JOURNAL OF SPATIAL INFORMATION SCIENCE
Number 20 (2020), pp. xx–yy

doi:10.5311/JOSIS.2020.20.603

RESEARCH ARTICLE

Service quality monitoring in confined spaces through mining Twitter data

Mohammad Masoud Rahimi[1], Elham Naghizade[2], Mark Stevenson[3], and Stephan Winter[1]

[1]Department of Infrastructure Engineering, The University of Melbourne, Australia
[2]Centre for Information Discovery and Data Analytics (CIDDA), RMIT University, Australia
[3]Transport, Health, and Urban Design Research Lab, The University of Melbourne, Australia

Received: November 30, 2019; returned: February 6, 2020; revised: March 9, 2020; accepted: May 4, 2020.

Abstract:
Promoting public transport depends on adapting effective tools for concurrent monitoring of perceived service quality. Social media feeds, in general, provide an opportunity to ubiquitously look for service quality events, but when applied to confined geographic area such as a transport node, the sparsity of concurrent social media data leads to two major challenges. Both the limited number of social media messages—leading to biased machine-learning—and the capturing of bursty events in the study period considerably reduce the effectiveness of general event detection methods. In contrast to previous work and to face these challenges, this paper presents a hybrid solution based on a novel fine-tuned BERT language model and aspect-based sentiment analysis. BERT enables extracting aspects from a limited context, where traditional methods such as topic modeling and word embedding fail. Moreover, leveraging aspect-based sentiment analysis improves the sensitivity of event detection. Finally, the efficacy of event detection is further improved by proposing a statistical approach to combine frequency-based and sentiment-based solutions. Experiments on a real-world case study demonstrate that the proposed solution improves the effectiveness of event detection compared to state-of-the-art approaches.

Keywords: service quality, public transport, event detection, fine-tuned BERT, aspect-based sentiment analysis, statistical analysis

1 Introduction

The rigid monitoring of Service Quality (SQ) of public transport services is a key challenge of the service operators, and yet a pre-requisite for the responsive management of their transport services, including the management of the confined and often crowded spaces of public transport nodes. Responsive management can contribute to public safety and to sustainable demand for public transport services. Sustainable use of public transport, in turn, can diminish the negative impact of private motorized mobility in urban space, such as congestion, parking pressure, noise pollution, fossil fuel depletion, and air pollution—with impact on climate and health [12]. Moreover, promoting SQ in public transport can lead to customer loyalty, passenger retention and service recommendation [9].

A transport node, e.g., a bus/tram stop, station, terminal or multi-modal transport hub, is one of the main components of a public transport system and can be defined as any confined geographic area in which a public transport transfer activity takes place. On a regular day, a transport node can be in charge of transferring a large number of people. Therefore, continuous assessment of perceived SQ inside a transport node is a critical task that promotes and increases the usage of public transport [14].

This task requires the development of an effective and efficient tool for measuring and monitoring the perceived SQ. Concurrent monitoring of different SQ aspects and characteristics should especially focus on the detection of unpredictable events impacting SQ, since constant drains on SQ or regular events impacting SQ are generally known. The tool will enable transport node operators to provide prompt responses to customers' expectations.

Conducting surveys is the typical approach for monitoring perceived SQ. This approach has been investigated extensively during the last decades [11, 12, 14, 15, 36]. Surveys, however, are limited to a predefined study period. This limitation makes them unable to constantly observe and understand the passengers' responses to the surrounding environment. Consequently, surveys may miss shorter, longer or frequent events affecting public transport SQ, e.g., delays, termination and overcrowding of services.

The emergence of social media allows service customers to report events and express their opinions about them, including the events which happen inside transport nodes. As a result, to mitigate the limitation of traditional approaches, event detection on social media feeds can be leveraged for monitoring perceived SQ inside transport nodes.

Recently, several studies [20, 21] have been conducted on developing various approaches for event detection from social media feeds. Most of the existing methods consist of two consecutive tasks: First, topic modeling, clustering or classification approaches are leveraged for grouping semantically-related tweets [20]. This paper calls this task *aspect extraction*. Second, machine-learning approaches are employed to detect uncommon high burstiness scores in the study period. Here, the burstiness score indicates the relative frequency of a group of semantically-related tweets appearing in a certain time window, as compared with their average frequencies over all time windows [22]. Thus, burstiness detection is an approach to capture occasions where a topic is mentioned more frequently than its average occurrences during a study period. In this paper, we call this task *detecting events of interest*. However, in a confined geographic area such as a transport node concurrent social media data is inevitably sparse. Over this confined spatial context, general event detection approaches would face two major challenges:

- *Low Variability*: Low numbers of tweets in a dataset can lead to a less frequent occurrence of features in machine-learning approaches. This limitation leads to lower vari-

ability in the dataset, while, based on the bias-variance trade-off philosophy, decreasing the variability will increase the bias of a trained machine-learning-based model. Therefore, typical machine-learning approaches for event detection do not achieve their highest potential as they face the problem of underfitting in the training phase.
- *Scarce Burstiness*: As the number of observations in a sparse dataset is limited, burstiness detection is a challenging task. This issue has been discussed in the literature [54, 63, 64], where general approaches cannot effectively catch events. There have been few attempts for strengthening these approaches by concurrent analysis of spatial characteristics of tweets along with burstiness detection. Nonetheless, these attempts are mainly on the basis of analysing geo-tagged tweets, while precise spatial information is no longer provided by Twitter. Moreover, as only 1% to 2% of the Twitter stream is geo-tagged [18] anyway, using merely geo-tagged data for event detection can intensify the sparseness in the dataset significantly.

To face those challenges, this paper leverages language modeling, sentiment analysis and a hybrid solution for detecting events of interest. Specifically, first a language representation model is utilized to perform the aspect extraction task, i.e., grouping semantically-related feeds. Such models, which are pre-trained on a large corpus of data such as all Wikipedia[1] articles, can be fine-tuned for various downstream applications including event detection in the context of public transport. Therefore, they can bring extra semantic features into the process of text classification which increases the variability and reduce the bias in the classifier. As a language model, a state-of-the-art deep language representation model, namely Bidirectional Encoder Representations from Transformers (BERT) [13] is employed. BERT is leveraged since it applies bidirectional training of a transformer, where this ability provides a deeper understanding of the language context and improves the learning capacity. This capability is especially important for our case, where there is a limited number of tweets, and learning from this limited context can lead to the low variability challenge. Therefore, in this paper, a pre-trained BERT model is fine-tuned by a range of transport-related Twitter feeds extracted from a transport hub in order to transform tweets into vector features. Next, these features are used for multi-label text classification. Finally, the effectiveness of the classification is investigated using tweets extracted from another hub.

Furthermore, the second task, i.e., detecting events of interest and the analysis of passengers' sentiment information as another class of observations, is utilized to face the challenge of scarce burstiness. Sentiment information brings external knowledge into the problem of event detection, which can be derived by leveraging domain-specific and general lexicons and assigning a prior sentiment score to each word in tweets. This paper proposes that sentiment information can enrich sparse datasets, thus improving the sensitivity of detection as an intuitive feature. For extracting passengers' sentiments, Aspect-Based Sentiment Analysis (ABSA) [40], a novel computational approach to understand the perception of a user about specific entities and their corresponding aspects, is employed. This approach is chosen as it provides the opportunity to evaluate customers' perception regarding different aspects of public transport SQ, which leads to a more granular understanding of their perceived service quality. Here, ABSA is used to estimate daily sentiment scores and classify them into different aspects of SQ. To know these aspects, this paper follows the classification of Eboli and Mazzulla [15] by categorising SQ aspects into seven discrete

[1]http://www.wikipedia.org

classes, namely “Safety”, “View”, “Information”, “Service Reliability”, “Comfort”, “Personnel”, and “Additional Services”.

Finally, this paper proposes a novel Frequency-based and Sentiment-based Event Detection (FSED) approach. Using a statistical approach for this hybrid solution, FSED can combine and integrate event detection solutions based on frequency and sentiment information in order to overcome low variability and scarce burstiness challenges in a fine-grained SQ detection.

In summary, this research makes the following contributions:

- It presents a novel aspect extraction approach by fine-tuning a BERT model using a range of transport-related feeds. This approach enables event detection in a confined geographic area, where traditional approaches such as Latent Dirichlet Allocation (LDA) [6] or skip-gram [34] would fail.
- It shows that aspect-based sentiment analysis can contribute to the detection of events of interest as a complementary solution by improving the sensitivity of detection.
- It proposes a novel Frequency and Sentiment-based Event Detection (FSED) approach that combines and integrates frequency-based and sentiment-based event detection methods to overcome existing challenges and detects events of interest in a fine-grained geographic area. The proposed method significantly outperforms state-of-the-art event detection approaches.

The remainder of this paper is organized as follows. Section 2 provides an overview of recent related works. Section 3 introduces the proposed methodology for event detection. Section 4 discusses the experimental details including the dataset and settings. Section 5 discusses the results and findings. Finally, Section 6 concludes and summarizes the paper.

2 Related works

2.1 Event detection

In the context of social media, an event is defined as an occurrence at a specific time and place which prompts related discussions on social media [20]. During the last decade, numerous studies [21, 23, 24, 32, 41, 53, 58] have been conducted on developing different methods for event detection from social media feeds.

Topic-modeling-based event detection relies on estimating the probability distributions of latent topics from social media feeds. LDA [6], as a prominent topic modeling approach, can link terms and documents based on latent topics, thus, can present each document as a combination of multiple topics or as a set of representative words. This approach has been extensively utilized and adopted for the task of aspect extraction during the last decade [50,60,65]. A key limitation of LDA topic-modeling-based approaches is their need for prior knowledge of model parameters which are hard to determine [23]. Several attempts have been conducted to face LDA’s demonstrated drawbacks. As an example, Aiello et al. [2] propose a topic-modeling-based event detection approach BNgram, which considers co-occurrences of n-grams in order to detect topics of interest. They compare their proposed method with two other state-of-the-art event detection methods. Then, they compare term frequency on each day and its preceding days in order to detect any anomaly using the well-known word embedding approach Term Frequency-Inverse Document Frequency (TF-IDF). However, due to the short length of tweets, capturing significant topics from a limited context is still a challenge [20].

Other attempts leverage incremental clustering strategies to alleviate the challenge of predicting the number of clusters in traditional clustering approaches and create event-centric clusters and catch their uncommonly high burstiness scores. For instance, Hu et al. [23] leverage word embedding to improve the efficiency and accuracy of event detection. They integrate skip-gram [34] with an incremental clustering approach as well as k-means in order to represent documents and categorise news documents into event-centric clusters. They also adopt a metric similar to F-score in order to assess the performance of an unsupervised clustering with labelled data. Their work shows better results compared to LDA and a simple incremental clustering approach in terms of recall, F-score, and time efficiency. Recently, Hasan et al. [21] used $tf-idf$ word embedding along with an incremental clustering approach to reduce the computational complexity of event detection from high volume Twitter social media feeds.

One of the main challenges in extracting aspects from multi-labelled data is the lack of balance in the dataset. This challenge arises when labels are unevenly distributed over the dataset, which affects the performance of classification algorithms negatively. As the learning method may overfit the majority class and underfit the minority class, this leads to a biased model [7, 8]. Such a model tends to predict labels as the majority aspect, thus is missing important but less frequent events like Melbourne's car attack at Flinders Street Station in 2017. To avoid such situations, Synthetic Minority Oversampling TEchnique (SMOTE) [8, 29] can be employed to balance the dataset. SMOTE is based on a greedy algorithm which tries to make the distribution of labels as similar as possible and produce the best cross-validation folds. It does this by synthetically increasing the samples of each minority aspect. SMOTE can reduce classifier's risk of overfitting and improve the effectiveness of the classification method and prediction accuracy, thus is applicable to our application. Recently, this approach has also been successfully employed by Zahra et al. [61], where the authors incorporate domain-knowledge along with textual features and SMOTE for balancing the classes to achieve better classification performance.

