The Effectiveness of Sentiment Analysis for Detecting Fine-grained Service Quality

M. M. Rahimi*[1], E. Naghizade[1], S. Winter[1], and M. Stevenson[2]

[1]Department of Infrastructure Engineering, The University of Melbourne, Australia

[2]Transport, Health, and Urban Design Research Hub, The University of Melbourne, Australia

*Email: mmrahimi@student.unimelb.edu.au

Abstract

Improving public transport relies on the development of appropriate tools for measuring and monitoring service quality. While there are several studies exploring perceived service quality, these works are mainly based on *surveys*, which are limited to a pre-defined period and do not enable to constantly observe and analyze the respond of passengers to different events. As a result, they may miss events affecting service quality in a transport hub. In contrast, *social media feeds* provide an opportunity to constantly look for events that may affect the service quality. However, in a confined geographic area like a transport hub, the sparsity of data considerably reduce the effectiveness of straight-forward event detection methods. In contrast to earlier works and in order to face these challenges, in this paper, we explore the effectiveness of sentiment analysis to detect events impacting perceived service quality in a confined location. Southern Cross Station, a major transport hub in Melbourne, Australia, is selected as a case study. Moreover, a statistical approach to combine and integrate sentiment-based and frequency-based event detection is proposed. Main findings confirm the effectiveness of sentiment analysis for detecting events affecting service quality in a transport hub. Additionally, the results prove that the proposed method can improve the sensitivity of event detection for events affecting service quality.

Keywords: Sentiment Analysis, Event Detection, Time-series analysis, Statistical analysis.

1 Introduction

Public transportation can be considered as the most efficient way of transferring a large number of people, especially in crowded urban areas. One of the most important components of public transportation infrastructure is a transport hub, which can be defined as a transit link for transferring passengers between different modes of transportation (e.g. railway, aviation, tram, bus) (Zhong et al., 2017). On a typical day, a large number of passengers may use the transport hub. As a result, continues assessment of passengers experiences and their perceived Service Quality (SQ) in a transport hub is critical to promote the usage of the services and increase public transport share (Eboli and Mazzulla, 2012).

*Corresponding author

While there are several studies exploring perceived SQ, these works are largely based on surveys (Eboli and Mazzulla, 2012, 2015; Monsuur et al., 2017; de Oña et al., 2014b,a, 2015). However, surveys are mainly limited to a pre-defined period and do not enable us to constantly observe and analyze the respond of passengers to different events. Therefore, they may miss longer, shorter or regular events (e.g. delays, termination, overcrowding) which cause an abrupt change in passengers' perception about the received SQ in a transport hub. To mitigate this limitation, event detection approaches can be applied on social media feeds to detect these events impacting transport hub SQ.

During the last decade, numerous studies (Marcus et al., 2011; Xie et al., 2016; Wei et al., 2018; Cai et al., 2016; Hasan et al., 2019) have been conducted on developing different methods for event detection from social media feeds. Most of these approaches are based on finding unusual changes in the frequency of terms (Paltoglou, 2016). While frequency-based approaches seem efficient in general-purpose applications, in case of quality assessment of transport hub services, they face two challenges. First, the geographic extent covered by a transport hub is significantly smaller than the ones considered in past studies (i.e. global, country or city scale). This leads to a sparse dataset, which will reduce the effectiveness of event detection approaches. Second, detected events by common frequency-based approaches are not necessarily related to the perceived SQ of transport hub. Consequently, in this case, the sensitivity of event detection tool will be reduced. Leveraging passengers' sentiment information, as another class of observations, can be used to enrich sparse datasets and improve the sensitivity of detection as an intuitive feature.

In this research, we aim to examine the effectiveness of sentiment analysis as a complementary solution for identifying events affecting the SQ in a transport hub. The main contributions in this article are summarized as follows:

First, sentiment information, as an additional class of observations, is leveraged to detect events affecting SQ in a fine-grained geographic area like a transport hub. In this case, daily sentiment scores are classified and aggregated to three discrete classes (i.e. negative, positive and overall score). Then, an investigation is made to determine which class will have better effectiveness of detection of events of interest.

Second, a novel Frequency and Sentiment-based Event Detection (FSED) approach is proposed. FSED is a hybrid statistical solution which could combine and integrate sentiment and frequency-based event detection methods in order to overcome existing challenges and detect events of interest in a fine-grained geographic area like a transport hub. Findings are supported by empirical evidence extracted from a large real social media dataset. Results of FSED are compared with three time-series-based event detection methods. Findings confirm the effectiveness of sentiment analysis for detecting events affecting SQ in a fine-grained geographic area like a transport hub. Moreover, FSED improves the sensitivity of event detection successfully.

The rest of the paper is organized as follows. Section 2 provides an overview of recent related works. Section 3 introduces our dataset and methodology. Then, Section 4 discusses our results and findings. Finally, Section 5 concludes and summarizes the paper.

