Journal of Informatics and Web Engineering
Vol. 3 No. 1 (February 2024) eISSN: 2821-370X

Term Standardisation with LDA Model to Detect Service Disruption Events using English and Manglish Tweets

Noraysha Yusuf [1], Maizatul Akmar Ismail[2], Tasnim M. A. Zayet[3*], Kasturi Dewi Varathan[4], Rafidah MD Noor[5]

[1,2,3,4,5] Universiti Malaya, 50603 Kuala Lumpur, Wilayah Persekutuan Kuala Lumpur, Malaysia.

**corresponding author:(wva180007@siswa.um.edu.my; ORCiD: 0000-0001-5755-5953)*

Abstract - Rapid transit is one of Malaysia's most important transportation modes, where commuters use public transportation to travel. Any disruption in the rapid transit service affects their daily routines. Therefore, detecting such service disruption has become fundamental. In this study, the disruption in Malaysia's rapid transit service was assessed using English and Manglish (a combination of English and Malay) tweets through Latent Dirichlet Allocation (LDA). The gathered tweets were classified into event and non-event tweets and LDA was applied to the event tweets. Manglish event tweets were pre-processed using the proposed term standardisation technique. As a result, LDA has proved its efficiency in topic detection for both English and Manglish tweets with better performance for Manglish tweets; The best event detection rate of the LDA_English model was at the likelihood of 80% while the best detection rate of the LDA_Manglish model was at a likelihood of 60%.

Keywords— Rapid Transit, Latent Dirichlet Allocation, Manglish, Multilingual, Twitter

Received: 21 June 2023; Accepted: 9 August 2023; Published: 16 February 2024

I. INTRODUCTION

Rapid transit is described differently in different countries. It typically includes metros, subways, rail lines, tubes, etc. It usually exists in urban areas of the developed and developing countries. Rapid transit has gained its importance due to its ability to reduce traffic and pollution and provide fast, low-cost, and safe traveling service. Hence, any disruption in the rapid transit service may affect many sectors. Disruption in the rapid transit service is common and typically involves delays or cancellations of trips. Open Data Institute analysed 11-week data of 16 trains and hubs in the U.K. and found that, on average, delays or cancellations of trips affected 42.5% of train services during the morning rush hour[1] . Such disruption generally causes a delay of fewer than 30 minutes.

[1] Available at: https://theodi.org/project/visualising-rail-disruption-travel-smarter/

Journal of Informatics and Web Engineering
https://doi.org/10.33093/jiwe.2023.3.1.1

Published by MMU Press. URL: https://journals.mmupress.com/jiwe

One of the significant effects of the disruption in the rapid transit service is that it affects other interconnecting rapid transit lines; thus, affecting even more commuters [1],[2]. This increases the number of commuters waiting at the stations and their waiting time [2],[3]. Such disruption, clearly, affects the smoothness of the planned transit network schedules, which requires backup solutions and more transit modes, especially when the disruption takes up a long time.

One of the ways to discover, minimise and manage the disruption in rapid transit service involves the use of the Internet of Things (IoT), which is a technology with big data analysis [4],[5]. The use of IoT in transit systems, has made some developments in the transit system, including rail systems [5],[6]. IoT can activate smarter rails. Certain studies employed IoT to predict the need for maintenance by monitoring the transit system and getting timely alerts about the stress and condition of the system [4],[6]. For example, IoT was employed to discover any faulty doors in the Singapore Mass Rapid Transit (SMRT) [7],[8]. However, the implementation of IoT sensors has many drawbacks, such as it needs experts and conversion from the legacy transit system [6]. Furthermore, IoT sensors are unable to discover the reasons for outdoor system disruptions, such as worker strikes and weather conditions. Hence, other data sources are needed to discover the disruption events.

Another way to predict service disruption is by analysing historical data. [1] developed a Bayesian model and analysed the tap-in and tap-out transport smart card data to discover the disruption events. The study found a stable data pattern in the number of commuters, which differed significantly in the case of service disruption. However, this method is only effective in studying commuters' traveling behaviour and not necessarily effective in detecting real-time disruptions, since this method needs the collection of data beforehand. [9] built a machine learning model using a transit incident dataset to predict delays. [10] also used a similar dataset with a parametric log-logistic-based accelerated time failure model to estimate the occurrence of transit delay events.