Although the above approaches investigate the occurrence of events on social media feeds, most of these methods are evaluated in the case of larger numbers of observations, while in a confined geographic area, such as a transport node, sparse observations can lead to lower variability in categorization and event detection models. In addition, whereas leveraging balancing methods like SMOTE can mitigate the effect of imbalanced classes on the machine-learning-based models, they are unable to address problems of high similarity between features of inputs with different labels and lack of diversity among them. Thus, based on the bias-variance trade-off philosophy, these problems will increase the bias of a trained machine-learning-based model. A possible solution is embedding sentiment information as another class of observations that can be helpful for enriching the dataset and identifying human-related events, which is clearly neglected by the aforementioned approaches.

Additionally, in such a case, capturing uncommonly high burstiness scores is also challenging. This issue has been reported frequently in the literature [54, 63, 64], where general approaches fail to achieve expected effectiveness in the event detection. There have been a few attempts for strengthening these approaches by concurrent analysis of spatial characteristics of tweets along with burstiness detection. Nonetheless, these attempts are mainly on the basis of the analysis of geotagged tweets, while precise spatial information is no longer provided by Twitter. Moreover, as only 1% to 2% of Twitter streams are geotagged [18], limiting an event detection approach to this data can intensify the sparseness in the dataset.

2.2 Sentiment analysis in event detection

Over the past decade, few studies have used sentiment information for detecting minor or major events. Popescu and Pennacchiotti [43] used sentiments analysis as well as other linguistic and structural features to catch controversial events that led to public discussion on the Twitter platform. The research leveraged a 7590-word lexicon to estimate and quantify sentiment polarity score associated with each snapshot. Results indicated the feasibility and effectiveness of leveraging sentiment analysis as a contributing factor to the problem of event detection.

Marcus et al. [32] developed a system that uses a graphical interface to visualize and track events. They also leveraged an analytics engine and a classifier based on Naive Bayes theory to estimate and analyse crowd sentiments on those events. Their research revealed that while the sentiment analysis component was working properly, its polarity did not necessarily reflect the general feeling about an event. Nevertheless, they did not explore the connection between the strength of sentiments and real-world events. Thelwall et al. [51], on the other hand, found that while positive sentiments can also be used to detect events, the efficacy of negative sentiments showed better results.

Unlike Thelwall et al. [51], Paltoglou [39] has shown that both negative and positive sentiments could yield robust results. In addition, he presented a comparative study with data with different sample sizes between frequency-based and sentiment-based solutions for event detection. He proposed that sentiment-based solutions in specific environments, where data collection has been done by keyword-based queries, can have unique advantages compared to other frequency-based solutions. While his study showed the capabilities of sentiment analysis on Twitter data as a solution for event detection, it did not discuss how to integrate these solutions to enhance the detection method's precision.

Likewise, Nguyen et al. [37] employed sentiment analysis as a feature in the process of detecting real world events. In their research, each human-being is considered as a sensor. Therefore, sentiments can be considered as sensor measurements. The authors proposed their event detection method on the basis of a psychological behaviour framework along with a novel indexing method for temporal sentiment analysis and a basic anomaly detection approach based on time-series analysis.

Xiaomei et al. [57] recently introduced a corpus-based sentiment analysis method for micro-blog streams to detect breaking events. In this model, a classification and burst detection method were used to identify significant events after building a corpus-based dictionary and running sentiment analysis. The hashtags were then used to show the description of the events. The results demonstrate the potential to use sentiment analysis as an event detection solution. Nevertheless, in the case of the detection method, they missed the frequency of words as a significant source of data. Additionally, their approach takes only tweets that contain hashtags into consideration, whereas a large number of valuable tweets do not contain hashtags.

While these studies have reported promising results and provided valuable information, to the best of our knowledge none of them assessed the sensitivity of the instrument—the efficacy of event detection using sentiment analysis to detect shorter, longer, or regular events affecting SQ in public transport. Another research gap is the lack of an effective method for combining frequency-based and sentiment-based approaches in event detection. In addition, these studies have looked at public transport systems as a whole, while this study is interested in monitoring SQ in a single transport node. Due to the confined geographic area that can result in data sparsity, challenges constitute extracting significant aspects from a limited context, scarce burstiness in sparse feeds, and effective integration of sentiment-based solutions with general approaches.

A preliminary version of this work has been presented before [45]. Compared to that paper, this work employs BERT, as a cutting-edge language model, to classify tweets into semantically-related aspects. Moreover, this paper performs Aspect-Based Sentiment Analysis to capture passengers' perception regarding different aspects of SQ in public transport, whereas the previous research is based on a binary classification of sentiment classes. The major advantage of this work over the preliminary version is that this paper considers aspects in the process of event detection, where aspects can bring a more in-depth view into passengers' perception regarding different angles of a provided service. In addition, this work frees the model from the assumption that each day can contain at most one SQ event.

3 Event detection using Twitter data

In this section, we discuss our proposed solution for detecting events within a confined geographic area such as a transport node. As Figure 1 illustrates, the proposed method comprises of two main phases.

In the first phase, namely aspect extraction[2], multi-label tweets need to be grouped into semantically-related categories. To do so, this paper employs BERT to transform tweets into a vector of words. Then, using a binary classification approach, multi-label tweets can be classified into semantically-related groups, i.e., SQ aspects in our application. The second phase is called detecting events of interest[3], which conducts burstiness detection for each aspect and captures occasions where an aspect is mentioned more frequently than its average occurrence. With this aim, this paper proposes a statistical approach, called FSED, to combine frequency-based and sentiment-based event detection solutions. FSED aggregates candidate events captured by both methods in order to improve the sensitivity of the approach and tackle the challenge of Scarce Burstiness, which is discussed before.

In what follows, first, our proposed approach for extracting aspects based on the BERT language representation model is discussed. Next, a brief review on aspect-based sentiment analysis is presented, followed by the proposed statistic solution FSED for detecting events of interest.

3.1 Aspect extraction using a BERT language model

Recently, Google released a novel bidirectional language representation model BERT [13], which is known as a key innovation in the Natural Language Processing (NLP) tasks. On the one hand, similar to word embedding approaches, BERT leverages a large unlabeled corpus for understanding the semantic structure of a sentence and adopting feature representation of the sentence for many NLP tasks. On the other hand, unlike any other

[2]https://github.com/mmrahimi/SQ_monitoring_AE
[3]https://github.com/mmrahimi/SQ_monitoring_DEOI

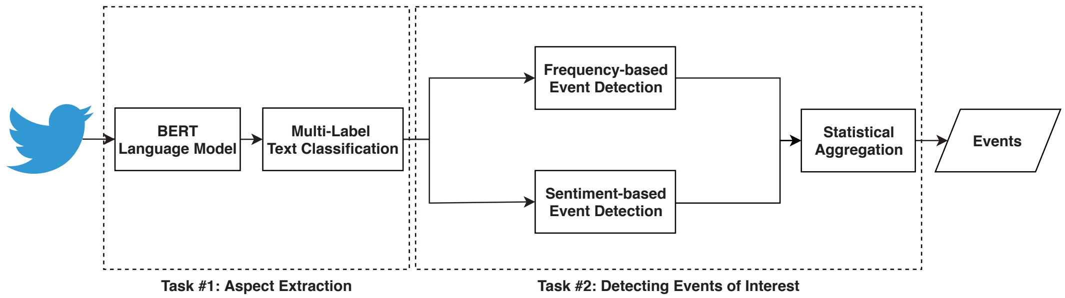

Figure 1: The proposed approach for event detection.

approach, BERT provides a fine-tuning step that minimizes the usage of prior contextual knowledge in the model design by learning such knowledge from the input data [59]. To do so, the last few layers of the language model will be modified to adapt the model to specific downstream applications. Moreover, in contrast to previous language models that read inputs sequentially either from left-to-right or right-to-left, BERT reads the entire sentence at once. This characteristic enables bidirectional training of the model, thus, letting BERT have a more profound sense of language context and a higher learning capacity. All these capabilities are especially important for our case, where there is a limited number of tweets, and learning from this limited context can be difficult due to the low variability challenge. Therefore, in this paper, a pre-trained BERT model is fine-tuned by a range of transport-related Twitter feeds extracted from a transport hub in order to transform tweets into vector features.

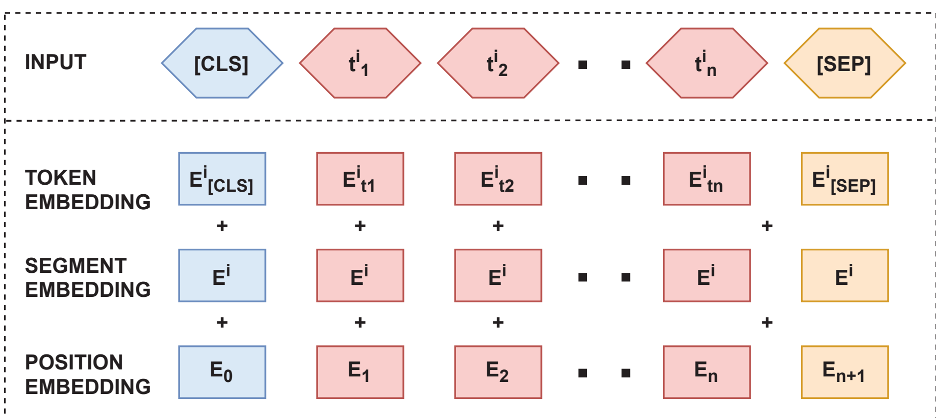

Figure 2: Overview of BERT's input for the task of multi-label text classification. Here i represents an identifier in tweets, t denotes tokens, and A is an aspect of SQ. [CLS] and [SEP] are twos special tokens which are needed to be added to each tweet.

In order to leverage the BERT language model, first, the format of the input data should be constructed. As Figure 2 shows, each tweet T^i should be broken down into smaller chunks, namely tokens, where a token can be a word, a punctuation mark, a number, an n-gram or even a symbol (e.g., '&' and '$'). Using a WordPiece tokenizer [56], each tweet can be chopped into a sequence of tokens $\{t_1^i, ..., t_n^i\}$. Secondly, two special tokens should be padded to each sentence. [CLS] is a special symbol which needs to be added at the beginning of every input token sequence. Moreover, [SEP] is another special token which is required at the end of the sequence [13]. Each input token, then, will be passed to a token embedding component where they are transformed into a vector representation, i.e., $E_{t_1}^i$. Next, segment embedding is used to label each token with its corresponding tweet. Finally, a position embedding is also employed for each token in order to indicate its position in the sequence.

After fine-tuning the model using collected tweets, a final linear classification layer is employed to receive BERT final hidden vectors and classify those which are semantically related. To do so, a sigmoid function is trained to obtain the probability of classification and compute the loss value for a standard binary classification. Thus, a sequence of labels as $[A_1^i, ..., A_6^i]$ is provided for each tweet which indicates whether the tweet belongs to each of the SQ aspects. Figure 3 depicts an example of our proposed aspect extraction method.

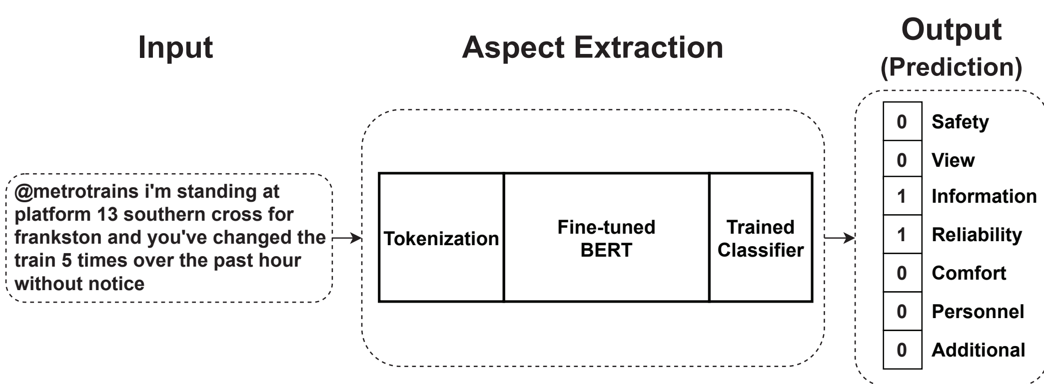

Figure 3: The process of aspect extraction using the fine-tuned BERT model and a trained binary classifier.