2 Related Works

During the last decade, few studies have employed the sentiment analysis for minor and major event detection. Popescu and Pennacchiotti (2010) utilized sentiments, along with other linguistic

and structural features, in order to detect controversial events that provoked public discussion on Twitter. A 7590 sentiment lexicon was employed to measure polarity strength score associated with each snapshot. Findings of this research revealed the effectiveness of using sentiment analysis as a factor in event detection model.

Marcus et al. (2011) developed a system to visualize and track events using a graphical interface. They also employed an analysis engine and a Naive Bayes-based classifier to explore and aggregate crowd sentiments about those events. The authors found that while the sentiment analysis engine was working correctly, its polarity did not necessarily reflect general feeling regarding an event. Nevertheless, they did not investigate the relationship between sentiment strength and real-world events. In contrast, Thelwall et al. (2011) discussed that while positive sentiments can also be used for event detection, the effectiveness of negative sentiments shows better results.

In contrast to Thelwall et al. (2011), Paltoglou (2016) showed that both negative and positive sentiments can provide robust results. Moreover, this study provided a comparison between frequency-based and sentiment-based solutions for event detection with data with different sample sizes. The author argued that sentiment-based solutions can have unique advantages compared to other frequency-based solutions in specific environments where data collection has been done by keyword-based queries. While this study revealed the potential of using sentiment analysis as a solution for event detection on Twitter, it did not discuss how to combine these solutions to improve the precision of the detection method.

Similarly, Nguyen et al. (2013) used sentiments as a feature in the process of real-world event detection. They considered each human-being as a sensor. Consequently, sentiments can be considered as measurements of the sensor. Using a psychological behavior framework, a novel temporal sentiment indexing method and a simple anomaly detection approach based on time-series analysis, they proposed a method for extracting important events.

Recently, Xiaomei et al. (2018) proposed a corpus-based sentiment analysis approach to detect breaking events in micro-blog streams. In this model, after constructing a corpus-based dictionary and running sentiment analysis, a classification and a burst detection method were used to identify important events. Then, hashtags were used to reveal the event description. The results of this study show the potential of using sentiment analysis as a solution for event detection. However, they missed the frequency of terms as an important source of data in event detection methods. Moreover, their method considers only tweets which contain hashtags, while a lot of important tweets do not contain any hashtags.

While these studies have revealed promising results and provided valuable information, to the best of our knowledge, none of them measured the sensitivity of the instrument – effectiveness of event detection using sentiment analysis for detection of longer, shorter, or regular events impacting SQ of public transportation. Another gap relates to the lack of an effective method to combine frequency-based and sentiment-based event detection approaches. Moreover, these studies have looked at public transport systems as a whole, while this research is interested in the SQ of a single transport hub. Due to the small geographic area which can lead to the sparsity of data, challenges constitute anomaly detection of sparse feed and effective combination of sentiment-based solutions with other features.

3 Event Detection Using Twitter Data

The aim of this research is to study the correlation between events detected using a geo-tagged Twitter dataset corresponding to perceived service quality in a fine-grained geographic area like a transport hub.

3.1 Data collection and data cleansing

In this research, we consider Southern Cross Station, a major transport hub in Melbourne, Australia, as a case study. First, a Twitter dataset comprising of more than 32 million tweets is collected. These tweets are posted within the Melbourne area from June 2017 to May 2018. This data is obtained from the Australian Urban Research Infrastructure Network (AURIN[1]). We used keywords and spatial proximity to the station to detect relevant tweets, which resulted in a total of 3456 tweets. We further performed a number of pre-processing steps including language filtering, removing the re-tweets and repetitive tweets and converting all emojis into textual equivalent words.

While keyword and geographic search approaches can be used to select tweets related to Southern Cross Station, only a fraction of these feeds are expected to be related to public transportation. To improve the effectiveness of our event detection, we utilize a transport-specific dictionary developed by Kuflik et al. (2017) to filter unrelated tweets. Finally, 1976 tweets are selected for this study.

3.2 Frequency-Based Event Detection

In the context of SQ, we define an event as anything causing an unusual change in passengers' perception of the received SQ. It is argued that these events can be detected using changes in the volume of social media feeds, the sentiments in these feeds, or both.

Since events happen rarely in our dataset, this approach assumes that each day can contain at most one SQ event (where this does not hold further topic modelling is required). In this case, term-frequency can be simplified to daily-frequency. First, the frequency of tweets at a time step τ is considered as F_τ. Figure 1 shows the variation of F_τ over the study period. Next, considering F_τ as time-series, anomaly detection methods can be applied to find outliers, where outliers are interpreted as candidate dates for events. Seasonal Hybrid Extreme Studentized Deviate (S-H-ESD), Auto-Regressive Integrated Moving Average (ARIMA) and Long Short-Term Memory Network (LSTM) are some of the state-of-the-art approaches considered in this paper (Tonon et al., 2017; Aaron et al., 2018; Badjatiya et al., 2017).

