frontiers in Built Environment

PERSPECTIVE
published: 09 July 2018
doi: 10.3389/fbuil.2018.00030

Role of Pre-processing in Textual Data Fusion: Learn From the Croydon Tram Tragedy

Mohd H. Bin Osman[1,2]* *and Sakdirat Kaewunruen*[1]

[1] School of Civil Engineering, University of Birmingham, Birmingham, United Kingdom, [2] Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, Bangi, Malaysia

Tram/train derailment subject to human mistakes makes investments in an advanced control room as well as information gathering system exaggerated. A disaster in Croydon in year 2016 is recent evidence of limitation of the acquired systems to mitigate human shortcoming in disrupted circumstances. One intriguing way of resolution could be is to fuse continuous online textual data obtained from tram travelers and apply the information for early cautioning of risk discovery. This resolution conveys our consideration regarding a resource of data fusion. The focal subject of this paper is to discuss about role of pre-processing ventures in a low-level data fusion that have been distinguished as a pass to avoid time and exertion squandering amid information retrieval. Inclines in online text data pre-processing is reviewed which comes about an outline suggestion that concede traveler's responses through social media channels. The research outcome shows by a case of data fusion could go about as an impetus to railway industry to effectively partake in data exploration and information investigation.

Keywords: data fusion, text processing, social media, alert system, risk mitigation, disruption, tram accident, microsleep

OPEN ACCESS

Edited by:
Min An,
University of Salford, United Kingdom

Reviewed by:
Grigorios Fountas,
University at Buffalo, United States
Ivo Haladin,
Faculty of Civil Engineering, University of Zagreb, Croatia

***Correspondence:**
Mohd H. Bin Osman
mxo574@student.bham.ac.uk

Specialty section:
This article was submitted to Transportation and Transit Systems, a section of the journal Frontiers in Built Environment

Received: *02 February 2018*
Accepted: *12 June 2018*
Published: *09 July 2018*

Citation:
Bin Osman MH and Kaewunruen S (2018) Role of Pre-processing in Textual Data Fusion: Learn From the Croydon Tram Tragedy. Front. Built Environ. 4:30. doi: 10.3389/fbuil.2018.00030

INTRODUCTION

Data is enormous in railway industry, covering both train/tram operations and infrastructure management dimensions. Acknowledging many successful stories from outside railway domain associated with an adaptation of data-driven innovation, British railway infrastructure manager has launched the "Challenge Statements" program series which recognize data exploration/exploitation is one of their agenda (Network Rail, 2017). One element which has a great potential to accelerate the journey of worthwhile state is data fusion.

Data fusion is famously defined as (White, 1987): "A multi-process dealing with the association, correlation, combination of data and information from single and multiple sources to achieve refined position, identify estimates and complete and timely assessments of situations, threats and their significance." The definition grasps a spirit of encouragement and inspirations for data analysts to continuously conduct data exploration/exploitation to appropriately enrich the value of existing information about an object of interest. In regard to application-wise, data fusion is nowadays beyond the dominant and matured domain; remote sensor and signal processing, as case studies can be found in condition monitoring (Raheja et al., 2006), crime analysis (Nokhbeh Zaeem et al., 2017), forest management (Chen et al., 2005), and engineering (Steinberg, 2001). The key of data fusion being applied in diverse research domains is about the way the problem is formulated and the choice of methods (Hannah et al., 2000; Starr et al., 2002) recognize the capacity of identifying a parallel between problems under consideration with data fusion models is the key

to success. These influential factors drag our regard to the classical data fusion framework, known as the JDL (an acronym for Joint Directors of Laboratories) framework.

The JDL framework put major elements of data fusion definition into five successive steps (a pre-processing, object refinement, situation assessment, threat assessment, and process refinement) to aid the encouragements expressed in the definition in a systematic and efficient way. The framework could be conveniently attended in hierarchy sense upon user's capacity to accommodate the current problem at hand. For example, the low-level data fusion comprises of a pre-processing step is significant for a new application (or project) in a context of data refinement. Hence, this study is carried out to address design issues of low-level data fusion, especially pre-processing steps regarding textual data. The discoveries are relied upon to serve the future application of online text data-driven alerting system inspired from the tragedy of Croydon tram derailment.

