IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

A Text Mining Approach to Identify and Analyse Prominent issues from Public Complaints

Febina Merson[1], Roseline Mary[2]

Student, Department of Computer Science, Christ University, Bangalore, India [1]

Assistant Professor, Department of Computer Science, Christ University, Bangalore, India [2]

Abstract: The emergence of Social Media has given a platform to the public to articulate their interests and opinion. Social Media has a great impact on the public's thoughts and opinion and the public does not hesitate to express their thoughts through Social Media. Lately, Social Media is increasingly used in Political Context. Through Social media the public have got a platform to voice their opinion about anything and everything, this also include politics. Due to the transparency of the social media, any issues raised will reach out to the whole world. Hence the participation of the citizens in government related issues will make the authorized officials accountable to look into the issues immediately. These opinions pertaining to political institutions and government are gathered from micro blogging services and social networking sites to identify the efficiency of the government. This paper discuss an approach where data collected from social media are preprocessed and classified to identify the major issues encountered by the public in their daily life. In this paper we propose a method to identify the most important government related issues from the perspective of the public.

Keywords: Social Media**,** Text Categorization, Text Mining, Data Pre-processing, Opinion Mining**.**

I. INTRODUCTION

In the past few years, the world has seen a great evolution due to social media. With each passing day, more people are being a part of this new world emerged due to social media. As per the scientific analysis, there is a speedy growth in the number of users joining the social media. People use Social media to voice their opinion and to get information about materials of their interest and about everything happing in this world. Until few years back, the platform to reach out to a large number of people was through Television and Print media. Using this platform to voice ones opinion was costly and this was the major barrier that the public experienced in the past. Also, this platform was not easily available to all. Due to these reasons, public did not have much power in their hands. With the emergence of Social media, the public is more powerful. Today a platform is available to everyone around the globe, which is widespread and easily accessible to everyone. Social Networking sites like Facebook are used by people across the globe. Using this medium, one can locate individuals who share common interests and thus they can share ideas on various topics of their choice. Today there are social networking sites available exclusively for a particular subject. People who are interested in the particular subject associate with such sites. Opinions that are posted on these sites reflect the individual's state of mind and this Information become a major source for social media analytics. Today there are SNS available that are used in political context, which are used by the citizens to voice their opinions and issues. Any information about government posted on social networking sites are crucial information. Political Institutions and Government services make use of this information to improve the services and communications with citizens. Using techniques such as text mining, this valuable information's are extracted for analysis.

The objective of this research is to mine these opinions with the intent to uncover the major issues that are faced by the citizens. This would in turn help the government to discover the major issues and make essential changes to eliminate the problem. In this paper we explain how the complaints from a SNS which is used in political context, is mined and preprocessed and then examine the effectiveness of identifying the major issues reported by the citizens. The complaints that are gathered goes through several preprocessing steps, to get the relevant contents for training a model.

II. RELATED WORK

The birth of social media and micro blogging sites has given the world the freedom to express their opinion and this has resulted in an abundance of information available on web. These opinions are used in various research and analysis which would benefit diverse areas like marketing, government, business, sports, entertainment and so on. Before the emergence of social media, feedback was gathered manually, by asking individuals about their experience of using a product or about their opinion on a product [5]. This feedback was then used by the marketing team and business personals to improve their product or service. The evolution of social media made the process of collecting feedback easier. Marketing and Business areas made extensive use of the abundant data available through social media and wide range of open- source and social media analytical tools were developed to analyze the data [5]. Lately, social media is seen to be used a lot in political context. Politicians across the globe are using social media to converse with citizens and for political discussion [5].