LSTM outperforms other frequency-based event detection methods considered in this paper, hence, we merely discuss this approach in this section. This study defines prediction error as:

$$E_\tau = F_\tau - \hat{F_\tau} \tag{1}$$

where $\hat{F_\tau}$ is the predicted value for F_τ. Then, a τ will be considered as a candidate if:

$$|E_\tau - \mu_{E_\tau}| < k * \sigma_{E_\tau} \tag{2}$$

[1] www.aurin.org.au

where k is a pre-defined threshold and μ_{E_τ} and σ_{E_τ} are the mean and standard deviation of E_τ, respectively. Similar to Wei et al. (2018), the threshold of anomaly detection is set to $k = 3$.

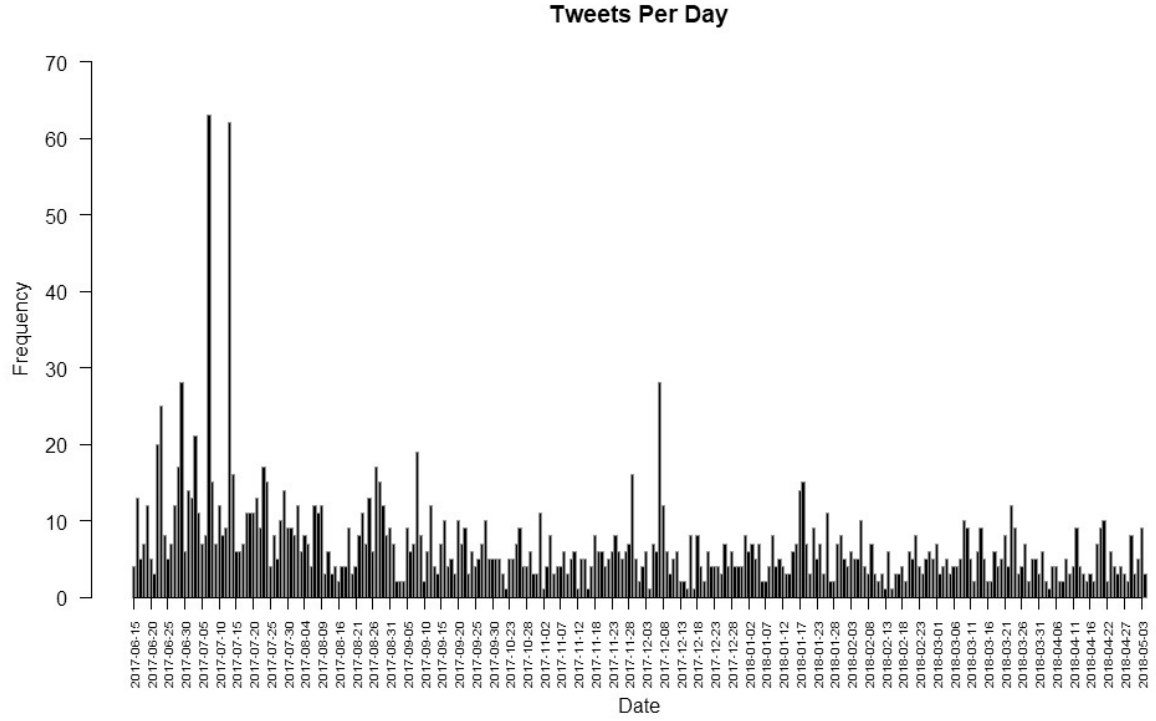

Figure 1: Frequency of tweets per day during the study period.

3.3 Sentiment Analysis

Sentiment analysis (or opinion mining) is a computational approach to understand the perception of authors of text about specific entities (Feldman, 2013). The aim of sentiment analysis is to measure positivity, negativity or neutrality of a text, which can be denoted as polarity.

In this research, we leverage a sentiment analysis approach named SentimentR (Rinker, 2017), which also has been used in other works (Ikoro et al., 2018; Weissman et al., 2019). SentimentR is chosen because the method leverages a dictionary look-up along with considering valence shifters (i.e. negators, amplifiers, deamplifiers and adversative conjunction). The dictionary comprises of 11709 words, where each individual score can take a value between -2 and 1 (Naldi, 2019). First, the sentiment score for each tweet is calculated. Then, the sentiments scores on a daily basis are averaged (S_τ). Figure 2 shows the variation of S_τ during the study period. In the process of daily aggregation, it is observed that, in many cases, tweets with negative and positive polarity can contain different topics. Consequently, their aggregation may neutralize daily sentiments (Figure 3). As a result, it is decided to classify and aggregate daily sentiment scores to three discrete classes (i.e. negative, positive and overall score) and investigate which class will have better effectiveness of the detection of SQ-related events.