Apart from that, disruption events can also be detected using a low-latency and low-cost approach by employing social media data [11]. Unlike tweets about scheduled events, tweets about unexpected events vary and are less structured [12]. In addition, in case of unexpected events, there will be a sudden increase in the number of tweets [13,14], which has propelled certain studies to extract and analyse this type of tweets [15]. Certain studies even attempted to synchronically identify the most used terms over a period of time [13,14]. This logic was deemed sufficient to identify the occurrence of disruption events in the public transport network [16,17].

For this study, Latent Dirichlet Allocation (LDA) was used to detect the disruption in Malaysia's rapid transit service based on English and Manglish tweets. As Malaysia is a multicultural and multilingual country [18] and the Malay language uses English letters, Manglish, which is a combination of Malay and English words, is popular. However, multilingual tweets pose several issues. Therefore, the following research questions were addressed in this study:

- What are the forms of tweet content that suggest the occurrence of service disruption events?
- How to prepare the multilingual tweets for event detection?
- How to detect tweets that are related to the disruption in the rapid transit service (event tweets)?

Hence, the main contribution of this paper is the preparation of Manglish text for the detection of rapid transit service disruption events.

II. LITERATURE REVIEW

A. Scope of Detecting Service Disruption Events in Tweets

Tweets are a form of user-generated content. In other words, the content of tweets is informal. It does not describe events in a matter-of-fact manner. Hence, there is a need to understand the scope of the content of event tweets.

[19] identified disruption events using a pre-defined dictionary of commuters' common terms used to complain about any service disruption. The proposed approach was aided by identifying the names of the stations and lines. However, the use of a pre-defined dictionary seems to limit the discoverability of disruption events. Commonly, the overall quality of the transport service is the target of social media-based research [16],[20]. These researches are classified under the general category, of "reliability".

With regard to rapid transit service disruption, multiple forms of tweets can be identified from the literature. Firstly, tweets can be in the form of opinion expression towards a particular event [21]. Commuters tend to express their opinion toward the service providers or public transit service [16],[20]. Secondly, commuters also express the shortage of service, resulting in the formation of complaint tweets [22],[23]. Thirdly, inquiry tweets are formed when commuters use tweets to seek information. Usually, commuters need updates on service delays and transport choices [20],[24]. Fourthly, tweets can also be in the form of descriptions or comments of a particular disruption event. For instance, [17] used tweets to identify service disruption events, such as accidents, crowded situations, long queues, lengthy waiting time, slow buses, and breakdowns of buses. Meanwhile, [25] classified tweets related to bus services according to disruption types, such as accidents and obstructed trips.

Intuitively, the topic of tweets that are related to rapid transit service includes the causes or indications of service disruption. Apart from accidents, the ticketing system's failure or technical problems in the rapid transit system may disrupt the rapid transit service [10],[26].

B. Topic Modelling Using LDA

Topic modeling or categorisation can be performed using supervised and unsupervised techniques. There is one main drawback to the use of supervised techniques, which is their need for pre-labeled data. LDA is one of the unsupervised topic modelling techniques, which has been widely used in literature. LDA assumes that each document is formed from a mixture of topics and each topic is formed from a set of words [27]. The general steps of LDA are as follows:

Assume that M is the number of documents, α is the prior distribution for topics in document í, then, θi is the topic distribution of document i that is generated from a Dirichlet distribution with parameter α (see Equation (1)).

$$\theta_i \sim \mathrm{Dir}(\alpha) \text{ for } 1 \leq i \leq M . \quad (1)$$

Assume that K is the number of topics and β is the prior distribution of words in a topic. Then, φk is the word distribution of topic k generated from a Dirichlet distribution with parameter β (see Equation (2)).

