Smart Twitter Analysis on Location Using Kali Linux

Nishant Sikri

Abstract Twitter is a rich source for mining views and opinions that are often connected with habits or show a pattern. Social media platforms add geo-location as a parameter in tweets or posts, making it accessible for marketing tactics and focus on a user base. A similar scenario is when a local business owner aims to open a new store or a cafe in a neighborhood. To address this problem, location-based tweets help to keep track of post timings, interests, number of active users, and opinions. The extraction of analyzed data can help achieve targeted advertisements, open restaurant chains, skin clinics, and set up shops. This results in overcoming the trial and error method that utilizes a lot of the investment and often goes unnoticed by the audience.

Keywords Trend analysis · Twitter · Opinion mining · Pushpin · Social media

1 Introduction

Social media like Twitter and Instagram are valuable sources that not only contain images and text, but they are a rich source of the geographical location of users, each getting updated constantly. This data can be extracted from social networking sites and can be used to identify trends among areas of interest. Using a radius around a location, the number of users can be identified with their corresponding location and time on the map. This form of data can further be used for identifying the type of user base in an area or their likes and dislikes, places they visit, interests as well. This extracted data can be used to target advertisements, open stores of similar interests, theme cafes, and much more.

Unlike other social platforms, Twitter is opinion based that establishes a level of openness and honesty. Making it an excellent platform to obtain personal views and harness intelligence. With advancements in search tracking to generate search

N. Sikri (✉)
Master of Computer Applications Department,
Sardar Patel Institute of Technology, Mumbai, India

trends with the help of Google Trends, one can only keep track of keywords up to a particular state or a country as a whole.

Looking up for trends not only ensures the interest of people but reveals how engaging people are. There are reports for the top five coffee enthusiast states in India, but it is hard to tell if people in your neighborhood love to drink coffee.

This paper aims to identify tweets according to pinpoint location and within a radius of a few kilometers. Tools like Recon-ng offer tracking of Instagram, YouTube, and Twitter posts with their geolocation coordinates and time. Once satisfied with the location, one can keep track of post timings, interests, number of active users, and their views.

2 Related Work

Social media analysis over Facebook posts, Twitter tweets, Instagram likes talks about the interests of people. Analyzing certain aspects of each platform gives off different results such as customer retention, opinions, or type of demographic. Likewise, [3] focuses on the popularity of KFC and McDonald's with over 7000 tweets, each using a lexicon-based model to classify tweets positive, negative, and neutral [5].

According to [2], Instagram users can be distributed among segments of their preferences that help in building content marketing strategy according to organizations needs. Focusing on a certain group of people ensures customer engagement while maximizing reach. Taking a step forward, [10] emphasizes using Latent Dirichlet Allocation (LDA) for location-specific trends on Twitter with the help of an android application and accuracy of 84% in best case scenario.

Twitter opinions and views come forward as a look inside the mind [8] of the individual user as well as exhibits difference in behavior with the help of Natural Language Processing (NLP) presenting variation in general sentiment across the industry sector [12]. However, leveraging the location parameter [7] from tweets to predict general election results [6] for a country and monitoring events like natural disasters [9] with heat maps of affected zones [11]. Although most of the related trends are easily accessible through Google Trends, providing a great opportunity to check what range of people are interested in, by tracking search habits across the globe. While most algorithms focus only on uni-gram or bi-gram models, thus considering a single word or two words from the entire sentence. [1] implements the n-gram model with features to support emoticons, synonyms, and acronyms making this hybrid approach a reliable source to predict trends through tweets spread geographically.

The ever-growing platform Instagram where posts are made up of pictures and captions that are partially meaningful and bombarded with misleading hashtags, [4] takes an approach to tag images with a modified version of hyperlink-induced topic search (HITS) algorithm and combining it with crowd-tagging to produce a content-based image retrieval system leading to a maximum achievable recall value of 0.931.

3   Proposed Methodology

This paper focuses on pinpointing a geographical location, obtaining tweets for that area, analysis to generate interesting tweet topics, and identifying a neighborhood for the business of interest. Thus, improving reach and brand recognition among individuals. Figure 1 shows the flow of the process.

3.1   Kali Linux

Kali Linux is a linux based operating system that offers a wide variety of tools for reconnaissance, penetration testing, and other offensive security measures. It is a bundle of prebuilt packages that are easily accessible to its users that provides out-of-the-box experience.

3.2   Location Retrieval

Our focus lies in a demanding area that could help generate optimal revenue. Using Google Maps, the latitude, and longitude of a geographical location can be identified and added into Recon-ng.

