Twitter review mining and analysis

Vismay Golwala
NC State University
vjgolwal@ncsu.edu

Dharmang Bhavsar
NC State University
dmbhavsa@ncsu.edu

Srija Ganguly
NC State University
sgangul2@ncsu.edu

Tushita Roychaudhury
NC State University
troycha@ncsu.edu

ABSTRACT

Reviews, might it be from peers or from the internet, is the first thing that a person looks for while making a choice between different places. Reviews are also helpful when you consider the organizational facade of businesses, because these reviews gives the management team a chance to look back and evaluate themselves and improve their services in the areas according to the feedback from the reviews. There are too many websites where a person can look for reviews of places. But, it is equally important nowadays to get a feel of how that place is perceived on social media. This report is about our effort to do exactly the same by finding reviews about a particular location from Twitter and as an extended effort, we also recommend places surrounding the searched place having the most positive social media reviews.

Keywords

Reviews, Tweets, Rating, Sentiment analysis, Aggregated reviews, Twitter, Yelp, Google

1. INTRODUCTION

People, wherever they go, look for reviews online and sometimes there are a lot of polarizing reviews about the same place and people just aren't sure what to do. During our research, we came across a strange positivity in reviews of places on portals such as Yelp and Zomato. To put things in perspective, about 68% of the reviews were 4 or 5 starred on Yelp [5]. This data is from the day it was established till September 2017. This seemed eerily positive for a website which has most of the restaurants in the world listed; most of them cannot be that good right? Another thing that we came across during out research was that users were particularly more inclined to post negative reviews on social media than anywhere else [6]. That is the reason that we are trying to get reviews that are disguised as Tweets.

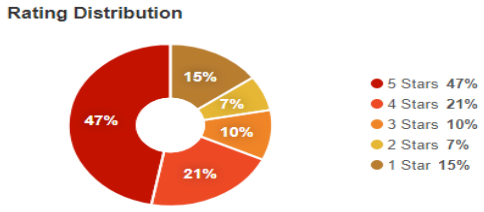

Figure 1: Yelp rating distribution

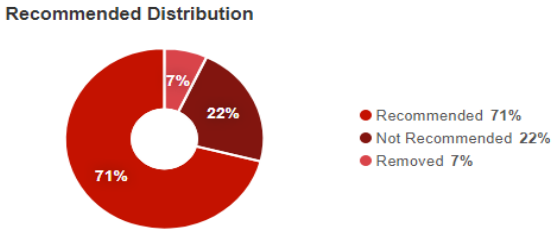

Figure 2: Yelp recommendation distribution

Restaurant reviews form an important part in the lives of people especially because they don't want to compromise on the quality of food and the reputation of the restaurant [4]. It is also an important factor in their decision making process when they suggest places to their close ones. The reviews are a direct expression of the thoughts that people have about the restaurant and categorize the rating into three parts - positive , negative and neutral - and it uses heavy involvement of users to filter data regarding a restaurant and performs prediction analysis by cross-checking the current data with other users' data [4]. The overall rating is based on an equation and takes into account all the categories of rating.

Following are the responses to certain questions asked to general users:

