International Journal of Geo-Information

MDPI

Article

Understanding Users' Satisfaction Towards Public Transit System in India: A Case-Study of Mumbai

Rahul Deb Das [1,2]

[1] Department of Geography, University of Zurich, CH-8006 Zurich, Switzerland; Rahul.Deb.Das@ibm.com or das.rahuld@gmail.com
[2] IBM, Mies-van-der-Rohe-Strasse 6, 80807 Munich, Germany

Abstract: In this work, we present a novel approach to understand the quality of public transit system in resource constrained regions using user-generated contents. With growing urban population, it is getting difficult to manage travel demand in an effective way. This problem is more prevalent in developing cities due to lack of budget and proper surveillance system. Due to resource constraints, developing cities have limited infrastructure to monitor transport services. To improve the quality and patronage of public transit system, authorities often use manual travel surveys. But manual surveys often suffer from quality issues. For example, respondents may not provide all the detailed travel information in a manual travel survey. The survey may have sampling bias. Due to close-ended design (specific questions in the questionnaire), lots of relevant information may not be captured in a manual survey process. To address these issues, we investigated if user-generated contents, for example, Twitter data, can be used to understand service quality in Greater Mumbai in India, which can complement existing manual survey process. To do this, we assumed that, if a tweet is relevant to public transport system and contains negative sentiment, then that tweet expresses user's dissatisfaction towards the public transport service. Since most of the tweets do not have any explicit geolocation, we also presented a model that does not only extract users' dissatisfaction towards public transit system but also retrieves the spatial context of dissatisfaction and the potential causes that affect the service quality. It is observed that a Random Forest-based model outperforms other machine learning models, while yielding 0.97 precision and 0.88 F1-score.

Keywords: Tweet; sentiment analysis; public transit system; machine learning; natural language processing; georeferencing; users' satisfaction

Citation: Das, R.D. Understanding Users' Satisfaction Towards Public Transit System in India: A Case-Study of Mumbai. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 155. https://doi.org/10.3390/ijgi10030155

Academic Editor: Wolfgang Kainz

Received: 26 December 2020
Accepted: 7 March 2021
Published: 10 March 2021

1. Introduction

Public transport is often considered as the primary mode of mobility in developing cities due to its affordability compared to private automobiles and privately operated shared-ride transportation systems [1,2]. It has been observed that the service provided by the public transit systems in developing cities often do not meet the users' demand and their expectations [3]. The quality of public transport service has been modeled and attributed by various criteria, for example, network coverage, accessibility, affordability, safety, and cleanliness, to name a few [3,4]. However, the criteria are region-specific, which are also influenced by specific socio-economic and cultural setup, topography, and the level of governance. If the criteria do not fulfill users' need, then that affects their perceived satisfaction level. Understanding users' perceived satisfaction towards public transit system can help in improving the quality of service and performance measures. Information about users' perceived satisfaction is often unavailable which leads to mismanagement and improper allocation of transportation supply owing to drop in quality of service. As a consequence, users experience difficulty in commuting with limited activity space. Users' satisfaction information is primarily collected through manual survey [3,5], which often suffers from quality issues. The satisfaction information can also be collected through

offline paper-based or online web-based questionnaire [6]. However, manual surveys are expensive, time consuming, and often suffer from under-reporting or miss reporting. For example, respondents may not provide all the detailed travel information in a manual travel survey. The survey may have sampling bias. Due to close-ended design (specific questions in the questionnaire), lots of relevant information (related to users' perceived experience) may not be captured in a manual survey process.

With the emergence of ubiquitous mobile platform and easy access to internet, people can share their opinions and reactions towards various events in the form of user-generated contents (UGC) [7]. UGC are generally produced by the users on social media platforms in the form of text, image, or video. In this paper, we explore the potential of UGC to understand users' perceived satisfaction towards public transit system. To understand this, we chose Mumbai suburban railway system in Greater Mumbai in India.

According to a report [8] in 2018, the total number of social media users is 326.10 million in India, out of which Facebook accounts for 73% usage, followed by Instagram (20%) and Twitter (2.37%). In India, more than half (52.30%) of the social media users are young generation with an average age group of 27, followed by generation Z (28.40%), and the age group of 35–44 (15.80%). As a matter of fact, older generations account for only 4.20% [8]. Although the socio-economic status of young commuters who are active on social media platforms is not readily available in Greater Mumbai, based on a study of (young) consumer behavior in Mumbai [9], it has been observed that majority of the young generation are graduate to postgraduate with 32.70% has no monthly income, 19% earns less than 25,000 INR, 22% earns in the range of 25,000 to 50,000 INR, and 26% earns more than 50,000 INR. Eventually, a vast majority of the low income population will prefer public transit due to its affordability [1]. Although there is a prominent gender inequality on Twitter (16% being female, 84% being male) [8], there is a growing use of this platform to disseminate news and personal experience [10]. For example, people can post about various events on Twitter in the form of micro-blogs, also known as tweets. The events can be related to political agenda [11], disaster incidents [12], or transportation systems [13,14]. Recently, there has been a growing trend to use Twitter to share traffic conditions, road conditions, and transport service quality in major metro cities in India. This helps users to adapt their travel plans in a more effective way. On the other hand, transport authority also gets to know current traffic conditions and infrastructure issues perceived by the users. Given that the majority of social media users are mostly young generation, there is a user bias on Twitter. However, as there is a growing use of sharing travel experience on Twitter, we aimed to investigate how users express their (dis)satisfaction towards public transit systems in Greater Mumbai by analyzing Twitter data to support transport authorities to improve the transport service quality.

In Greater Mumbai, people primarily use public transport service (bus service or suburban railway service) for 75% of the motorized trips [15], with a higher mode share on railway system [16]. Mumbai Suburban railway system caters to almost 7.5 million users on daily basis. On yearly basis, Mumbai railway service caters to 2.64 billion people in Mumbai, which makes it one of the busiest railway services in the world [17,18]. The travel demand for the railway service is generally very high, making the trains and platforms overcrowded, which often poses strain on the mobility supply and infrastructure. To meet such high travel demand, it is important to understand quality of service. Since Mumbai railway service has more users compared to its bus counterpart, we chose the railway service to understand how users react to public transit system in Mumbai. In this paper, the author possesses local (geographical) knowledge about Mumbai suburban areas. That said, past studies showed importance of local knowledge helps in more effective (geographic) information extraction and validation [19]. Factors related to high travel demand for Mumbai railway system and the author's local knowledge motivated the choice of Greater Mumbai as a study area and the Mumbai suburban railway service as the primary public transport mode system to be analyzed in this research. Due to very high travel demand, the Mumbai suburban railway system often experiences overcrowded compartments, mismanagement of service, and fatal accidents [20].

A recent study shows India has a growing number of Twitter users (https://www.statista.com/statistics/381832/twitter-users-india/, accessed on 7 March 2021). Thus, in this paper, we investigated if tweets can be analyzed to understand users' perception about Mumbai railway system in an automated manner.

Previous works investigated the feasibility of UGC to model travel demand [21], usage of urban space [22], and road traffic incident detection [23]. To the best of our knowledge, only a few works investigated the feasibility of Twitter data to understand users' satisfaction towards public transit system [24]. That said, if it is known from the tweets that people are not satisfied due to delay or frequent cancellation of trains at a given location, then the authority can take proper measures to improve the service. However, the majority of the tweets do not have geotag information [10]. People often mention a location context while reporting about a mobility issue. In the following tweets, we show how users report about various issues at different locations.

- *Tweet 1:* Issue is not only at Parsik Tunnel for UP fast trains but at 1. before arrival at Kalyan station 2. Before arrival at Thane station 3. Before arrival at Dadar station. Most of the fast trains are stopping at above places & at Parsik tunnel which is causing delay. Please look into...
- *Tweet 2:* Escalator of platform no 6 & 7 of BORIVLI Mumbai not working since long. Mail n express trains comes here. Passengers with luggage have to climb foot over bridge causes problems. Authorities look in to the matter.
- *Tweet 3:* @se_railway 1802 from CST now crammed with passengers for 1823. Dangerously overcrowded train not yet left CST. Fucking nightmare.

In the first tweet (Tweet 1), the user reported about the delay of trains at *Parsik tunnel*. In the second tweet (Tweet 2), the user mentioned the malfunctioning of the escalator at *Borivli* with a wrong spelling. The location *Borivli* should be correctly spelled as *Borivali*. Sometimes a tweet may be grammatically incorrect or may have abbreviations. Such informal and noisy text leads to poor performance in simple lexicon-based lookup while retrieving the place name. In the third tweet (Tweet 3), a user reported a risk of boarding on an overcrowded train at *CST*, which is an abbreviation of *Chhatrapati Shivaji Maharaj Terminus*. The user also mentioned railway authority in the tweet to bring this situation into their notice.

In a previous work, Collins and colleagues [24] primarily investigated users' sentiments towards public transit system in Chicago. However, they did not attempt to extract the location information from the tweets. In this work, we go beyond the existing works by not only investigating the sentiments in a tweet but also retrieve the location contexts by combining machine learning models and knowledge-based approaches. Sentiment analysis classifies a text into negative, positive and neutral sentiment. In this research, we considered that a tweet with negative sentiment (also called negative tweet) expresses dissatisfaction of the user. When done at an aggregate level, the model will also provide information of spatial distribution of service quality. This will help transport authority to pinpoint which locations need more attention based on their frequency in negative tweets.

The primary motivation of this research is to enrich existing travel survey process by developing a UGC-based machine learning technique that can retrieve users' perception towards transport service along with the spatial context. Although UGC is often informal, does not follow proper syntactic and lexicographical structure, and poses challenges in location retrieval, there are a number of advantages of using UGC that can add value to the existing travel survey process. The value additions are as follows.