$$\varphi_k \sim \mathrm{Dir}(\beta) \text{ for } 1 \leq k \leq K . \quad (2)$$

Assume that N is the number of words in a document, I is the index of documents and j is the index of words in a document, then, zi,j is the topic of word j in document I that is generated from a multinomial distribution with parameter θi (see Equation (3)).

$$z_{i,j} \sim \mathrm{Multinomial}(\theta_i) \text{ for } 1 \leq i \leq M \text{ and } 1 \leq j \leq N_i . \quad (3)$$

Generate the word j in document i from a multinomial distribution with parameter as depicted in Equation (4).

$$w_{i,j} \sim \mathrm{Multinomial}\left(\varphi_{z_{i,j}}\right) \text{ for } 1 \leq i \leq M \text{ and } 1 \leq j \leq N_i . \quad (4)$$

The model that incorporates the previously listed steps is graphically presented in Figure 1, where the joint probability of the model is in Equation (5).

$$p(w,z,\theta,\varphi \mid \alpha,\beta)=p(w \mid \varphi,z)p(\varphi|\beta)p(z|\theta)p(\theta|\alpha)\ . \quad (5)$$

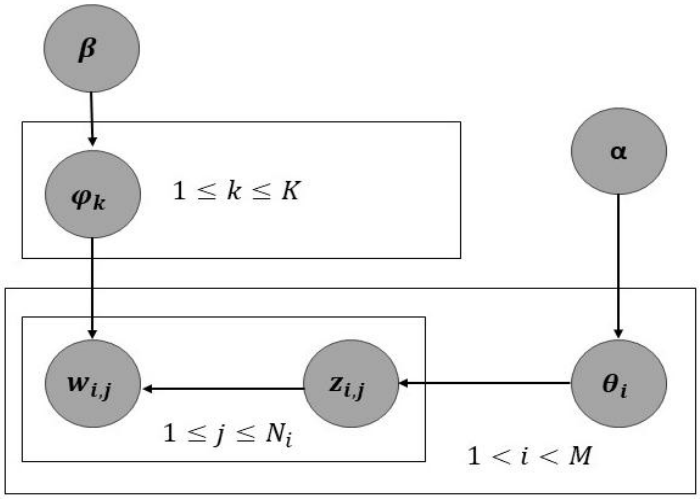

Figure 1. Graphical Representation of The LDA Model

Using social media data, LDA has been used for topic modeling purposes [13],[14],[15]. Pre-defined topics are not needed in LDA; this advantage makes LDA a useful technique in the case of identifying unexpected events. Cosine similarity and LDA were previously utilised for event detection in Twitter and news articles [27]. LDA was also used to detect real-time events in the transport network through Twitter [28],[29]. Meanwhile, the support vector machine (SVM) and a supervised version of LDA (sLDA) were used to identify traffic events [30]. Besides that, LDA was used to detect real-time traffic events [29],[31].

However, considering the mechanism of LDA which mentioned earlier, LDA is usually efficient in the case of a large number of long texts [14,15, 30,32]. Using LDA with short texts may not be efficient; as short texts, mostly, revolve around one single topic with very concise and few words. Tweets are typically short and with pre-processing, they become even shorter. Employing LDA with short texts may cause different issues such as words being assigned to an unrelated topic [14], a small number of topics being identified [16] or different events being classified under the same topic [15].

Several prior studies addressed the short text problem by combining similar tweets into one document to increase the reliability of the discovered topics [14],[33]. This process is called the pooling technique. However, in this study, a simpler technique was used, which involved term standardisation. Term standardisation assures homogeneity and harmonisation throughout the texts and decreases the needed efforts for further text processing [34],[35],[36]. In addition, through the standardisation process, synonyms, slang, abbreviations, and other related aspects can be standardised, which potentially enhances the ability of LDA in identifying topic terms as LDA considers the distribution and frequency of words in the documents.