3.3   Recon-Ng

Reconnaissance is gathering information regarding a particular topic. Recon-ng is a web reconnaissance tool that can track location-based tweets, Instagram posts, YouTube uploads, interacting with databases, and managing API keys.

1. Load Twitter Module: Taking a modular approach to focus on the problem at hand makes it easier to break down the problem. A location-based pushpin module is just what is required to track tweets of a neighborhood.
2. Set Radius: The geolocation coordinates added earlier defines our target, but don't want to focus on just one point. Thus, a radius is set around the location to widen our search for tweets.
3. Capture Tweets: Once done with setting up parameters, simply run the module to collect all the tweets around the location.
4. Load Pushpin Report Module: Creates an organized report for captured tweets for visual representation of tweets with their location.
5. Set Location and Radius: Geolocation coordinates with radius play a significant role in the visual look and feel over a map.

Fig. 1 System flow diagram
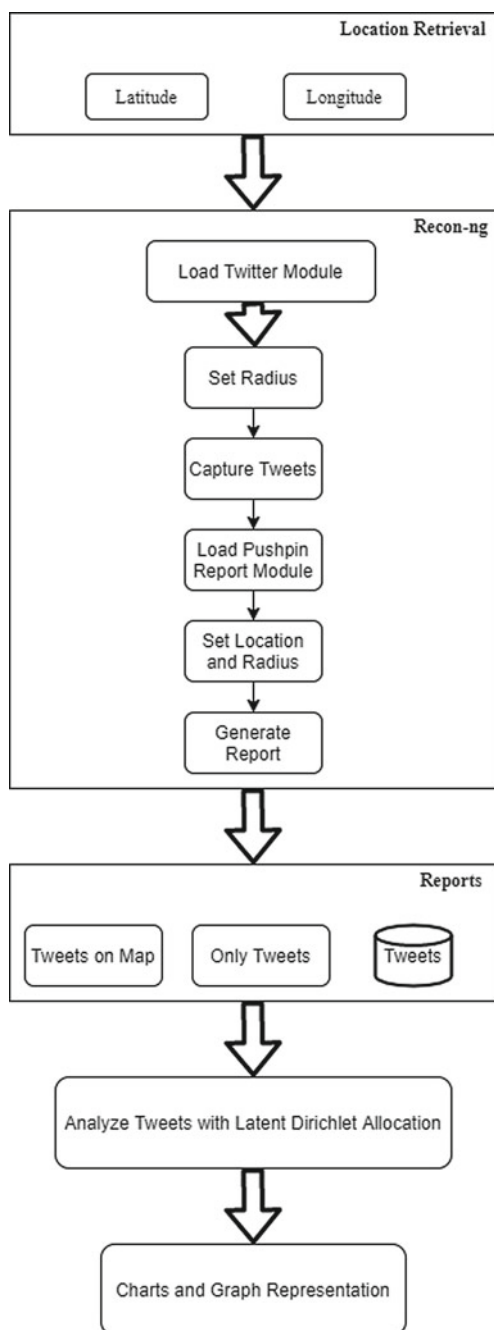
6. Generate Report: The reporting module generates a media file with a description of all the tweets, a map file that contains the pinpoint location of each tweet, and a database file with details of all tweets.

3.4 Dataset

While working with real-time tweets, it is quite significant to understand that the number of tweets in an area fluctuates. Seven days worth of tweet data retrieved from each location that may consist of roughly 250 tweets on average for happening places. Since the dataset is real-time, this number varies a lot and that is why a higher number like 500 tweets is also noticeable in some areas of Los Angeles, California.

3.5 Data Loading

The reports generated in the previous step provide a dataset of real-time tweets of a particular location for a whole week. An SQLite file generated from the pushpin reporting module from which it is required to select the pushpin table.

3.6 Pre-processing

Our primary focus is to highlight popular topics among a list of tweets, so it is better to drop other columns such as source, screen_name, profile_name, profile_url, media_url, thumb_url, latitude, longitude, time, and module while retaining only tweets content column that is named as the message. Tweets content also consists of punctuation and links to other pages that can dramatically alter our results with the most common word to be "https" due to its presence in all the tweets. It is better to remove them from the message column for reliable results.