Copyright to IJARCCE

DOI 10.17148/IJARCCE.2017.6313

The increased use of social media in political context has benefitted the government to understand the people's mindset and to make imperative advancement. Political Institutions and Government agencies frequently congregate and analyze the data from Social media environment to better the communication with citizens. Government officials make use of these data to upgrade the facilities provided to citizens [1]. Political practices such as calibrating and administering public opinion through voting and media are the foundation of liberal democracies. With the availability of Text mining and Natural Language Processing, these methods are going through a conversion [6]. Text Mining is even used in digital newspaper to evaluate the political sentiment [7]. Frameworks are available that propose a text mining approach that would study the media broadcasting on matters related to politics [10]. There are several applications that are available today, which are used by citizens to ask questions to the members of the parliament and they also get their answers to those questions from the parliament members [9]. Availability of platforms to express public opinion has attracted political parties to pay attention to the public attentively. This has heightened the clarity and responsibility in establishing the democracy [9].

Analysts need only relevant information to be extracted from the data collected from the different networking sites, and they could personalize based on their operational needs [8].There are considerable amount of text categorization methods and algorithms that are developed in many years [2-3]. The task of classifying text documents into predefined classes is called Text categorization. The approach of text categorization, benefit in boosting the performance. Several techniques of statistical classification are tested on text categorization. The prominent advantages of the derived classifier are its dependence on the data that are provided by the users, the smooth design, flexible to customization based on personal interest. [11]. One of the most popular methods for text categorization is the bag-of-words method [4]. Terms in a sentence can be assigned a weight based on its occurrence in the entire document [4] A form of linguistic content shows a fair segregation of the words that are nouns, verbs and adjectives from the other set of words [8]. Words that are of most significance can then be extracted from this set of words. easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

III. DATA PREPROCESSING

The opinions and complaints posted on social networking sites are straight from people's minds, due to which the nature of the language could vary. Along with useful information, most of the posts contain unwanted information as well. These posts could be incomplete or inconsistent and would have grammatical errors, spelling mistakes and even slangs. Hence it is necessary to institutionalize certain tokens on these posts. This is achieved by performing pre-processing of the collected data and then use this preprocessed data for training a model. The following steps had to be performed to order and clean the data, to use this for analysis.

A. *Tokenization*

This is the first step of pre-processing where the steam of text is split into words or meaningful elements called as tokens. The purpose of tokenization is the examination of words in a string. This is done by splitting the words with white spaces resulting in a list of words or tokens that are given as input for further processing.

B. *Stop Words Removal*

Stop words are words that do not have any important significance and that are frequently used in a language. Some of the stop words include "a", "and", "as", "the", "that", "is", "was", "or", "also", "then". Since these words don't have special meaning or significance, they can be removed from the list of words received after tokenizing. Removing the stop words are done by comparing the list of words against a stop list which contains a list of stop words, and then removing the matching words.

C. *Lower Casing*

Opinions or comments posted on social media are found to have uneven casing (Such as UnEvENCAsinGToLOweRCaSINg). It is important to have consistency in the casing of the words in order to make sure that tokens are mapped to the respective feature.

D. *Stemming*

For grammatical reasons, affixes are added to words to form a proper meaningful sentence (such as achieve, achieving, achieves). These words, with or without affix still have the same meaning. The list of words generated after tokenization would have several words that are nouns, verbs and adjective of the same word. This would increase number of words in the token list and also cause redundancy. For efficiency, the affixes of these words can be cropped and structure the words to a one common form. This process is called stemming. We used Porter Stemmer to implement the stemming operation.

IV. METHODOLOGY

This section would explain about the techniques used to extract the useful information from the extensive amount of data.

A. *Data Collection Approach*

As there are only few public complaints available on facebook and twitter on government related issues, most of the data was collected manually from an online social networking platform that is used by the citizens, exclusively to voice their issues. Issues of 3 major cities (i.e.; Bangalore, Delhi and Mumbai) are collected for model construction.

DOI 10.17148/IJARCCE.2017.6313

B. *Feature Extraction*

List of words that are useful for model construction are selected as features and the remaining large set of words that are not useful are removed. This process is called feature extraction. This process will remove the noise from the sentence and will retain only the necessary information for the analysis. The reduced representation will then be given as the input to execute the desired task. This process will streamline the models and will make it effortless to understand.