On 9th November 2016, a commuter community particularly in London, was stunned when the unfortunate tram No. 2551 derailed and killed seven passengers. The tragedy can be delegated as a *black swan* in the clean track records of the tram operator company following 15 years of re-operations in London. According to early investigation report, a human-error is identified as a primary contributor to the derailment. The tram driver endured *black out* which prompt to an inability to reduce the tram speed as far as possible before entering the accident area. The report proclamation sparks our interest to discover a data-powered solution to mitigate the derailment risk root from human weakness.

Having text as data source creates a unique data pre-processing challenge, as it is unstructured data. Text appears in non-uniform length of words that does not reside in fixed row-column database. This causes a dimensionality in terms of words space of each text is vary from one sender to another which requires a robust algorithm for text classification. Consequently, assigning conventional pre-processing tools that are designed for structured data, to textual data might fail to optimally exploit hidden information.

In the following section, role of pre-processing steps in a data fusion model is disclosed to highlight adaptability of the steps for different research motives. An investigation report of the Croydon Tram tragedy was analyzed in parallel to author's viewpoint to underpinning a motivation of a passenger-participation warning system proposal. For a set of identified system requirements, a suitable design of online text pre-processing steps to be embedded in the system under purview is presented. A conclusion remark about synergy among low-level data fusion, text pre-processing and online data source is stated in the Conclusion section.

ROLE OF PRE-PROCESSING

Pre-processing is the lowest level activities in the process of combining input data from multiple sources to gather information in order to achieve inferences. Some authors call it as source processing due to high workload is applied on data source. Depending on the problem under consideration, data sources can be sensors, a prior knowledge, databases, or human input. The step is metaphorically viewed as a bridge connecting external elements with data fusion framework that must be traversed before subsequent data fusion steps, more complex in nature, can be performed. By imposing barriers to free-flow of data streaming, a compact representation of raw data is sufficient to produce brief but reliable decision making process (Eggers and Khuon, 1990). In addition, an efficiency and scalability of an object refinement; a subsequent step applied to processed data, can be improved through a proper pre-processing (Mitali et al., 2003).

However, this particular benefit is challenging to gain when online text as a data source. Online text is delivered from vast points of network to the servers at different arrival time and in various styles which is very informal in nature. Language dependent factors which do not have an impact to information retrieval are also identified obstacle (Singh and Kumari, 2016; Nokhbeh Zaeem et al., 2017) points out the challenge in dealing with social media text data for fortification of sentiment classification especially in terms of short length and internet slang word. In addition, an existence of uninformative text such as HTML tags, scripts, internet abbreviations and advertisements rises the computational complexity to an upper level as compared to a well-presented text (Petz et al., 2012; Dos Santos and Ladeira, 2014) underline the significant of having a language detector as an additional component to standard text pre-processing process in case of multilingual responses are eligible to a system of response. In (Hamouda and Ben Akaichi, 2013), major text processing issues when dealing with non-English data source was a topic of discussion.

CROYDON TRAM DERAILMENT

The Tragedy

More than 30 million passengers use the lime green-and-blue Croydon trams in 2013/2014 and this number is expected to increase to up to 60 million in 2030 (Transport for London, 2014). The projection is expected to face a small degradation in a period of December 2016 to February 2017 due to a fatal tram derailment which killed seven passengers and more than 51 suffered from various degree of injury. The tragedy took place at the first track junction (diverging track) which is located approximately 200 m from the Sandilands stop toward the Wimbledon station (refer to a star marker in **Figure 1**).