- A UGC is more free flowing. Users can express their (travel) experience, which is sometimes difficult to capture through manual surveys due to their close-ended design.
- Through UGC-based approach, transport authority can monitor service quality, along with location information, either in real time or in historical manner.
- In contrast to manual survey, a UGC-based approach does not need any field staff, or longer time for survey design and implementation. The UGC-based approach is more automated and can be deployed on streaming data on a specific city or multiple

cities simultaneously. Due to better scalability, the UGC-based approach can save time and effort when a larger geography or multiple cities need to be monitored.

- The success of manual survey is highly dependent on the ability of the surveyors to design the questionnaire to collect the most relevant information or field staff to motivate the respondents. In order to keep the field staffs focused and motivated, regular team meetings and trainings are held. This is often tedious when performed periodically. On the other hand, a UGC-based approach can be conducted continuously or in a periodical manner or around a specific time frame or an event. This will provide more insight to the authority how users perceive (transport) service quality at different time periods (different seasons), or during a new policy implementation (e.g., implementing new transport service or rise in travel fare).

Thus, UGC can be a valuable source of information which can complement manual travel survey approach. However, as UGC is informal and unstructured, we developed a machine learning-based model that can retrieve transport service quality perceived by the users along with location information that may need attention from the authority. Since Twitter has published their application programming interface (API) to retrieve tweets, in this research, we used that API to collect tweets for further analysis.

In this paper, we aimed to address the following research questions.

- Can Twitter be used to understand users' satisfaction towards public transit system in Greater Mumbai?
- How can we automatically identify tweets that are relevant to public transit authority in the context of users' satisfaction study?
- How do people characterize service criteria related to mobility and infrastructure issues?

We hypothesize that, with the growing Twitter usage in Greater Mumbai, it is possible to understand users' satisfaction and spatial distribution of the service quality by analyzing ungeotagged tweets.

The remaining paper has been organized in the following manner. Section 2 provides state-of-the-art. We explain our model in Section 3. In Section 4, we discuss data preparation, experimental setups and results. In Section 5, we address the research questions. Section 6 contains concluding remarks, limitations, and key findings.

2. Related Work

Most of the developing cities face mobility issues with public transportation systems primarily due to budget constraints and lack of coordination and knowledge gap in users' perceived quality of service and operators' perception [4,25]. In India, the main problem of poor public transportation system is lack of financial resources and proper planning [25]. To improve the service strategically with a restricted budget, it is important to prioritize the issues that need immediate attention. This can be done by understanding users' perception about different attributes of service quality at different locations [4].

Users' perception towards service quality of transportation system reflects their satisfaction level. Users' satisfaction can be estimated either at a global level (aggregate analysis over a public transit system) or at a specific level (individual analysis over a given service criteria) [26]. In literature, estimating users' satisfaction at a global level is known as global satisfaction, whereas satisfaction for a given service criteria is known as specific satisfaction [26]. Currently, users' satisfaction is studied through manual travel survey process where users are asked about their socio-demographic profile, commuting behavior, and their satisfaction level (on a Likert scale or alike) towards the overall transit system or some specific criteria [3,26,27].

Many researchers have studied the most relevant service criteria specific to a given geography [3]. For example, Vuchic highlighted a number of service criteria, e.g., accessibility, availability, travel time, reliability of service, comfort, safety, environmental impact [28]. Eboli and Mazzulla conducted a study on users' satisfaction towards public bus service in Europe based on a number of criteria, e.g., service reliability, availability, comfort, cleanliness, safety and security, environmental impact, access to trip information [29]. Ngoc

and colleagues studied the most relevant criteria that influence users' satisfaction towards public transit system in Hanoi using factor analysis and linear regression [3]. Ngoc and colleagues found the most relevant service criteria in Hanoi are safety and security, service coverage, and comfort [3]. Dube conducted a manual survey to understand users' satisfaction towards Indian railway service using a close-ended questionnaire [30]. Three field staffs were deployed to interview 700 participants (including 100 railway officials) over 10 days. Out of all the participants, 72% were male, and 28% were female. The results showed users expressed their dissatisfaction towards cleanliness in toilets and platforms, delay of trains, and unauthorized vendors on the train. On the other hand, users expressed their satisfaction towards the waiting service, seating and water facility on the platform, pricing of railway food, and fans and lighting arrangement in the trains. While understanding the most relevant service criteria for users' satisfaction is an active area of research [31], the manual travel survey process which is used to collect users' satisfaction information involves quality issues and budget constraint [32]. For example, a manual travel survey process involves financial constraints, lack of field staffs (interviewers), long gestation periods, and quality issues (correctness and bias in the response). To address the shortcomings of manual travel survey process, there is a need to automate the process of understanding users' satisfaction towards public transit system, especially in developing cities where financial resource is scarce and lack of communication exists between service providers and users. The automation can also complement the existing manual process.

With the emergence of ubiquitous information and communication technologies (ICT) and social media platforms (e.g., Twitter), people can share various information in a more dynamic way. Recent studies show people share their perception and feelings about different objects and their attributes in the form of user-generated contents [33]. The feelings about any entity is called sentiment towards the given entity (or any of its attributes) [33]. Sentiments are subjective and can be expressed in terms of positive, negative, or neutral polarity. In the context of quality assessment of a product or service, a negative sentiment means dissatisfaction towards the product or service, whereas a positive sentiment reflects satisfaction of the user. While sentiment analysis using UGC is yet to be practiced at a larger scale in users' satisfaction study, it has been already used in other domains, e.g., in business management [34], movie reviews [35], socio-political study [36], and spam detection in product reviews [37], to name a few. In the literature, sentiment analysis has been performed primarily by two ways, e.g., unsupervised approach and supervised learning technique [33]. In the unsupervised approach, a sentiment lexicon (also known as affective lexicon) is used to estimate the word level sentiment in the document and thereby compute the document level sentiment through an aggregator function [38,39]. Generally, an average or maximum sentiment value is computed [39]. The overall sentiment type is then detected based on the aggregated sentiment value. An unsupervised approach is more suitable where the data is not annotated. Since most of the sentiment lexicons are manually crafted or developed for general applications, they show different inter-annotator agreement in a specific domain [40]. On the other hand, a supervised approach is used when an annotated data exists to train a machine learning model. Once the model is trained with annotated sentiment data, then the model can be deployed to detect sentiment of test data [33].

Limsopatham and colleagues used 1700 tweets to understand the temporal patterns of users' reactions towards disruption of railway service in Glasgow [41]. They found users mostly react at a rush hour on weekdays, whereas users react during the late evening on weekends. Congosto and colleagues used tweets generated from subway users in Madrid to detect micro events related to cleanliness and delay, to name a few, using a handcrafted transport event lexicon [42]. Collins and colleagues explored tweets to understand users' global satisfaction in Chicago using 557 tweets through sentiment analysis [24]. Collins and colleagues used a word level sentiment analysis approach using a sentiment lexicon, namely SentiStrength [39]. Collins showed the number of negative tweets are generally more than positive ones. This implies users are more likely to share negative sentiments

compared to positive sentiments. Anastasia and Budi used 2500 annotated tweets to detect users' satisfaction towards two popular transportation providers (e.g., Go-JEK and Grab) in Indonesia in terms of net sentiment score using supervised learning techniques [43]. Jurdak and colleagues studied human mobility patterns in Australia from geotagged tweets, and they found that the majority of human mobility is centered around metropolitan cities [44]. Zornoza and colleagues analyzed human mobility patterns in Valencia, Spain, using geotagged tweets. They developed a model to detect users' home location [45]. Most of the mobility-based research has used geotagged tweets, which accounts for only 0.1% to 3% of total tweet volume [10,46]. Thus, a lot of ungeotagged tweets containing significant information do not provide explicit location information.

To the best of our knowledge, no work has been done to understand users' satisfaction and the location mentions with different sentiment types from ungeotagged tweets. Since, in this research, we used ungeotagged tweets, which do not have any explicit location information, we developed a model that will detect tweets that contain negative sentiments and the concerned location mentions from the tweet contents. Since tweets are informal and contain region-specific peculiarities in location mentions, we combined a supervised approach and a knowledge-based approach similar to Gelernter and colleagues [47] and developed a hybrid georeferencing model, which is tuned to perform at Greater Mumbai.

3. Methodology

To understand users' satisfaction from UGC, the first task is to retrieve tweets that contain users' perception and any other information relevant to railway service. Once the relevant tweets are retrieved, a sentiment analysis is performed to extract tweets that are relevant and contain negative, positive, or neutral sentiment. A tweet containing negative sentiment can be called as negative tweet. Similarly, a tweet containing positive or neutral sentiment is called a positive or neutral tweet, respectively. To classify a relevant tweet into a specific sentiment type, a number of machine learning models are evaluated.

Once a tweet is identified containing negative sentiment, it is also important to understand the spatial context of that negative sentiment. Since, in this research, we aim to explore the potential of ungeotagged tweets to understand users' (dis)satisfaction, we developed a novel georeferencing module to retrieve location information from the tweet content. The retrieved location information will provide spatial context of service quality in terms of users' sentiment type. The entire workflow has been depicted in Figure 1. The workflow can be deployed as a batch service on a cloud platform or as a standalone application. The model first retrieves a raw tweet, which is then pre-processed. Then, the model detects if a tweet is relevant or not. If the tweet is irrelevant, the model retrieves the next tweet from the repository. In case of a relevant tweet, a sentiment detection is performed followed by georeferencing. The process continues to iterate over the entire repository containing historical tweets. The process can also handle real time tweets collected in streaming mode.

3.1. Sentiment Analysis Using Supervised Learning Technique

Once the data is manually annotated by the volunteers, the annotated data is used to build a supervised machine learning model that can automatically identify the sentiment type in a text. As negative sentiments signify users' dissatisfaction towards a service, we specifically focused on extracting negative tweets, which are critical to the transport authority.