I. *Unigram*

This is an approach of Feature extraction where words are sampled independently of each other. In a Unigram, one word from a sequence of text or speech is fragmented independently. Unigram can be performed on characters, texts and even sentences.

II. *N- Gram*

In an N-gram Feature, words are sampled as adjoining sequence of n items (word, character, syllables). 'N' could be any number. The value of 'N' will decide the number of word or characters to form a contiguous sequence. If N= 1 then it is a unigram and in a unigram, sequence of text are fragmented independently. If N=2, then it is a bigram and it generates sequence of two adjacent words or characters. (e.g. "This is an example of bigram" will become "This is", "is an", "an example", "example of", "of bigram". In this research, N grams of size N= 1 and N= 2 are used.

C. *Feature Filtering*

To construct a model with only relevant information for the analysis, the volume of corpora has to be reduced further. This is done by extracting the nouns from the collection of words. This would remove all the words that are verbs/adjectives. After this a weighing scheme called as a Term Frequency is used to ensure that the resulting dataset will have only words that are significant to analysis.

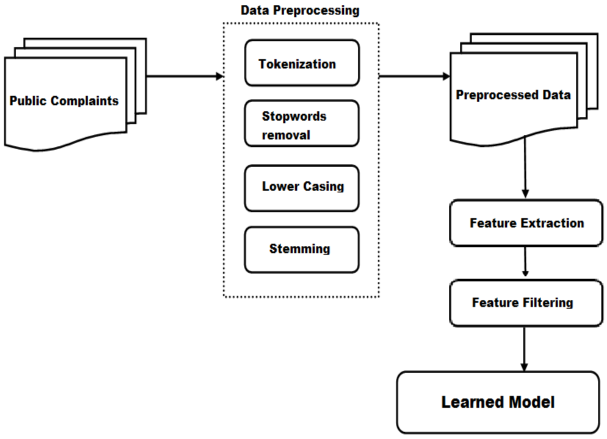

Fig 1: Process flow for Model Construction

V.RESULTS AND DISCUSSION

After reducing the corpora by performing several preprocessing steps, feature extraction and feature filtering, the resulting dataset would have only the issue filtered out from each complaint in the document.

After analyzing the dataset for the 3 cities, the major issues for each of the cities were detected. This will help the government and political institutions to identify the major concern experienced by each city and to prioritize and work towards fixing these issues. If a particular issue was reported the most compared to other issues, then it is likely that other citizens in the city would also be experiencing the same issues. Fixing this issue will not only appease the ones who complained but also the other citizens who are experiencing the same issue. This will in turn reduce future complaints on the same issue.

The result of the observation for all the 3 cities (Bangalore, Mumbai and Delhi) is shown below.