To face the challenge of multi-label learning in case of an unbalanced dataset, SMOTE is employed. SMOTE, as discussed in Section 2, can reduce a classifier's risk of overfitting and improve the effectiveness of the classification method and prediction accuracy, thus applicable to our application.

3.2 Detecting events of interest using frequency and sentiment analysis

As shown in Figure 1, our proposed approach for the second task can be split into three steps, namely Frequency-based Event Detection, Sentiment-based Event detection, and statistic aggregation. Details on frequency-based event detection can be found in the initial version of this paper [45]. In this section, first we present the details of aspect-based sentiment analysis, and then the proposed approach for statistical aggregation of both methods.

3.2.1 Aspect-based sentiment analysis (ABSA)

Sentiment analysis is defined as automated mining of opinions and emotions through the processing of text, speech, images and other kinds of data sources [26]. The aim of sentiment analysis is to measure positivity, negativity or neutrality, i.e., polarity of texts, and classify them into separate sentimental categories. Most of the early studies on sentiment analysis were conducted on predicting overall sentiments of a document while evaluating different aspects of different entities were neglected. Recently, Aspect-Based Sentiment Analysis [40] has been suggested as a novel computational approach to understand the perception of a user about specific entities and their corresponding aspects [42]. Due to the specific domain of this research, i.e., monitoring SQ in public transport, this paper uses aspect-based sentiment analysis for improving the effectiveness of general event detection and face the challenge of Scarce Burstiness.

Typically, ABSA comprises of two major tasks: aspect extraction and sentiment classification [42]. In the first phase, the goal is to extract entities "E" and attributes "A" from a given text. By analyzing the training data, it is observed that a single tweet can be related to various aspects of SQ. Therefore, the problem is defined as a multi-label text classification, where the only entity is defined as the SQ of public transport and attributes are Eboli and Mazzulla's [15] pre-defined aspects of SQ, i.e., "Safety", "View", "Information", "Service Reliability", "Comfort", "Personnel", and "Additional Services". Using the developed aspect extraction and fine-tuned language model in the last section, tweets can be classified and predicted based on their corresponding labels.

In the second task, a lexicon-based sentiment analysis approach named SentimentR [47] is employed to measure sentiment polarity of each aspect. SentimentR, which also has been used in other recent work [25, 55], is a tool that uses a dictionary look-up together with the consideration of valence shifters, i.e., terms that intensify or diminish emotions significantly. The dictionary includes 11709 words, where each individual score takes a value ranging from -2 to 1. In this task and for each aspect, first, the sentiment score for each tweet is estimated. Next, daily sentiments scores are calculated using an averaged (S_τ). Figure 4 shows the variation of S_τ during the study period.

During daily aggregation, it is observed that tweets with negative and positive polarities can neutralize daily sentiments in many cases. Therefore daily sentiment scores are classified and aggregated here into three distinct classes, i.e., negative, positive, and overall score, and the most effective class is explored in the process of detecting SQ-related events.

3.2.2 Frequency and sentiment-based event detection (FSED)

Due to the sparse observations within a confined space, general time-series-based approaches would not be able to detect the events of interest effectively. To alleviate this limitation, this research proposes a Frequency-based and Sentiment-based Event Detection (FSED) method. Using a statistical approach, FSED can combine and integrate event detection solutions based on sentiment and frequency. The proposed method employs the probability distribution of frequency and sentiments of tweets to estimate the probability of obtaining each possible value for daily frequency and sentiment score. Finally, a single probability value can be obtained that can be used to detect candidate events.

To begin with, first the best-fitting distribution functions for the frequency of tweets are explored. As the frequency has a discrete distribution over time, a series of well-known discrete distribution functions, namely Geometric, Binomial, Negative Binomial and Poisson

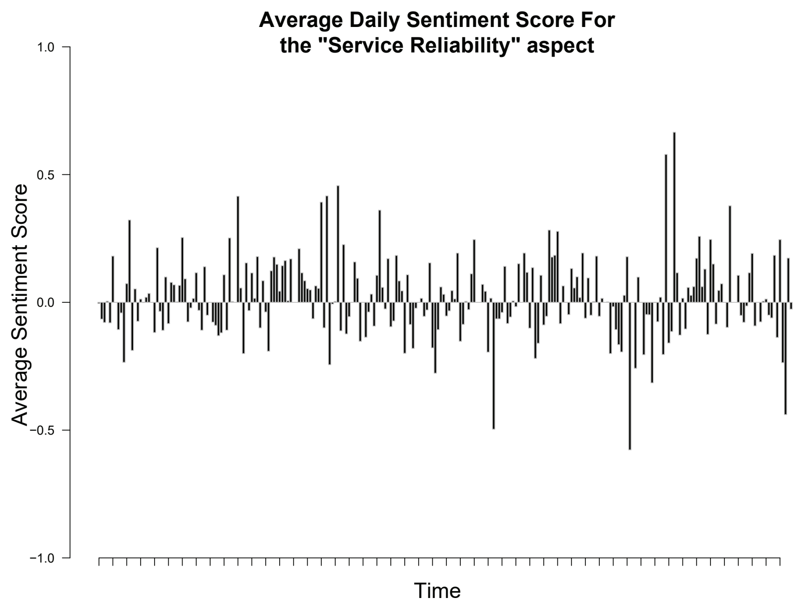

Figure 4: Daily variation of average sentiment scores for the “Service Reliability” aspect during the study period.

distributions, are considered to be fitted to the dataset. These distributions have been used quite effectively to analyze count data in the literature [10]. To find the best fitting distribution function, Maximum Likelihood Estimation (MLE) is employed to explore models' parameters. Next, the function with the highest log-likelihood value (goodness of fit) is leveraged as the best fitting distribution function for each aspect. For instance, this measure reveals that negative binomial has better fitness to the frequency of tweets, i.e., number of tweets per day, compared to other distribution functions for the aspect of “Service Reliability” (Figure 5). Here, Figure 5 indicates the accordance between the empirical (Observations) and the theoretical distributions (Expected values). Finally, for each aspect, the vector of probabilities is determined using empirical values of probability of success and mean.

On the other hand, due to the continuous nature of sentiment information, a series of commonly used univariate continuous distributions, namely Normal, Exponential, Gamma, Beta, Lognormal and Weibull distributions are considered for different classes of sentiments. Such functions have been used extensively in data analysis [28]. Similar to the last step, the best fitting distribution function is selected based on MLE and goodness of fit for each aspect of SQ. For example, for the aspect of “Service Reliability”, a Beta distribution is employed as the best-fitting distribution function for the negative class of sentiments (Figure 6). The vector of probabilities is calculated for each aspect and class of sentiments using empirical values for its shape parameters.

Finally, for each aspect a, these two solutions are combined by defining the overall distribution of probability as:

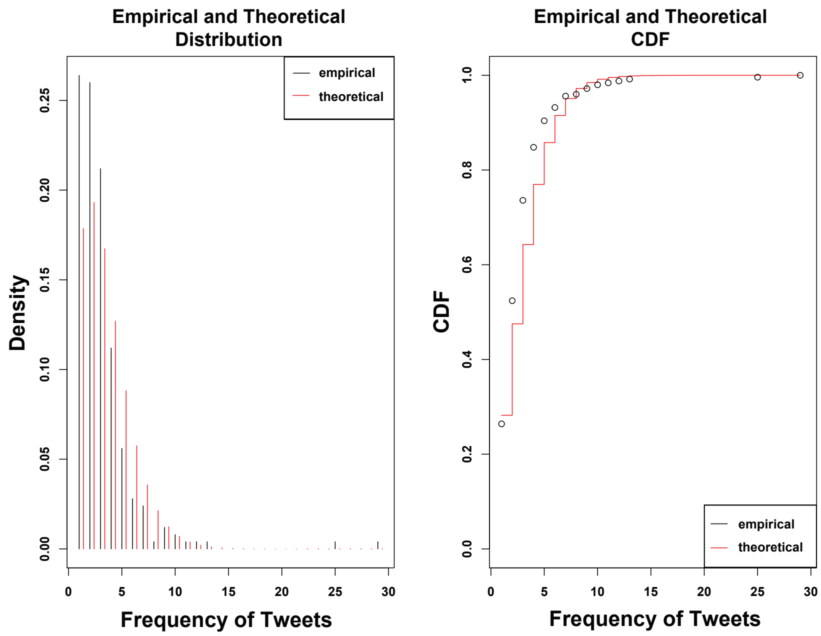

Figure 5: Empirical and theoretical distribution (left) and Cumulative Distribution Function (CDF) (right) of frequency data for the aspect of "Service Reliability" using negative binomial distribution.

$$P(F_\tau^a < \alpha, S_\tau^a < \alpha) = P(F_\tau^a < \alpha) * P(S_\tau^a < \alpha) \tag{1}$$

where P is the variable's probability value and α is the level of significance. Therefore, for each day, if the overall probability value is smaller than the level of significance, the day is selected as a candidate for an event.

4 Experiments

This section presents and discusses details of experiments including the used dataset, parameter settings, evaluation schema and final results.

4.1 Dataset and ground truth

In this research, two major transport hubs are considered as case studies: Victoria's two busiest public transport interchanges. These two hubs are Southern Cross Station (SCS) and Flinders Street Station (FSS) in the central business district of Melbourne. The experiment is based on a Twitter dataset comprising of more than 32 million tweets, posted within the metropolitan area of Melbourne between June 2017 to May 2018. This data is obtained from the Australian Urban Research Infrastructure Network (AURIN[4]). To detect relevant

[4]www.aurin.org.au

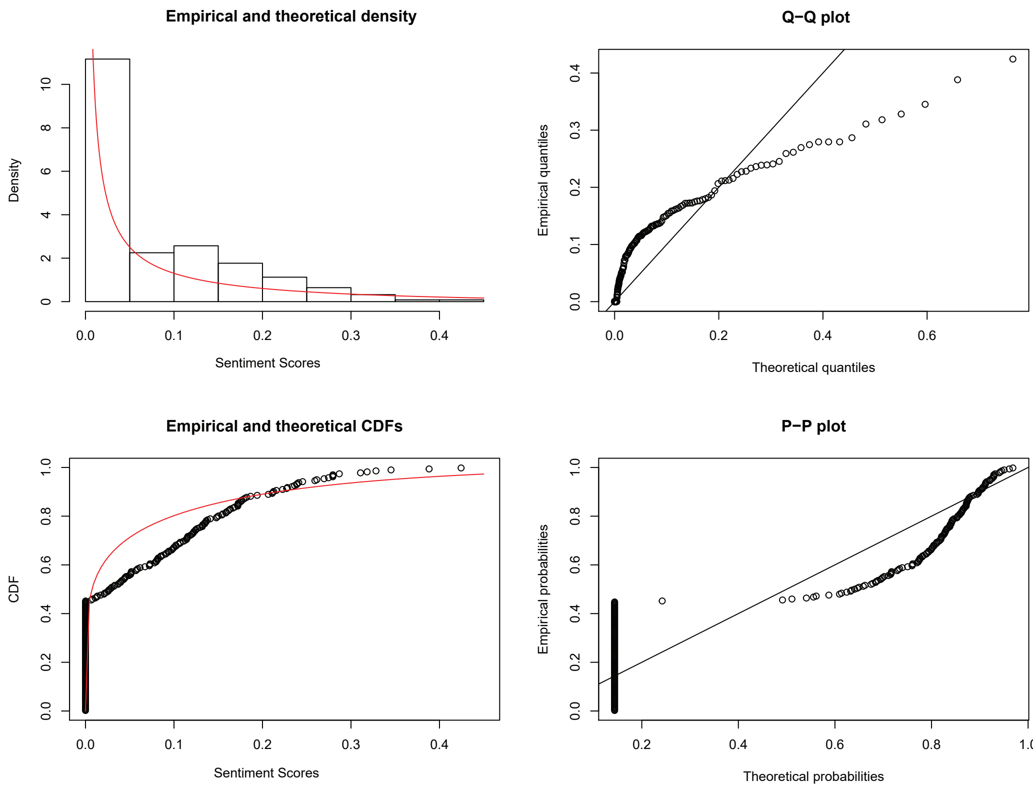

Figure 6: Empirical and theoretical densities (top-left), Q-Q plot (top-right), Cumulative Distribution Function (CDF) (bottom-left) and the P-P plot (bottom-right) of exponential distribution of the negative class of sentiment scores in the aspect of "Service Reliability".

