A Project Report on

Citizen problem detection from Twitter data using Naive Bayes classifier

Submitted in partial fulfillment of the requirements for the award
of the degree of

Bachelor of Engineering

in

Computer Engineering

by

Amruta Koshe(18102019)
Nikita Sarode(18102063)
Nidhi Vanjare(18102066)
Rakshita Tantry(18102054)

Under the Guidance of

Prof. Ramya RB

Department of Computer Engineering
NBA Accredited
A.P. Shah Institute of Technology
G.B.Road,Kasarvadavli, Thane(W), Mumbai-400615
UNIVERSITY OF MUMBAI

Academic Year 2021-2022

Approval Sheet

This Project Report entitled ***"Citizen problem detection from Twitter data using Naive Bayes classifier"*** Submitted by ***"Amruta Koshe"(18102019),"Nikita Sarode"(18102063),"Nidhi Vanjare"(18102066),"Rakshita Tantry"(18102054)*** is approved for the partial fulfillment of the requirement for the award of the degree of ***Bachelor of Engineering*** in ***Computer Engineering*** from ***University of Mumbai***.

(Prof. Ramya RB)
Guide

Prof. Sachin Malave
Head Department of Computer Engineering

Place:A.P.Shah Institute of Technology, Thane
Date:

CERTIFICATE

This is to certify that the project entitled ***"Citizen problem detection from Twitter data using Naive Bayes classifier"*** submitted by ***"Amruta Koshe" (18102019), "Nikita Sarode" (18102063), "Nidhi Vanjare" (18102066), "Rakshita Tantry" (18102054)*** for the partial fulfillment of the requirement for award of a degree ***Bachelor of Engineering*** in ***Computer Engineering***,to the University of Mumbai,is a bonafide work carried out during academic year 2020-2021.

(Prof. Ramya RB)
Guide

Prof. Sachin Malave
Head Department of Computer Engineering

Dr. Uttam D.Kolekar
Principal

External Examiner(s)

1.

2.

Place:A.P.Shah Institute of Technology, Thane
Date:

Acknowledgement

We have great pleasure in presenting the report on **Citizen problem detection from Twitter data using Naive Bayes classifier.** We take this opportunity to express our sincere thanks towards our guide **Prof. Ramya RB** Department of Computer Engineering, APSIT thane for providing the technical guidelines and suggestions regarding line of work. We would like to express our gratitude towards his constant encouragement, support and guidance through the development of project.

We thank **Prof. Sachin Malave** Head of Department, Computer Engineering, APSIT for his encouragement during progress meeting and providing guidelines to write this report.

We thank **Prof. Amol Kalugade** BE project co-ordinator, Department of Computer Engineering, APSIT for being encouraging throughout the course and for guidance.

We also thank the entire staff of APSIT for their invaluable help rendered during the course of this work. We wish to express our deep gratitude towards all our colleagues of APSIT for their encouragement.

Student Name1: Amruta Koshe
Student ID1: 18102019

Student Name2: Nidhi Vanjare
Student ID2: 18102066

Student Name3: Nikita Sarode
Student ID3: 18102063

Student Name4: Rakshita Tantry
Student ID4: 18102054

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, We have adequately cited and referenced the original sources. We also declare that We have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

_____

(Signature)

_____

(Amruta Koshe - 18102019)
(Nidhi Vanjare - 18102066)
(Nikita Sarode - 18102063)
(Rakshita Tantry - 18102054)

Date:

Abstract

Urbanization has increased dramatically in recent years, resulting in significant changes in people's lives as they relocate to cities in large numbers. Many city issues, such as pollution, accidents, and traffic, have burgeoned as a result of the increased population. Many people daily face such problems and share about them on social media platforms like Twitter, Facebook, etc. However, it is difficult to fetch the right information at the right time. In this project, the problems faced by residents of a city or their requests are detected by analyzing tweets posted by users including the different localities in a city and made available to the authorities so that proper action can be taken. Tweets are extracted from twitter that contain keywords denoting areas, places or localities present in a city. (Thane) A Naive Bayes model is trained to detect whether a tweet signifies a problem faced by a citizen or not. The problems are analyzed to get information of any city event, citizens' complaints, or requests, and stored in the database along with the location. The tweets denoting citizens' problems and the locations are obtained from the database and displayed on the front-end interface to make it available to the authorities, to form a smart city management system.