3.7 Process Tweets

Aim is to find trending topics that rise in popularity. Say, a food chain plans to open a new branch in a neighborhood, the intention is to focus on areas that offer increased demand. Concentrating on tweets that mention keywords like food, recipe, cooking, ice-cream, healthy, sweet, and smoothie shows that the area is favorable for the food chain. Manual analysis of tweets is a time-consuming task and quite tedious when there are hundreds of tweets within a 3 km radius. Since our focus is on

word frequency from a set of documents, Latent Dirichlet Allocation (LDA) provides reasonably accurate mixtures of topics within a given document. The implementation consists of:

1. Exploratory Analysis: To make sure our pre-processing works as expected, where it generates a word cloud to provide a visual representation of most common words.
2. Preparing data for LDA Analysis: Tweets are converted into a bag of words to count their occurrence in the database and plot the top 10 common words which should occur in the word cloud as well.
3. LDA model training: Using prebuilt libraries for Latent Dirichlet Allocation (LDA), it can tweak the parameters of the number of topics and words to achieve optimal results.
4. Analyzing LDA model results: To help us understand individual topics and relationships between the topics, it can use a visualizing package like pyLDAvis.

3.8 Charts and Graph Representation

Visualizing the relevant keywords that repeat over multiple instances. Not only tweet content can be analyzed to represent over graphic visuals, but it visualizes the time people usually post tweets or if there is any variation of timings on some days.

4 Result

Reports generated while collecting tweets help in analysis and visualization. Geographically pinpoint tweets from nearby locations on maps is the evidence to see active users in the area from Fig. 2. Stored records of each tweet in a database file for the pushpin location mentioned in Fig. 3.

Four locations were selected randomly to find out trending topics for the area with different search radius to make sure the search is optimized. To make sure that LDA is giving favorable results, real-time tweets are fetched again for Juhu within a duration of two months to find the change in trends. Initial results for Juhu, Mumbai dated April, 2020 are represented by Fig. 4 consisting of a word cloud that represents the most common words. Further, the top ten most common words are plotted in Fig. 5 and after LDA model training, the results consist of five topics in Fig. 6 with ten words associated with a particular topic. For better understanding, the results are visualized as shown in Fig. 7. Words like divine, guidance, mindseed, and dabbooratnani are among the top 30 most important terms where the λ parameter can be used to adjust relevance among different terms. The relationships between the topics are understood by Intertopic Distance Plot that reveals how different topics relate to each other.
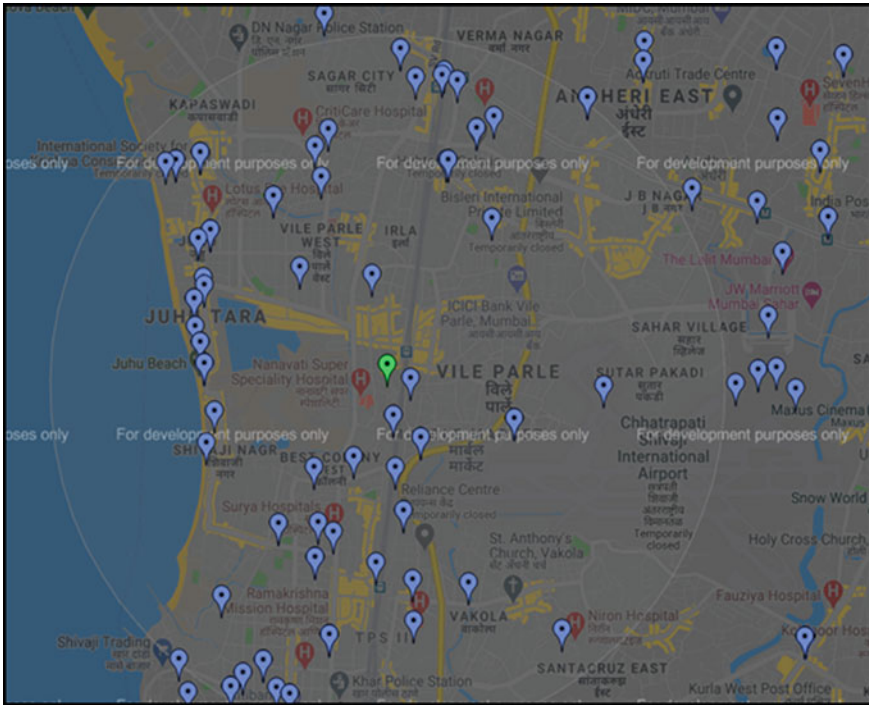
Fig. 2 Pushpin for Juhu, Mumbai