tweets, a two-phase search, i.e., a textual keyword search followed by a spatial search, is conducted. First, the following keywords are queried and all tweets containing any of these keywords are selected for further processing:

- Different combinations of station names (as extracted from Wikidata[5] and tweets themselves) such as "Southern Cross", "SouthernCross", "Spencer Street Station" (station's old name), "SCS" or "SSS", "Flinders", "Flinder Street Station" (without S).
- Names of public transport operators such as Metrotrains, V/Line, PTV

Next, after extracting the geographic extent of each transport hub using OpenStreetMap Nominatim web service[6], a point-in-polygon operation is performed to retrieve other corresponding tweets which are possibly missed by the keyword search process but were send

[5]https://www.wikidata.org/wiki/Q801455
[6]www.nominatim.org

from within the geographic extent of the stations (this twitter dataset still had precise geotagging). A final aggregation based on Tweet-IDs result in a total of 2038 tweets for SCS and 1386 tweets for FSS.

While keyword and geographic search approaches can be used to select feeds related to both transport hubs, it is expected that only a fraction of these tweets will be related to the public transport SQ. Moreover, in order to categorize tweets into semantically-related groups, it is required to know the corresponding aspect of SQ that each of these tweets belong to. To this aim, this paper uses Eboli and Mazzulla's [15] proposed taxonomy where they characterize SQ of public transport by seven different aspects namely "Safety", "View", "Information", "Service Reliability", "Comfort", "Personnel", and "Additional Services". This taxonomy is leveraged here to map selected tweets to the corresponding aspect(s) of SQ. To this end, one of the authors manually annotated tweets with one or more aspects of SQ. Those tweets that do not fall into any of these aspects are considered as irrelevant to the SQ of public transport, and therefore, are discarded. Finally, 1375 and 1190 tweets are selected for further analyses in SCS and FSS, respectively. Table 1 shows the distribution of tweets over different aspects of SQ for two case studies, SCS and FSS.

Node	Safety	View	Information	Reliability	Comfort	Personnel	Additional	Discarded	Total Labelled
SCS	9.7%	9.6%	10.6%	53.7%	14.3%	4.8%	14.9%	663	1375
FSS	6.3%	1.7%	16.3%	79.6%	9.9%	8.9%	1.3%	196	1190

Table 1: Distribution of selected and manually-annotated tweets with different aspects of service quality. Each tweet can belong to more than one aspect of SQ.

In addition, in order to evaluate the effectiveness of the proposed approach, a list of events happened inside of SCS are manually-labelled for the study period by the aforementioned annotator. Here, we record an event in the list if there are at least two tweets supporting the occurrence of the event. The resulted events then get verified using a list of events provided by SCS authorities as a ground-truth. Table 2 shows the number of verified manually labelled events for each aspect during the study period.

Node	Safety	View	Information	Reliability	Comfort	Personnel	Additional	Total
SCS	18	8	16	73	19	2	14	150

Table 2: The number of manually-extracted events for different aspects of service quality.

4.2 Limitations on the dataset

In order to know the limitations of the sparse dataset used in this research, tweets corresponding to SCS are compared with their superset, i.e., the whole set of tweets posted in Melbourne during the study period. To this aim, a simple comparison of the distribution of words in the two datasets is performed. This is done by fitting a few well-known discrete probability distribution functions to both datasets. Here, discrete distribution functions (Geometric, Binomial, Negative Binomial and Poisson), are leveraged due to the nature of both datasets. To solve the parameters that best fit each distribution function and find the best-fitting distribution, maximum likelihood estimation is employed. Finally, Poisson regression is chosen as it has better fitness to the distribution of word frequencies (in terms of log-likelihood value) compared to other discrete distributions. Table 3 illustrates the results of the Poisson regression. In this table, rate (μ) is the mean rate of the occurrence of tweets in the dataset per unit of offset (a particular unit of observation). This paper sets the offset value to 3 due to a previous assumption that at least three tweets are required to reflect the occurrence of an event (as discussed in Section 4.1).

Place	Offset	Rate (μ)	STD
SCS	3	14.7	29.9
Melbourne	3	93.3	162.7

Table 3: Parameters used for Poisson regression along with the corresponding standard deviation (STD).

As Table 3 shows, the standard deviation of the tweets in SCS is significantly smaller than the corresponding value for the whole of Melbourne, where a small standard deviation reflects small variability in the dataset. This smaller variability will lead to a more biased trained model. As a result, normal event detection approaches may fail to achieve expected precision and recall values (discussed in Section 4.4) due to the data sparseness, and this can considerably reduce the effectiveness of straight-forward event detection approaches.

4.3 Experimental settings

As Figure 1 shows, the event detection task can be split into two consecutive tasks. For each task, the performances of the proposed approaches are compared with other state-of-the-art methods with ground-truth results and on a sparse dataset. In this section, experimental settings considered for each of these approaches are discussed.

4.3.1 Aspect extraction approaches

For the aspect extraction phase, the performance of BERT, the proposed approach, is compared with two other state-of-the-art baseline approaches along with multiple classifiers:

- **LDA**, as a frequently-used state-of-the-art approach [3,19,31], is implemented for the comparison with the language model-based classification. The first step in finding latent topics in LDA is setting the number of topics. A comparative approach found in the literature [3, 44] is used to extract the optimum number of topics. This method applies a quantitative comparison on a single model by training it with a different number of topics. Then, models will be evaluated, as their corresponding topics should be distinguished based on the subject while their interpretability should be preserved. The other hyper-parameters of LDA, α and β, are empirically set to 50/K and 0.1 respectively.
- **Skip-gram** [34], a novel word embedding-based approach, is also applied here to evaluate the performance of the proposed method. Similar to Hu et al. [23], the dimension of features is set to 100 and δ is set to 0.4.

After transforming terms into multi-dimensional vectors using LDA or skip-gram, three state-of-the-art classification approaches (Support Vector Machines (SVM) with a Gaussian Radial Basis Function (RBF) kernel, Logistic Regression (LR), and Multi-Layer Perceptron (MLP)) are employed for grouping tweets into semantically-related categories. These classifiers are selected based on their demonstrated effectiveness in analysis of high-dimensional datasets with multiple classes. Classifiers are trained using one-vs-rest strategy, where we fit one classifier for each aspect of SQ. For the MLP classifier, similar to Lenc et al. [30], an architecture with one layer and 512 neurons are considered.

In the proposed method, a fine-tuned BERT model along with one final layer of a sigmoid function is used to apply a multi-label text classification and predict the corresponding sequence of labels for each tweet. To do so, a pre-trained uncased BERT$_{base}$ with 12 layers is utilized as the base platform of the experiments. Similar to Devlin et al. [13], a batch size of 32 and three epochs were used for the fine-tuning of the model. Here, two epochs of training are applied to reduce the possibility of "over-fitting". In order to face the imbalance in the dataset, random minority oversampling with replacement is used [29].

4.3.2 Detecting events of interest

For the second phase, the proposed method is compared against three state-of-the-art time-series anomaly detection approaches: Auto-Regressive Integrated Moving Average (ARIMA), Long Short-Term Memory Network (LSTM), and Seasonal Hybrid Extreme Studentized Deviate (S-H-ESD) [1,4,52,53]. For time-series-based approaches, the outlier detection is applied to sentiments scores (S) and frequency of tweets (F) to detect candidate events. Therefore, two sets of candidate dates are extracted for each criterion. We also consider the union and intersection of these sets to assess the effectiveness of sentiment analysis for detection of SQ event. The significance level (α) is set as 5% as it leads to the best event detection results in the case of F-score. Other parameter values chosen for each baseline method are:

- S-H-ESD: Similar to Tonon et al. [52], the significance level (α) is set as 5%.
- ARIMA: Similar to Khongsrabut and Waiyamai [27], ARIMA's hyper-parameters, i.e., p, d, and q, are defined using an approach named "AutoARIMA", where ARIMA's hyper-parameters are extracted using an optimization approach based on the Bayesian Information Criterion (BIC).
- LSTM: Similar to Saeidi et al. [49], the model uses a batch size equal to the number of observations (number of days in the study period in our case) and the number of hidden units of size 50. In addition, the anomaly detection threshold is set at $k = 3$, similar to Wei et al. [53].

4.4 Evaluation metrics

As illustrated in Figure 1, the event detection comprises of two consecutive stages. For each stage, various metrics can be adopted for the assessment of the aforementioned approaches. Aligned with these stages, this section can be divided into two subsections. First, adopted metrics for aspect extraction is presented. Next, the second subsection elaborates chosen metrics for detecting events of interest.

4.4.1 Metrics for aspect extraction

In order to evaluate the obtained results and find the optimum number of topics in LDA, the semantic coherence metric UMass [35] is used. This method, which has been used in the literature extensively [3,31,48], is a co-occurrence measure that uses a smoothed conditional probability in order to quantify the coherence in each topic. Moreover, in order to investigate the resulting topics, a common way of listing them is looking at their proportions in the dataset and manually exploring their top-k words. Nonetheless, it is not sufficient to list the most frequent words for each topic, as several topics may share some very common and uninformative terms. An alternative approach is using FRequency-EXclusivity (FREX) [5], which can balance the frequency of terms with the exclusivity of them to each topic. FREX uses a harmonic mean of an empirical Cumulative Distribution Function (CDF) of each the frequency of each word along with an empirical CDF of specificity to that topic. In this research, FREX is used to explore top words within each topic in order to have a better understanding of the resulting topics.

In order to assess the effectiveness of the proposed method in the aspect extraction task, this paper conducts a comparative study between BERT, skip-gram and LDA in terms of their performance in the multi-label classification. To do so, SVM, LR and MLP are used on top of two baseline methods and a sequence of labels predicted for each tweet by each method is compared with corresponding annotations from the ground-truth. Models are trained based on the tweets from SCS. Next, trained models are tested on FSS using a 10-fold cross-validation. Similarly, in the proposed method, the BERT model is fine-tuned on tweets from SCS and tested based on data from FSS to ensure an out-of-sample evaluation. Here, for each aspect, the following well-established metrics for the evaluation of NLP tasks can be adopted:

- Precision$_{\text{AE}}$: The fraction of relevant tweets which are correctly classified as relevant to the aspect from the total number of cases that the model is classified as relevant.

$$Precision_{AE} = \frac{TP_{AE}}{TP_{AE} + FP_{AE}} \tag{2}$$

- Recall$_{\text{AE}}$ (Sensitivity): The fraction of relevant tweets which are correctly classified as relevant to the aspect from a total number of real relevant tweets to the aspect in the dataset.

$$Recall_{AE} = \frac{TP_{AE}}{TP_{AE} + FN_{AE}} \tag{3}$$

- F-score$_{\text{AE}}$: The harmonic mean of the Precision and Recall values which indicates the effectiveness of the classification method.

$$F - score_{AE} = 2 * \frac{Precision_{AE} * Recall_{AE}}{Precision_{AE} + Recall_{AE}} \tag{4}$$

where TP_{AE} (true positives) is the number of relevant tweets which are properly classified as relevant, FP_{AE} (false positives) is the number of irrelevant tweets which are improperly classified as relevant and FN_{AE} (false negatives) is the number of relevant tweets which are improperly classified as irrelevant. Based on the values for Precision$_{\text{AE}}$, Recall$_{\text{AE}}$ and F-score$_{\text{AE}}$ of each aspect, the overall performance of the classification approach can be obtained. To do so, two well-known metrics, Micro-average and Macro-average, can be employed. While Micro-Average aggregates the contributions of all classes to compute the average metric, Macro-Average computes the metric independently for each aspect and then takes the average.