Motivation for Pursuing the Research

An earlier investigation conducted by U.K Rail Accident Investigation Branch found that the tram was speeding excessively when entering the junction (Rail Accident Investigation Branch, 2016). Theoretically, a significantly large difference in speed between traveling speed and the speed limit increases the risk of train or tram to over-tuning at a transition curve (Bearfield and Marsh, 2005). Routine tram passengers who survived the tragedy agree with the initial finding. The

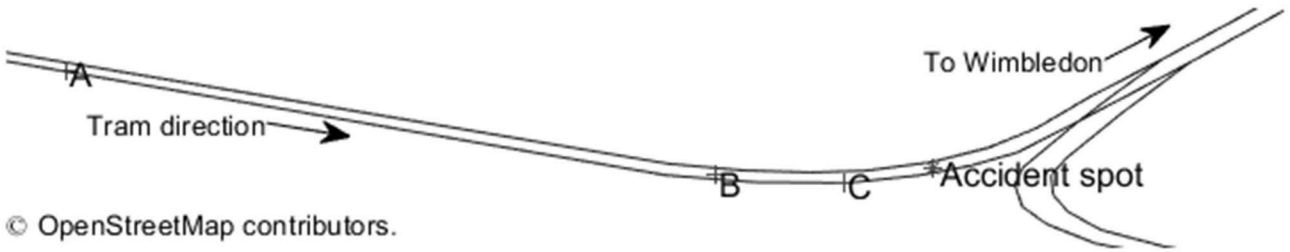

FIGURE 1 | Three points of interest to a maximum speed and braking distance analysis. The OpenStreetMap representation of the accident area is downloaded from www.openstreetmap.org.

passengers felt that the tram traveled at unusual speed along a tangent track- and are possibly, a track segment that connects a point A and B shown in **Figure 1** (Christodoulou, 2016). Their latter claim may make sense if the tram speed was over the speed restrictions applied to that section. Regardless of whether the London Arms, the Croydon Tramlink operator, implements a speed restriction to the tangent line or not, technically speaking, a tram needs to reduce an initial speed when approaching the point B. Failure to do so will cause a tram to make a sharp/hard turn at the transition curve, point C In extreme cases (i.e., when trams travel too fast), a tram derailment is possible to occur. One may query how an experienced tram driver was unable to control the over speeding tram. Unfortunately, the tram driver was reportedly suffering *black out* which could be grouped in the same category with explosion, sabotage and sinkhole as a disruption i.e., rare events but have high consequences.

Speaking about technology, Croydon Tramlink equips their network with a sophisticated computer-based system known as SCADA (Supervisory Control and Data Acquisition) (Parascandolo, 2007). Technically speaking, controllers in the control room should have no problem stopping or reducing the speed of any misbehavior/suspicious trams under their supervision remotely by reducing or cutting-off electric supply. Indeed, we believe the company has a standard procedure to trigger that action. In respect to the recent tram crash were over-speeding is greatly believed to be the cause of the tram derailment, the system or its associated system should detect an unusual power usage by the trams only seconds before the crash (particularly in the braking zone) unless the system is not programmed to execute that kind of proactive action. This situation motivates us to introduce a system patch which capable to warn a control office in the event of hazard detection. Interestingly, tram passengers participation and social media will be the core of the proposed data-driven innovation.

TEXTUAL DATA PRE-PROCESSING

Four components are proposed to increase quality of online text raw data for the use of an alerting system. **Figure 2** depicts their sequence in a basic data fusion model i.e., input, pre-processing and decision.

Input

The system is only function with a participation of tram passengers. Limited Wi-Fi access and/or traveler's phone apps should be provided at no cost to all passengers in order to channel the system with various sources of online textual data.