source	screen_name	profile_name	profile_url	media_url	thumb_url	message	latitude	longitude	time	module
Twitter	foramshah6550	XTRA_SACHET_SU...	https://twitter.com/for...	https://twitter.com/for...	https://pbs.twimg.co...	Dreams... @foramsh...	19.11483	72.83736	2020-04-23 14:06:32	twitter
Twitter	divyadesaililil	Divya Desai	https://twitter.com/div...	https://twitter.com/div...	https://pbs.twimg.co...	#पुलायस_अंधेरी_पूर्...	19.11727198	72.85923576	2020-04-23 13:59:47	twitter
Twitter	SDutta555	Sanjay Tarun Dutta	https://twitter.com/SD...	https://twitter.com/SD...	https://pbs.twimg.co...	Just posted a photo ...	19.0821	72.8363	2020-04-23 13:50:24	twitter
Twitter	SDutta555	Sanjay Tarun Dutta	https://twitter.com/SD...	https://twitter.com/SD...	https://pbs.twimg.co...	It doesn't matter whic...	19.0821	72.8363	2020-04-23 13:49:48	twitter
Twitter	ChArMiNg_ArU	arunima das	https://twitter.com/Ch...	https://twitter.com/Ch...	https://pbs.twimg.co...	Blast from the past ...	19.1193	72.84715	2020-04-23 13:25:37	twitter
Twitter	hive_home	Hive Home	https://twitter.com/hiv...	https://twitter.com/hiv...	https://pbs.twimg.co...	Everything progressi...	19.1135714	72.8362468	2020-04-23 12:26:35	twitter
Twitter	DiabolikAllen	Allen Claudius	https://twitter.com/Di...	https://twitter.com/Di...	https://pbs.twimg.co...	What's the one pair y...	19.1124904	72.8475441	2020-04-23 11:59:51	twitter
Twitter	sharhaansingh	sharhaan singh	https://twitter.com/sh...	https://twitter.com/sh...	https://pbs.twimg.co...	#wishuveryhappybirt...	19.09328744	72.82776586	2020-04-23 11:14:36	twitter
Twitter	Rochatter	Rowena Dsouza	https://twitter.com/Ro...	https://twitter.com/Ro...	https://pbs.twimg.co...	In an attempt to grow...	19.1124904	72.8475441	2020-04-23 10:27:22	twitter
Twitter	ksgunjan	Gunjan Sharma	https://twitter.com/ks...	https://twitter.com/ks...	https://pbs.twimg.co...	Door HUMAN FLAG ...	19.0969033	72.8268983	2020-04-23 09:44:10	twitter
Twitter	rafflesmumbai	Raffles Mumbai	https://twitter.com/raf...	https://twitter.com/raf...	https://pbs.twimg.co...	What is your passion...	19.10814945	72.8795886	2020-04-23 08:33:15	twitter
Twitter	tmj_inm_jobs	TMJ-INM Jobs	https://twitter.com/tmj...	https://twitter.com/tmj...	https://pbs.twimg.co...	Have you ever had a ...	19.0759837	72.8776559	2020-04-23 07:43:03	twitter
Twitter	DrRoshani	Simple talk doc	https://twitter.com/Dr...	https://twitter.com/Dr...	https://pbs.twimg.co...	#Repost aasaanhealt...	19.06991901	72.83866153	2020-04-23 06:08:10	twitter
Twitter	DrRoshani	Simple talk doc	https://twitter.com/Dr...	https://twitter.com/Dr...	https://pbs.twimg.co...	🐾😊Getting busy ...	19.06991901	72.83866153	2020-04-23 05:26:20	twitter
Twitter	tarot_cards	Roop Lakhani - Tarot	https://twitter.com/tar...	https://twitter.com/tar...	https://pbs.twimg.co...	📷 Today's Mindsee...	19.07214402	72.8291443	2020-04-23 05:09:17	twitter
Twitter	tarot_cards	Roop Lakhani - Tarot	https://twitter.com/tar...	https://twitter.com/tar...	https://pbs.twimg.co...	❤️Thought for the d...	19.07214402	72.8291443	2020-04-23 05:08:16	twitter
Twitter	tarot_cards	Roop Lakhani - Tarot	https://twitter.com/tar...	https://twitter.com/tar...	https://pbs.twimg.co...	🌷Divine Guidance ...	19.07214402	72.8291443	2020-04-23 05:06:25	twitter
Twitter	taher07	TAHER SM	https://twitter.com/tah...	https://twitter.com/tah...	https://pbs.twimg.co...	The doctors responsi...	19.0742207	72.8318634	2020-04-23 05:05:36	twitter
Twitter	JobsAtEricsson	Ericsson Jobs	https://twitter.com/Jo...	https://twitter.com/Jo...	https://pbs.twimg.co...	Join the Ericsson tea...	19.0759837	72.8776559	2020-04-23 05:01:11	twitter