Moreover, in order to confirm the performance of the classification approach, a probabilistic measure can also be leveraged. With this aim, the Receiver Operating Characteristic (ROC) curve and the Area Under the ROC Curve (AUC) are employed in this research. Generally, the ROC curve demonstrates the ability of the classification method to distinguish labels and the AUC provides the opportunity to quantify the ROC curve. AUC has been widely used for evaluating various classification approaches recently [17].

4.4.2 Metrics for detecting events of interest

To evaluate the effectiveness of the proposed method for event detection, the final step of event detection needs to be performed (Figure 1). Here, FSED is compared with three state-of-the-art baseline approaches based on precision, recall and F-score. Here, for each aspect of SQ, the following adopted metrics can be adopted (equations are similar to what is discussed in Section 4.4.1):

- $\text{Precision}_{\text{DEOI}}$: The fraction of eventful days that are correctly detected from the total number of days that the model has detected as eventful.
- $\text{Recall}_{\text{DEOI}}$ (Sensitivity): The fraction of eventful days that are correctly detected from the total number of eventful days in the dataset.
- $\text{F-score}_{\text{DEOI}}$: The harmonic mean of the $\text{Precision}_{\text{DEOI}}$ and $\text{Recall}_{\text{DEOI}}$ values, which indicates the effectiveness of the event detection method.

where TP_{DEOI} is defined as the number of eventful days that are correctly detected as eventful, FP_{DEOI} is the number of uneventful days that are incorrectly detected as eventful, and FN_{DEOI} is the number of eventful days that are incorrectly detected as uneventful.

5 Results and discussion

As Figure 1 shows, our event detection framework is divided into two consecutive tasks, Aspect Extraction and Detecting Events of Interest, which are discussed in more detail in this section. Section 5.1 compares the effectiveness of a conventional aspect extraction method, i.e., LDA, against language modeling techniques in the case of the sparse dataset. In Section 5.2, the effectiveness of the proposed FSED method for detecting events of interest is compared with other state-of-the-art baseline approaches.

5.1 Aspect extraction

For the aspect extraction phase, the performance of two state-of-the-art baseline approaches, namely LDA and skip-gram, are compared with BERT.

5.1.1 LDA topic modeling

LDA is one of the most common techniques to extract aspects in a set of tweets and group them accordingly [23]. As a result, we first consider the effectiveness of LDA to extract the aspects that can be used to group semantically-related tweets in our dataset.

To convert each tweet into the vector of words, we use Bag of Words and Term Frequency-Inverse Document Frequency (TF-IDF) as the two most common word embedding approaches. Gensim [46] and MALLET tools (MAchine-Learning for LanguagE Toolkit) [33] are employed to train the model based on the provided word embedding. The output of the training phase is a set of topics, where each topic comprises a sequence of words.

The first step in finding latent topics in LDA is setting the number of topics. As discussed in Section 4.4.1, similar to [3, 31, 48], this paper uses UMass [35] coherence to evaluate the obtained results and find the optimum number of topics in LDA. Using UMass coherence, a quantitative comparison between models trained with a different number of topics can be conducted [3,44]. Figure 7 shows the variation of the coherence over different numbers of topics. The optimum number of topics can be found at the end of rapid growth in the value of coherence [3]. As Figure 7 demonstrates, here the optimum number of topics can be either 6 or 10 where corresponding coherence values are almost 74%.

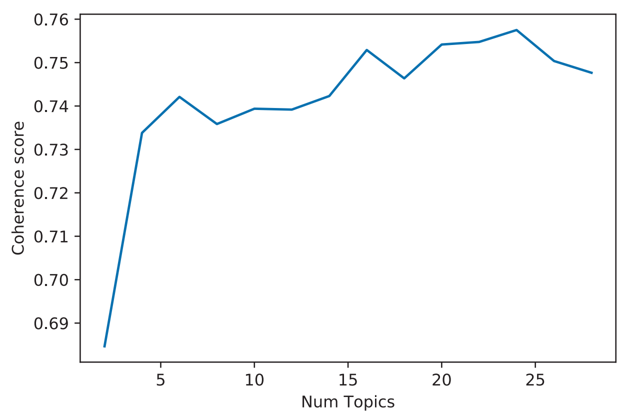

Figure 7: The variation of coherence score for a different number of topics.

Figure 8 illustrates inter-topic distance map extracted by LDA topic models with six (right panel) and ten (left panel) as the optimum number of topics. In these plots, the size of topics (circles) shows their coherence while the distance between them reveals their inter-topic correlations. As the figure demonstrates, when the number of topics is equal to 10 (left figure), some of resulting topics, i.e., Topics 2, 4, 5, 9, and 10, are neighbouring and overlapping. The same pattern can be observed in the right figure, where Topics 1 and 2 and Topics 3, 4, and 6 are overlapping. In addition, it can be observed that most of these topics are being clustered in the two top quadrants instead of being scattered throughout the distance map. This shows that there is a high correlation between these topics and they are not significantly segregated. Thus, it can be argued that the trained model cannot provide meaningful and interpretable topics.

As discussed in Section 4.4.1, the FREX score is employed to explore top significant words within each topic. Table 4 and Table 5 show the top eight words assigned to each topic based on the FREX score for LDA models with ten and six topics, respectively. As these two tables show, although UMass coherence of topic modeling is more than 70%, the obtained topics do not completely match with individual SQ aspects. In other words,

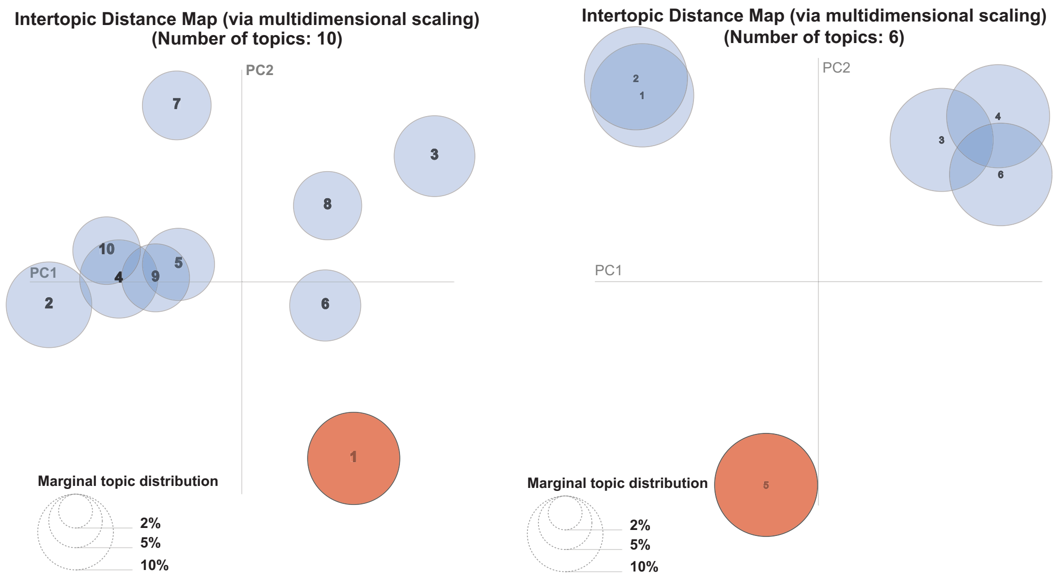

Figure 8: Visualization of modeled topics with ten topics (left) and six topics (right).

while some topics should be merged to match with a specific SQ aspect, others need to be split. For instance, in Table 4, Topic 1 contains keywords related to the "View", "Information" and "Service Reliability" categories, and Topic 2 comprises of "Service Reliability", "Comfort" and "Additional services" aspects. The same pattern can be observed in Table 5, where Topic 4 comprises of "Service Reliability" and "Safety", whereas Topic 5 contains keywords related to "Safety", "Information" and "Comfort".

	1	2	3	4	5	6	7	8
Topic 1	look	name	**early**	**friend**	**warning**	**meet**	**bin**	current
Topic 2	**stand**	yet	toward	**timetable**	**earlier**	name	**delay**	looker
Topic 3	full	bollard	aboriginal	yet	look	sign	star	toward
Topic 4	toward	name	look	yet	clear	starbucks	info	staff
Topic 5	board	operation	toward	notice	name	look	yet	request
Topic 6	yet	name	look	power	footy	starwars	fault	guy
Topic 7	wait	racist	name	straight	toward	extra	look	minute
Topic 8	tomorrow	outside	connect	name	station	look	stick	passenger
Topic 9	name	life	sky	youtube	look	toward	escalator	issue
Topic 10	free	yet	look	police	name	fan	final	classic

Table 4: Top eight words assigned to each topic based on the FREX score for an LDA model with ten topics. Terms in bold are used for interpreting the SQ aspect.

Similarly, we observe that the model groups tweets belonging to different SQ categories into a single cluster. For instance, Table 6 shows three different examples of tweets which are categorized in a single topic using LDA model with ten topics (Topic 1).

	1	2	3	4	5	6	7	8
Topic 1	stop	bus	photo	notice	vline	move	year	fix
Topic 2	walk	night	track	free	start	trip	short	lift
Topic 3	meet	sign	place	fault	ticket	weekend	close	work
Topic 4	run	leave	**minute**	**cancel**	metrotrain	operate	**early**	**police**
Topic 5	**bollard**	**display**	**escalator**	request	pay	x-wing	taxi	long
Topic 6	line	passenger	depart	staff	travel	tram	late	direct

Table 5: Top eight words assigned to each topic based on the FREX score for an LDA model with six topics. Terms in bold are used for interpreting the SQ aspect.

Tweet	Aspect
"the frankston service from southern cross we were all **waiting** as per the **screens** and then when the time came it just changed to flinders"	**Service Reliability**
"band **playing** in front of the southern cross steps this morning were soothing?"	**Additional Services**
"omg help, i just passed a huge **crowd** of people in fursuits at southern cross."	**Comfort**

Table 6: Example of tweets categorized into the first topic using LDA.

As a result, it can be argued that LDA topic modeling has demonstrated drawbacks in categorizing sparse tweets into semantically similar groups.

5.1.2 Word embedding-based clustering

To evaluate the effectiveness of our proposed aspect extraction method, a state-of-the-art word embedding-based approach, skip-gram, is also evaluated in this paper.

Here, the effectiveness of the method proposed by Hu et al. [23] as a state-of-the-art skip-gram-based event detection approach is evaluated, where we first train the model on tweets from SCS and then employ their proposed adaptive online clustering for grouping tweets into semantically-related categories. This approach sequentially processes inputs, one at a time, and grows the clusters incrementally. It uses cosine similarity to find the most similar cluster generated previously to add the tweet to it, otherwise, a new cluster will be generated and the tweet will be assigned to that.