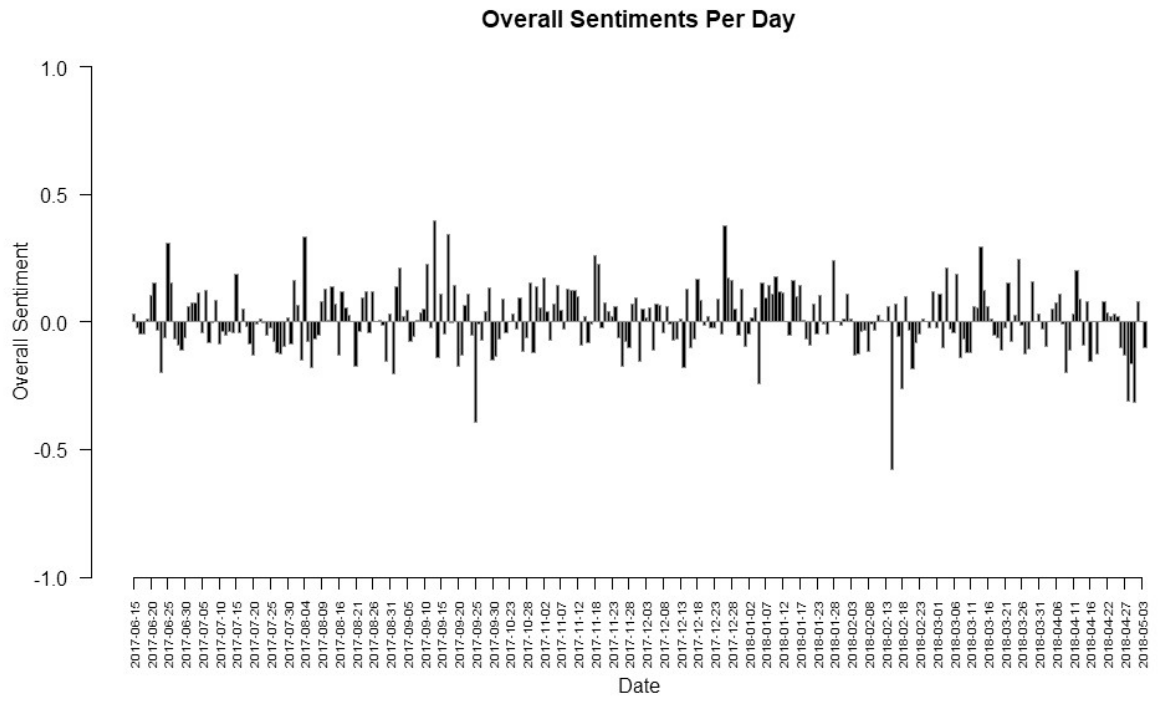

Figure 2: Overall sentiment variation for each day during the study period.

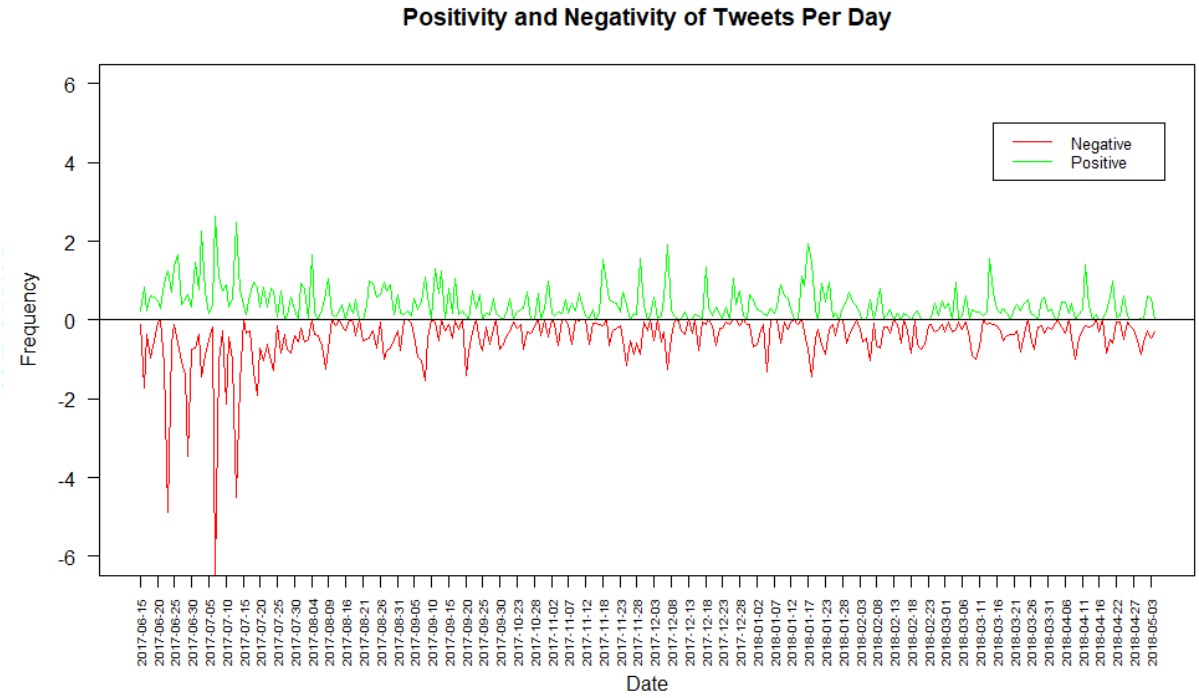

Figure 3: Variation of negative and positive sentiment classes for each day during the study period.