Fig. 3 Reporting database for pushpin

Fig. 4 Word cloud for Juhu, Mumbai dated April, 2020

Fig. 5 10 most common words for Juhu, Mumbai dated April, 2020

Let's look at the data for Juhu, Mumbai area dated mid-June, 2020, quite a dramatic change in the word cloud is noticed in Fig. 8 with some new trends among the top ten most common words in Fig. 9. LDA identifies a new set of topics with different sets of words such as dabbooratnani, happy, sandesh, and a few other words in one topic while words such as carpool, offered, and rideshare come under topic #2 in Fig. 10. When visualized using the library pyLDAvis in Fig. 11, words like carpool, offered, rideshare, and seats stood among the top 30 most important terms helping with our analysis that there is a dramatic change in trends in the same area within a matter of months. Similarly, the tweets are analyzed for Powai and Churchgate in Mumbai and Los Angeles in California. Considering the area of Churchgate in Fig. 12, there

```
Topics found via LDA:

Topic #0:
vegetable juice love lunch नसभ journey mumbai like drink 30

Topic #1:
mumbai mh click job quarantine india juhu just apply hiring

Topic #2:
mumbai challenge beautiful time make throwback jumping know thank juhu

Topic #3:
dabbooratnani manishadratnani amp april know day 20th like live look

Topic #4:
divine guidance mindseed day follow today research decks cards card
```
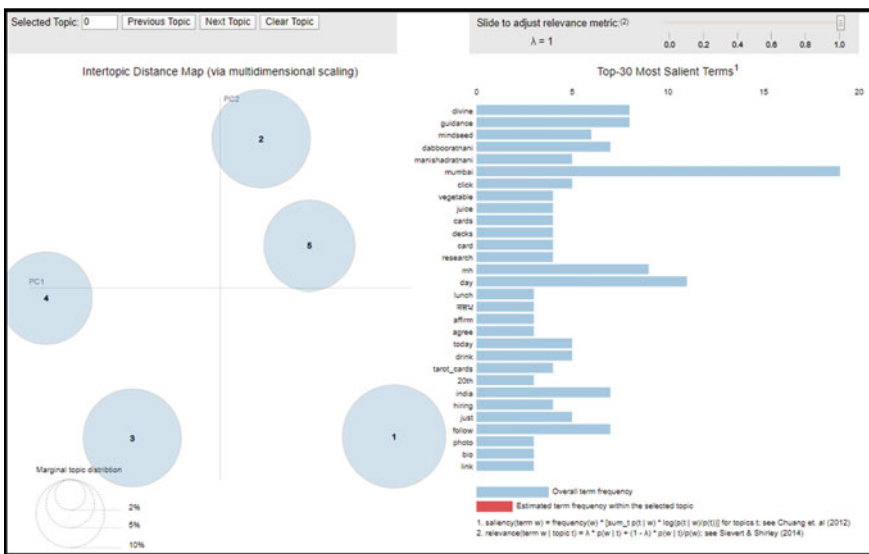
Fig. 6 Topics found via LDA for Juhu, Mumbai dated April, 2020