However, we observe that this approach categorizes all the tweets into a single cluster. This is due to the fact that more than 98% of vectors, i.e., tweets transformed using the skip-gram method, have a similarity score of more than 0.9. This can be interpreted such that due to the limited number of tweets and their short length, skip-gram fails to capture the semantic and syntactic relations between terms. This issue, which also has been observed by Nguyen et al. [38], highlights the low variability challenge of our problem.

5.1.3 A comparative study

This study proposes to use a multi-label text classification method using BERT to maximize the usage of prior contextual knowledge in the model design and bring extra semantic features into the process of text classification. To investigate the effectiveness of the proposed method, this paper conducts a comparative study between BERT and two other baseline approaches, i.e., skip-gram and LDA. Note that unlike the conventional way of using the output of LDA and skip-grams to cluster tweets, we use the LDA extracted features and the word embedding representations of the skip-gram to classify tweets into the prespecified SQ categories. Hence, we use three classifiers, namely SVM, LR, and MLP to compare the performance of LDA, skip-grams, and BERT in the multi-label classification. As discussed in Section 5.1.1, the two LDA models with six and ten topics are chosen for further evaluations.

First, classification models are trained based on the tweets from SCS. Then, using the trained models, labels for FSS are predicted for cross-validation. Here, according to resulting values for precision, recall, and F-score of each aspect, the overall performance of each classification approach can be obtained. To do so, Micro-Average and Macro-Average can be employed. While Micro-Average aggregates the contributions of all classes to compute the average metric, Macro-Average computes the metric independently for each class and then take the average. In addition, the AUC metric is leveraged to confirm the findings.

Aspect	**BERT**	**SG + MLP**	**LDA 6 + MLP**	**LDA 10 + MLP**
Safety	56% ± 8%	11% ± 9%	1% ± 1%	0%
View	34% ± 4%	6% ± 7%	0%	1% ± 3%
Information	75% ± 3%	55% ± 3%	2% ± 2%	1% ± 2%
Reliability	92%	86% ± 2%	69% ± 2%	70% ± 2%
Comfort	51% ± 4%	9% ± 4%	1% ± 1%	2% ± 2%
Personnel	59% ± 10%	12% ± 1%	2% ± 3%	1% ± 3%
Additional	45% ± 19%	2% ± 3%	0%	2% ± 2%
Micro Avg	82% ± 1%	71% ± 2%	44% ± 1%	45% ± 2%
Macro Avg	59% ± 2%	26% ± 1%	11%	11%
AUC	74% ± 1%	57%	52%	52%

Table 7: Classification results for different aspects of SQ and different aspect extraction methods in terms of F-score and AUC metrics. Here, Avg stands for average F-score and SG stands for Skip-Gram. Standard deviation values smaller than 1% are not mentioned in the table.

Table 7 illustrates the results of text-classification using various classification models for different aspects of SQ. Due to the limited space, results of approaches with majority of zeros in aspects of SQ are excluded from the table (mainly SVM and LR-based classification approaches). As Table 7 demonstrates, our proposed method outperforms all other state-of-the-art baseline approaches in terms of Micro-Average F-score (82%), Macro-Average F-score (59%) and AUC (74%). To be more specific, it can be observed that BERT slightly improves the Micro-Average value up to 82% compared to the skip-gram-based classification approach, where it can achieve maximum 71% with a standard deviation of 2%. Here, similar Micro-Average values of all methods can be justified by the fact that all of baseline approaches achieve at least 70% F-score values in the “Service Reliability” aspect, where there are enough observations to train the models (Table 1).

However, in case of Macro-Average F-score values, it can be observed that BERT’s performance is more than two times better than the second-best approach, the skip-gram-based MLP. In other words, when it comes to aspects with fewer observations, it can be observed that baseline approaches are unable to detect tweets’ associations with minor aspects of SQ. This can be justified by the fact that a small number of tweets means fewer occurrences of input features in word-embedding or topic models, which can lead to the low variability challenge mentioned before. Table 8 provides a more detailed comparison between our proposed method and skip-gram-based MLP classifier as the second-best approach in terms of F-score and AUC values. As Table 8 demonstrates, while skip-gram-based MLP achieves considerable precision values in "Comfort" and "Personnel" aspects, i.e., 43% and 63% respectively, unlike BERT, it fails to keep the trade-off between the corresponding recall values. This shows that the model is not sensitive enough to different kinds of feature vectors that can show up.

In contrast to all baseline methods, BERT is pre-trained on a large corpus of data and can be fine-tuned for the context of public transport, thus, it can bring extra semantic features into the process of text classification, which increases the variability and reduces the bias in the classifier. Additionally, BERT's bidirectional training of a transformer can also provide a deeper understanding of the language context and improves the learning capacity. As Table 7 illustrates, BERT outperforms all baseline approaches in all other aspects of SQ significantly. This discussion can also be confirmed by the resulting AUC metric, where BERT can achieve 17% higher AUC values compared to the second-best approach, the skip-gram-based MLP. Moreover, it can be seen that other baseline approaches have an AUC value around 0.5, which reflects their incapacity to effectively separate different classes. After BERT, skip-gram-based classifiers perform better than LDA based approaches. In particular, skip-gram-based MLP has a significantly better performance especially in minor aspects where it achieves an F-score of 55% in the "Information" aspect.

Aspect	**Metric**	**BERT**	**SG + MLP**	**Support**
	P	86% ± 10%	40% ± 32%	
Safety	R	41% ± 14%	6% ± 5%	75
	F	56% ± 8%	11% ± 9%	
	P	43% ± 1%	13% ± 17%	
View	R	29% ± 4%	4% ± 5%	21
	F	34% ± 4%	6% ± 7%	
	P	68% ± 1%	70% ± 5%	
Information	R	83% ± 5%	45% ± 3%	194
	F	75% ± 3%	55% ± 3%	
	P	92% ± 1%	87% ± 1%	
Reliability	R	92%	86% ± 5%	948
	F	92%	86% ± 2%	
	P	63% ± 3%	43% ± 16%	
Comfort	R	43% ± 5%	5% ± 2%	117
	F	51% ± 4%	9% ± 4%	
	P	74% ± 4%	63% ± 7%	
Personnel	R	50% ± 13%	7% ± 1%	107
	F	59% ± 10%	12% ± 1%	
	P	83% ± 13%	2% ± 2%	
Additional	R	31% ± 6%	2% ± 3%	16
	F	45% ± 4%	2% ± 3%	

Table 8: Precision, recall and F-score for two best approaches, BERT compared with skip-gram-based MLP classifier, for different aspects of SQ. Standard deviation values smaller than 1% are not mentioned in the table.

It can be observed in Table 8 that BERT and skip-gram-based MLP models manage to achieve their best results in the "Service Reliability" and "Information" aspects. This is due to the fact that there has been more support in these two aspects compared to the others in the dataset. Focusing on the aspect of "Service Reliability", it can be observed that, having more than 700 tweets in the fine-tuning phase and 900 tweets in the evaluation phase, the fine-tuned BERT model achieves a recall value of 92%, while the precision value is also preserved. A similar observation is made for skip-gram-based MLP where it achieves 87% precision and the recall value of 86%. Similarly, looking at the aspect of "Information", 83% of recall (45% for skip-gram-based MLP) and 75% (55% for skip-gram-based MLP) of F-score proves the impact of bias-variance trade-off where higher variability of data can lead to a less biased model.

On the other hand, focusing on the "Additional Services" aspect, it can be seen that while the precision value of the classification is 83%, the recall value is 31%. This situation means that the model is not sensitive enough to the different kinds of feature vectors that can occur. This observation can be justified by two arguments. First, having 16 tweets in the aspect of "Additional Services" and 21 tweets in the aspect of "View" in FSS, it can be argued that the variability of tweets is not large enough for the model to obtain enough sensitivity. In other words, low variability in features leads to a more biased model. Again, this highlights the main limitation of event detection on sparse datasets. Second, it is a fact that the nature of additional services inside or around the SCS differs with FSS. For example, SCS contains a well-known outlet shop, multiple cafés, a food court and a grocery supermarket, while FSS has none of these. Moreover, SCS is next to a stadium, whereas FSS is close to an arts venue and cultural and public events. This means that the type of activities that happen inside/around SCS and FSS are different and therefore, tuning the model based on one of them, may not fully cover the other. Thus, in this case, the model will suffer from under-fitting. The same arguments hold for the aspect of "View".

In summary, by achieving 82% F1 Micro-Average and 74% AUC, it can be argued that text classification using state-of-the-art language models can be a promising approach for grouping a sparse dataset of tweets into semantic-related categories.

5.2 Detecting events of interest

In this section, our proposed method in Section 3.2.2 is compared to three state-of-the-art time-series anomaly detection approaches, namely S-H-ESD, ARIMA, and LSTM [1, 4, 52]. For time-series-based approaches, outlier detection is applied to sentiment scores (S) and frequency of tweets (F) to detect candidate events (Figure 9). Therefore, two sets of candidate dates are extracted for each criterion. Here, the union and intersection of these sets are also considered in order to assess the effectiveness of sentiment analysis for detection of SQ event.

Using the list of events as ground-truth, candidate events are compared and evaluated. In this section, the results of three aspects of SQ, namely "Service Reliability", "Information" and "Comfort" are discussed. These aspects are chosen since they have the most number of support, the most significant F-score values in the aspect extraction phase and the maximum number of events in the study period. Table 9 compares the results of different event detection methods with regard to the sentiment class, anomaly detection method, and criteria for the aspect of "Service Reliability".

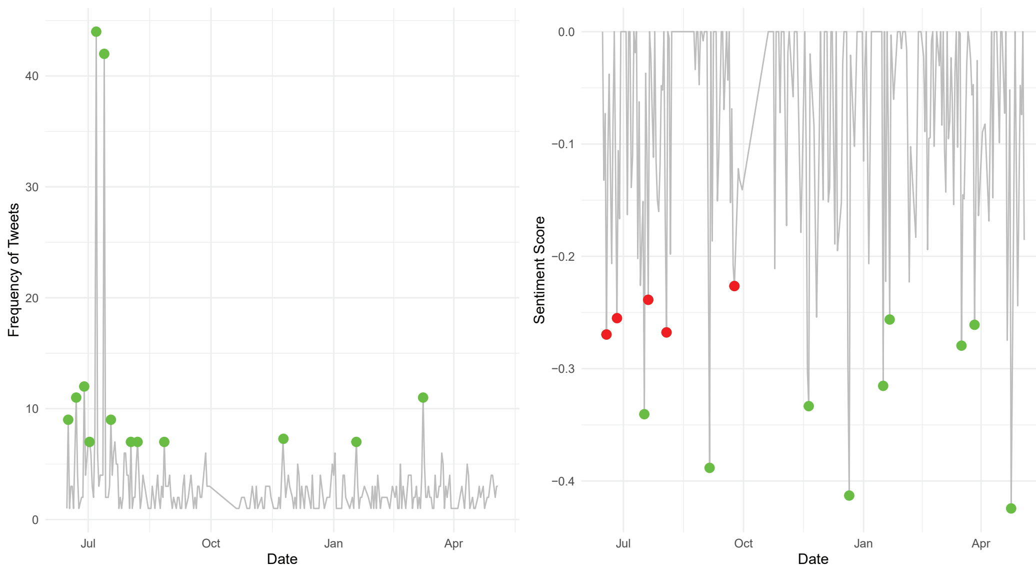

Figure 9: S-H-ESD anomaly detection for the frequency of tweets (left) and sentiments scores (right) for the aspect of "Service Reliability". Here, green dots indicate correctly detected candidate dates as events while red dots show uneventful days which are incorrectly detected as events. A large number of tweets in July relates two major service terminations caused by a police operation and a system failure in 2017.