C. Handling Multilingual Tweets

The analysis of multilingual texts has not been widely explored due to the lack of lexical and labeled resources in non-English languages [37],[38]. LDA was utilised for English and Spanish tweets and other texts where dictionaries were used to associate each language topic with a topic from the other language [39]. Also, in the context of English and Spanish tweets, [40] used Support Vector Machines Sequential Minimal Optimization (SVM SMO) for sentiment classification along with standardisation of emotion words, negation, intensifiers, punctuation marks, and others to words such as 'NEGATIVE', 'POSITIVE', 'MULTISTOP', 'NEGATION', etc. The standardisation was done using general and machine-translated dictionaries. The standardisation helped in discovering the sentimental patterns and improved the result of multilingual tweets classification.

[41] analysed Singlish (Singapore's mixed informal language) tweets and discovered event topics by discovering the candidate event day and then feed the tweets published on this day to the LDA model. However, the studies did not

distinguish event tweets from non-event ones. Distinguishing event tweets can increase topic identification efficiency by applying the topic identification technique to the event tweets only. Non-event tweets may cause disambiguation in the extraction of service disruption events.

However, till the time that this paper was written and further to the author's knowledge there is no work considered the Manglish tweets for event disruption events in Malaysia. In Manglish tweets, different words in Malay may indicate the same word in English, hence, a method for preparing the Manglish text for event detection is needed.

III. RESEARCH METHODOLOGY

This section presents the experiment and settings of the models. Two LDA models were constructed in this study as tweets in two languages, specifically English and Manglish were examined. Figure 2 shows the flow of the method.

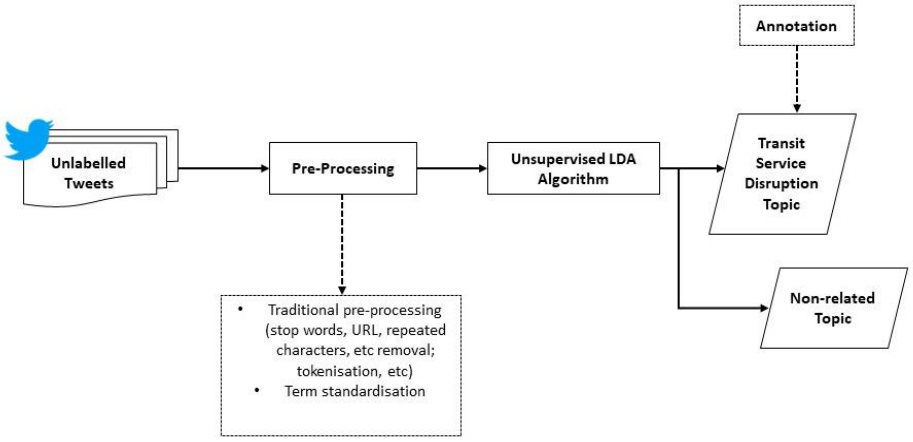

Figure 2. The Flow of The Method

A. Data Collection

Twitter data was crawled using the Twitter scraper package in Python. Three filters were used to retrieve the related tweets, namely keywords, date, and language filters. The keyword was set to "@MyRapidKL", the official Twitter account of Malaysia's rapid transit service provider. According to [16], commuters tend to tag the service provider's account when they post their complaints.

After applying the keyword filter, the raw tweets were then filtered to retrieve tweets from 2014 to 2019. As a result, 152,833 tweets were retrieved. To identify the language of the tweets effectively, multinomial-based language prediction from the MALAYA[2] Python package was used. Subsequently, the raw tweets were filtered to include only Manglish and English tweets.

B. Data Pre-Processing

As for this stage, RegEx[3] and NLTK[4] Python packages were used to perform the following processes: (1) cleaning, (2) term standardisation, (3) term correction, stemming, and lemmatisation. The following describes these processes in detail.

(1) Cleaning of Raw Tweets

The cleaning process in this study incorporated the following steps:

[2] https://pypi.org/project/malaya/
[3] https://pypi.org/project/regex/
[4] https://www.nltk.org/

- Special characters, non-English characters, stop words, URLs, and emoticons were removed.
- Duplicate tweets were removed.
- Words that contain repeated letters were shortened.