Complaints reported from Bangalore

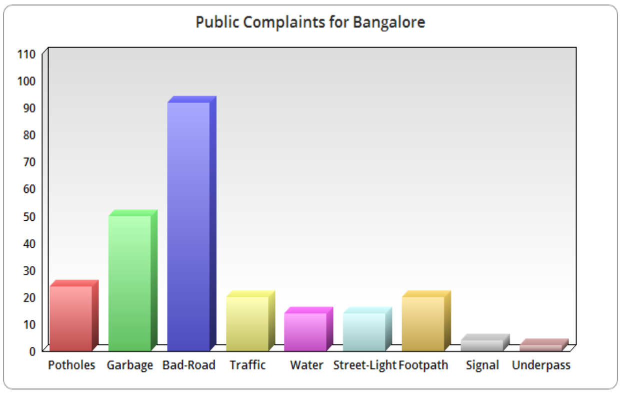

Fig 2: Major Issues reported for Bangalore

*

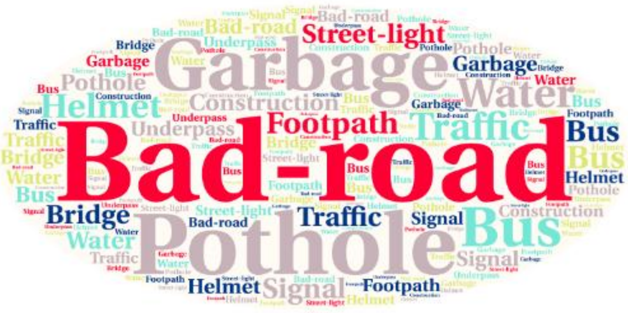

Fig 3: Tag Cloud of Issues reported for Bangalore

DOI 10.17148/IJARCCE.2017.6313

Table 1: Issues and the number of issues reported for Bangalore

Issue	City - Bangalore
Potholes	24
Garbage	50
Bad-Road	92
Traffic	20
Water	14
Street-Light	14
Footpath	20
Signal	4
Underpass	2

Analysis on complaints from Bangalore as shown in Fig 2 shows that the major issue faced by the people in Bangalore is about Bad Roads. Around 38% of the complaints that are collected are about bad roads. The ratio of complaints for road issues is almost twice as the second major issue which is garbage. Around 20% of the complaints are regarding garbage. The third major issue is about potholes, which can be combined with bad roads. By fixing the first major issue, which is bad roads, the complaints on bad roads and potholes will subsequently reduce.

Complaints reported from Mumbai

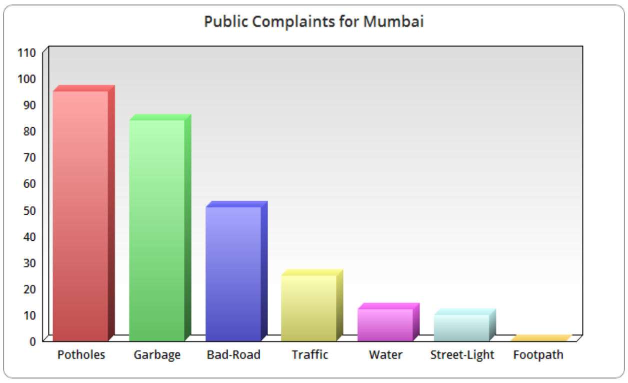

Fig 4: Major Issues reported for Mumbai

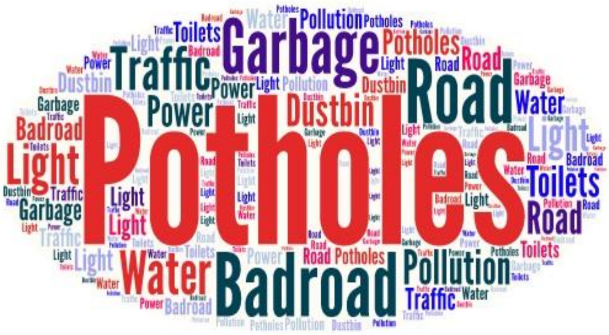

Fig 5: Tag Cloud of Issues reported for Mumbai

Table 2: Issues and the number of issues reported for Mumbai

Issue	City - Mumbai
Potholes	95
Garbage	84
Bad-Road	51
Traffic	25
Water	12
Street-Light	10
Footpath	0
Signal	0
Underpass	0

Analysis on complaints from Mumbai as shown in Fig 4 shows that the major issue faced by the people in Mumbai is about Potholes. From the complaints collected from the Social Networking Site, 34% of the complaints are about Potholes. The second major issue is garbage which constitute to 30% of the complaints. The third major issue is about bad roads. By fixing the first major issue, which is potholes, the complaints on bad roads and potholes will subsequently reduce.