Contents

List of Figures

Chapter 1

Introduction

With the advancement of web technology and its growth, there is a huge volume of data present in the web for internet users and a lot of data is generated daily. Social networking sites like Twitter, Facebook, Google+ are rapidly gaining popularity as they allow people to share and express their views about topics,have discussion with different communities, or post messages across the world. Many such posts contain problems or complaints faced by citizens in a city. This project focuses on citizen problem detection from twitter data with an objective to support smart city management. It is aimed to form a smart system, which detects problems of citizens and extracts the problems' exact locations from tweet texts. Tweets are extracted from Twitter using Tweepy API based on location keywords of Thane city. A Naive Bayes Classifier is trained to detect if tweets contain city problems or not. All the identified city problems are displayed on the web interface.

Chapter 2

Project Concept

2.1 Objectives

The objective of this project is to help keep track of problems in cities. Tweets will be extracted based on locations present in the text of these tweets. A website will be created to help keep a record of all the problems using tweets from twitter that are extracted via twitter API.

2.2 Literature Review

Gizem Abalı et al published a paper under IEEE in 2018 based on a Turkish city and uses Turkish language tweets for the study. The aim was to form a smart system, which detects problems of citizens and extracts the problems' exact locations from tweet texts. Naive Bayes classifier is trained on the tweets and tested on a separate set of tweets [8-10]. The high accuracy which is obtained by the classifier indicates that it is desirable to use this classifier for our study. This project will be based on these research paper findings and is mainly focused on Thane city.

When the past studies using Twitter data were observed, it was seen that many of them are about analyzing traffic tweets [1-3], some of them were done to understand the happiness in a city [4], finding the locations of where a user tweets[6], analyzing disaster tweets [5, 6], and detecting place names that are passed on a tweet text [7].

2.3 Problem Definition

In the past few years, technology has developed a lot and brought about a big change in the lives of people. Many people started to migrate to the cities. The increment in the number of citizens brought along many city problems like pollution, accidents, traffics, etc. Plenty of information is available over the internet, but it is very difficult to get the right information. Thus, an application will be created that will fetch data from one of the largest social media platforms i.e. Twitter. User tweets will be extracted to get information about citizens' problems and their locations. This application will display these problems with their locations which will be obtained by analyzing the tweets. All this information will be made available to everyone.

2.4 Scope

There are many sources present for obtaining information about events, accidents and various problems in citizens' day to day lives. People post about these problems on different social media platforms. Thus, an application will be created that will detect such citizens' problems with twitter. Tweets will be fetched with the locations for the dataset. Subsequently, this dataset will be used to train and test the machine learning model. As for frontend, a website will display all the tweets that mentions the citizens' problems along with the location. This website will enlist almost all the problems faced by citizens in the city. Several filters will be added to this website for the authorities to get the information with ease.

2.5 Technology Stack

Naive Bayes Classifier
- Train the system using Naive Bayes Classifier implemented with Sklearn to classify tweets containing problems in a city.

Google Colab
- Colab is used to write and execute python code through the browser.

Database - MongoDB
- MongoDB is used to store, retrieve and access the data.

Twitter API
- In order to collect data, Twitter API is used. The Twitter API provides nearly 1% of tweet data that streams from the selected area.

Programming language - Python
- Used for building machine learning model
- Used for building a code to recognize locations from the tweet data
- Tweepy - Python package used to access the Twitter API.
- Pandas - Python library for data manipulation and analysis.
- Sklearn - Python library used for machine learning and statistical modeling (here for naive Bayes classifier model)
- Flask - An API of Python used to build web-applications(front - end)

2.6 Benefits for Environment and Society

In the past few years, the number of people who live in cities and rural areas has changed and a big increment was observed in the cities' population. This brought along many city problems like pollution, accidents, traffic, etc.

This project can be used to detect these problems along with their locations which can be helpful in order to solve the problems quickly.

A smart city can be explained as the area where intelligent functions are used to collect the data and synthesize it to improve the efficiency of services, equity, sustainability and quality of life.
Some of the other features of smart city:

- A city that uses smart computing technologies to create all its infrastructures and services (including health-care, education, transportation).
- A city that watches all its critical transportation, communication, energy and water infrastructures and also major buildings.
- A city which is more effective, liveable, fair and sustainable.
This project can be used to develop a city into a smart city.