(2) Term Standardisation

To effectively identify tweets that are related to the disruption in rapid transit service, direct terminology indicators were detected such as "late" and "interrupted" [19]. Moreover, slang, abbreviations, and specific names of rapid transit, such as station or line names were detected [16]. As for the case of Manglish tweets, different words indicate service disruption. For example, the word "delay" in the Manglish tweets corpus refers to "lama" and "lambat". Hence, standardisation of key terms was necessary for this study. The term standardisation process incorporated four steps: standardisation of train-related terms, standardisation of abbreviations and slang words, standardisation of synonyms, and standardisation of emotional expression.

As for the first step, train-related terms, such as "koc" and "gerabak" were standardised as "carriage", while "rel" and "tren" were converted to 'train'. Additionally, station names were also standardised to formal names. In the second step, where Term Frequency-Inverse Document Frequency (TF-IDF) was involved, abbreviations and slang words were identified. Terms that were standardised in this step had the following features: (1) frequently used, (2) did not contain proposition or verb, (3), not the proper term, (4) informal term. Table 1 shows examples of standardised abbreviations and slang words. In the third step, synonyms were standardised to the most commonly used words. Table 2 shows examples of standardised synonyms. The final step involved standardising terms that indicate emotional expression. Examples of standardised emotional expressions are displayed in Table 3.

Table 1. Examples of Standardised Abbreviations and Slang Words

Abbreviations and slang words	**Standardised terms**
X	tak
x dak, xde xda, xdop, tadak, takda, takdak, tkde, takde	tiada
xper, xpe, xpa, tak pe, tk pe, takpe, takpa	tak apa
keje, kerje, kijo	kerja
dgn, ngan, dgan	dengan
apesal, apehal, nape, knpa, knape, bakpo, pehal	kenapa
mcm mana, mcm mne, cemana, cam mana, cane, camne, macam mana, mcmana	bagaimana
memanjang, manjang	asyik
pi, gi	pergi
tgu, tggu, tnggu	tunggu
jgn, jngn, jangn	jangan
prob, probz, probs	problem
tenkiu, tenqiu, tq	thank you
habaq, btahu, bagitau, kasi tau	beritahu
uolls, uoll, uollss, uolss	you all
mampuih, mampoih, mampus, mampos	mati
imho	in my honest opinion
smh	shake my head
ikr	i know right

Table 2. Examples of Standardised Synonyms

Terms	**Standardised terms**
lihat, perati	tengok
masalah, isu	problem
lewat	lambat
kerana	sebab

bagitahu	beritahu
hangat, bahang	panas
minit	minute
mengapa	kenapa
daripada	dari
aliran	line
pemandangan, pandangan. penlihatan	view
topup, topap, top up, tambah nilai	reload
selalu	kerap
penumpang	passenger
myr, rm	ringgit
netizen	orang

Table 3. Examples of standardised emotional expressions

Emotional Expressionism n	**Standardised terms**
hahahaha, huhuhu, hehehehe, lolololol, lol, haha	laugh
adui, oitt, woi, woit, ahh, siot, duhh	annoy
alaa, aduhai, haih, haishh, hurmm, hmmm	sigh
bengong, cilake, lahanat, hanat	celaka (meaning to curse)

(3) Term Correction, Stemming and Lemmatisation

Misspelt words are common mistakes in user-generated content. To detect misspelled words in multilingual texts, the language of the word should be first detected. Secondly, the spelling should be corrected. Malaya package was used to detect the language of the tokenised tokens. Following that, its spelling corrector was used to correct misspelled Malay words, while SymSpell's[5] spelling corrector (Python's package) was used to correct misspelled English words. The words were then stemmed and lemmatised to standardise the words to their roots. NLTK's and Malaya's stemmer and lemmatisation functions were used for the above purpose.

C. Topic Identification Model

LDA was used to identify the service disruption topics. Parameters were tuned to develop the most optimal LDA model. The Gensim[6] Python package was used for the LDA experiment. The LDA experiment settings are presented in the following subsections.

(1) Dataset Construction

Table 4 and Table 5 show four datasets that were used in the LDA experiment. The year span of the datasets was chosen to ensure an almost equal number of tweets in each set of the training and testing for English and Manglish. Additionally, tweets from different years were used in the testing set to ensure the variety of tweets. After that, the testing sets were manually annotated as event and non-event tweets.