Thus, we developed a two-stage classification models to categorize tweets using supervised machine learning approach. The first classification model distinguishes relevant tweets and irrelevant tweets. Then, a second classifier categorizes each relevant tweet into three different sentiment types (positive, negative, neutral).

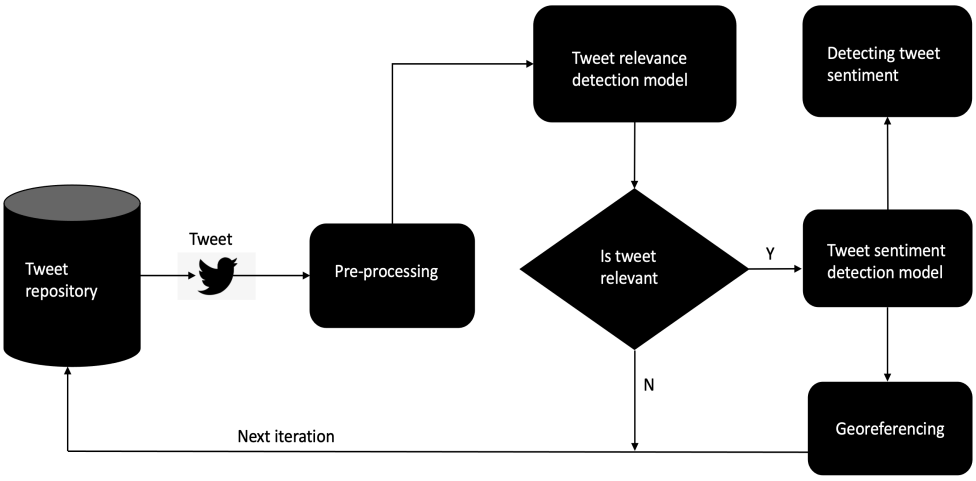

Figure 1. A schematic diagram showing the workflow.

Since any machine learning model cannot deal with raw text, we need to convert the text into a numerical representation that could be used by a predictive model. To do that, first, we cleaned the text to remove any white space, non-ASCII characters, and special symbols. This helps the model to not get overfitted on the training data. Following that, we removed a number of stop words, which do not bear any semantics, e.g., a, the, is, and was, to name a few. In the third step, each tweet is tokenized into a number of sentences, and each sentence is tokenized into a number of words. Since words with similar meaning can be used in different forms in a text, we used a Lovins Stemming algorithm to convert each word to its base form, thereby reducing the dimensionality of the feature space.

Following the pre-processing phase, a Bag-of-Word (BoW) approach is used to generate feature vectors for each tweet. In this case, each tweet is considered as a single document. A BoW is a classical representation of text, which is used in natural language processing (NLP) where each word is considered as a feature to be used by a machine learning model. In the BoW approach, the grammar and the order of the words in a document are not considered. Since the machine learning model requires a numerical input, the (word) features are usually converted to numerical representation. To do that, we computed a numerical weight for each word token in terms of its term frequency-inverse document frequency (*TF-IDF*) to weigh a word based on its occurrence in a document, and also in the whole corpus, as follows (Equations (1)–(3)).

$$TF = T_t, \tag{1}$$

$$IDF = log[\frac{N}{(1 + D_t)}], \tag{2}$$

$$TF - IDF = T_t \times log[\frac{N}{(1 + D_t)}], \tag{3}$$

where T_t is the total count of term 't' in document 'D'. N is the total number of document in the corpus, and D_t is the total number of documents containing the term 't'.

3.2. Affective Lexicons for Sentiment Analysis

An affective lexicon is a vocabulary of words with given sentiment types, e.g., SentiWordNet [38], NRCEmotion Lexicon (EmoLex) [48], and Sentistrength [39]. In this research, we explored these three state-of-the-art lexicons to detect sentiments in tweets related to railway service in Greater Mumbai. To detect sentiment of a document, the document is first tokenized into a number of constituent word tokens. Then, each word token is looked up in a given lexicon to find its sentiment type (or sentiment strength). Then, an average or maximum sentiment strength is computed over the entire document. Sometimes, a docu-

ment can also contain both positive and negative sentiment. However, in this research, we assume each tweet contains a single sentiment type. We used an aggregator function to compute an overall sentiment score for a given tweet.

3.2.1. SentiStrength

SentiStrength (SNS) is an affective lexicon which is primarily developed for detecting sentiment of a short text, e.g., tweets [39]. SentiStrength originally consists of 298 positive and 465 negative words. The sentiment strengths are assigned on a scale of 1 to 5 or -5 to -1, respectively, where 5 and -5 indicates maximum positive and maximum negative sentiment strength, respectively. Since SNS is designed to handle typos, colloquialisms, negations, punctuation, and emoticons and their associated sentiments in a text, it eventually performs better on shorter and informal text compared to other lexicon-based approaches. For example, a word *gooood* will be converted to *good* by SentiStrength. If a negation word, e.g., *shouldn't* or *don't* or *not*, appears before a specific sentiment word, then SentiStrength inverts the sentiment type of the given sentiment word. For example, the word *good* carries a positive sentiment, but, if SentiStrength encounters a negation word, for example, *not*, before *good*, it returns an overall negative sentiment. SentiStrength also increases sentiment strength of a word if it is preceded by a booster word, e.g., *very*, or followed by extra punctuation or repetition of same word to indicate the strong sentiment of the user. SentiStrength also supports expanding the lexicon by adding domain-specific terms with their respective sentiment strength. In this research, we aggregated the maximum positive (V_{pos}^{max}) and maximum negative (V_{neg}^{max}) sentiment value in a tweet as follows.

$$V^{resultant} = V_{pos}^{max} + V_{neg}^{max}. \tag{4}$$

If $V^{resultant} > 0$, we label the tweet as positive. If $V^{resultant} < 0$, we label the tweet as negative. If $V^{resultant} = 0$, we label the tweet as neutral.

3.2.2. SentiWordNet

SentiWordNet (SWN) is based on a semantic lexicon, e.g., WordNet [49], where terms that share similar sense are clustered in a same group known as synset. The terms in a given synset have the same parts of speech. Since a given term can be used either in a positive or negative way, each term is assigned values for both positive (V_{pos}) and negative sentiment (V_{neg}). Based on the context, the neutrality or objective value (V_{obj}) of a term can also be computed using the following formula.

$$\begin{aligned} V_{obj} &= 1 - (V_{pos} + V_{neg}) \\ \ni 0 \leq V_{pos} \leq 1, V_{neg} &\leq 0; V_{pos} * V_{neg} \leq 0 \end{aligned}. \tag{5}$$

Thus, the resultant sentiment value (V_i) of any term (W_i) in a tweet can be expressed as follows.

$$V_i = (V_{pos} + V_{neg}) : -1 \leq V_i \leq 1. \tag{6}$$

In this research, when using SWN, we computed the overall sentiment ($V^{resultant}$) of a given tweet by averaging the sentiment values of all the sentiment bearing terms (W_1,.....,W_N) in a tweet (T_i) as follows.

$$V^{resultant} = \frac{1}{N} \sum_{i=1}^{i=N} V_i. \tag{7}$$

In this research, we used SentiWordNet 3.0, which consists of total 117,660 terms with different senses. Unlike SentiStrength, SentiWordNet does not support spelling correction or negation, nor boosting sentiment strength in presence of certain booster words or punctuation.

3.2.3. EmoLex

EmoLex (NRC) is primarily an emotion lexicon which consists of 14,182 terms with approximately 25,000 senses. Each term is assigned either a positive or negative sentiment type and a specific emotion from eight pre-defined emotion types, e.g., anger, anticipation, disgust, fear, joy, sadness, surprise, trust. In this research, we used NRC to detect sentiment type of a tweet by matching each word in the tweet against the lexicon. We considered the most frequent sentiment type found in the tweet is the overall sentiment of the given tweet. Let us assume a '+' symbol denotes a positive sentiment, whereas a '−' denotes a negative sentiment. Now, if a tweet (T_i) consists of three word tokens (W_i) with their respective sentiment types such that T_i : = {$W_1(+)$, $W_2(-)$, $W_3(+)$}, then the overall sentiment of T_i is positive sentiment as the number of positive sentiment words are more than the negative ones.

3.3. Georeferencing Module

To extract location mentions in tweet content, we developed a hybrid georeferencing module which consists of two layers. The first layer (Layer 1) uses supervised model, whereas the second layer (Layer 2) uses a knowledge base. The knowledge base is developed using a number of spatial rules and local geographical aspects adapted to Indian context.

Since tweets contain different degree of informality, we used two supervised models in first layer, e.g., a linear chain Conditional Random Field (CRF) and a Maximum Entropy (MaxEnt) model. We observed tweets generated by news channels or transport authority is more formal than tweets generated by common users. To handle such diverse informality and unstructuredness in the text, we used transfer learning by using a pre-trained CRF model trained on formal texts (CoNLL-2003 data set). This pre-trained CRF model is provided in StanfordNER [50]. On the other hand, we retrained MaxEnt model on informal tweets collected in Greater Mumbai. Thus, the two models deal with different degree of informality in the text.

To further strengthen the performance of the model, we constructed spatial rules based on spatial prepositions and a number vernacular names used in Greater Mumbai. These rules are used to develop a knowledge base in the second layer of the georeferencing module. A location entity is generally a proper noun or common noun and appears after spatial prepositions, e.g., *at*, *near*, *towards*, and *from*, to name a few.