Fig. 7 LDA analysis for Juhu, Mumbai dated April, 2020

seems to be a variation in most salient terms such as hope, reading, acts, and gospel emphasizing different likes of people. In Fig. 13, i.e. Powai, there seems to be an enormous trend for real estate. The results are quite revealing to the extent that some areas are more inclined towards workout or crossfit as seen in Fig. 14 while some areas offer quite a lot of flats or properties.
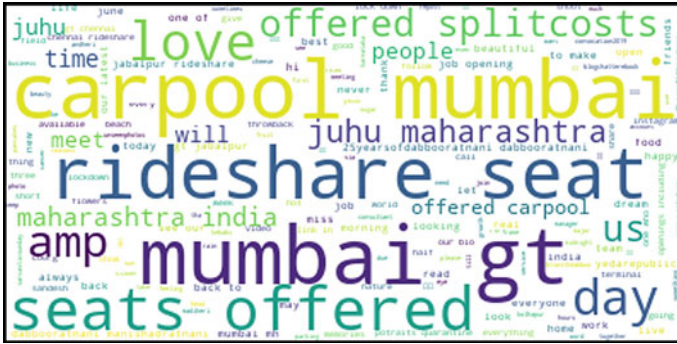
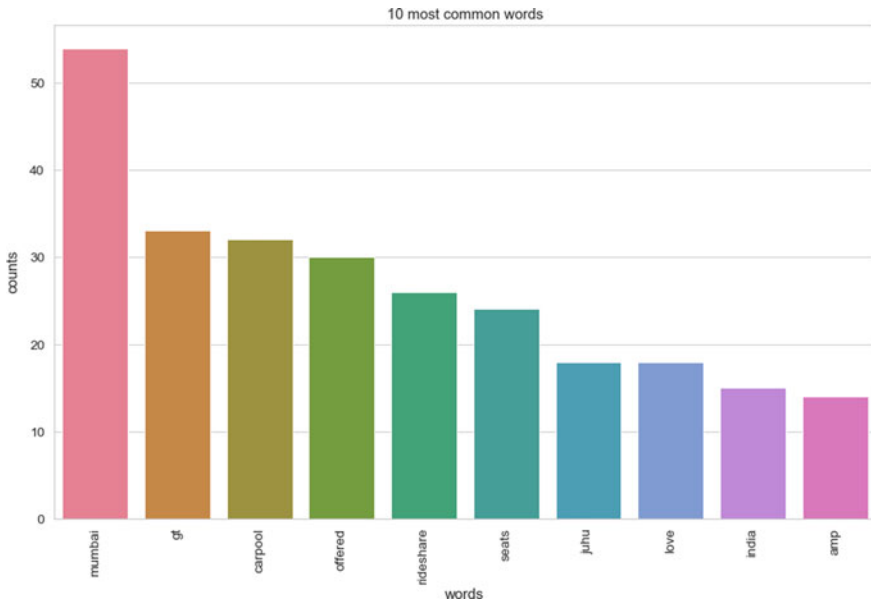**Fig. 8** Word cloud for Juhu, Mumbai dated June, 2020

Fig. 9 10 most common words for Juhu, Mumbai dated June, 2020

5 Conclusion and Future Work

This research revolves around finding tweets that are fine-grained to a few kilometers from a pinpoint location providing a great opportunity to organizations or other businesses to identify their hot-selling areas. Analyzing a social platform that is filled with views and opinions of people provides a glimpse of their habits. Tracking the tweets time of people and approximately how many tweets per day are from an area results in activity tracking for a location. The core focus is to identify keywords

```
Topics found via LDA:

Topic #0:
25yearsofdabbooratnani dabbooratnani love birthday manishadratnani happy sandesh time
days today

Topic #1:
juhu india maharashtra love like job home morning time openings

Topic #2:
mumbai gt carpool offered rideshare seats splitcosts people seat कम

Topic #3:
day amp miss make open meet happy mumbai juhu आज

Topic #4:
mumbai dabbooratnani love best amp mh job looking latest bio
```
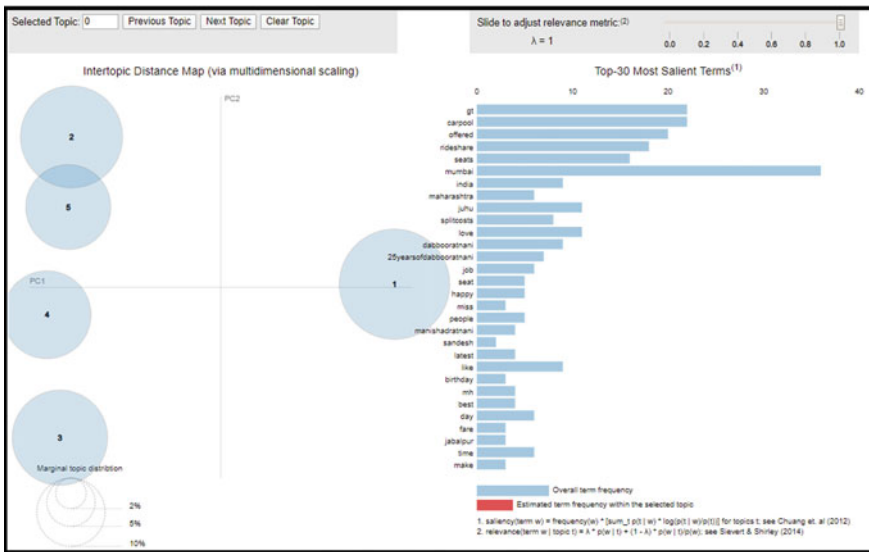
Fig. 10 Topics found via LDA for Juhu, Mumbai dated June, 2020