Table 4. Training Sets of LDA Models

	Description	**Year**	**Total**
LDA_training_Eng	The training set for English tweets	2018	4253
LDA_training_Manglish	The training set for Manglish tweets	2018 - 2019	3336

[5] https://pypi.org/project/symspellpy/
[6] https://pypi.org/project/gensim/

Table 5. Testing Sets of The LDA Models

	Description	Year	Event	Non-Event	Total
LDA_test_Eng	Test set for English tweets	2015 - 2019	260	82	342
LDA_test_Manglish	Test set for Manglish tweets	2014 - 2019	227	77	304

(2) Pre-processing

The pre-processing of tweets was further performed according to the following steps:

- Stop words were removed.
- Tweets with less than four words were deleted.
- Words related to highly frequent rapid transit, which can affect the topic-term distribution, were deleted. For example, station names, line names, and mode names.
- Texts were tokenised into unigrams and bigrams.
- Documents were converted into the Bag-of-Word (BOW) model.

D. Tuning LDA Parameters

For this study, four parameters were tuned, namely: the number of iterations, the number of topics, and two hyper-parameters, α and β. The number of iterations was set at 100 to ensure the convergence of the model, while the remaining parameters were tuned by constructing different models and using the training sets.

E. Model Evaluation

The optimal LDA model was selected based on topic coherence and perplexity values. For the coherence measure, the C_V measure, Equation (7) was utilised, which refers to the arithmetic mean of the pairwise of the top resulted in terms in a sliding window. In the C_V measure, the normalised version of pointwise mutual information (NPMI) Equation (8) was utilised. Meanwhile, as for the perplexity measure (see Equation (9)), the log-likelihood of untested sets was measured. A lower perplexity value of the LDA model was preferred. In addition, a human evaluation was performed and LDA models were visualised using the pyLDAvis tool [38].

$$Coherance = \sum_{w_q, w_s \in W} Score(w_q, w_s). \tag{7}$$

$$NPMI(w_q, w_s) = \frac{log \frac{P(w_q, w_s) + \epsilon}{P(w_q).P(w_s)}}{-\log P(w_q, w_s) + \epsilon}. \tag{8}$$

$$Perplexity(D) = exp\left(-\frac{\sum_{d=1}^{M} log\, P(w_d)}{\sum_{d=1}^{M} N_d}\right). \tag{9}$$

Where W is the set of vocabulary that represents the topic; ϵ is added due to the case of zero; P(wq,ws) is the joint probability of word wq and word ws; D represents the testing set; Nd is the number of words in document d; M is the number of documents; P(wd) is the probability of word wd in the document.

IV. RESULTS AND DISCUSSION

Two LDA models, specifically LDA_English and LDA_Manglish models, were constructed in this study, as tweets in English and Manglish were examined. The following subsections discuss the results of each model.

A. LDA_ENGLISH Model

For the training set of English tweets, the related parameters were tuned accordingly. Figure 3 shows the number of topics and the corresponding coherence value for each topic. The best coherence scores are recorded when the number of topics is equal to 5, 7, and 9.