Chapter 3

Project Design

3.1 Proposed System

We propose to create a web application that will display citizen problems occurring in Thane region along with their location based on Twitter data along with other filtering features.
Tweets will be extracted from Twitter using location keywords.
A Naive Bayes model will be trained using SK-learn which will be used to identify tweets containing citizen problems.
These tweets will then be stores in the back-end for front-end requirements.
This project will allow people to view the problems in a locality at one stop.

3.2 Design(Flow Of Modules)

- Extracting tweets from Twitter using Twitter API i.e. Tweepy.
- Creating a dataset manually that is classifying tweets as related (1) and Non-related (0).
- Cleaning and pre-processing the dataset.
- Training and Testing the model that is classifying tweets as related or unrelated using Naive Bayes.
- Storing the related tweets along with their location.
- Displaying citizen problems with respect to location on the front end interface.

3.3 Class Diagram

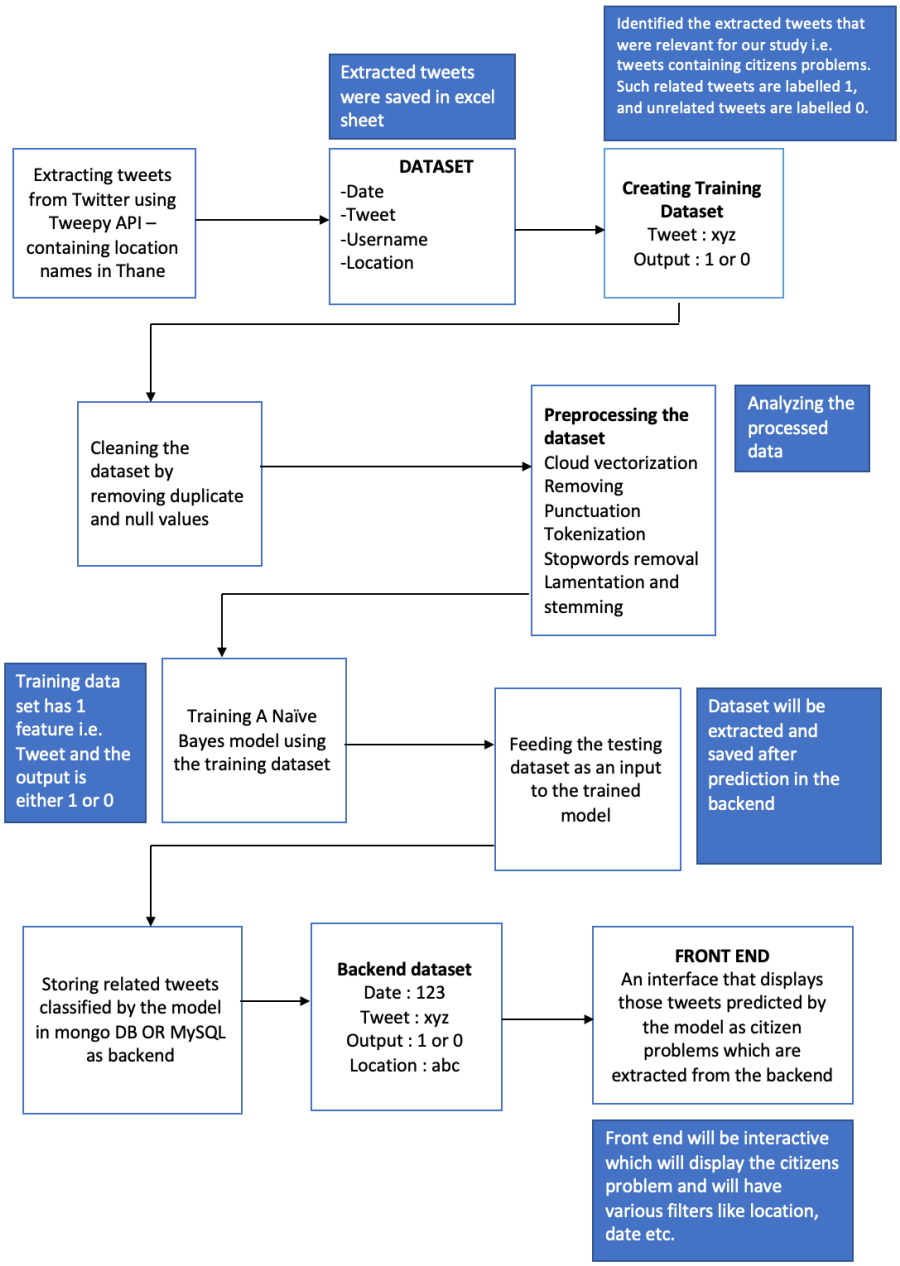

Figure 3.1: Class Diagram

3.4 Modules

3.4.1 Tweet Extraction:

Tweepy is an open source Python package that gives you a very convenient way to access the Twitter API with Python. After filling in a detailed questioner Twitter allows us to used the API and provides with a secret and API key. Then tweets were extracted that contained selected location names from Thane region.