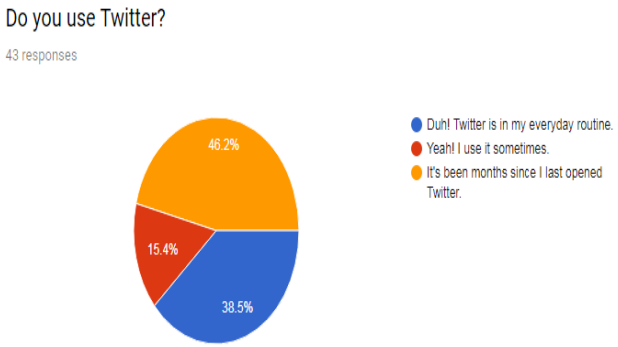

Figure 3: Twitter usage feedback

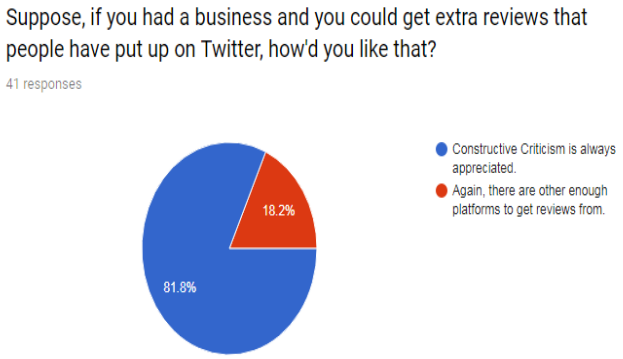

Figure 4: Need evaluation from feedback

Twitter being one of the most used social platform and the fact that Constructive feedback (including criticism) is always helpful to the customers; this idea can be effectively turned into a mechanism to aid in their decision process.

2. BACKGROUND LITERATURE

Substantial efforts have been spent on location identification and geo-tagging of documents and web pages. Social Networking on the other hand is still a very new field of computer science and little or no previous work has been done towards identifying the location of users based on their social activity.

The proposed method assumes that a dialog between users on Twitter has a constant topic and uses this model of social interaction in the Twitter network, along with content-derived location information, to employ a probabilistic framework to estimate the city-level location of a Twitter user [1]. This is done purely based on the content of the tweets, which may include reply-tweet information, without the use of any external information, such as a gazetteer, IP information etc.

Another work we read predicts a Twitter user's location based on his/her social network as opposed to the geo-tagging of tweets [2]. In the approach described in this paper, implicit attributes associated with the user in his/her social network are mined and user's location is predicted based on them. This paper does not estimate location based purely on content but uses additional knowledge to evaluate the user's location.

The work in geographic lexical variation [3] studies the variation of language usage on Twitter. Their multi-level generative model reasons jointly about latent topics and geographical regions allowing them to render each topic differently in each geographic region. This can be used to improve the accuracy of our location mining process.

These works can be used to augment our work in estimating if the user is really talking about the geographic location in consideration, thereby extracting the tweet content.

3. IMPLEMENTATION

3.1 Use Cases

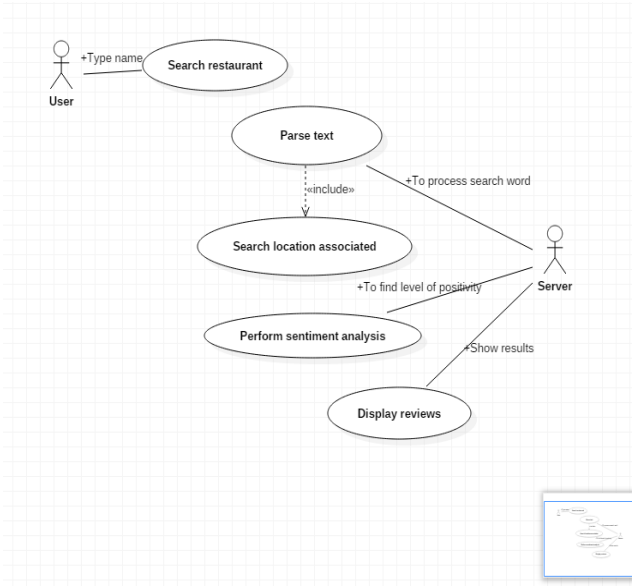

Figure 5: Use case diagram

3.1.1 Use case 1

Use Case Name	Search restaurant
Actors	User (of the software)
Description	The user will enter name of a restaurant and hit search
Trigger	Use case is triggered when the user initiates the software to lookup a restaurant and selects the search feature
Preconditions	1. User has a PC 2. User has active internet connection
Post conditions	1. User interface shows list of reviews or error
Normal Flow	1. User starts software 2. User selects search feature and enters a name 3. User hits search button 4. Server takes up user's request
Exceptions	1. Software doesn't start appropriately - Restart software and try it again 2. Server remains unresponsive - Restart server - Try it after sometime 3. Interface is unresponsive - Refresh page

	- Try it after sometime
Includes	-
Special Requirements	The software must have a robust interface that can handle user input effectively and is easy for the user to understand it's functionality.
Assumptions	1. User understands English 2. User wants to get feedback on a restaurant within USA
Notes and Issues	Interface is basic and provides a trigger to the main backend work of the software