3.4 Frequency and Sentiment-based Event Detection (FSED)

Due to the sparsity of data, it was expected that general time-series-based approaches would not be able to effectively detect the events of our interest. To mitigate this limitation, in this research, we propose a Frequency and Sentiment-based Event Detection (FSED) approach for event detection. FSED utilizes a statistical approach to combine and integrate sentiment-based and frequency-based solutions. The idea is to use the probability distribution of frequency and sentiments of tweets. Using these distribution functions, the probability of obtaining each possible value for daily frequency and sentiment can be measured. By multiplying these probabilities, a single probability value is obtained for each day which is used to detect candidate events.

First, the best-fitting distribution function for the frequency data is investigated. As the frequency has a discrete distribution over time, negative binomial, Poisson and binomial distributions are considered. A Log-likelihood function, which is a measure of the goodness of fit, reveals that negative binomial has a better fitness compared to the others (Figure 4). Finally, the vector of probabilities is calculated using empirical values of probability of success and mean.

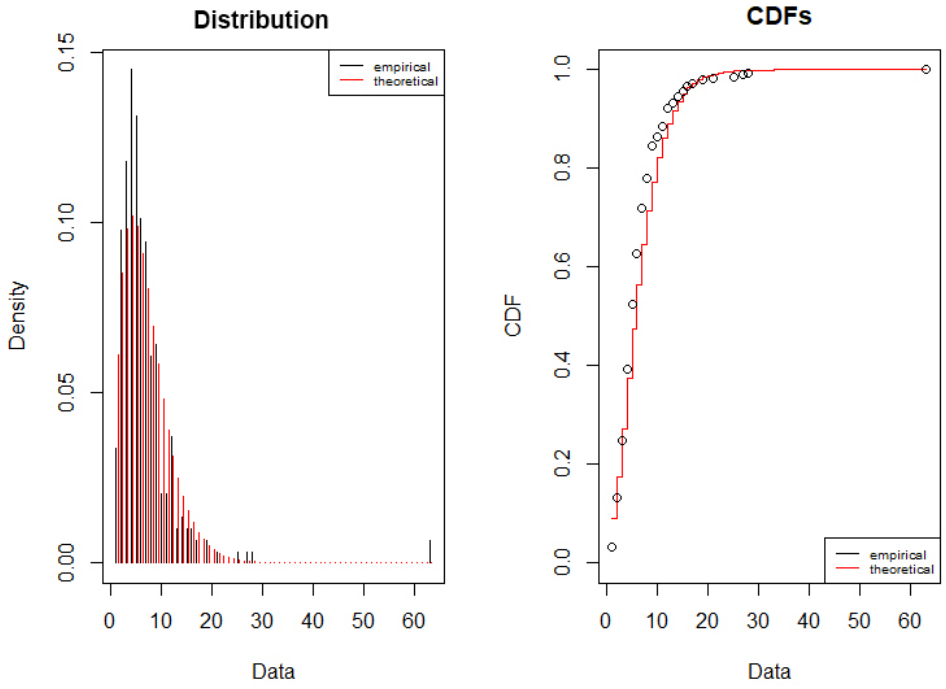

Figure 4: Empirical and theoretical distribution (left) and Cumulative Distribution Function (CDF) (right) of frequency data using negative binomial distribution.

Due to the continuous nature of sentiments, Normal, Exponential, Gamma, Lognormal and Weibull distribution functions are considered for different classes of sentiments. Then, based on the log-likelihood function, exponential distribution is chosen for the fitness. Consequently, the exponential distribution is successfully fitted to the negative class of sentiment scores (Figure 5). Last, the vector of probabilities is calculated for each class using empirical values for λ. Finally, these two