We noticed, in Greater Mumbai, people mention place names in different ways. For example, people use different abbreviations or multiple tokens to refer a same place name, e.g., *chhatrapati shivaji terminus* or *lokmanya chhatrapati shivaji terminus* or *CSMT* or *CST*, all refer to the same place. We also observed, in Greater Mumbai, while mentioning place names, people use lots of vernacular names, for example, Tilak *nagar*, Raj and *bhavan*, to name a few. Generally, such vernacular names occur after a proper noun. Most of these vernacular names refer to some local geographical objects. For example, *nagar* in Hindi means suburb or city in English.

The parts of speech (POS) of these vernacular names are generally proper noun or common noun when detected by a pre-trained POS tagger. However, due to the peculiarities of these vernacular place names, sometimes they are not detected as proper noun or common noun. To make sure the georeferencing model finds the legitimate place names that end with vernacular names, a lexicon is developed that contains potential vernacular names used in Greater Mumbai. This helps to identify if a word is a potential place name based on the POS tag and spatial rules.

To retrieve the place names, each tweet is first fed into Layer 1, which extracts location mentions using CRF and MaxEnt models. Then, the tweet is fed into Layer 2, which consists of spatial rules. Based on the rules, the place names are further extracted from Layer 2. Then, a duplication check is performed to detect unique place names retrieved from both layers. Following that, the place names are geocoded using OpenStreetMap (OSM) Nominatim service (https://nominatim.openstreetmap.org/, accessed on 7 March 2021).

4. Data Preparation and Experiment

4.1. Data Collection and Preparation

Mumbai suburban railways system consists of mainly four routes, e.g., Western line, Central line, Harbor line, and Trans-Harbor line. The data has been collected using a Twitter Search API from 15 April 2019 to 24 April 2019. The sampling period was chosen randomly over a shorter interval to understand how users react to public transit system. The sampling period includes both weekdays and weekends. The choice of 10 days sampling period was motivated by a previous study by Dube [30] where a manual passenger satisfaction survey was conducted over 10 days in north India. The sampling period also overlapped the time when there was an ongoing metro construction work going on in Mumbai. We wanted to understand if that ongoing metro construction could impact on public transit system in Mumbai and how users express their reactions around that metro construction work. Interestingly, we found users posted a number of tweets around Mumbai metro construction during that time period (Section 4.2.4). Although the sampling is conducted over a shorter duration, the model is scalable enough that it can process even a longer sampling period. To retrieve the relevant tweets, we used a number of keywords pertaining to Mumbai suburban railway system. The keywords (https://github.com/rddspatial/mumbai-railway_sentiment_analysis, accessed on 7 March 2021) are related to different routes and station names. Instead of choosing a particular urban section, we collected data for all the stations, along all the train routes, in Greater Mumbai using keyword search. Collecting data over an urban section would need more complex keyword search and data collection strategy. Through this search process, a total 878 original tweets were collected (after filtering retweets) over 10 days.

Once we collected the data, we recruited volunteers from Figure-Eight Crowdflower (CRW) platform (https://www.figure-eight.com/, accessed on 7 March 2021) to manually annotate the tweets. The annotation was done in the following stages. In the first stage, if the tweet is related to Mumbai railway service, then the tweet is labeled as relevant. In the next stage, the volunteers are asked to label the relevant tweets into three sentiment categories (negative, positive, and neutral) based on the following criteria.

If the given tweet contains information about an issue related to mobility (delay, cancellation, long waiting time, etc.), infrastructure (sitting facility, maneuvering, cleanliness, comfort, service not working), or safety (chance of accident due to negligence, slippery floor, etc.), then it is labeled as negative tweet. On the other hand, if the tweet contains positive information about mobility (e.g., trains are on time), infrastructure (e.g., clean platform, lift or escalator facility), or safety aspects, then that tweet is labeled as positive. Otherwise, if there is no specific sentiment mentioned in the (relevant) tweet, then that tweet is labeled as a neutral tweet.

4.2. Experiment and Results

4.2.1. Tweet Classification

To identify if a tweet is relevant in an automated manner, we used five different machine learning models, i.e., a Naive Bayes (NB), a Support Vector Machine (SVM), a Logistic Regression (LR), a Decision Tree (DT), and a Random Forest (RF). The model is selected based on existing literature [14,51,52]. The models are trained with 70% tweets and tested with remaining 30% tweets. It has been observed that an LR-based model provides maximum recall accuracy and F1-score while detecting relevant tweets with an accuracy of 0.75, whereas an RF yields highest precision accuracy (0.81) (Table 1). In terms of detecting irrelevant tweets, an LR-based model provides maximum precision accuracy and F1-score of 0.81, and an RF-based model provides highest recall accuracy of 0.91 (Table 1).

Table 1. Tweet classification accuracy: Relevant vs. Irrelevant.

	Relevant			Irrelevant		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
LR	0.75	**0.75**	**0.75**	**0.81**	0.82	**0.81**
SVM	0.75	0.68	0.71	0.78	0.83	0.80
RF	**0.81**	0.54	0.65	0.73	**0.91**	**0.81**
DT	0.63	0.66	0.66	0.75	0.70	0.72
NB	0.63	0.73	0.68	0.77	0.68	0.73

To detect sentiment types of the tweets, we used four classifiers, i.e., NB, DT, RF, and an SVM. In the first stage, we developed 3-class classifiers using 383 relevant tweets. The models have been evaluated using a 3-fold cross validation.

In terms of tweet distribution, the majority of tweets are negative (55.88%), followed by neutral (28.46%) and positive tweets (15.66%).

In the next stage, to improve the accuracy of the classifiers, we used a systematic oversampling of minority class (SMOTE) algorithm to make a balanced data set in an iterative way.

In the first step, we counted the number of tweets that fall in positive, negative, and neutral category as C_{pos}, C_{neg}, C_{obj}, respectively. Then, we used a max-min operator to retrieve the maximum (C_i) and minimum count (C_j) of two given categories. Then, the difference (α) between the maximum count and minimum count is computed as follows.

$$\alpha = C_i - C_j. \tag{8}$$

Then, we checked if α is less than or equals to a given threshold (δ). A small δ ensures almost equal number of instances in all the classes. If $\alpha > \delta$, then an oversampling is performed by incrementing the counts of minority class (C_j) by a factor of (α/C_j). Since the counts are integers, we took the smallest nearest integer of (α/C_j) as the sampling factor. The process iterates until all the classes satisfy the condition $\alpha < \delta$ (Figure 2). By doing such an iterative oversampling strategy, it generates a balanced data set. In this research, we assume $\delta = 5$.

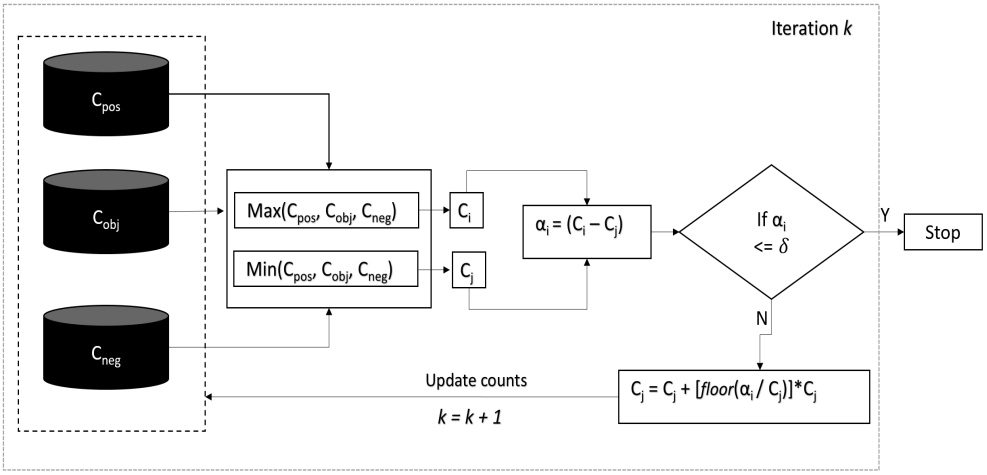

Figure 2. Systematic oversampling of minority class in an iterative manner.

When using a balanced data set, the accuracy has significantly increased. While detecting positive tweets, an RF yields the highest precision accuracy (0.97), and an DT and SVM perform equally well in terms of recall accuracy (0.82) (Table 2). In terms of F1-score while detecting positive tweets, an RF performs best, with 0.88 accuracy. For detecting negative tweets, an RF provides the highest recall accuracy (0.95), whereas an SVM provides 0.77 precision accuracy and 0.80 F1-score.

Table 2. Tweet sentiment classification accuracy: 3-class, balanced data set.

	Positive			**Negative**			**Neutral**		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
SVM	0.91	**0.82**	0.86	**0.77**	0.84	**0.80**	0.74	**0.75**	**0.75**
NB	0.95	0.81	**0.88**	0.66	0.94	0.78	**0.88**	0.64	0.74
DT	0.89	**0.82**	0.86	0.68	0.74	0.71	0.71	0.71	0.71
RF	**0.97**	0.81	**0.88**	0.67	**0.95**	0.79	0.87	0.67	**0.75**

4.2.2. Lexicon-Based Models

In order to understand how well a lexicon-based sentiment analysis could be performed to understand users' satisfaction towards public transit system, we used three different affective lexicons to detect overall sentiment of each relevant tweet by aggregating all the sentiment bearing word tokens in a given tweet. The inter-annotator agreement (Kappa statistic) shows SWN and NRC agree slightly with CRW (Tables 3–5). This can be justified as SWN and NRC are mainly developed from formal English vocabulary which is domain independent and generic in nature resulting similar performance. Since SNS is specifically designed to handle short and informal text, we also tested with SNS with its original vocabulary (SNS1), which shows slight agreement (0.18) with CRW (Table 5). To interpret the significance of agreement, we used the framework of Reference [53]. In all the cases (Tables 3, 4, and 6), most of the negative tweets are detected as neutral. A close observation revealed that lots of (transport) domain-specific sentiment bearing nouns, adjectives, verbs, and adverbs do not bear any sentiment value in these three lexicons. Some of the words are *slippery*, *wastage*, *slowly*, *jam*, *stuck*, and *working*, to name a few. In Greater Mumbai, users complain about lack of cleanliness and bad smell on the platform toilet. Hence, some of the tweets that contain the word *smell*, *derail*, or *toilet* are labeled as negative by the volunteers; however, as those words do not bear any sentiment in the lexicons, those tweets are miss-classified as neutral category owing to high Type II error for negative types. To address this issue, we added some additional transportation-specific sentiment bearing words with sentiment values from -5 to 5 and updated SNS1 to SNS2, and tested on all the relevant tweets. In the second instance, SNS2 shows a performance improvement in detecting positive and negative tweets with a fair agreement (0.31) with CRW (Tables 5 and 7).