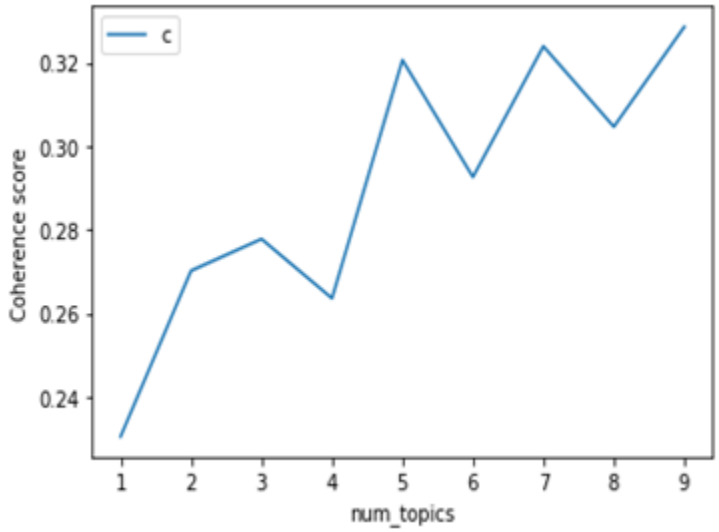

Figure 3. The Number of Topics and The Associated Coherence Values

In order to select the best number of topics, the perplexity value was calculated. Table 6 shows the perplexity value for each number of topics. In particular, 5 topics recorded the highest coherence value and lowest perplexity value, which represented the best number of topics. This result was achieved with asymmetric α (a fixed normalised asymmetric prior to 1.0, divided by the number of topics) and β = 0.91 at 100 iterations.

Table 6.The Number of Topics and The Corresponding Coherence and Perplexity Values

Number of topics	**Coherence value**	**Perplexity value**
5	0.396	-6.456
6	0.358	-6.482
7	0.355	-6.514
9	0.376	-6.560

Table 7 shows the topic terms and the associated labels. Topics appeared to be distinctly distributed, except for a slight overlapping between Topic 2 and Topic 4. This indicates a relationship between Topic 2 and Topic 4. Besides that, topic 1 is the largest, which implies the highest spread of terms at 49.8% in LDA_training_Eng data. After the topics and related terms were identified, each topic was assigned a label associated with rapid transit services.

Table 8 displays the detection rate of the LDA_English model on the service disruption event tweets for unseen LDA_test_Eng data. The detection rate of the model was assessed on three different likelihood measures 60%, 70%, and 80%. A likelihood measure (Pt,k) refers to how likely a particular event or non-event tweet belongs to a specific topic. In the ideal case, all the service disruption event tweets should be classified under topic 1 (service disruption).

Table 7. The Topic Terms and Associated Labels

Topic	Top 30 Terms	Label (Manual)	Prevalence (%)
1	Station, time, what, service, minute, now, why, go, stop, people, problem, wait, delay, move, like, get, issue, today, work, more, hour, still, morning, stuck, announce, happen, when, long, know, every	Service disruption	49.8
2	station, ringgit, only, tng (Touch & Go), parking, pass, reload, park, when, ride, staff, people, pay, one, then, where, which, free, charge, who, fare, card, counter, well, phone, machine, ticket, passenger, instead, car	Fees & payment method	15.8
3	Ac (air-conditioning), carriage, function, check, hot, door, head, thank, now, number, crowd, morning, fix, help, like, people, inside, thanks, stuffy, air, good, passenger, light, break, currently, open, mana, leak, switch, track	Train carriage conditions	13.6
4	Card, concession, student, thank, apply, lose, online, still, good, renew, day, rapid, get, know, application, one, people, more, public, new, already, provide, ask, process, everyone, reply, public_transport, again, long, try	Application & Renewal	12.1
5	okay, seat, naik, apa, morning, thank, head, location, orang, point, balik, bag, left, give, even, action, return, space, passenger, kenapa, people, woman, situation, annoy, lama, where_current, jalan, thanks, see, very	n/a	8.7

Table 8. The Detection Rate of The Lda_english Model

Likelihood (percentage)	Tweets class	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
80%	Event	137 (53%)	0	0	0	1 (0.4%)
	Non-Event	13 (16%)	8 (10%)	3 (4%)	2(2%)	2(2%)
70%	Event	176 (68%)	0	2 (0.8%)	2 (0.8%)	2 (0.8%)
	Non-Event	15 (18%)	9 (11%)	5 (6%)	2 (2%)	5 (6%)
60%	Event	210 (81%)	3 (1%)	2 (0.8%)	3 (1%)	2 (0.8%)

However, the best detection rate was at the likelihood measure of 60%, where the LDA_English model was able to infer 81% (210) of 260 event tweets under topic 1. Meanwhile, only, 10 tweets were wrongly classified under other topics, and 40 tweets were not classified under any topic due to their low likelihood.

$$P_{t,k} = \frac{1}{N_t} \sum_{i=1}^{N_t} z_{t,i} \; for \; 1 \leq t \leq T \; and \; 1 \leq k \leq K. \tag{10}$$

B. LDA_MANGLISH Model

The hyperparameters were tuned and the best number of topics that recorded the highest coherence value and lowest perplexity value was identified. The best number of topics was Topic 8 with asymmetric α and β of 0.91 as shown in Figure 4.