3.1.2 Use case 2

Use Case Name	Parse text
Actors	Application server
Description	The server will parse the text requested by the user to find reviews related to the restaurant
Trigger	When the user sends a request by hitting search button , the use case is triggered
Preconditions	- There should be a valid search request
Post conditions	- The server should be able to recognize the word effectively to search for location.
Normal Flow	1. Search word is recognized by the server 2. Using certain method, parsing happens
Exceptions	1. Software doesn't start appropriately - Restart software and try it again 2. Server remains unresponsive - Restart server - Try it after sometime 3. Interface is unresponsive - Refresh page - Try it after sometime
Includes	Search location associated
Special Requirements	The software must have a robust interface that can handle user input effectively and is easy for the user to understand its functionality.
Assumptions	1. Server is modeled to understand English 2. User wants to get feedback on a restaurant within USA
Notes and Issues	The first step towards contacting the server

3.1.3 Use case 3

Use Case Name	Search location associated
Actors	Application Server
Description	The server will parse the data and search the twitter feeds for any data related to the text
Trigger	Triggered right after server receives text to parse
Preconditions	1. Text is valid 2. It is possible to search for the text
Post conditions	1. State of software that has reviews available
Normal Flow	1. Server takes parsed text 2. The twitter feeds are searched via the text 3. Finds data
Exceptions	1. Software doesn't start appropriately - Restart software and try it again 2. Server remains unresponsive - Restart server - Try it after sometime 3. Interface is unresponsive - Refresh page - Try it after sometime
Includes	-
Special Requirements	The software must have a robust interface that can handle user input effectively and is easy for the user to understand it's functionality.
Assumptions	1. User understands English 2. User wants to get feedback on a restaurant within USA
Notes and Issues	Multiple locations can be returned based on the area being searched

3.1.4 Use case 4

Use Case Name	Perform sentiment analysis
Actors	Application server
Description	This module is dedicated to estimate the level of positivity in the feedbacks concerning the particular restaurant
Trigger	This can be initiated upon searching for a restaurant by the user
Preconditions	1. A block of text related to a restaurant on twitter 2. Software is still active
Post conditions	1. Interface shows if the feedback sounds positive or not
Normal Flow	1. Server finds data in Twitter 2. Using sentiment analysis, language is processed to judge the level of positivity 3. Results generated
Exceptions	1. Software doesn't start appropriately - Restart software and try it again 2. Server remains unresponsive - Restart server - Try it after sometime 3. Interface is unresponsive - Refresh page - Try it after sometime 4. Sentiments vary from person to person.
Includes	-
Special Requirements	The software must have a robust interface that can handle user input effectively and is easy for the user to understand it's functionality.
Assumptions	1. A valid block of tweet is present that can be analyzed 2. Display monitor is active 3. English as a language can be analyzed by the system
Notes and Issues	The most sensitive part of the system that needs a lot of attention

3.1.5 Use case 5

Use Case Name	Display reviews
Actors	Application server
Description	After finding data related to the restaurant, server displays the reviews on the interface
Trigger	Server searches for data related to the location and state of the software changes to display
Preconditions	- Match for text found - Software is still active
Post conditions	- Interface shows the reviews or error
Normal Flow	- Server finds data in Twitter - Displays results
Exceptions	Software doesn't start appropriately - Restart software and try it again Server remains unresponsive - Restart server - Try it after sometime Interface is unresponsive - Refresh page - Try it after sometime
Includes	-
Special Requirements	The software must have a robust interface that can handle user input effectively and is easy for the user to understand it's functionality.
Assumptions	- User understands English - Display monitor is active
Notes and Issues	Interface is basic that tells us the effective usage of the system

3.2 Technologies used

- Twitter API - Twitter API will be used to get the tweets of the users which will be analyzed to find a certain location, be it restaurant or a recreational area. We will get tweets making a specific location a center point and draw a radius of 10 miles around it. To get the location of a place, we will use the Google Places API.
- Google Places API - Google Places API will be used to find the location of a place. The location will be obtained as a latitude and longitude of a place. Then Google Places API will also give us surrounding places within a radius of 10

miles. We will be including Tweets regarding this place as well.