Complaints reported from Delhi

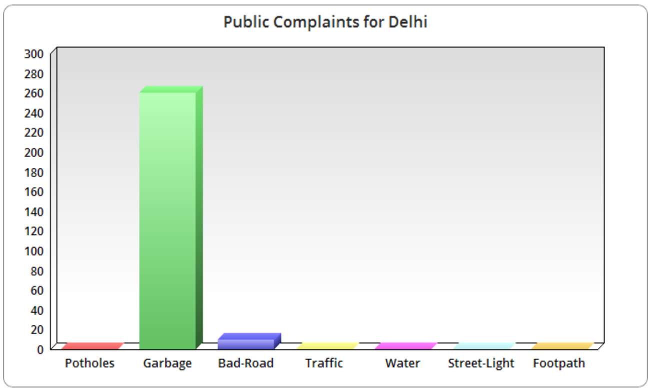

Fig 6: Major Issues reported for Delhi

Table 3: Issues and the number of issues reported for Delhi

Issue	City - Delhi
Potholes	0
Garbage	260
Bad-Road	10
Traffic	0
Water	0
Street-Light	0
Footpath	0
Signal	0
Underpass	0

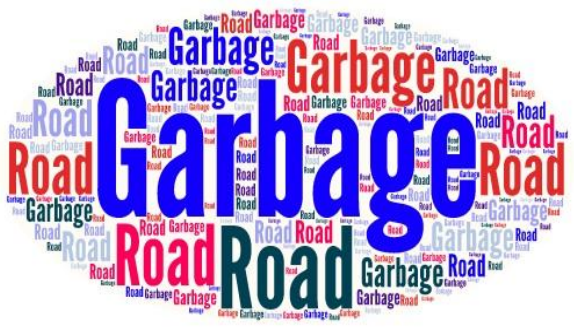

Fig 7: Tag Cloud of Issues reported for Delhi

Analysis on complaints from Delhi as shown in Fig 6 shows that the major issue faced by the people in Delhi is about Garbage. There were only few complaints reported on the second issue which is about roads. In Delhi 96% of the complaints are reported for garbage. Fixing this issue will eradicate the public's major concern and will also reduce the future complaints from Delhi.

Complaints Report from Bangalore, Delhi and Mumbai

Table 3: Issues and the number of issues reported for Mumbai, Bangalore and Delhi

Issue	Bangalore	Mumbai	Delhi
Potholes	24	95	0
Garbage	50	84	260
Bad-Road	92	51	10
Traffic	20	25	0
Water	14	12	0
Street-Light	14	10	0
Footpath	20	0	0
Signal	4	0	0
Underpass	2	0	0

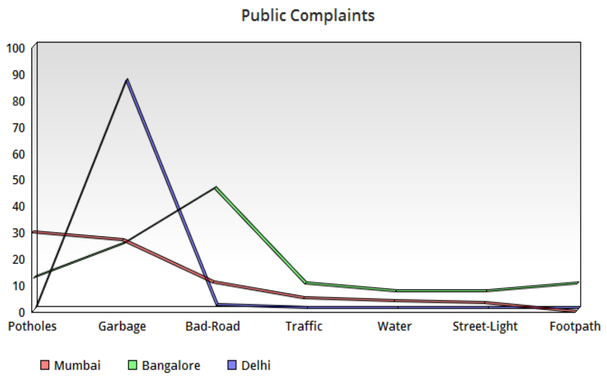

Fig 8: Complaints from Mumbai, Bangalore and Delhi

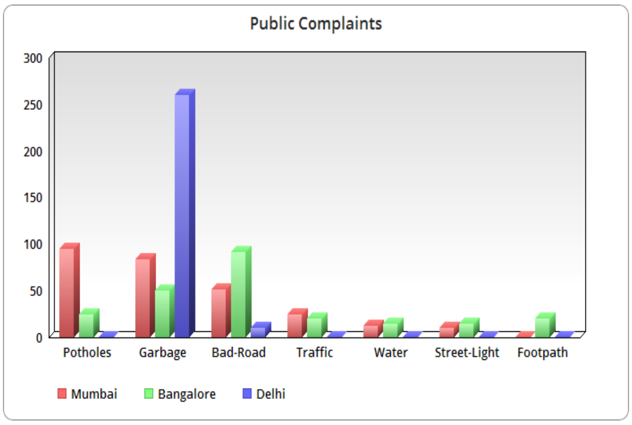

Fig 9: Consolidated report from Mumbai, Bangalore and Delhi

This report shows that, there are several issues for which complaints are reported for Bangalore and Mumbai, out of which the major issues found for both Mumbai and Bangalore, are bad road and potholes. Whereas for Delhi, complaints are majorly for one issue and that is garbage.