Fig. 11 LDA analysis for Juhu, Mumbai dated June, 2020

or topics that people are repetitively talking about, allows to visualize where the target audience lies on a map and strategize accordingly. Henceforth, a way around choosing a favorable area for finally deciding the exact area for business to earn maximum profit with increased reach.
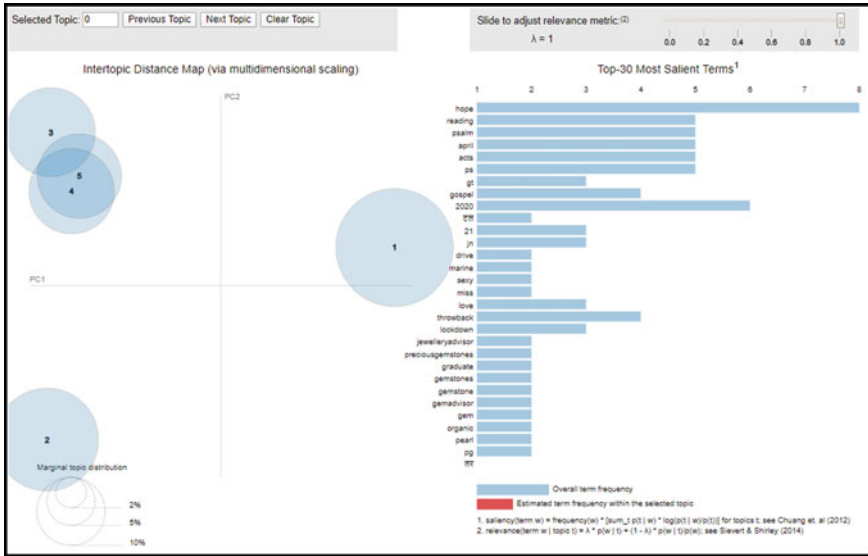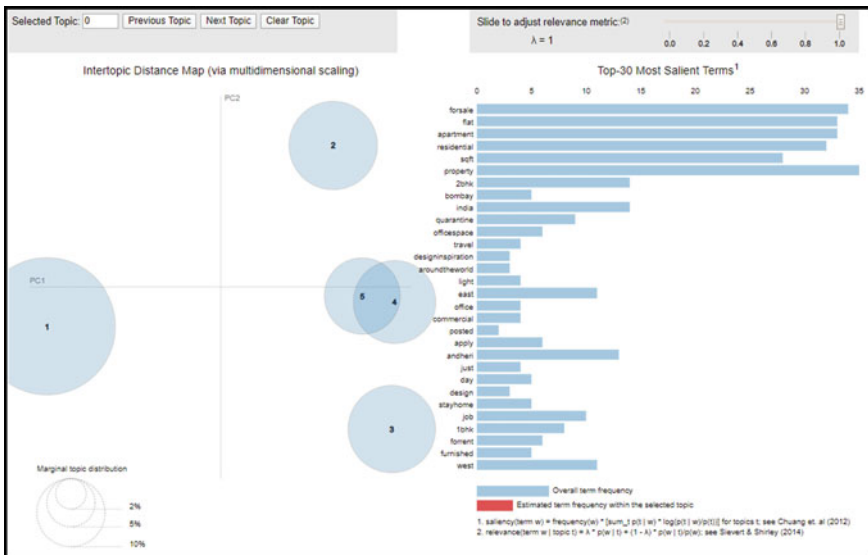
Fig. 12 LDA analysis for Churchgate, Mumbai
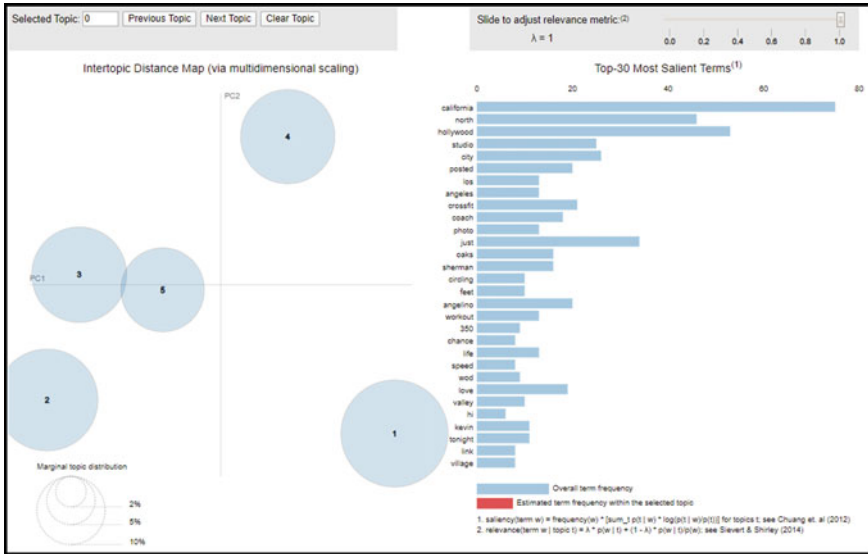