Table 3. Confusion matrix: SentiWordNet (SWN) vs. Crowdflower (CRW).

	SWN_Positive	**SWN_Negative**	**SWN_Neutral**
CRW_Positive	**19**	9	32
CRW_Negative	42	**45**	129
CRW_Neutral	17	12	**78**

Table 4. Confusion matrix: NRC vs. CRW.

	NRC_Positive	NRC_Negative	NRC_Neutral
CRW_Positive	**29**	10	21
CRW_Negative	77	**62**	77
CRW_Neutral	38	12	**57**

Table 5. Observed accuracy and kappa statistic.

Lexicon	Observed Accuracy (%)	Kappa	Interpretation
SWN	37.07	0.09	Slight agreement
NRC	38.64	0.12	Slight agreement
SNS1	46.47	0.18	Slight agreement
SNS2	**55.09**	**0.31**	**Fair agreement**

Table 6. Confusion matrix: SentiStrength (SNS)1 vs. CRW.

	SNS1_Positive	SNS1_Negative	SNS1_Neutral
CRW_Positive	**15**	14	31
CRW_Negative	40	**95**	81
CRW_Neutral	26	13	**68**

Table 7. Confusion matrix: SNS2 vs. CRW.

	SNS2_Positive	SNS2_Negative	SNS2_Neutral
CRW_Positive	**42**	6	12
CRW_Negative	38	**109**	69
CRW_Neutral	31	16	**60**

4.2.3. Georeferencing

Using the georeferencing module, a total of 508 locations are retrieved which contain some ambiguous place names, e.g., *platform*, *metro*, and *Mumbai metro*. This can be justified as some of these terms appear after spatial prepositions; these words are picked by the rule-based layer in the georeferencing module leading Type I error. After removing such 127 ambiguous place mentions, we retrieved total 381 reasonable place names, which can be categorized into 43 different place types, e.g., station or platform, city, suburb or administrative, supermarket, bus stop, theatre, hotel, and cafe, to name a few. The place types are defined by OSM place hierarchy with the highest number of location mentions as place type *station*, followed by *suburb*, where the majority of the negative sentiment has been expressed (Figure 3). Some of the top-5 station names mentioned in the relevant tweets are *Andheri*, followed by *Bandra*, *CST*, *Ghatkopar*, and *Borivali*. On the other hand, the top-5 station names mentioned in negative tweets are *Andheri*, *Bandra*, *Borivali*, *CST*, and *Dadar*. Thus, these stations require more attention from the Mumbai railway authority to improve the service quality.

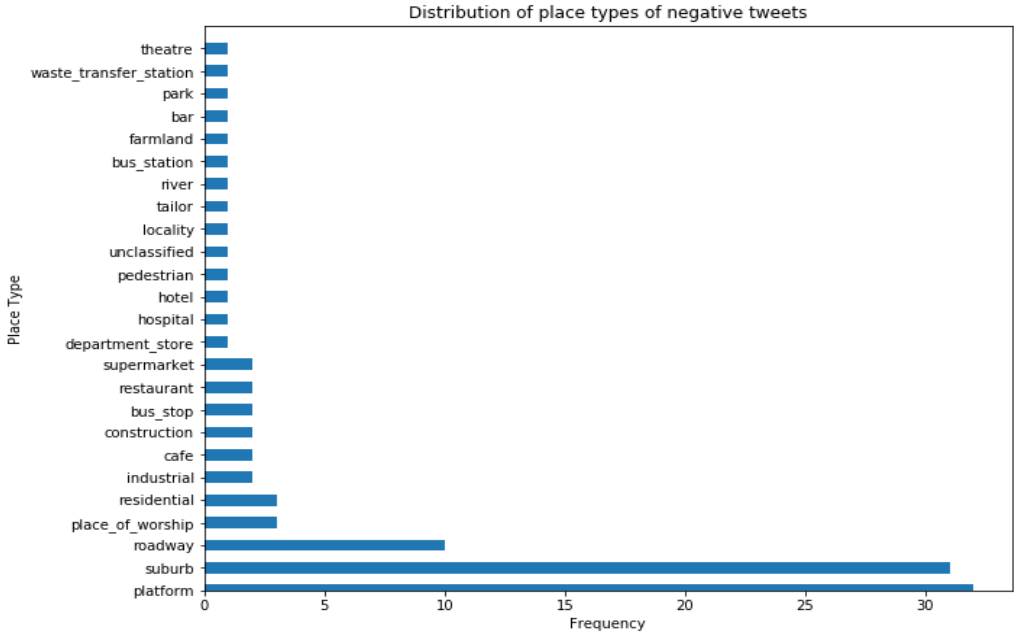

Figure 3. Distribution of negative tweets over different place types.

4.2.4. Exploratory Analysis

In terms of temporal pattern, people mostly express negative sentiments or dissatisfaction during peak hours in the morning at 10 a.m., and in the evening at 4 p.m., and again in the night at 9 p.m. (Figure 4). Although positive tweets peak at 2 p.m., however, there are multiple short peaks, which indicates there is no discernible pattern for expressing positive sentiments. It should be also noted that, similar to the previous studies [24], in this study, we also observe people generally post more negative tweets to express their dissatisfaction compared to their positive counterpart. Figure 5 shows the location mentions in relevant tweets. Figure 6 shows the number of negative tweets at different locations. Users express dissatisfaction mainly at *Andheri*, *Bandra*, *Borivali*, and *Mira-Bhayandar* along the Western railway line. On the Central line, users mainly express their dissatisfaction at *Thane* and *Mulund*. However, the maximum dissatisfaction can be viewed at *Andheri*, *Bandra*, and *Bhayandar*.

We observed 25% of the positive tweets are related to ongoing metro construction in Mumbai (Figure 7a). People expressed their satisfaction with the progress of the work as it is deemed to supplement Mumbai suburban railway service and reduce traffic congestion during peak hours. Some of the positive tweets are as follows.

- *Tweet 4:* #Mumbai's new Metro lines offer an opportunity to create a modern public transport system with innovative technological features @MumbaiMetro3 @MumMetro #urbantransport @ADBTranport
- *Tweet 5:* Platform widening work at #Andheri Metro station inches towards completion ensuring more space for commuters in the coming months. #MumbaiMetro @Lokhandwala_Bom

On the other hand, the majority of the negative tweets are related to railway service in the Western and Central line. When expressing negative sentiments, users often mention the transport authority (e.g., piyushgoyaloff, drmbct, GMNWRailway) to bring the issue into their notice (Figure 7b). Most of the negative tweets contain information about infrastructure and mobility issues. For example, users often express their dissatisfaction regarding the quality of drinking water on the platform and malfunctioning of the fans in the train or indicators on the platform. Users also express their concerns for lack of sitting space on the platform. Users tweet about delay and cancellation of trips. Some of the negative tweets are as follows.

- *Tweet 6:* @GMNWRailway @WesternRly @Srdcmbct12 @srdcmmb This stall on platform No2/3 Goregaon station does not have Municit Water Connection. They are using filthy tanker water for preparation and serve the same water to passengers to drink. For eye wash they have kept a water filter.
- *Tweet 7:* Sad day for Nature Lovers Have many sweet memories of Aarey Colony Please try to save whatever possible of this Green cover. Tagging...

Most of the time tweets are self-explanatory. But, sometimes, some tweets may need contextual or background information to interpret their meaning. For example, Tweet 6 directly reflects some issue with the water connection on the station. However, in hindsight, Tweet 7 may not seem to be relevant to public transit system in Mumbai, but it is actually a relevant one. This tweet (Tweet 7) is posted in the context of constructing a parking space by cutting trees in Aarey Colony. With the progress of the construction of Colaba-Bandra-Seepz metro line, the authority started constructing a large carshed in the green hub of Aarey colony by cutting down a large number of trees. Although the local people and activists pleaded to stop the construction of the carshed by cutting down the trees, the Supreme Court rejected their plea (https://timesofindia.indiatimes.com/city/mumbai/sc-rejects-plea-by-activist-for-alternative-to-aarey-in-mumbai-for-metro-car-shed/articleshow/68887845.cms (accessed on 7 March 2021)). Many users expressed negative sentiments against that decision.