3.4.2 Preparing training dataset:

For training dataset preparation, the extracted tweets were manually bifurcated into related and unrelated tweets. Related tweets are defined as those tweets that are relevant for our study and contain citizen problems in Thane region along with keywords like traffic, potholes, issue etc.

3.4.3 Cleaning the collected dataset:

Extracted tweets contained irrelevant symbols like emojis, extra characters like *,&,@,# and links. Such irrelevant data was removed. Duplicate tweets were discarded as they were repeated in the dataset.

3.4.4 Pre-processing:

- Removing stopwords - Stopwords are those words that do not add much meaning to a sentence.
- Remove punctuations
- Tokenization - Converting a sentence into list of words
- Lammetization / Stemming - Transforming any form of a word to its root word

3.4.5 Model Training and Testing :

Naive Bayes models are a special kind of classification machine learning algorithms. They are based on a statistical classification technique called 'Bayes Theorem'. Naive Bayes model are called 'naive' algorithms because they make an assumption that the predictor variables are independent from each other. In other words, that the presence of a certain feature in a dataset is completely unrelated to the presence of any other feature. We will be using the multinomial Naive Bayes implementation. This particular classifier is suitable for classification with discrete features (such as in our case, word counts for text classification). It takes in integer word counts as its input. We will use the sklearn's CountVectorizer method which counts the occurrence of each token.

3.5 References

[1] N. Wanichayapong, W. Pruthipunyaskul, W. Pattara-atikom, and P. Chaovalit, "Social-based traffic information extraction and classification," in Proc. 2011 International Conference on ITS Telecommunication (ITST), pp. 107-112.

[2] M. Hasby and M. L. Kodra, "Optimal path finding based on traffic information extraction from Twitter social-based traffic information," in Proc. 2013 International Conference on ICT for Smart Society (ICISS), pp. 1-5.

[3]I. L. B. Liu, C. M. K. Cheung, and M. K. O. Lee, "Understanding Twitter usage: What drives people to continue to tweet," in Proc. 2010 Pacific Asia Conference on Information Systems (PACIS), pp. 928-939.

[4] S. B. Marupudi, "Framework for semantic integration and scalable processing of city traffic events," M.Sc. Thesis, Wright State University, 2016.

[5] W. Guo, N. Gupta, G. Pogrebna, and S. Jarvis, "Understanding happiness in cities using Twitter: Jobs, children and transport," in Proc. 2016 IEEE International Smart Cities Conference, pp. 1-7.

[6] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes Twitter users: Real-time event detection by social sensors." in Proc. 2010 International Conference on World Wide Web, pp. 851-860.

[7] A. Acar and Y. Muraki, "Twitter for crisis communication: Lessons learned from Japan's tsunami disaster," International Journal of Web Based Communities, vol. 7.3, pp. 392-402, 2011.

[8] M. Pennacchiotti and A. M. Popescu, "A machine learning approach to Twitter user classification," in Proc. 2011 International AAAI Conference on Weblogs and Social Media, pp. 281-288.

[9] A. Z. H. Khan, M. Atique, and V. M. Thakare, "Combining lexiconbased and learning-based methods for Twitter sentiment analysis," International Journal of Electronics, Communication and Soft Computing Science & Engineering, vol. 4(4), pp. 89-91, 2015.

[10] G. Sidorov, and et al., "Empirical study of machine learning based approach for opinion mining in tweets," in Proc. 2012 Mexican International Conference on Artificial Intelligence, Springer Berlin Heidelberg, pp. 1-14.

Chapter 4

Planning for next semester

- Streaming tweet data to the model to make it real time.
- Storing the related tweet data which is detected by the naive bayes model in mongoDB/MySQL along with the location.
- Building the front end Web application to display the citizen problems with respect to the location.
- Rank the seriousness of the problem and categorise it.