System users, who initially agree to give permission for the system to extract his/her online data, could feel their privacy is threatened when the exclusivity as an individual deteriorates. In this context, Van Wel and Royakkers (2004) describes such circumstance as "de-individualisation"—an occurrence when group profiles are often used to judge, treat and possibly discriminate people instead of gratifying them as per individual characteristics. However, there tends to exist a group of users who are voluntarily responsive toward deals and promotions offered by companies—such as fast speed internet connection, ticket discounts or meal vouchers— at the expense of their own data privacy. Surprisingly, even though they are informed of losing their own privacy, this doesn't stop them from blindly responding and agreeing to the so-called "online privacy-policies protection."

To reduce unnecessary stress related to online privacy protection between stakeholders of the system, privacy policies are normally equipped with an instrument used for online data collection, processing and storage (Van Wel and Royakkers, 2004; Dean et al., 2016) highlights the need for a policy that is short and precise, easy to understand and consistent in its contents. In fact, an enterprise can avoid a myopic policy by adopting a sound policy which included legal, consumer and social, business perspectives into consideration (Dean et al., 2016). On the other hand, transparency in the privacy policy can be promoted by letting users to have control over data distribution and data protection. Achieving these features is possible from the users' side which is only sending encrypted data to the web-based system. Furthermore, to facilitate textual data encryption technology, a browser plugin called ShadowCrypt (He et al., 2014), is recommended for user's consideration.

ShadowCrypt performs the so-called diplomatic role between the user and the web application. User input is captured and will be encrypted either by using a deterministic or random encryption before the application is allowed to access the data. Encrypted data is only accessible to the web apps

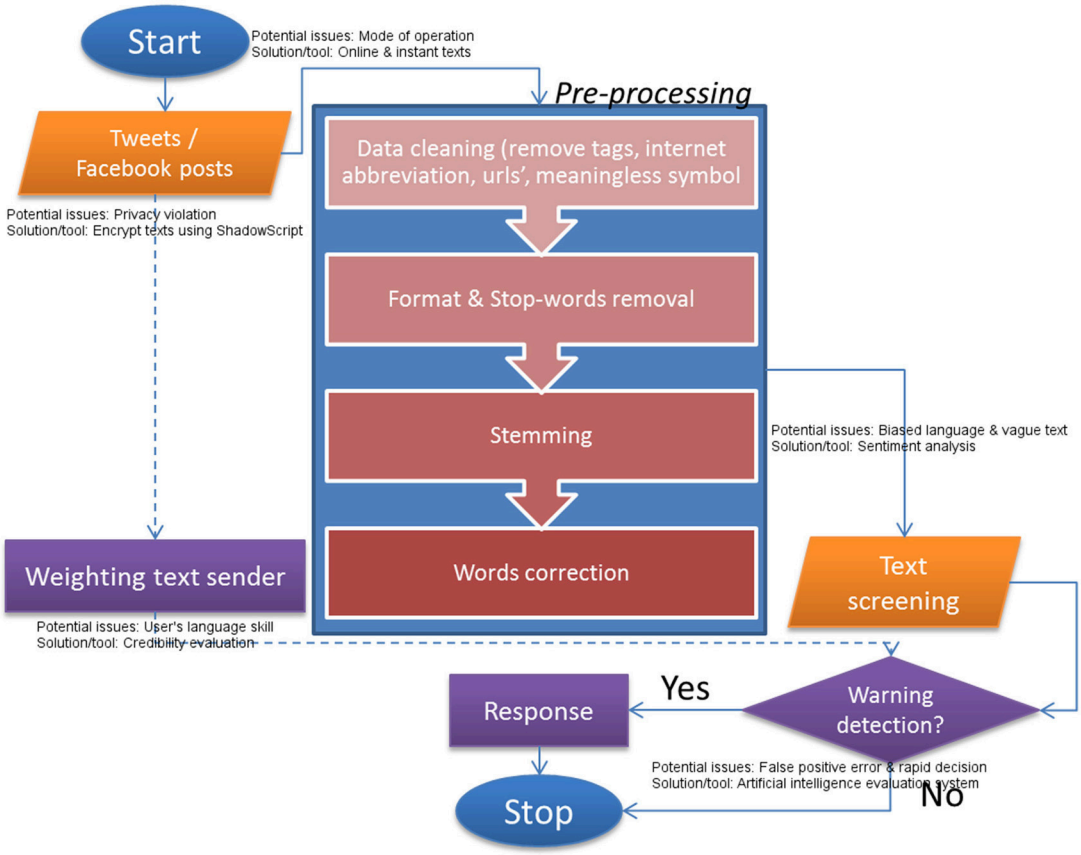

FIGURE 2 | A position of the proposed online text pre-processing steps in the passenger-participation warning system.