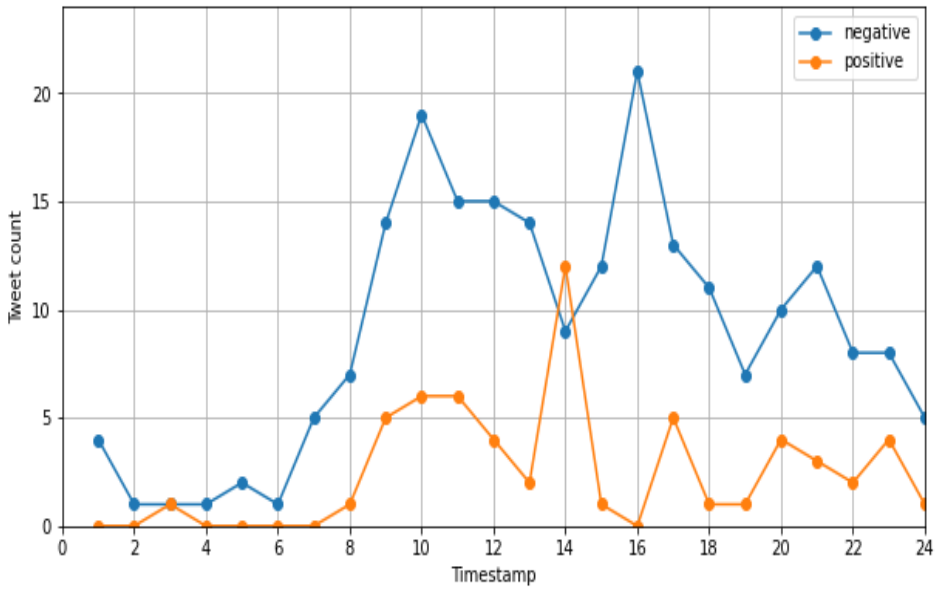

Figure 4. Tweet distribution over 24-h time period.

Tweets are either detected as positive or negative. However, in many cases, tweets contain mixed sentiments. For example, in the following tweet (Tweet 8), the first part contains positive sentiment, whereas the subsequent part contains negative sentiment of the user. Understanding the inherent sentiment from such mixed connotation is a challenge, and this requires a quantitative measure of sentiment expressed in a tweet, instead of detecting a discrete class label (positive/negative).

- *Tweet 8:* @RailMinIndia @WesternRly @PiyushGoyal Thanks and Congrats for building new bridge at Naigaon stn. But slipery floor tiles are dangerous in rain.Pls don't use shining, marble tiles on platform. Once I fall while running to catch train and my leg came between platform and train.

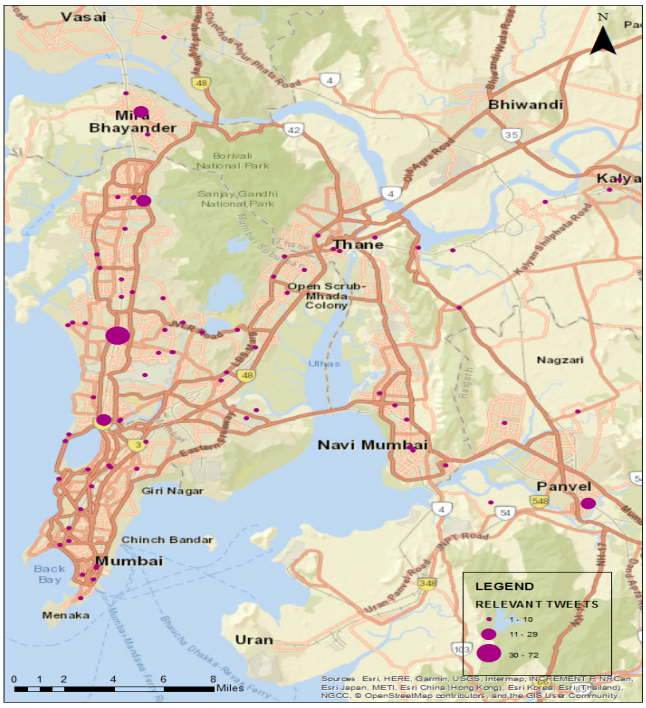

Figure 5. Locations of relevant tweet intensity.

Figure 6. Map showing the distribution of negative tweets at different locations. The size of the blob is directly proportional to the number of negative tweets at a given location.

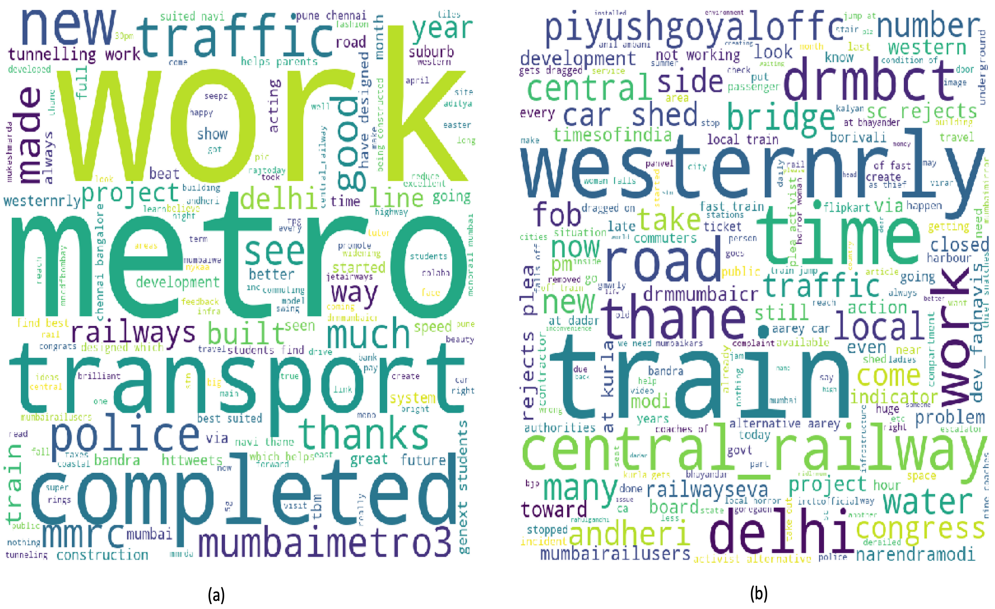

Figure 7. (**a**) Positive word cloud. (**b**) Negative word cloud.

To understand which words co-occur more often than random chance, we computed an n-gram (n = 2) collocation measure using likelihood ratio on all the relevant tweets. We found the following top-15 word pairs that co-occur more frequently represent the context in the corpus: [(mumbai, metro), (nine, coaches), (tunnelling, work), (anil, ambani), (home, tutors), (rejects, plea), (railway, station), (fast, train), (car, shed), (train, jump), (bhayander, station), (central, line), (platform, number), (jump, platform), (tbm, manages)].

The first and last word pairs, (mumbai, metro), (tbm, manages), respectively, relate to the ongoing metro rail project in Mumbai. Word pairs, such as (nine, coaches), (fast, train), (bhayander, station), relate to an incident where nine coaches of a fast train derailed at Bhayander Station on 20 April 2019 in Mumbai (https://www.india.com/news/india/mumbai-9-coaches-of-fast-train-overshoots-platform-at-bhayander-station-3636387/, accessed on 7 March 2021). Other co-located words, for example, (railway, station), (central, line), co-occur more often than their individual counter-parts. Word pairs, such as [(rejects, plea), (car, shed)], relate to the legal verdict of carshed in Aarey colony.

Likewise, when measuring the trigram collocation (n = 3) likelihood, we got the following top-5 trigrams, e.g., [(metro, car, shed), (car, shed, controversy), (railway, station, platform), (thief, snatches, roz), (fast, train, jump)].

The trigrams reflect the contexts even at a finer granularity. The trigrams relate to the setting up the new carshed in the light of ongoing metro project, miscellaneous events on railway platform, stealing personal properties on the train, and derailment of a train at Bhayander station.

Other approaches, for example, topic modeling, can be explored to understand the inherent topics in the corpus of the relevant tweets. However, as we used a comparatively smaller data set with a limited number of word distribution, a collocation measure explores the intricacies in the corpus. Through this analysis, we demonstrated UGC can be useful to understand users' perceptions towards public transit system in Greater Mumbai. This will provide an insight about the issues pertaining to the service quality and major locations that need further attention from the authority.

5. Discussion

In most of the developing cities, there is a gap between users' perceived quality of mobility service and service providers' perception towards the mobility service. The level of service (LOS) can be viewed either from a public transit operator or from a user's

perspective. While users create the demand, operators provide the supply. This research bridges the gap between such supply and demand by analyzing how users perceive the LOS provided by the railway operators in Greater Mumbai. It is often difficult for the authority to realize users' needs and deterioration of certain service criteria at a given place. This knowledge gap leads to quality degradation and loss of patronage. To bridge this gap, in this research, we explored if Twitter, a social-media platform, can be used to complement current (manual) survey practice for understanding users' satisfaction and performance measures. Our study shows Twitter can be used to understand users' satisfaction towards public transit system. In this research, we developed a framework that consists of a number of modules. The first module collects Twitter data followed by detecting relevant tweets and sentiment analysis. Most of the tweets are ungeotagged [10] without any explicit location information, which requires a georeferencing module to extract location information. So, we used all ungeotagged tweets in our study. We observed an RF classifier outperforms other models in terms of detecting negative tweets (0.95 recall) (Table 2). When tested on knowledge driven models, among three lexicon-based models, SentiStrength (SNS2) works best to detect (negative) sentiments from the tweets with Kappa = 0.31 (Table 5).

In this paper, we developed a novel approach to retrieve transport service quality information using Twitter-based user-generated content. The presented model can complement existing manual travel survey process in a more adaptive and scalable way. Since the proposed approach does not need a field staff or a surveyor, it can be conducted any time, on streaming data or on historical data. This approach can be used over a multiple cities simultaneously or at different temporal interval or to understand the impact of any policy change on transport service quality.