VI. CONCLUSION/ FUTURE WORK

In this paper, we analyzed the complaints reported from three major cities of India (i.e. Bangalore, Mumbai and Delhi) to identify the major issue faced by the people in the respective cities. We performed several preprocessing steps, feature extraction and feature filtering on the data that was collected. After filtering out all the noise, the resulting dataset had only the Issues reported by the public (e.g. "Pothole", "Garbage"). After analyzing the dataset for Bangalore, we found out that 38% of the complaints that are collected are about bad roads, which is the major concern of the people in Bangalore. In Mumbai, 34% of the complaints are about Potholes and 18% of the complaints are about Bad roads. In Delhi 96% of the complaints are reported for garbage. Fixing the major issues identified in the respective cities will eradicate the public's major concern in the respective cities and also will diminish the future complaints.

In Future, the results could be further analyzed by filtering the issue along with the locality (within the city) of the issue. By filtering the locality along with the issues, we can identify the area where most of the issues are reported from and if the issues are common to the people in that locality. If there are many complaints coming from the same locality on the same issue, then it would indicate the criticality of the issue and it will help the government in prioritizing and fixing the issue immediately.

REFERENCES

[1] A.L Kavanaugh, E.A Fox, S. Yang and L.T Li, “Social media use by government: from the routine to the critical”, Digital Government Innovation in Challenging Times, College Park, Maryland, pp. 480- 491, June. 2011.

[2] P. S. Jacobs, “Joining statistics with NLP for text Categorization," in Proceedings of the third conference on Applied natural language processing, GE Research and Development Center, New York, pp. 178-185, 1992S.

[3] R. E. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization," Machine learning, AT&T Labs, Shannon Laboratory, Florham Park, NJ , pp. 135-168, 2000.

[4] C. D. Manning, P. Raghavan and H. Schutze, "Introduction to Information retrieval," Cambridge University Press Cambridge, England, pp. 117, April. 2009.

[5] B Gokulakrishnan, P. Priyanthan, T. Ragavan, N Prasath and A. S. Perera, “Opinion mining and sentiment analysis on a twitter data stream”, Department of Computer Science and Engineering, University of Moratuwa, Moratuwa, Sri Lanka, pp. 182-188, December. 2012.

[6] L. Ampofo, S. Collister, B. O’Loughlin and A. Chadwick, “Text mining and social media: when quantitative meets qualitative, and software meets humans”, New Political Communication Unit Working Paper, pp. 1- 67, October. 2013.

[7] Y. E. Soelistio and M. R. Sigit Surendra, “Simple text mining for sentiment analysis of political figure using naive bayes classifier method” Faculty of Information and Communication Technology, Multimedia Nusantara University, Indonesia, pp. 99-104, 2013.

[8] R. Basili, A. Moschitti and M. T. Pazienza, “Language sensitive text classification” Department of Computer Science, Systems and Production, University of Rome Tor Vergata Italy.

[9] Y. A. Wu and S. K. Hsieh, “Public opinion toward CSSTA: A text mining approach”, The Association for Computational Linguistics and Chinese Language Processing, pp.19-28. December. 2014, Vol. 19.

[10] E. J. Fortuny, T. D. Smedt, D. Martens and W. Daelemans, “Media coverage in times of political crisis: a text mining approach”, Expert systems with Applications, Tarrytown, NY, pp. 1-18, October. 2012, Vol. 39.

[11] S. Dumais, J. Platt, D. Heckerman and M. Sahami “Inductive learning algorithms and representations for text categorization” One Microsoft Way Redmond, WA.

[12] M. Anjaria, R. M. Guddeti “Influence factor based opinion mining of twitter data using supervised learning “, National Institute of Technology Karnataka, Mangalore, 2014.

DOI 10.17148/IJARCCE.2017.6313