Anomaly Detection Method	Criteria	Sentiment Class	Precision	Recall	F-score
FSED	-	**Negative**	**0.78**	**0.57**	**0.66**
FSED	-	Positive	0.79	0.5	0.61
S-H-ESD	**FUS**	**Positive**	**0.81**	**0.26**	**0.39**
S-H-ESD	**F**	**-**	**1**	**0.18**	**0.31**
ARIMA	FUS	Negative	1	0.14	0.24
ARIMA	F	-	1	0.13	0.23

Table 9: Summary of the event detection approaches for the aspect of "Service Reliability". Due to the limited space, only the top-six approaches with higher F-score are presented in this table.

As Table 9 illustrates, leveraging sentiment analysis in time-series-based event detection can increase the overall recall scores in the aspect of "Service Reliability". Results suggest that combining frequency-based and sentiment-based solutions increases the recall score of S-H-ESD by 8%, while the precision score slightly decreases but remains relatively acceptable. This decrease can be justified by the fact that sentiment-based event detection approaches observe negative or positive peaks in sentiments scores on a daily scale. However, daily tweets corresponding to each aspect of SQ may not reflect a common topic, i.e., correspond to a specific SQ event. Thus, although such a day can be labelled as a candidate

date by the detection approach, it will not be considered as an eventful day in a real-world scenario. These observations can also be further investigated in Figure 9. As the figure shows, sentiment information clearly contributes to catching events which are missed by the frequency-based solutions. This figure also reveals that the majority of resulting false positives in the sentiment-based solution occurs around July. This pattern can be due to the effect of weather on travel-related mood and travel satisfaction consequently [16], where cold weather can affect the mood of public transport users and this can lead to exhibiting more negative emotions when facing similar SQ issues. Nonetheless, based on the F-score values, it can generally be argued that using sentiment information as an additional class of observations improves the model's ability to find more events of interest in the sparse dataset.

Similar results can be observed for the two other aspects. As Table 10 shows, combining sentiment-based and frequency-based event detection increases the recall value of S-H-ESD by 10%. Similarly, as Table 11 illustrates the recall value of Anomalize is increased by 10% when sentiment information is also considered in the event detection process. This research confirms the effectiveness of sentiment analysis in detecting events impacting SQ in a fine-grained geographic area as a complementary solution.

Anomaly Detection Method	Criteria	Sentiment Class	Precision	Recall	F-score
FSED	-	**Negative**	**0.78**	**0.7**	**0.74**
FSED	-	Positive	0.57	0.7	0.62
S-H-ESD	**FUS**	**Negative**	**0.67**	**0.4**	**0.5**
S-H-ESD	FUS	Positive	0.57	0.4	0.47
Anomalize	F	-	1	0.3	0.46
S-H-ESD	**F**	-	**0.57**	**0.3**	**0.39**

Table 10: Summary of the event detection approaches for the aspect of "Information". Due to the limited space, only the top-six approaches with higher F-score are presented in this table.

It is also observed that the proposed statistical combination method is significantly better than the other time-series-based approaches in both aspects. As it is shown in Table 9, the proposed method increases the recall of the frequency-based S-H-ESD by 40%. Moreover, FSED outperforms the union of frequency-based and sentiment-based S-H-ESD by a 32% increase in the recall score.

As can be seen in Table 10, where FSED increases the recall value of a frequency-based S-H-ESD by 40%. This method also outperforms the union of frequency-based and sentiment-based S-H-ESD by 30% in both sentiment classes. Aligned with other two aspects, Table 11 illustrates FSED increases the recall value of frequency-based Anomalize by 19% for both sentiment classes, while the method offers a meaningful trade-off between precision and recall value of the event detection. The reason can be that, due to the sparsity of data, it is generally more difficult to robustly find significant changes between pairs of records in frequency or sentiment time-series. As a result, our statistical approach outperforms common time-series-based event detection approaches in the case of recall and F-score. It can also be observed that there is an acceptable trade-off between recall and precision of FSED, hence implying that FSED can improve the sensitivity of event detection approach for longer, shorter or regular events.

Finally, the sentiment class of best approaches in these aspects shows that there is a correlation between negative sentiments scores and higher F-score in the detection of SQ-related events. This finding is aligned with the findings of Thelwall et al. [51] who highlighted the correlation between important events on Twitter and negative sentiment score.

Anomaly Detection Method	Criteria	Sentiment Class	Precision	Recall	F-score
FSED	-	**Negative**	**0.54**	**0.64**	**0.59**
FSED	-	Positive	0.5	0.64	0.56
Anomalize	**FUS**	**Positive**	**0.4**	**0.54**	**0.46**
Anomalize	FUS	Negative	0.37	0.54	0.44
Anomalize	**F**	**-**	**0.38**	**0.45**	**0.41**
LSTM	FUS	Positive	0.67	0.18	0.28

Table 11: Summary of the event detection approaches for the aspect of "Comfort". Due to the limited space, only the top-six approaches with higher F-score are presented in this table.

6 Conclusion and future work

In this paper, a novel approach for detecting events for monitoring a fine-grained public transport SQ is provided. In the proposed approach, a state-of-the-art language model, namely BERT, is employed to bring extra semantic features into the process of extracting aspects in event detection, which increases the variability and reduce the bias in the classifier. Moreover, the potential of using sentiment analysis as another class of observation to improve the effectiveness of event detection in a limited context is highlighted. Finally, a statistical approach called FSED is presented to combine and integrate event detection solutions based on sentiment and frequency.

Experiments on a real-world dataset indicate the limitations of data and how BERT manages to handle the aspect extraction task, whereas topic modeling or word-embedding approaches, as common solutions for handling this task, fail in the limited context of the dataset. Our evaluations also prove that FSED can significantly improve the sensitivity of SQ event detection from tweets for longer, shorter or regular events. Moreover, results confirm the correlation between negative sentiments scores and improved detection of events affecting SQ.

Although the proposed approach with current configuration may not be able to reflect events from other contexts, for example, public transport in a developing country, training the aspect extraction model in one station and testing it on another station reflects the potential of the proposed method in addressing the problem of model transferability from one geographic region to another [62]. This aligns with the demonstrated potential of BERT in transfer learning, where an effective recipe is to fine-tune models with other datasets from different contexts [13].

Further steps can be improving the performance of the proposed approach by fine-tuning the aspect extraction model by a more number of tweets with more diversity of locations and context, which may increase the effectiveness of the method for aspects with smaller support. Such a model can be used for real-time monitoring of SQ in confined spaces of public transport. Moreover, methods will be developed for learning patterns in dynamic events, when events can travel through the public transport network.

Acknowledgments

The authors acknowledge assistance and advice from Amir Khodabandeh on statistical analysis.

References

[1] AARON, S., HICKMAN, T.-T., RAY, S., WRIGHT, A., AND MCEVOY, D. S. Using statistical anomaly detection models to find clinical decision support malfunctions. *Journal of the American Medical Informatics Association 25*, 7 (5 2018), 862–871. doi:10.1093/jamia/ocy041.

[2] AIELLO, L. M., PETKOS, G., MARTIN, C., CORNEY, D., PAPADOPOULOS, S., SKRABA, R., GOKER, A., KOMPATSIARIS, I., AND JAIMES, A. Sensing trending topics in twitter. *IEEE Transactions on Multimedia 15*, 6 (10 2013), 1268–1282. doi:10.1109/TMM.2013.2265080.

[3] ARGYROU, A., GIANNOULAKIS, S., AND TSAPATSOULIS, N. Topic modelling on Instagram hashtags: An alternative way to Automatic Image Annotation? In *Proceedings - 13th International Workshop on Semantic and Social Media Adaptation and Personalization, SMAP 2018* (9 2018), IEEE, pp. 61–67. doi:10.1109/SMAP.2018.8501887.

[4] BADJATIYA, P., GUPTA, S., GUPTA, M., AND VARMA, V. Deep Learning for Hate Speech Detection in Tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion* (Geneva, Switzerland, 2017), pp. 759–760. doi:10.1145/3041021.3054223.

[5] BISCHOF, J. M., AND AIROLDI, E. M. Summarizing topical content with word frequency and exclusivity. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012* (Edinburgh, Scotland, 2012), vol. 1, Omnipress, pp. 201–208.

[6] BLEI, D. M., NG, A. Y., AND JORDAN, M. I. Latent Dirichlet allocation. *Journal of Machine Learning Research 3*, 4-5 (2003), 993–1022. doi:10.1162/jmlr.2003.3.4-5.993.

[7] CHARTE, F., RIVERA, A. J., DEL JESUS, M. J., AND HERRERA, F. MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation. *Knowledge-Based Systems 89* (11 2015), 385–397. doi:10.1016/j.knosys.2015.07.019.

[8] CHAWLA, N. V., JAPKOWICZ, N., AND KOTCZ, A. Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter - Special issue on learning from imbalanced datasets 6*, 1 (6 2004), 1–6. doi:10.1145/1007730.1007733.

[9] CHOU, P.-F., LU, C.-S., AND CHANG, Y.-H. Effects of service quality and customer satisfaction on customer loyalty in high-speed rail services in Taiwan. *Transportmetrica A: Transport Science 10*, 10 (11 2014), 917–945. doi:10.1080/23249935.2014.915247.

[10] DAVIS, R. A., AND DUNSMUIR, W. T. M. State Space Models for Count Time Series. In *Handbook of Discrete-Valued Time Series*. Chapman and Hall/CRC, 1 2016, ch. 6, pp. 121–144. doi:10.1201/b19485.

[11] DE OÑA, J., AND DE OÑA, R. Quality of Service in Public Transport Based on Customer Satisfaction Surveys: A Review and Assessment of Methodological Approaches. *Transportation Science 49*, 3 (8 2015), 605–622. doi:10.1287/trsc.2014.0544.

[12] DE OÑA, J., DE OÑA, R., AND CALVO, F. J. A classification tree approach to identify key factors of transit service quality. *Expert Systems with Applications 39*, 12 (9 2012), 11164–11171. doi:10.1016/J.ESWA.2012.03.037.

[13] DEVLIN, J., CHANG, M.-W., LEE, K., AND TOUTANOVA, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Minneapolis, Minnesota, 6 2019), Association for Computational Linguistics, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[14] EBOLI, L., AND MAZZULLA, G. Structural Equation Modelling for Analysing Passengers' Perceptions about Railway Services. *Procedia - Social and Behavioral Sciences 54*, 1 (10 2012), 96–106. doi:10.1016/j.sbspro.2012.09.729.

[15] EBOLI, L., AND MAZZULLA, G. Relationships between rail passengers' satisfaction and service quality: a framework for identifying key service factors. *Public Transport 7*, 2 (2015), 185–201. doi:10.1007/s12469-014-0096-x.

[16] ETTEMA, D., FRIMAN, M., OLSSON, L., AND GÄRLING, T. Season and Weather Effects on Travel-Related Mood and Travel Satisfaction. *Frontiers in Psychology 8* (2017), 140. doi:10.3389/fpsyg.2017.00140.

[17] FAWCETT, T. An introduction to ROC analysis. *Pattern Recognition Letters 27*, 8 (6 2006), 861–874. doi:10.1016/J.PATREC.2005.10.010.

[18] GRAHAM, M., HALE, S. A., AND GAFFNEY, D. Where in the World Are You? Geolocation and Language Identification in Twitter. *The Professional Geographer 66*, 4 (10 2014), 568–578. doi:10.1080/00330124.2014.907699.

[19] HAGHIGHI, N. N., LIU, X. C., WEI, R., LI, W., AND SHAO, H. Using Twitter data for transit performance assessment: a framework for evaluating transit riders' opinions about quality of service. *Public Transport 10*, 2 (2018), 363–377. doi:10.1007/s12469-018-0184-4.

[20] HASAN, M., ORGUN, M. A., AND SCHWITTER, R. A survey on real-time event detection from the Twitter data stream. *Journal of Information Science 44*, 4 (2017), 443–463. doi:10.1177/0165551517698564.

[21] HASAN, M., ORGUN, M. A., AND SCHWITTER, R. Real-time event detection from the Twitter data stream using the TwitterNews+ Framework. *Information Processing & Management 56*, 3 (5 2019), 1146–1165. doi:10.1016/j.ipm.2018.03.001.