In 2018, Ola Mobility Institute conducted a survey on 43,000 participants across 20 different cities in India. According to that survey, half of the commuters in Mumbai use public transport system to save time or money. However, when measured based on Ease of Moving Index, given a number of other commuting options, only 12% of the commuters preferred public transport system in the first place [54]. In 2012, Dube conducted a manual passenger satisfaction survey on Indian railway in northern part of India and showed users often express their dissatisfaction towards cleanliness in toilets and platforms, delay of train service, and unauthorized vendors on the train [30]. Our findings align with their work especially on cleanliness and delay. This suggests these problems are still prevalent and common in many different Indian cities. This requires an attention from the authority to address these issues. We found people generally tweet more negative sentiments than positive ones, which also conforms earlier study by Reference [24]. This indicates that Twitter can be used as a source of understanding users' (dis)satisfaction. Figure 4 shows users mostly express their negative sentiments at the peak hours in the morning and evening. In this research, we assume a tweet which contains any complaint with some words related to malfunctioning of railway infrastructure or mobility service expresses user's dissatisfaction, thus carrying negative connotation. For example, *wasting water*, *smell from the toilet*, *lack of sitting facility*, and *malfunctioning of the fans* has been used by the users to indicate their lack of satisfaction with the railway service. Our study shows that users generally provide spatial context while mentioning any issue. They also mention railway ministers, or other key administrative personnel, in their tweets to bring the problem into their notice. The georeferencing module shows users are concerned with various infrastructure and mobility related issues at *Andheri*, *Bandra*, *Borivali*, *CST*, and *Dadar*.

In this paper, it is assumed that, if a tweet contains a negative sentiment related to transport service, then that is of interest to the authority, irrespective of the user characteristics, as it indicates a quality issue in the transport service. Although we did not study user bias in this research, with more data, the model can be strengthen to infer users' reaction in a more comprehensive way. Understanding the bias may infer the effect of different socio-demographic and economic segment on public transport usage and their concerns towards

various service aspects. The bias may also refer to over-estimation or under-estimation of satisfaction level about certain service criteria by a specific socio-demographic group.

This research conforms to some of the earlier findings of References [24,41]. For example, people generally share more negative tweets than positive tweets related to mobility services. People are generally active in tweeting during the peak hours. However, previous works [24] used only a lexicon-based approach without any validation. In our research, we compared the accuracy of different lexicon-based and supervised machine learning models, while detecting users' (dis)satisfaction. Previous works [3,24,41] did not use ungeotagged tweet; thus, they did not address the spatial context of negative sentiments. In this paper, we developed a novel georeferencing module that can retrieve locations from informal tweets in Greater Mumbai and can geocode them on a map which can further help in informed decision-making and policy implementation.

6. Conclusions

In this paper, we developed a framework that can understand users' (dis)satisfaction towards public transit system, in particular Mumbai railway service from ungeotagged tweets. Understanding users' (dis)satisfaction can help public transit authority to prioritize different service criteria or locations that require immediate improvements and thus increase patronage on public transport modes. Currently, users' satisfaction information is collected through manual travel survey process or through indirect means, which often involves high investment, longer gestation periods, and quality issues. In this research, we developed a novel framework that can leverage UGC harvested from online social-media platform(s) and can infer users' (dis)satisfaction through sentiment analysis.

Since most of the tweets do not have any explicit geolocation, we explored the potential of ungeotagged tweets to understand users' satisfaction. We compared the performance of supervised machine learning models to understand users' sentiments. If the aim is to detect all the negative tweets to retrieve transport service related issues as much as possible, an RF-based model should be used, which provides 0.95 recall (Table 2).

In terms of knowledge-driven techniques, an updated SentiStrength-based lexicon performs best compared to other two lexicons (SentiWordNet, EmoLex). Due to inherent nature of learning the pattern from the data, supervised models work better than lexicon-based models. However, when there is scarcity of annotated data, a lexicon-based approach can be used. Other advanced machine learning techniques based on word-embeddings, for example, Transformer, can also be evaluated in future to understand people's perception towards transportation service given the larger data set. Since the tweets that we used in this study are ungeotagged, we developed a novel georeferencing model that can retrieve location mentions in the text which can be used as a spatial cue for the negative sentiments towards railway service in Mumbai.

Based on the research, we present some key points, which are critical to the public transport system, particularly the railway service in Greater Mumbai.

- Ungeotagged tweets can be used to understand people's perception about railway service quality. A negative tweet indicates the quality issue with the transport service.
- An RF-based model outperforms other models to detect negative sentiments in the transportation related tweets with 0.95 recall accuracy.
- People tend to tweet more negative sentiments compared to positive ones and express their (dis)satisfaction towards railway service in Greater Mumbai. That said, negative sentiments are more critical to understand transport service quality.
- We observed that people often express their dissatisfaction in *Andheri*, *Bandra*, *Borivali*, and *Mira-Bhayandar* railway station along Western railway line. On the other hand, most of the dissatisfied tweets are reported in *Thane* and *Mulund* railway station along the Central line.
- Most of the complaints are related to infrastructure and mobility issues in Greater Mumbai.
- We used comparatively a smaller data set. Based on our findings, most of the transport related tweets are reported at 10 a.m. and at 4 p.m. in Greater Mumbai.

- Although, in this research, the user bias is not investigated, further research should be undertaken to study what kind of user bias exists towards understanding public transport quality in Greater Mumbai. It is also important to understand how the bias affects overall findings at different granularity, and in terms of different socio-demographic aspects.

6.1. Limitations and Future Outlook

Despite the model proving its efficacy in understanding users' (dis)satisfaction towards Mumbai railway system, there are some limitations in this research. First of all, the data set used in this research is collected over a shorter duration. A future study should look into a longer duration, including seasonal variation and impact of various events (e.g., political rally, accidents, change in rail fare, change in petrol price, etc.) on the usage of public transit system. The data set may have user bias as most of the social-media users are young generation with an uneven gender distribution. A further analysis can be carried out in that direction.

Although, in this study, we detected users' sentiments and the location associated with it, there are a number of other service quality parameter to be considered [3,26]. One of the limitations of the proposed approach is that it is difficult to understand if user is rider or non-rider in contrast to manual survey approach. This could potentially create a user bias in the data. To distinguish a user as a rider or non-rider, future research can extract user mobility patterns from the tweets.

From the perspective of text analysis, a classical BoW model is used without considering any context and order of the words. This may affect the classification performance. A future study should compare a BoW model with more sophisticated word embedding models (e.g., Word2Vec or Glove) that capture context and word semantics. Sometimes, users can also tweet in a sarcastic manner, for example, *The service is "great"!*, which could actually mean a negative sentiment. Users can also attach emojis while tweeting to express their feelings. In this study, we did not perform any sarcasm detection or emoji analysis. This may introduce some bias in the prediction. A follow up study should address these limitations.

6.2. Recommendations

The model presented in this research can complement existing manual survey approach. Using the proposed model the users' perceived quality information can help the authority to better manage their infrastructure and supply. The majority of the negative tweets are associated to Western and Central railway lines, especially at *Andheri*, *Bandra*, and *Bhayandar*. Authority should pay more attention to improve timeliness, cleanliness, and safety aspects at these locations.

Based on the frequency of (dis)satisfaction level at a given location, alternative services, e.g., para-transit (auto-ricksaw) or shared-ride, can also be deployed as a gap filler, especially in the locations where there is a frequent delay of connecting services. Based on the frequency of safety-related tweets, police and safety departments can also take proper measures. The model developed in this research is scalable and adaptive to similar UGC, e.g., Facebook posts. This will also support urban planners and other public departments in various policy implementation related to revising the transport fare, allocating budget for new railway services, and estimating travel demand, to name a few.

Although the model has been developed and tested for railway service in Greater Mumbai, the same approach can be used for other transport mode(s) in other cities. Thus, in this research, we demonstrated that Twitter can provide fine grained information on users' travel experience in a more adaptive manner. The georeferencing module further demonstrates the potential of ungeotagged tweets for understanding mobility issues at various locations. The model presented in this paper can complement existing transport infrastructure to bridge the gap between supply and demand. The insights retrieved by the model can improve the existing service quality.

Funding: This research was funded by Swiss National Science Foundation (SNSF) grant number 166788.