- Spacy - Spacy is a NLP library based on Python and supports features such as parts of speech tagging, sentiment analysis among others. We will be using Spacy to get the tweets that mention a certain place.
- Web Stack - A web application will be developed where a user can search for a particular location by writing the name of the location in the text box. The UI will then display all the tweets of that location and the percent amount of tweets that are positive. Similarly, other locations and their positive percent of tweets will be displayed.

3.3 Plan of Action

- During the first week, we will try to integrate the Google Places API and Twitter API in the application. The application will be such that a person searches for a location and from Google API we will show him the results.
- After the user selects a particular location, we will then collect locations in 10 miles radius of the selected locations.
- After getting the latitude and longitude of the location that the user selected, we will get the Tweets for a radius of 10 miles from that location.
- From the tweets, we will find the tweets that were about the location selected by the user. On those tweets we will perform sentiment analysis categorizing tweets into positive or negative according to the content.
- Then we will get the percentage score of tweets that were positive from the total tweets found about that place. That score will be displayed too. All the tweets for the location selected by the user will be displayed.
- Same process will be done for the other locations which surrounds the user selected location except all the tweets about those locations will not be displayed. Only the positive tweets score will be displayed.

4. EVALUATION PLAN

The aim of the project is to obtain tweets for a given location, analyze their sentiment and aggregate all the responses in one place. For this problem, we plan to use an algorithm which can accurately gather tweets for a given location and identify the polarity of these tweets. We plan to test the system using previously classified text data sets and after that crowd-sourcing user responses to test correctness. Evaluation also includes asking humans whether a tweet retrieved is indeed related to the location that the user had searched about.

- Analyzing - We will use already classified Twitter data sets from UCI ML Repository or Kaggle and the algorithm will analyze the sentiment of each of these tweets and classify whether the review is positive or negative.
- Validating - We will validate these tweets using predefined class labels and test the accuracy of our system. The average accuracy obtained from the datasets will be our threshold for next level validation.
- Crowdsourcing - After setting a base threshold, we aim to validate the system again using human responses. We will ask users questions regarding whether a tweet is correctly classified. If these responses exceed the minimum accuracy threshold set in the previous part, the system succeeds.

5. TIMELINE

- Week 1 - Integrating Google Places API and Twitter API with the application
- Week 2 - Making the UI of the application and coding up the NLP part of the project (finding the tweets).
- Week 3 - Full working model ready and getting prepped up for the evaluations part.
- Week 4 - Sending in surveys and getting scores from the Kaggle datasets. Comparing them and including them in the report.

6. CHALLENGES

We expect challenges based on n the semantic complexities of the natural language used in tweets.

- Some users may not use standard vocabulary in their tweets.
- They may refer to the same location using different names or different locations using the same name. For example, a user can refer to New York as NY or N.Y. or City of New York, etc. Or the user can say Washington for either the state or the city.
- The tweets may not always contain location-specific reviews and thus such tweets will be considered noisy.
- Users may use different languages to tweet since twitter allows the use of several languages.

7. ACKNOWLEDGMENTS

Our special thanks to our Teaching Assistant, Amritanshu Agrawal and Prof. Timothy Menzies for helping us in coming up with the idea and a concrete plan to successfully implement this project.

8. REFERENCES

[1] Chandra, Swarup, et al. “Estimating Twitter User Location Using Social Interactions--A Content Based Approach.” 2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing, 2011, doi:10.1109/passat/socialcom.2011.120. ↩

[2] Abrol, Satyen, and Latifur Khan. “Tweethood: Agglomerative Clustering on Fuzzy k-Closest Friends with Variable Depth for Location Mining.” 2010 IEEE Second International Conference on Social Computing, 2010, doi:10.1109/socialcom.2010.30. ↩

[3] Jacob Eisenstein , Brendan O'Connor , Noah A. Smith , Eric P. Xing, A latent variable model for geographic lexical variation, Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, p.1277-1287, October 09-11, 2010, Cambridge, Massachusetts. ↩

[4] Sanjukta Saha and Dr. AK Santra, “Restaurant Rating Based on Textual Feedback”, IEEE 2017, Available: ↩

[5] Yelp Factsheet, September 2017. ↩

[6] Consumers increasingly turning to social media to share negative reviews. ↩