[22] HOONLOR, A., SZYMANSKI, B. K., AND ZAKI, M. J. Trends in computer science research. *Communications of the ACM 56*, 10 (2013), 74–83. doi:10.1145/2500892.

[23] HU, L., ZHANG, B., HOU, L., AND LI, J. Adaptive online event detection in news streams. *Knowledge-Based Systems 138* (12 2017), 105–112. doi:10.1016/j.knosys.2017.09.039.

[24] HUANG, Y., SHEN, C., AND LI, T. Event summarization for sports games using twitter streams. *World Wide Web 21*, 3 (5 2018), 609–627. doi:10.1007/s11280-017-0477-6.

[25] IKORO, V., SHARMINA, M., MALIK, K., AND BATISTA-NAVARRO, R. Analyzing Sentiments Expressed on Twitter by UK Energy Company Consumers. In *Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)* (2018), pp. 95–98. doi:10.1109/SNAMS.2018.8554619.

[26] KHARDE, V., AND SONAWANE, S. Sentiment Analysis of Twitter Data: A Survey of Techniques. *International Journal of Computer Applications 139*, 11 (1 2016), 5–15. doi:10.5120/ijca2016908625.

[27] KHONGSRABUT, I., AND WAIYAMAI, K. Outliers detection in time series data: Case study: Provincial waterworks authority. In *ECTI DAMT-NCON 2019 - 4th International Conference on Digital Arts, Media and Technology and 2nd ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering* (4 2019), Institute of Electrical and Electronics Engineers Inc., pp. 234–238. doi:10.1109/ECTI-NCON.2019.8692257.

[28] LEE, C., FAMOYE, F., AND ALZAATREH, A. Y. Methods for generating families of univariate continuous distributions in the recent decades. *WIREs Computational Statistics 5*, 3 (5 2013), 219–238. doi:10.1002/wics.1255.

[29] LEMAITRE, G., NOGUEIRA, F., AND ARIDAS, C. K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *The Journal of Machine Learning Research 18*, 1 (9 2016), 559–563. doi:10.5555/3122009.3122026.

[30] LENC, L., AND KRÁL, P. Deep neural networks for Czech multi-label document classification. In *Computational Linguistics and Intelligent Text Processing* (Cham, 2018), A. Gelbukh, Ed., vol. 9624 LNCS, Springer International Publishing, pp. 460–471. doi:10.1007/978-3-319-75487-1_36.

[31] LI, C., WANG, H., ZHANG, Z., SUN, A., AND MA, Z. Topic modeling for short texts with auxiliary word embeddings. In *SIGIR 2016 - Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, New York, USA, 2016), ACM Press, pp. 165–174. doi:10.1145/2911451.2911499.

[32] MARCUS, A., BERNSTEIN, M. S., BADAR, O., KARGER, D. R., MADDEN, S., AND MILLER, R. C. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada, 2011), ACM, p. 227. doi:10.1145/1978942.1978975.

[33] MCCALLUM, A. K. MALLET: A Machine Learning for Language Toolkit., 2002.

[34] MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G., AND DEAN, J. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* (10 2013), Curran Associates, Inc., pp. 3111–3119.

[35] MIMNO, D., WALLACH, H. M., TALLEY, E., LEENDERS, M., AND MCCALLUM, A. Optimizing semantic coherence in topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2011), 262–272.

[36] MONSUUR, F., ENOCH, M., QUDDUS, M., AND MEEK, S. Impact of Train and Station Types on Perceived Quality of Rail Service. *Transportation Research Record: Journal of the Transportation Research Board 2648*, 1 (2017), 51–59. doi:10.3141/2648-06.

[37] NGUYEN, T., PHUNG, D., ADAMS, B., AND VENKATESH, S. Event extraction using behaviors of sentiment signals and burst structure in social media. *Knowledge and Information Systems 37*, 2 (11 2013), 279–304. doi:10.1007/s10115-012-0494-9.

[38] NGUYEN, T. V., NGUYEN, A. T., PHAN, H. D., NGUYEN, T. D., AND NGUYEN, T. N. Combining Word2Vec with revised vector space model for better code retrieval. In *Proceedings - 2017 IEEE/ACM 39th International Conference on Software Engineering Companion, ICSE-C 2017* (Buenos Aires, Argentina, 6 2017), Institute of Electrical and Electronics Engineers Inc., pp. 183–185. doi:10.1109/ICSE-C.2017.90.

[39] PALTOGLOU, G. Sentiment-based event detection in Twitter. *Journal of the Association for Information Science and Technology 67*, 7 (2016), 1576–1587. doi:10.1002/asi.23465.

[40] PAVLOPOULOS, I. *Aspect Based Sentiment Analysis*. PhD thesis, Athens University of Economics and Business, Athens, Greece, 2014.

[41] PETROVI, S., OSBORNE, M., AND LAVRENKO, V. Streaming First Story Detection with application to Twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (Stroudsburg, PA, USA, 2010), HLT '10, Association for Computational Linguistics, pp. 181–189.

[42] PONTIKI, M., GALANIS, D., PAPAGEORGIOU, H., MANANDHAR, S., AND ANDROUTSOPOULOS, I. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (6 2015), Association for Computational Linguistics (ACL), pp. 486–495. doi:10.18653/v1/s15-2082.

[43] POPESCU, A.-M., AND PENNACCHIOTTI, M. Detecting controversial events from twitter. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (New York, NY, USA, 2010), ACM, p. 1873. doi:10.1145/1871437.1871751.

[44] PRABHAKARAN, V., ARORA, A., AND RAMBOW, O. Staying on Topic: An Indicator of Power in Political Debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Stroudsburg, PA, USA, 2014), Association for Computational Linguistics, pp. 1481–1486. doi:10.3115/v1/D14-1157.

[45] RAHIMI, M. M., NAGHIZADE, E., WINTER, S., AND STEVENSON, M. The Effectiveness of Sentiment Analysis for Detecting Fine-grained Service Quality. In *GeoComputation 2019* (Queenstown, New Zealand, 2019), University of Auckland. doi:10.17608/k6.auckland.9848132.v2.

[46] REHUREK, R., AND SOJKA, P. Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (2010), 45–50. doi:10.13140/2.1.2393.1847.

[47] RINKER, T. W. SentimentR: Calculate Text Polarity Sentiment, 2017.

[48] ROBERTS, M. E., STEWART, B. M., TINGLEY, D., LUCAS, C., LEDER-LUIS, J., GADARIAN, S. K., ALBERTSON, B., AND RAND, D. G. Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science 58*, 4 (10 2014), 1064–1082. doi:10.1111/ajps.12103.

[49] SAEIDI, M., BOUCHARD, G., LIAKATA, M., AND RIEDEL, S. SentiHood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. In *26th International Conference on Computational Linguistics* (Osaka, Japan, 12 2016), The COLING 2016 Organizing Committee, pp. 1546–1556.

[50] STILO, G., AND VELARDI, P. Efficient temporal mining of micro-blog texts and its application to event discovery. *Data Mining and Knowledge Discovery 30*, 2 (3 2016), 372–402. doi:10.1007/s10618-015-0412-3.

[51] THELWALL, M., BUCKLEY, K., AND PALTOGLOU, G. Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology 62*, 2 (2011), 406–418. doi:10.1002/asi.21462.

[52] TONON, A., CUDRÉ-MAUROUX, P., BLARER, A., LENDERS, V., AND MOTIK, B. ArmaTweet: Detecting Events by Semantic Tweet Analysis. In *The Semantic Web*, vol. 10250 LNCS. Springer International Publishing, Cham, 2017, pp. 138–153. doi:10.1007/978-3-319-58451-5_10.

[53] WEI, H., ZHOU, H., SANKARANARAYANAN, J., SENGUPTA, S., AND SAMET, H. Detecting latest local events from geotagged tweet streams. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (2018), ACM, pp. 520–523. doi:10.1145/3274895.3274977.

[54] WEI, H., ZHOU, H., SANKARANARAYANAN, J., SENGUPTA, S., AND SAMET, H. DeLLe: Detecting Latest Local Events from Geotagged Tweets. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Analytics for Local Events and News - LENS'19* (New York, New York, USA, 2019), ACM Press, pp. 1–10. doi:10.1145/3356473.3365188.

[55] WEISSMAN, G. E., UNGAR, L. H., HARHAY, M. O., COURTRIGHT, K. R., AND HALPERN, S. D. Construct validity of six sentiment analysis methods in the text of encounter notes of patients with critical illness. *Journal of Biomedical Informatics 89* (2019), 114–121. doi:10.1016/j.jbi.2018.12.001.

[56] WU, Y., SCHUSTER, M., CHEN, Z., LE, Q. V., NOROUZI, M., MACHEREY, W., KRIKUN, M., CAO, Y., GAO, Q., MACHEREY, K., KLINGNER, J., SHAH, A., JOHNSON, M., LIU, X., KAISER, L., GOUWS, S., KATO, Y., KUDO, T., KAZAWA, H., STEVENS, K., KURIAN, G., PATIL, N., WANG, W., YOUNG, C., SMITH, J., RIESA, J., RUDNICK, A., VINYALS, O., CORRADO, G., HUGHES, M., AND DEAN, J. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144.* (9 2016).

[57] XIAOMEI, Z., JING, Y., AND JIANPEI, Z. Sentiment-based and hashtag-based Chinese online bursty event detection. *Multimedia Tools and Applications 77*, 16 (8 2018), 21725–21750. doi:10.1007/s11042-017-5531-y.

[58] XIE, W., ZHU, F., JIANG, J., LIM, E., AND WANG, K. TopicSketch: Real-Time Bursty Topic Detection from Twitter. *IEEE Transactions on Knowledge and Data Engineering 28*, 8 (2016), 2216–2229. doi:10.1109/TKDE.2016.2556661.

[59] XU, H., LIU, B., SHU, L., AND YU, P. BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* (Stroudsburg, PA, USA, 4 2019), Association for Computational Linguistics, pp. 2324–2335. doi:10.18653/v1/N19-1242.

[60] YOU, Y., HUANG, G., CAO, J., CHEN, E., HE, J., ZHANG, Y., AND HU, L. GEAM: A general and event-related aspects model for Twitter event detection. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2013), vol. 8181 LNCS, pp. 319–332. doi:10.1007/978-3-642-41154-0_24.

[61] ZAHRA, K., IMRAN, M., AND OSTERMANN, F. O. Automatic identification of eyewitness messages on twitter during disasters. *Information Processing & Management 57*, 1 (1 2020), 102107. doi:10.1016/j.ipm.2019.102107.

[62] ZAHRA, K., OSTERMANN, F. O., AND PURVES, R. S. Geographic variability of Twitter usage characteristics during disaster events. *Geo-Spatial Information Science 20*, 3 (7 2017), 231–240. doi:10.1080/10095020.2017.1371903.

[63] ZHANG, C., LEI, D., YUAN, Q., ZHUANG, H., KAPLAN, L., WANG, S., AND HAN, J. GeoBurst+: Effective and Real-Time Local Event Detection in Geo-Tagged Tweet Streams. *ACM Trans. Intell. Syst. Technol. 9*, 3 (1 2018), 34:1–34:24. doi:10.1145/3066166.

[64] ZHANG, C., ZHOU, G., YUAN, Q., ZHUANG, H., ZHENG, Y., KAPLAN, L., WANG, S., AND HAN, J. GeoBurst: Real-time local event detection in geo-tagged tweet streams. In *SIGIR 2016 - Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (7 2016), Association for Computing Machinery, Inc, pp. 513–522. 10.1145/2911451.2911519.

[65] ZHOU, D., CHEN, L., AND HE, Y. An Unsupervised Framework of Exploring Events on Twitter: Filtering, Extraction and Categorization. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (Austin, Texas, USA, 2015).