with decryption keys supplied by the data owner who is also the system user. Interestingly, that keys are only stored in the ShadowCrypt of the user's computer. This implies that the administrator of the web apps cannot simply access the data as he must be physically present at the location of the user's computer. This smart mechanism shifts the threat level of data breach and privacy violation from high to low risk (possible but not expected) while still participating in the alert system.

Weight Assignment

When the system is opened to every tram passengers, it is necessary to take credibility of text sender into consideration. Users annotated with low credibility might post vague texts and this drives the system to trigger a false alarm. Fortunately, user's credibility can be measured quantitatively by applying a reputational ranking model to user's profile (Alrubaian et al., 2017). User's profile features such as number of followers, number of positive and negative posts, etc. being feed to the model in order to generate credible scores. The scores weighs user's texts and it will be used to prioritize resources when verifying the content of text.

Pre-processing

Data Cleaning

Not only regular computer program expressions, html characters, banners, and graphical images but non-text data will be rid out from the messages. However, emoticons are given exemption at this phase as it has been recognized as a trend in non-verbal expressions. An effective data cleaning significantly reduce data dimensionality of term space which has a great impact on information extraction efficiency.

Stop-Words Removal

At this stage, an input vector will consist of words and emoticons (if any) only. Stop-words which are a bunch of words that insignificant for analysis i.e., non subject keywords, including high frequency word categories such as articles, prepositions, and pro-nouns will be removed from the text vector. As the designed system expects text data which illustrates non-positive side of passenger's reactions about their on-board experiences, positive words and emoticons are also labeled as stop-words.

Stemming

Further number of words reduction can be gained from stemming the remaining inflected words to its stem or base form. Stemming the words with similar meaning also offers

time savings and larger memory space. Note that, this process should not be applied to words that are not semantically related. Hence, words that do not have the same meaning should be kept separate. In fact, a text classifier can be sometimes negatively affected from text stemming (Nokhbeh Zaeem et al., 2017).

In the context of alert/warning system, biases in language of texting should be treated with care for the system to have a low rate of false positive. False positive case occurred when the system triggered an alert for no apparent reason. To address this issue, it is necessary to equip the system with a stemming algorithm that is capable to deal with biased language in a framing situation e.g., a specific train journey. Referred to as framing bias in literature of linguistic subjectivity, this class of bias is considered by the fact that people use subjective word, phrase or sentence to express his/her opinion toward a particular frame of event. For instance, a phrase "train makes me dizzy" unlike "train is wavering," seems to be an excessive expression to describe uncomfortable feeling of a train journey.

Subjective expression which is commonly found in people's experiences and opinions can be treated by performing sentiment analysis. The analysis is an automatic process which begins with a classification of word, phrase, emoticon, and slags that occurs in a text into a positive or negative category. Following the classification, sentiment of each text entry is calculated with respect to the semantic orientation or polarity of the text. In the case of Twitter messages, adjectives would be the indicators of the orientation (Taboada et al., 2011). For a topic of interest in single domain, lexicon-based approach that need for a dictionary of (positive and negative) word rankings is preferably adopted to calculate sentiment value for any given text. Hence, a dictionary of (positive and negative) word rankings will be built from a list of adjectives and corresponding scores of semantic orientation. A list of ready-to-use dictionaries can be found in (Taboada et al., 2011).

Words Correction

Before leaving the pre-processing phase, any abbreviations or misspelling words are replaced. Association rules and online dictionary could be applied here to increase a transformation success rate.

