Evaluation of Geotagging Twitter Data Using Sentiment Analysis During COVID-19

V. Ajantha Devi and Anand Nayyar

Abstract Late advances in innovation have empowered individuals to add area data to informal communities’ geolocations where individuals share their correspondence and whereabouts in their day by day lives, yet in addition during anomalous circumstances, for example, emergency occasions. Be that as it may, since the volume of the information surpasses the limits of human logical capacities, it is nearly inconceivable to play out a direct subjective investigation of the information. The emerging field of visual investigation has been acquainted with tackle such difficulties by integrating the methodologies from factual information examination and human computer into exceptionally intelligent visual situations. The researcher extracts valuable hidden data from the huge volume of unstructured social media data and model the extracted information for visualizing meaningful information along with user-centered interactive interfaces. This paper proposes sentiment analytics of geo-tagged for anomalous social movement during COVID-19 by incorporating a novel classification technique and by a visual analytics approach for forecasting the overall flow of human crowds.

Keywords Sentiment analysis · Twitter · Geotagging · Social sensing · Machine learning

1 Introduction

The coronavirus disease of 2019, known as COVID-19, is a quickly spreading illness. The COVID-19 is currently viewed as a pandemic that has influenced nations in completely occupied continent. Since the primary instances of COVID-19 announced in Wuhan, China, in December 2019, the quantity of fatalities worldwide
has expanded quickly. Because of its high contamination and passing rate, governments have executed a wide scope of strategies planned for relieving the spread of this infection and its effect. Such activities started with the Chinese government request to isolate Wuhan on January 23, 2020, to, most as of late, different nations proclaiming highly sensitive situation and executing exacting isolate and social separating estimations (e.g., US, Italy, Argentina, Spain).

Most government pioneers have actualized measures to boost, and sometimes uphold, “social distancing” to decrease the spread of COVID-19. These measures have brought about the dropped amusement occasions, terminations of schools and universities, and organizations diminishing long periods of activity, actualize working from home, or close through and through. There is no uncertainty the pandemic and the measures set up to relieve it have and will keep on definitely sway the lives of millions. As this pandemic and the reactions to it are uncommon, be that as it may, we are probably going to be shocked by how individuals react.

Since the beginning times of the sickness, individuals have communicated their assessment and shared data, just as deception, about it by means of Internet-based life stages, for example, Twitter. As COVID-19 spreads to different nations and governments attempt to relieve its effect by actualizing countermeasures, individuals have additionally utilized Internet-based life stages to communicate their feeling about the measures themselves, the pioneers executing them, and the manners in which their lives are evolving. The utilization of Web-based life, for example, Twitter, as stages to communicate suppositions and offer data about COVID-19, will just keep on developing, exactly in view of the “social removing” measures set up to moderate it.

2 Proposed Methodology

Initially, the gathered Twitter information is pre-processed for information cleaning [1]. The significant features are separated from the pre-processed clean content, applying any of the feature selection techniques as shown in Fig. 1.

The segmented information is physically marked as positive, negative and neutral tweets to set up a training set. At last, the extracted features [2] and the marked training set are given as a contribution to the fabricated classifier to characterize and the rest of the information for test set. Especially the diseases affect people’s family life, business life and all kinds of social life. COVID is the most dangerous problem affecting people in recent period.

2.1 Data Resources

Decision of data resource leads the sentiment analysis [3–5] expect a tremendous activity. Web-based social networking stages as the information sources [6, 7] are
widespread as micro-blogging sites, blogs and review Web site. Micro-blogging (Twitter) has increased high notoriety because of its restricted quality of the substance and publicly accessibility of information, which utilize Twitter as the information hotspot for notion investigation.

**Twitter Studies**

According to the ongoing work, the examinations complete on Twitter information are in the area of marketing, governmental issues, health care, promoting market, sports and so forth. Furthermore, Twitter has been casted a ballot as the most encouraging hotspot for the investigations, for example, network or impact recognition, theme revelation, market and business expectations, proposal frameworks and tweet order.

**Tweets**

The message posted on Twitter which is restricted to 140 characters. Tweets are commonly made out of: content, joins, emojis and pictures. A 6 s video is even included as a tweet segment. In view of these parts, the mining is applied to arrange content, joins, pictures, emoticon or emojis and even recordings. The tweets contain three documentations including hashtags (#), retweets (RT) and record Id (@).

**Dataset Description**

The dataset presented is being continuously collected [8] using the Twitter API. The dataset presented here (v1) covers March 30, 2020–April 26, 2020 and contains 6,468,526 tweets. The keywords used for search tweets are: *virus*, *coronavirus*, *ncov19* and *ncov2019* since 30 March, and *COVID-19* since April 2020.
The average daily number of tweets collected on dataset v1 was 208,662.1 (SD = 100,448.7, Mdn = 243,087). The number of tweets collected increased every month from 724,877 in March, 3,084,729 in April.

### 2.2 Data Preprocessing

The gathered information is raw tweet data. So as to apply classifier [2], it is fundamental to pre-process or clean the raw data. The preprocessing task includes uniform packaging, evacuation of hashtags and other Twitter documentations (#, @, RT), emojis, URLs which is important to expel non-letter information and images, stop words (are, is, am and so on), decompression of slang words and pressure of lengthened words for Daddddyyyy as Daddy (Fig. 2).