Acknowledgments: The author would like to thank Ross Purves from University of Zurich, for his support and valuable feedback in this research. The author would like to appreciate the support extended by Ankita Biswas for helping with the map creation and visualizations. The author acknowledges the Swiss National Science Foundation (SNSF) grant number 166788 for supporting this work. The author also appreciates the anonymous reviewers for their useful feedback to improve the quality of this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pojani, D.; Stead, D. Sustainable Urban Transport in the Developing World: Beyond Megacities. *Sustainability* **2015**, *7*, 7784–7805. [CrossRef]
2. Pucher, J.; Peng, Z.; Mittal, N.; Zhu, Y.; Korattyswaroopam, N. Urban Transport Trends and Policies in China and India: Impacts of Rapid Economic Growth. *Transp. Rev.* **2007**, *27*, 379–410. [CrossRef]
3. Ngoc, A.M.; Hung, K.V.; Tuan, V.A. Towards the Development of Quality Standards for Public Transport Service in Developing Countries: Analysis of Public Transport Users' Behavior. *Transp. Res. Procedia* **2017**, *25*, 4560–4579. [CrossRef]
4. Guirao, B.; García-Pastor, A.; López-Lambas, M.E. The importance of service quality attributes in public transportation: Narrowing the gap between scientific research and practitioners' needs. *Transp. Policy* **2016**, *49*, 68–77. [CrossRef]
5. Andreassen, T.W. (Dis)satisfaction with public services: The case of public transportation. *J. Serv. Mark.* **1995**, *9*, 30–41. [CrossRef]
6. Sammer, G.; Gruber, C.; Roeschel, G.; Tomschy, R.; Herry, M. The dilemma of systematic underreporting of travel behavior when conducting travel diary surveys—A meta-analysis and methodological considerations to solve the problem. Transport Survey Methods in the era of big data:facing the challenges. *Transp. Res. Procedia* **2018**, *32*, 649–658. [CrossRef]
7. Krumm, J.; Davies, N.; Narayanaswami, C. User-Generated Content. *IEEE Pervasive Comput.* **2008**, *7*, 10–11. [CrossRef]
8. Pragati. Social Media Statistics in India. 2019. Available online: https://www.talkwalker.com/blog/social-media-statistics-in-india (accessed on 10 June 2019).
9. Gonsalves, D. The Relevance of Mumbai Millennials Socioeconomic Background on Their Purchase Behavior of Lifestyle and Luxury Apparels and Accessories. *IOSR J. Bus. Manag.* **2018**, *20*, 65–79.
10. Gu, Y.; Qian, Z.; Chen, F. From Twitter to detector: Real-time traffic incident detection using social media data. *Transp. Res. Part C Emerg. Technol.* **2016**, *67*, 321–342. [CrossRef]
11. Jungherr, A.; Schoen, H.; Jürgens, P. The Mediation of Politics through Twitter: An Analysis of Messages posted during the Campaign for the German Federal Election 2013. *J. Comput. Mediat. Commun.* **2016**, *21*, 50–68. [CrossRef]
12. Sakaki, T.; Okazaki, M.; Matsuo, Y. Earthquake shakes Twitter users: Real-time event detection by social sensors. In Proceedings of the 19th International Conference on World Wide Web, Raleigh, NC, USA, 26–30 April 2010; pp. 851–860. [CrossRef]
13. Wang, D.; Al-Rubaie, A.; Clarke, S.S.; Davies, J. Real-Time Traffic Event Detection From Social Media. *ACM Trans. Internet Technol.* **2017**, *18*, 1–23. [CrossRef]
14. Das, R.D.; Purves, R.S. Exploring the Potential of Twitter to Understand Traffic Events and Their Locations in Greater Mumbai, India. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 5213–5222. [CrossRef]
15. Castillo, M.; Bhattacharya, S. Public Transport Subsidies and Affordability in Mumbai, India. *Urban Stud. Res.* **2012**, *865972*. [CrossRef]
16. Vaas, M. A public Transport System on a Knife-Edge: Mumbai. 2012. Available online: https://www.alphabet.com/en-ww/article/public-transport-system-knife-edge-mumbai (accessed on 25 March 2020).
17. Soans, I. All You Need to Know about Mumbai's Newly Launched Metro. **2014**. Available online: https://www.firstpost.com/india/all-you-need-to-know-about-mumbais-newly-launched-metro-1560805.html (accessed on 25 March 2020).
18. Rajesh, M. Rush Hour on the World's Busiest Railway. **2015**. Available online: https://www.shorturl.at/fK139 (accessed on 20 March 2020).
19. Ostermann, F.O.; Tomko, M.; Purves, R.S. User evaluation of automatically generated keywords and toponyms for geo-referenced images. *J. Assoc. Inf. Sci. Technol.* **2013**, *64*, 480–499. [CrossRef]
20. Agarwal, S.; Mullick, A.; Ray, G.G. An Observational Study on Usability Issues in Mumbai Local Trains. *Hum. Factors Ergon. Soc.* **2013**, *57*, 531–535. [CrossRef]
21. Cheng, Z.; Jian, S.; Maghrebi, M.; Rashidi, T.H.; Waller, S.T. Is Social Media an Appropriate Data Source to Improve Travel Demand Estimation Models? In Proceedings of the Transportation Research Board 97th Annual Meeting, Washington, DC, USA, 7–11 January 2018.
22. Hollenstein, L.; Purves, R.S. Exploring place through user-generated content: Using Flickr tags to describe city cores. *J. Spat. Inf. Sci.* **2010**, *1*, 21–48.
23. Pereira, F.C.; Rodrigues, F.; Ben-Akiva, M. Text analysis in incident duration prediction. *Transp. Res. Part C Emerg. Technol.* **2013**, *37*, 177–192. [CrossRef]

24. Collins, C.; Hasan, S.; Ukkusuri, S. A Novel Transit Riders' Satisfaction Metric: Riders' Sentiments Measured from Online Social Media Data. *J. Public Transp.* **2013**, *16*, 21–45. [CrossRef]
25. Pucher, J.; Korattyswaroopam, N.; Ittyerah, N. The Crisis of Public Transport in India: Overwhelming Needs but Limited Resources. *J. Public Transp.* **2004**, *7*, 1. [CrossRef]
26. Castillo, J.M.d.; Benitez, F.G.; Descubrimientos, C.d.l. Determining a public transport satisfaction index from user surveys. *Transp. A Transp. Sci.* **2013**, *9*, 713–741. [CrossRef]
27. Fellesson, M.; Friman, M. Perceived satisfaction with public transport service in nine European cities. *J. Transp. Res. Forum* **2008**, *47*, 93–103. [CrossRef]
28. Vuchic, V.R. *Urban Transit: Operations, Planning, and Economics*; Wiley: Hoboken, NJ, USA, 2005.
29. Eboli, L.; Mazzulla, G. Performance indicators for an objective measure of public transport service quality. *Eur. Transp. Eur.* **2011**, *51*, 1–21.
30. Dube, K. *Passenger Satisfaction Survey Report and Benchmarking of Performance Standards*; Report; Indian Railways Institute of Transport Management Lucknow: Uttar Pradesh, India, 2012.
31. TRB. *Transit Capacity and Quality of Service Manual, TRCP Report 100*, 2nd ed.; Transportation Research Board: Washington, DC, USA, 2004.
32. Clarke, M.; Dix, M.; Jones, P. Error and uncertainty in travel surveys. *Transportation* **1981**, *10*, 105–126. [CrossRef]
33. Liu, B. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotion*; Cambridge University Press: Cambridge, UK, 2015.
34. Liu, X. Target and position article - Analyzing the impact of user-generated content on B2B Firms' stock performance: Big data analysis with machine learning methods. *Ind. Mark. Manag.* **2019**, *86*, 30–39. [CrossRef]
35. Turney, P.D. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 417–424. [CrossRef]
36. Thomas, M.; Pang, B.; Lee, L. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia, 22–23 July 2006; pp. 327–335.
37. Jindal, N.; Liu, B. Opinion spam and analysis. In Proceedings of the 2008 International Conference on Web Search and Data Mining, Palo Alto, CA, USA, 11–12 February 2008; pp. 219–230. [CrossRef]
38. Baccianella, S.; Esuli, A.; Sebastiani, F. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, 17–23 May 2010.
39. Thelwall, M.; Buckley, K.; Paltoglou, G. Sentiment strength detection for the social web. *J. Am. Soc. Inf. Sci. Technol.* **2011**, *63*, 163–173. [CrossRef]
40. Goncalves, P.; Araujo, M.; Benevenuto, F.; Cha, M. Comparing and combining sentiment analysis methods. In Proceedings of the First ACM Conference on Online Social Networks, Boston, MA, USA, 7–8 Ocotber 2013; pp. 27–38. [CrossRef]
41. Limsopatham, N.; Albakour, M.D.; Macdonald, C.; Ounis, I. Tweeting Behaviour during Train Disruptions within a City. In *Digital Placemaking: Augmenting Physical Places with Contextual Social Data*; ICWSM Workshop: Atlanta, GA, YSA, 2015.
42. Congosto, M.; Fuentes-Lorenzo, D.; Sánchez, L. Microbloggers as Sensors for Public Transport Breakdowns. *IEEE Internet Comput.* **2015**, *19*, 18–25. [CrossRef]
43. Anastasia, S.; Budi, I. Twitter Sentiment Analysis of Online Transportation Service Providers. In Proceedings of the 8th International Conference of Advanced Computer Science and Information Systems, Malang, East Java, 15–16 October 2016.
44. Jurdak, R.; Zhao, K.; Liu, J.; AbouJaoude, M.; Cameron, M.; Newth, D. Understanding human mobility from Twitter. *PLoS ONE* **2015**, *10*, e0131469. [CrossRef] [PubMed]
45. Zornoza Gallego, C.; Salom Carrasco, J. Geolocalized Tweets for assessing daily mobility: Methodology to analyse and detect home location in the urban area of Valencia. *Boletín De La Asoc. De Geógrafos Españoles* **2018**, *79*. [CrossRef]
46. Cheng, Z.; Caverlee, J.; Lee, K. You are where you tweet: A content-based approach to geo-locating twitter users. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management, Galway, Ireland, 19–23 October 2010; pp. 759–768. [CrossRef]
47. Gelernter, J.; Balaji, S. An algorithm for local geoparsing of microtext. *GeoInformatica* **2013**, *17*, 635–667. [CrossRef]
48. Mohammad, S.; Kiritchenko, S.; Zhu, X. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. In *Second Joint Conference on Lexical and Computational Semantics, Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*; Association for Computational Linguistics: Atlanta, GA, USA, 2013; pp. 321–327.
49. Miller, G.A. WordNet: A Lexical Database for English. *Commun. ACM* **1995**, *38*, 39–41. [CrossRef]
50. Sang, E.F.T.K.; Meulder, F.D. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, Edmonton, AL, Canada, 31 May–1 June 2003; pp. 142–147. [CrossRef]
51. D'Andrea, E.; Ducange, P.; Lazzerini, B.; Marcelloni, F. Real-time detection of traffic from Twitter stream analysis. *IEEE Trans. Intell. Transp. Syst.* **2015**, *16*, 2269–2283. [CrossRef]
52. Klaithin, S.; Haruechaiyasak, C. Traffic information extraction and classification from Thai Twitter. In Proceedings of the 13th Intern3ational Joint Conference on Computer Science Software Engineering (JCSSE), Khon Kaen, Thailand, 13–15 July 2016.

53. Landis, J.R.; Koch, G.G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **1977**, *33*, 159–174. [CrossRef] [PubMed]
54. Report. *Ease of Moving Index;* Ola Mobility Institute: India, 2018. Available online: https://olawebcdn.com/ola-institute/ease-of-moving.pdf (accessed on 7 March 2021).