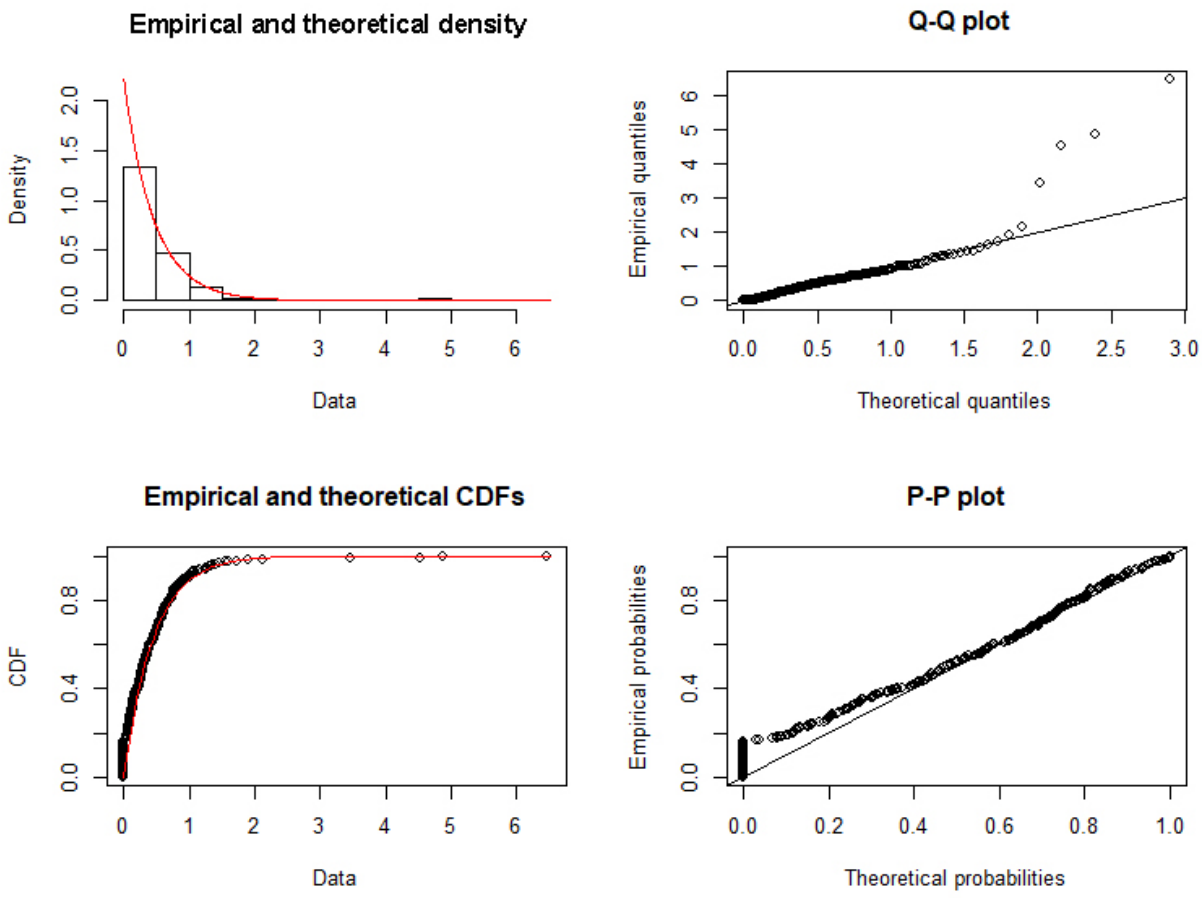

Figure 5: Empirical and theoretical densities (top-left), Q-Q plot (top-right), Cumulative Distribution Function (CDF) (bottom-left) and the P-P plot (bottom-right) of exponential distribution of negative class of sentiment scores.

solutions are combined by definition of the total probability distribution as:

$$P(F_\tau < \alpha, S_\tau < \alpha) = P(F_\tau < \alpha) * P(S_\tau < \alpha) \tag{3}$$

where α is the significance level and P represents the probability value of the corresponding variable. If the total probability of a day is smaller than the significance level, the day will be selected as a candidate for an event. Based on experiments and better effectiveness, we set $\alpha = 5\%$.

4 Results and Discussion

In this section, our proposed method is compared to three state-of-teh-art time-series anomaly detection approaches, namely S-H-ESD, ARIMA, LSTM (Tonon et al., 2017; Aaron et al., 2018; Badjatiya et al., 2017). For time-series-based approaches, the outlier detection is applied to sentiments scores (S) and frequency of tweets (F) to detect candidate events (Figure 6). Therefore, two