Text Screening (Decision)

An output of pre-processing can be treated as a cleaned, filtered and compact version of user's textual response. The final contents are then screened thoroughly to unsurfaced words or phrases related with ride comfort, motion, safety, noise and vibration and bad or suspicious feelings. The process loop is closed i.e., calling for subsequent texts, unless the system triggers an early warning in a specific identified tram. An expert assessment is called upon for human interpretation before any responsive action such as an immediate call to an involved tram driver is performed.

CONCLUSION

Trams in Britain are still manned-operated mode of transport, which means it is vulnerable to human errors as appeared in a recent Croydon tram derailment tragedy. Besides physical technologies adaption for human-related risk reduction in tram/train operations such as radio communication systems, one interesting aspect that has huge potential but not been explored extensively is passenger's participation in hazard detection. This study introduces a concept of data fusion in which processed passenger's negative reactions about real-time tram operations posted in media social medium is treated as complementary information to the existing safety system. A key element to successfully use text as data source is having a proper design of pre-processing steps which was a topic of discussion. Data cleaning, removal, stemming, and correction are necessary to refine raw data before high-level data fusion is considered. To analyse parameters sensitivity of the recommended design, a training dataset is required and its preparation has been identified as the next agenda. On top of that, an identification of suitable incentive-driven methodology to attract passenger's participation must be solved in parallel, otherwise the proposed system has nothing to improve.

AUTHOR CONTRIBUTIONS

MB developed the presented idea and wrote the first draft of manuscript with support from SK. All authors contributed to the final version of the manuscript. SK supervised the writing.

FUNDING

The authors are sincerely grateful to European Commission for the financial sponsorship of the H2020-RISE Project No. 691135 RISEN: Rail Infrastructure Systems Engineering Network, which enables a global research network that tackles the grand challenge in railway infrastructure resilience and advanced sensing.

ACKNOWLEDGMENTS

The first author would like to acknowledge scholarship from the Ministry of Higher Education of Malaysia and University Kebangsaan Malaysia. Map data copyrighted OpenStreetMap contributors and available from https://www.openstreetmap.org.

REFERENCES

Alrubaian, M., Al-Qurishi, M., Al-Rakhami, M., Hassan, M. M., and Alamri, A. (2017). "Reputation-based credibility analysis of Twitter social network users," in *Concurrency and Computation: Practice and Experience*, Vol. 29, 1–12.

Bearfield, G., and Marsh, W. (2005). "Generalising event trees using Bayesian networks with a case study of train derailment," in *Lecture Notes in Computer Science*, eds W. Rune, G. Bjørn Axel, and D. Gustav (Berlin; Heidelberg: Springer Berlin Heidelberg), 52–66.

Chen, L., Chiang, T., and Teo, T. (2005). "Fusion of LiDAR data and high resolution images for forest canopy modeling," in *Asian Conference on Remote Sensing* (Hanoi), 3–9.

Christodoulou, H. (2016). *RUNAWAY TRAM Croydon Tram was 'Speeding Excessively' After Arrested Driver 'Blacked Out' – As Passengers Tell*

of Terrifying Moment it Crashed Killing at Least Seven. London: The Sun.

Dean, M. D., Payne, D. M., and Landry, B. J. L. (2016). 'Data mining: an ethical baseline for online privacy policies'. *J. Enterprise Inform. Manage.* 29, 482–504. doi: 10.1108/JEIM-04-2014-0040

Dos Santos, F. L., and Ladeira, M. (2014). "The role of text pre-processing in opinion mining on a social media language dataset," in *Proceedings - 2014 Brazilian Conference on Intelligent Systems, BRACIS 2014* (São Paulo), 50–54.

Eggers, M., and Khuon, T. (1990). 'Neural network data fusion concepts and application', in *1990 IJCNN International Joint Conference on Neural Networks* (San Diego, CA), 7–16.