### 2.3 Feature Extraction

The pre-processed dataset has different discrete properties. In highlight extraction techniques [2, 6], which are classified as various viewpoints, for example, descriptive words, action words which are recognized as positive or negative to identify the extremity of the entire sentence. Followings are the broadly utilized feature extraction [2] strategies which highlights signify individual and particular words and their event tallies.

- **Negative phrases:** The negativity words can change the importance or direction of the feeling. So it is clear to take negative word in account.
- **Parts Of Speech (POS):** Finding things, action words, descriptors and so on as they are huge measures of sentiments.
2.4 Sentiment Classification Techniques

Information-based procedure is additionally called lexicon-based method. The dictionary put together procedure centers with respect to inferring the feeling-based vocabularies from the content and afterward recognizing the extremity of those dictionaries. Dictionaries are the assortment of known and precompiled notion terms. This methodology is additionally grouped into dictionary-based methodology and corpus-based methodology. In the dictionary-based methodology, the feeling focused words [6–10], and afterward look at the word reference are discovered to gather their equivalents and antonyms. Though in the corpus-based methodology, a rundown of feeling words and afterward dependent are made on their setting explicit directions, extra-related sentiment words [9] in an immense corpus are located. To lead vocabulary approach, a paltry arrangement of words portraying suppositions is gathered physically with their referred to directions as a mean of preprocessing task. The set is then developed step by step via looking in the recognized and broadly utilized vocabulary word reference instrument (Fig. 3).

![Tweet map on April 26, 2020](image-url)

**Fig. 3** Tweet map on April 26, 2020
2.5 Visualization

Visualization [11] gives a viable method for information investigation and information portrayal. This area presents some representation strategies for fleeting, spatial and literary examples, which are reasonable for showing our systematic outcomes.

Time is an important property of geo-tagged social media information. Outlines, for example, stacked charts and bar outlines, are a decent regular technique for envisioning straight time. A clock-like time hub could be embraced to underscore the cyclic character of time. Shading or association is commonly reasonable for the translation of the overall time of geo-tagged social media information, though the hub-based plan could be utilized to introduce supreme time (Figs. 4 and 5).

Fig. 4  Sentimental analysis of latest 10,000 tweets

Fig. 5  Sentimental analysis breakdown of the tweets
3 Geo-tagged Analysis for Detector

A client has an alternative to make their area open through a cell phone in the event that they wish to, or give area subtleties in their profile page. The area field is applicable to our exploration since it causes us in following the current/default area of a client. Geolocation codes are appended to the message in an area empowered portable post. Such geo-labeled substance will improve the accessibility and precision of the geographic information attached to COVID or coronavirus related Twitter API posts. For all different purposes, we accept the area quality inside the profile page to be his/her current location and pass it as a contribution to Google’s area-based Web administrations to get geo-area codes (i.e., scope and longitude) alongside the nation, state, city with a specific precision scale.

4 Conclusion

In summary, this research paper presents a nonexclusive structure for examining the effects of social events from geo-tagged social media. Following a conversation of the important information preprocessing, an occasion-related data extraction technique joining AI and hashtags is received to remove unlabeled occasion related data from the first dataset. Social sentiment analysis [3, 4] and social supposition mining are then used to investigate the open’s comprehension of and sentiments about the occasion COVID-19. The data visualization [11] has been given subsequent to estimating sentiment analysis on the data. The data curated on Twitter are based on two determined hashtag keywords, which are (“COVID-19, coronavirus”) and the user can be detected by using geotagging enforcing “social distancing.”

References

1. Bhuta S, Doshi A, Doshi U, Narvekar M (2014) A review of techniques for sentiment analysis of twitter data. In: Issues and Challenges In Intelligent Computing Techniques (ICICT), pp 583–591
2. Neethu MS, Rajasree R (2013) Sentiment analysis in twitter using machine learning techniques. In: 4th international communications and networking technologies (ICCCNT) conference of computing, pp 1–5
3. Ghag K, Shah K (2013) Comparative analysis of the techniques for sentiment analysis. In: International conference on advances in technology and engineering, pp 1–7
4. Medhat W, Hassan A, Korashy H (2014) Sentiment analysis algorithms and applications: a survey. Ain Shams Eng J 5(4):1093–1113
5. Khairnar J, Kinikar M (2013) Machine learning algorithms for opinion mining and sentiment classification. Int J Sci Res Publ 3(6)
6. Bahrainian S, Dangel A (2013) Sentiment analysis using sentiment features. In: International joint conference of web intelligence and intelligent agent technologies, pp 26–29
7. Kasture N, Bhilare, P (2015) An Approach for Sentiment analysis on social networking sites. In: Computing communication control and automation (ICCUBEA), pp 390–395
8. Dubey ADJAAS (2020) Twitter sentiment analysis during COVID19 outbreak
9. Gautam G, Yadav D (2014) Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In: 7th international conference on contemporary computing, pp 437–442
10. Gokulakrishnan B, Plavnathan P, Thiruchittampalam R, Perera A, Prasath N (2012) Opinion mining and sentiment analysis on a Twitter data stream. In: International conference on advances in ICT for engineering regions, pp 182–188
11. Ra M, Ab B, Kc S COVID-19 outbreak: Tweet based analysis and visualization towards the influence of coronavirus in the world