Fig. 13 LDA analysis for Powai, Mumbai

Fig. 14 LDA analysis for Los Angeles, California

References

1. Zala DK, Gandhi A (2019) A twitter based opinion mining to perform analysis geographically. In: 2019 3rd International conference on trends in electronics and informatics (ICOEI), Tirunelveli, India, pp 59–63. https://doi.org/10.1109/ICOEI.2019.8862548.
2. Arbi Siti RHM, Bismo A, Sutiyo L (2019) Segmentation analysis of Instagram Users Based on Preferences towards Forms and Types of Online Marketing Content. In: 2019 International conference on information management and technology (ICIMTech), Jakarta/Bali, Indonesia, pp 202–207. https://doi.org/10.1109/ICIMTech.2019.8843797.
3. El Rahman SA, AlOtaibi FA, AlShehri WA (2019) Sentiment analysis of twitter data. In: 2019 International conference on computer and information sciences (ICCIS), Sakaka, Saudi Arabia, pp 1–4. https://doi.org/10.1109/ICCISci.2019.8716464.
4. Giannoulakis S, Tsapatsoulis N (2019) Filtering instagram hashtags through crowdtagging and the HITS algorithm. IEEE Trans Comput Soc Syst 6(3):592–603. https://doi.org/10.1109/TCSS.2019.2914080
5. Guevara J, Costa J, Arroba J, Silva C (2018) Harvesting opinions in Twitter for sentiment analysis. In: 2018 13th Iberian conference on information systems and technologies (CISTI), Caceres, pp 1–7. https://doi.org/10.23919/CISTI.2018.8399226.
6. Joseph FJJ (2019) Twitter based outcome predictions of 2019 Indian General Elections using decision tree. In: 2019 4th International conference on information technology (InCIT), Bangkok, Thailand, pp 50–53. https://doi.org/10.1109/INCIT.2019.8911975.
7. Karan RS, Shirsat KK, Kasar PL, Chaudhary R (2018) Sentiment analysis on twitter data: a new approach. In: 2018 International conference on current trends towards converging technologies (ICCTCT), Coimbatore, pp 1–4. https://doi.org/10.1109/ICCTCT.2018.8551169.
8. Wagh R, Punde P (2018) Survey on sentiment analysis using twitter dataset. In: 2018 Second international conference on electronics, communication and aerospace technology (ICECA), Coimbatore, pp 208–211. https://doi.org/10.1109/ICECA.2018.8474783.

9. Zou Q (2019) A prototype system using location-based twitter data for disaster management. In: IGARSS 2019–2019 IEEE international geoscience and remote sensing symposium, Yokohama, Japan, pp 9514–9517. https://doi.org/10.1109/IGARSS.2019.8898545

10. Jayadharshini J, Sivapriya R, Abirami S, (2018) Trend square: an android application for extracting twitter trends based on location. In: 2018 International conference on current trends towards converging technologies (ICCTCT), Coimbatore, pp 1–5. https://doi.org/10.1109/ICCTCT.2018.8551056.

11. Albayrak MD, Gray-Roncal W (2019) Data mining and sentiment analysis of real-time twitter messages for monitoring and predicting events. In: IEEE integrated stem education conference (ISEC), Princeton, NJ, USA pp 42–43. https://doi.org/10.1109/ISECon.2019.8881956

12. Gupta N, Crosby H, Purser D, Javis S, Guo W (2018) Twitter Usage across industry: a spatiotemporal analysis. In: 2018 IEEE fourth international conference on big data computing service and applications (BigDataService), Bamberg, pp 64-71. https://doi.org/10.1109/BigDataService.2018.00018.

13. Smys S, Raj JS (2019) A stochastic mobile data traffic model for vehicular ad hoc networks. J Ubiquitous Comput Commun Technol (UCCT), pp 55–63. https://doi.org/10.36548/jucct.2019.1.006

14. Bhalaji N (2020) Efficient and secure data utilization in mobile edge computing by data replication. J ISMAC 2(01):1–12

15. Al Shammari AS (2018) Real-time Twitter Sentiment Analysis using 3-way classifier. In: 21st Saudi computer society national computer conference (NCC), Riyadh, pp 1–3. https://doi.org/10.1109/NCG.2018.8593205