Hamouda, S., Ben and Akaichi, J. (2013). Social networks' text mining for sentiment classification: the case of facebook statuses updates in the "Arabic Spring" Era. *Int. J. Appl. Innov. Eng. Manage.* 2, 470–478. doi: 10.1109/SocialCom.2013.135

Hannah, P., Starr, A., and Ball, A. (2000). "Decisions in condition monitoring-an examplar for data fusion architecture," in *Proceedings of the 3rd International Conference on Information Fusion, FUSION 2000*, Vol. 1 (Paris).

He, W., Akhawe, D., Jain, S., Shi, E., and Song, D. (2014). "Shadowcrypt: encrypted web applications for everyone," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security - CCS'14* (Arizona), 1028–1039.

Mitali, S., Garg, R., and Mishra, P. K. (2003). Preprocessing techniques in web usage mining: a survey. *Int. J. Comp. Appl.* 97, 1–9. doi: 10.5120/17104-7737

Network Rail, (2017). *Innovation and Suggestions: Data Quality, Confidence and Assurance*. Available online at: http://archive.nr.co.uk/Innovation_and_suggestions.aspx

Nokhbeh Zaeem, R., Manoharan, M., Yang, Y., and Barber, K. S. (2017). Modeling and analysis of identity threat behaviors through text mining of identity theft stories. *Comput. Sec.* 65, 50–63. doi: 10.1016/j.cose.2016.11.002

Parascandolo, S. (2007). *The Control Room, The Croydon Tramlink.* Available online at: http://www.croydon-tramlink.co.uk/info/infra/control.shtml (Accessed November 12, 2016).

Petz, G., Karpowicz, M., Fürschu,ß, H., Auinger, A., Winkler, S. M., Schaller, S., et al. (2012). "On text preprocessing for opinion mining outside of laboratory environments," in *Active Media Technology, AMT 2012,* Lecture notes in computer science, eds R. Huang, A. A. Ghorbani, G. Pasi, T. Yamaguchi, N.Y. Yen, and B. Jin (Berlin; Heidelberg: Springer-Verlag), 618–629.

Raheja, D., Llinas, J., Nagi, R., and Romanowski, C. (2006). Data fusion/data mining-based architecture for condition-based maintenance. *Int. J. Prod. Res.* 44, 2869–2887. doi: 10.1080/00207540600654509

Rail Accident Investigation Branch (2016). Fatal accident involving the derailment of a tram at Sandilands Junction, Croydon, 9 November 2016. (Derby).

Singh, T., and Kumari, M. (2016). Role of text pre-processing in twitter sentiment analysis. *Proc. Comput. Sci.* 89, 549–554. doi: 10.1016/j.procs.2016.06.095

Starr, A., Willetts, R., Hannah, P., Hu, W., Banjevic, D., and Jardine, A. K. S. (2002). "Data fusion applications in intelligent condition monitoring," in *Recent Advances in Computers, Computing and Communications*, eds N. Mastorakis and V. Mladenov (WSEAS Press), 110–115.

Steinberg, A. N. (2001). "Data fusion system engineering," in *IEEE Aerospace and Electronic Systems Magazine,* Vol. 16 (Pisa), 7–14.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Comput. Ling.* 37, 267–307. doi: 10.1162/COLI_a_00049

Transport for London (2014). *Trams Update*. London: Transport for London. Available online at: http://content.tfl.gov.uk/rup-20141113-part-1-item09-trams-update.pdf

Van Wel, L., and Royakkers, L. (2004). Ethical issues in web data mining. *Ethics Inform. Technol.* 6, 129–140. doi: 10.1023/B:ETIN.0000047476.05912.3d

White, F. E. Jr. (1987). *Data Fusion Lexicon, Joint Directors of Laboratories, Technical Panel for C3, Data Fusion Sub-Panel*. San Diego, CA: Naval Ocean Systems Center.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Bin Osman and Kaewunruen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.