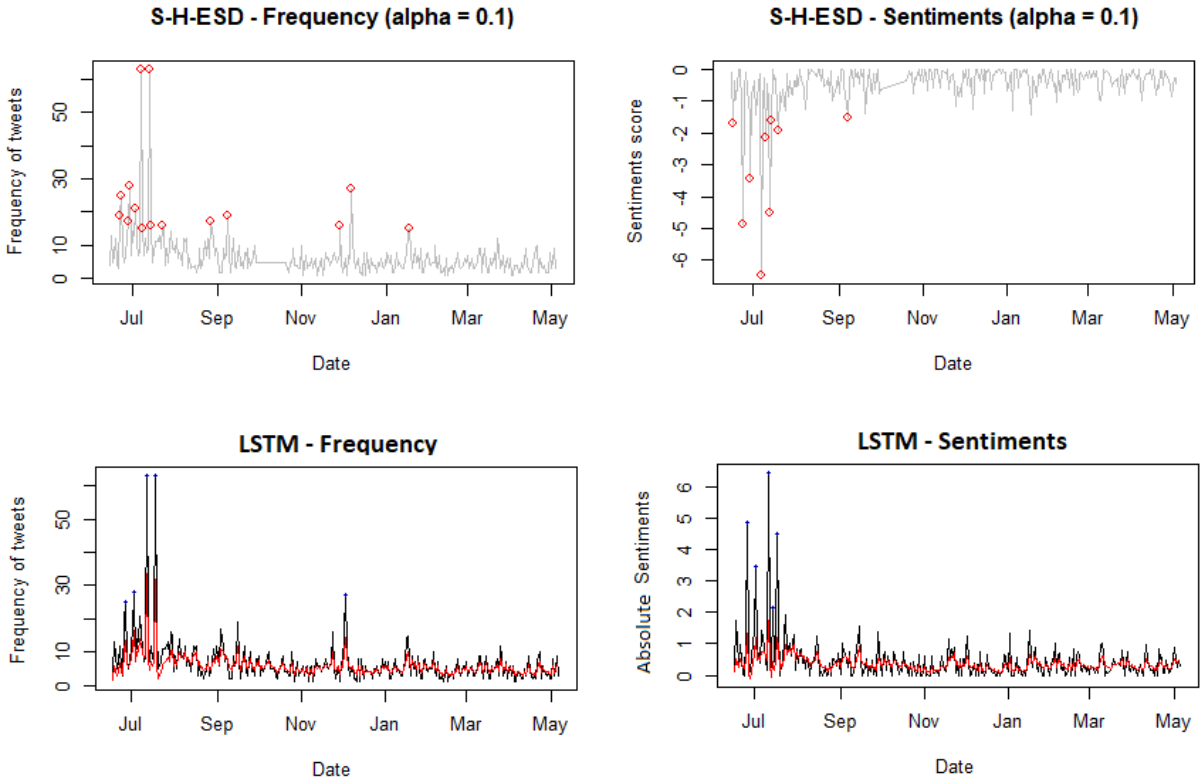

Figure 6: S-H-ESD anomaly detection for the frequency of tweets (top-left) and sentiments scores (top-right), LSTM anomaly detection for the frequency of tweets (bottom-left) and sentiments scores (bottom-right). The value of α is set based on experiments and better recall and precision values.

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.

To assess the effectiveness of these approaches, three well-established metrics in information retrieval, i.e., precision, recall, and F-measure, are used. First, days of our study period are manually

labeled to serve as our ground truth and then, the candidate events are compared against those labels. Table 1 compares the results of different event detection methods with regard to the sentiment class, anomaly detection method, and criteria.

As Table 1 illustrates, leveraging sentiment analysis in time-series-based event detection can increase the overall recall scores. Our results suggest that combining frequency-based and sentiment-based solutions increases the recall score of LSTM by 10%, while the precision score remained relatively stable at 63%. This can be because the new approach uses sentiment information as an additional class of observations, which will improve models ability to find more events of interest in the sparse dataset. The table also indicates the potential of using sentiment analysis as a plausible solution for event detection, as the sentiment-based LSTM outperforms the frequency-based LSTM by a 5% increase in F-Measure score. This analysis confirms the effectiveness of sentiment analysis for detecting events affecting SQ in a fine-grained geographic area.

It is also observed that the proposed statistical combination method is significantly better than the other time-series-based approaches. Our method increases the recall of the frequency-based LSTM by 33%. Moreover, FSED outperforms the union of frequency-based and sentiment-based LSTMs by a 25% increase in the recall score. 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 case of recall and F-measure. It can also be observed that there is an acceptable trade-off between recall and precision of FSED (60% approximately), hence implying that FSED can improve the sensitivity of event detection approach for longer, shorter or regular events.

Finally, the sentiment class of three best approaches shows that there is a correlation between negative sentiments scores and higher F-measure in the detection of SQ-related events. This finding is aligned with the findings of Thelwall et al. (2011) who highlighted the correlation between important events on Twitter and negative sentiment score.

Anomaly Detection Method	Criteria	Sentiment Class	Precision	Recall	F-Measure
FSED	**-**	**Negative**	**0.63**	**0.60**	**0.62**
LSTM	**F ∪ S**	**Negative**	**0.63**	**0.35**	**0.45**
LSTM	**S**	**Negative**	**0.71**	**0.31**	**0.43**
FSED	-	Positive	0.38	0.44	0.41
FSED	-	Overall	0.40	0.36	0.38
LSTM	**F**	**-**	**0.74**	**0.25**	**0.38**
LSTM	F ∪ S	Positive	0.41	0.35	0.38
LSTM	F ∩ S	Negative	0.87	0.24	0.37
S-H-ESD	F ∪ S	Negative	0.72	0.24	0.36
S-H-ESD	F ∪ S	Positive	0.54	0.25	0.35
S-H-ESD	F	-	0.80	0.22	0.34
LSTM	S	Positive	0.39	0.29	0.33
S-H-ESD	F ∪ S	Overall	0.71	0.22	0.33
LSTM	F ∪ S	Overall	0.35	0.31	0.33
LSTM	F ∩ S	Positive	0.83	0.18	0.30

Table 1: Summary of the event detection approaches. Due to the limited space, approaches with F-measure less than 30% are not presented (e.g. ARIMA).

5 Conclusion and Future Works

In this paper, we highlight the potential of using sentiment analysis as an independent observation to improve the effectiveness of event detection. A statistical approach, named FSED, is proposed to combine and integrate sentiment-based and frequency-based event detection on social media feeds. Our evaluation confirms that FSED can improve the sensitivity of instrument SQ event detection from tweets for longer, shorter or regular events. Our results also prove the correlation between negative sentiments scores and an improved detection of events affecting SQ. The authors see two main challenges for future work. First, to catch single but important SQ-related tweets, for which topic modeling methods can be applied. Second, applying a dynamic temporal window to improve the detection of overlapping events over time.

6 Acknowledgements

The authors acknowledge assistance and advice from Matthew Howe from Southern Cross Station Pty Ltd. They also thank Amir Khodabandeh (The University of Melbourne) for his advice on statistical analysis.

