Modelling (or Detection) of #privateschoolfeehikeatrocity:

The LDA returned below topic related to this hashtag:

'0.104*"privateschoolfeehikeatroc" + 0.024*"fee" + 0.020*"school" + 0.020*"educ"'

The calculated topic only partially describes the hashtag

c) Mapping the results:

Time of the Day	**Sentiment for the topic** (Based on Ratio of +/- tweets): "privateschoolfeehikeatroc "," fee ","school ","educ"'
Morning	Highly Negative (ratio < 0.5)
Evening	Highly Negative (ratio < 0.5)
Night	Negative (ratio < 1)
Mid - Night	Negative (ratio <1)

Figure 5 (h)

5.5 Hashtag: #bjymnec

This hashtag was trending because it's related to national executive meet of young ruling party members (BJP) and also to promote digital payment BHIM app.

a) Sentiment Analysis of #bjymnec:

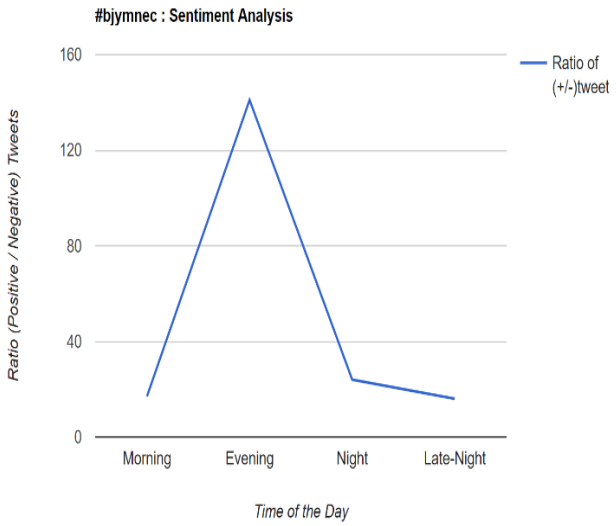

Figure 5 (i)

Observation:

1. The sentiment regarding the meeting event always remains positive, however it noteworthy that it increases in positive sentiment around evening, indicating the promotion / start of the of the event.

2. The positive sentiment decrease in magnitude towards the end of the day, possibly indicating the end of the event.

b) Topic Modelling (or Detection) of #bjymnec:

The LDA algorithm returned below topic, related to this hashtag.

'0.090*"bjymnec" + 0.082*"bjym" + 0.052*"amp" + 0.042*"bhim" + 0.041*"promot" + 0.041*"young" + 0.041*"leader" + 0.041*"shri" + 0.041*"payment" + 0.041*"address"'

The calculated topic describes the hashtag accurately

c) *Mapping the Results:*

Time of the Day	**Sentiment for the topic** (Based on Ratio of +/- tweets): bjymnec “,"bjym “,"amp “,"bhim “,"promot “,"young “,"leader “,"shri “,"payment “,"address"'
Morning	Positive (ratio > 1)
Evening	Extremely Positive (ratio > 100)
Night	Highly Positive (ratio > 20)
Mid - Night	Positive (ratio > 1)

Figure 5 (j)

CONCLUSION

In this project, a model is implemented which fills a communication gap between the government and the population, the model allows governing agencies to directly know about new problems faced by a larger section of the society almost instantly or the majority opinion on the issues crucial to the state without the need of any intermediary.

In the results, it has been observed that the population does respond to current events actively for instance in the case of Vijay Mallya the sentiment of the population changed abruptly to extreme negative as soon as the news of his bail came into the picture. Therefore, such systems are the need of the hour to create a real time self-check mechanism on government so that, they are aware of the people reaction and can modify their stand accordingly in public interest.

FUTURE WORK

In context of India, where there are more than twenty official languages, considering only English language for analysis leaves out a large portion of the population and their opinions are not considered, which could result in the wrong calculation of public sentiment or public opinion. In Future, efforts will be made to develop system for all the Indian languages supported by twitter and other social media platforms so that the analysis are more accurate and represents all sections of the society equally.

REFERENCES

[1]Training Dataset:https://docs.google.com/file/d/0B04GJPshIjmPRnZManQwWEdTZjg/edit

[2] http://www.nltk.org/api/nltk.tokenize.html

[3] http://www.nltk.org/api/nltk.stem.html

[4]https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[5] Latent Dirichlet allocation: David M. Blei, Andrew Y. Ng, Michael I. Jordan; 3(Jan):993-1022, 2003