7 References

Aaron, S., T.-T. Hickman, S. Ray, A. Wright, and D. S. McEvoy
2018. Using statistical anomaly detection models to find clinical decision support malfunctions. *Journal of the American Medical Informatics Association*, 25(7):862–871.

Badjatiya, P., S. Gupta, M. Gupta, and V. Varma
2017. Deep Learning for Hate Speech Detection in Tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, Pp. 759–760, Geneva, Switzerland.

Cai, H., Y. Yang, X. Li, and Z. Huang
2016. What are Popular. In *Proceedings of the 23rd ACM International Conference on Multimedia*, Pp. 89–98, New York, NY, USA. ACM.

de Oña, J., R. de Oña, L. Eboli, and G. Mazzulla
2014a. Heterogeneity in Perceptions of Service Quality among Groups of Railway Passengers. *International Journal of Sustainable Transportation*, 9(8):612–626.

de Oña, R., L. Eboli, and G. Mazzulla
2014b. Key Factors Affecting Rail Service Quality in the Northern Italy: A Decision Tree Approach. *Transport*, 29(1):75–83.

de Oña, R., J. L. Machado, and J. de Oña
2015. Perceived Service Quality, Customer Satisfaction, and Behavioral Intentions: Structural Equation Model for the Metro of Seville, Spain. *Transportation Research Record*, 2538(1):76–85.

Eboli, L. and G. Mazzulla
2012. Structural Equation Modelling for Analysing Passengers Perceptions about Railway Services. *Procedia - Social and Behavioral Sciences*, 54(1):96–106.

Eboli, L. and G. Mazzulla
2015. Relationships between rail passengers satisfaction and service quality: a framework for identifying key service factors. *Public Transport*, 7(2):185–201.

Feldman, R.
2013. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82.

Hasan, M., M. A. Orgun, and R. Schwitter
2019. Real-time event detection from the Twitter data stream using the TwitterNews+ Framework. *Information Processing & Management*, 56(3):1146–1165.

Ikoro, V., M. Sharmina, K. Malik, and R. Batista-Navarro
2018. Analyzing Sentiments Expressed on Twitter by UK Energy Company Consumers. In *Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, Pp. 95–98.

Kuflik, T., E. Minkov, S. Nocera, S. Grant-Muller, A. Gal-Tzur, and I. Shoor
2017. Automating a framework to extract and analyse transport related social media content: The potential and the challenges. *Transportation Research Part C: Emerging Technologies*, 77:275–291.

Marcus, A., M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller
2011. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, P. 227, Vancouver, BC, Canada. ACM.

Monsuur, F., M. Enoch, M. Quddus, and S. Meek
2017. Impact of Train and Station Types on Perceived Quality of Rail Service. *Transportation Research Record: Journal of the Transportation Research Board*, 2648(1):51–59.

Naldi, M.
2019. A review of sentiment computation methods with R packages. *CoRR*, abs/1901.0.

Nguyen, T., D. Phung, B. Adams, and S. Venkatesh
2013. Event extraction using behaviors of sentiment signals and burst structure in social media. *Knowledge and Information Systems*, 37(2):279–304.

Paltoglou, G.
2016. Sentiment-based event detection in Twitter. *Journal of the Association for Information Science and Technology*, 67(7):1576–1587.

Popescu, A.-M. and M. Pennacchiotti
2010. Detecting controversial events from twitter. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, P. 1873, New York, NY, USA. ACM.

Rinker, T. W.
2017. SentimentR: Calculate Text Polarity Sentiment.

Thelwall, M., K. Buckley, and G. Paltoglou
2011. Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418.

Tonon, A., P. Cudré-Mauroux, A. Blarer, V. Lenders, and B. Motik
2017. ArmaTweet: Detecting Events by Semantic Tweet Analysis. In *The Semantic Web*, E. Blomqvist, D. Maynard, A. Gangemi, R. Hoekstra, P. Hitzler, and O. Hartig, eds., Pp. 138–153, Cham. Springer International Publishing.

Wei, H., H. Zhou, J. Sankaranarayanan, S. Sengupta, and H. Samet
2018. Detecting latest local events from geotagged tweet streams. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Pp. 520–523. ACM.

Weissman, G. E., L. H. Ungar, M. O. Harhay, K. R. Courtright, and S. D. Halpern
2019. Construct validity of six sentiment analysis methods in the text of encounter notes of patients with critical illness. *Journal of Biomedical Informatics*, 89:114–121.

Xiaomei, Z., Y. Jing, and Z. Jianpei
2018. Sentiment-based and hashtag-based Chinese online bursty event detection. *Multimedia Tools and Applications*, 77(16):21725–21750.

Xie, W., F. Zhu, J. Jiang, E. Lim, and K. Wang
2016. TopicSketch: Real-Time Bursty Topic Detection from Twitter. *IEEE Transactions on Knowledge and Data Engineering*, 28(8):2216–2229.

Zhong, G., X. Wan, J. Zhang, T. Yin, and B. Ran
2017. Characterizing passenger flow for a transportation hub based on mobile phone data. *IEEE Transactions on Intelligent Transportation Systems*, 18(6):1507–1518.