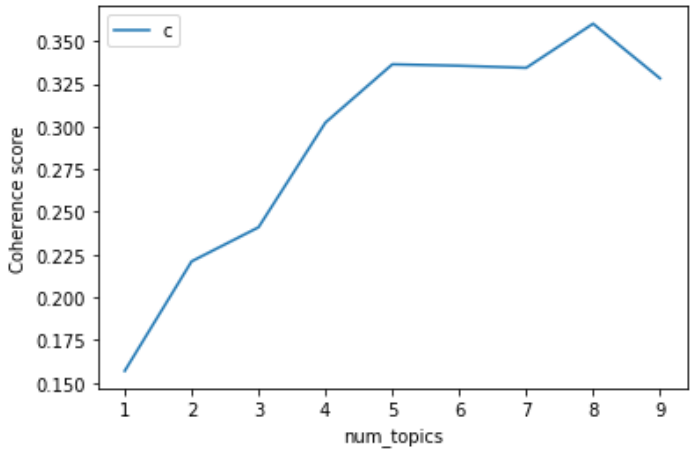

Figure 4. The Number of Topics and The Associated Coherence Values

Based on Table 9, Topic 1 to Topic 4 appeared prominent, while Topic 5 to Topic 8 were closely linked and not easily distinguished. Topic 1 and Topic 2 recorded prevalence values of 49.3% and 23.1%, respectively. This suggests that Topic 1 and Topic 2 had the most relevant terms in LDA_training_Manglish.

Table 9 displays the top 30 terms of each topic and the assigned manual labels. A total of eight topics were identified, where Topic 1 to Topic 4 were deemed to have the most meaningful and distinct terms. These results were further supported by the moderate to high prevalence values for Topic 1 and Topic 2. Although Topic 3 and Topic 4 recorded relatively lower prevalence values, their respective terms are still easily interpreted. Topic 5 to Topic 8 recorded extremely low prevalence values; as a result, these topics were not considered during labeling.

Table 9. The Topic Terms and The Associated Labels

Topic	**Top 30 Terms**	**Label**	**Prevalence (%)**
1	Station, kenapa, naik, hari, problem, orang, lama, min, pergi, apa, jalan, berapa, dalam, satu, tunggu, balik, kerja, lalu, masuk, ramai, lambat, rosak, baru, mana, pintu, bila, pukul, henti, lepas, minute	Service disruption	49.3
2	Orang, dalam, passenger, masuk, jangan, bag, naik, duduk, makan, lain, tolong, nampak, pihak, mungkin, bila, semua, bawa, letak, rasa, minta, kerap, beratur, terima_kasih, tengok, atas, pula, bukan, keluar, sila, tempat	Passengers' etiquette	23.1
3	ringgit, kad, guna, kaunter, bayar, tng (Touch & Go), staff, reload, beli, tanya, tambang, pass, hari, student, kad_concession, customer, card, bulan, bila, ic (Identification card), renew, cukup, bagaimana, kurang, pas, rapid, duit, mohon, tanpa	Fees & payment method	13.3
4	Ac (air-conditioning), carriage, panas, dalam, tuju, sejuk, buka, rasa, bau, sekarang, pengsan, tolong, rosak, arah, terima_kasih, mohon, peluh, kuat, fungsi, number, pasang, thank, lalu, check, beku, menitik, semua, help, selamat, nafas	Train carriage conditions	7.8
5	Panggong, lompat, panggil. sultan, polis, tulis, tepi, kedai, side, belah, terus, ltan, signboard, huruf, sia, sedap, lontong, ciara, zone, cabut, sama, mirror, resign, semenyih, bincang, member, semangat, huyung, hayang, drop	n/a	2.1
6	Durian, lanjut, khidmat, Bahasa, sempena, negara, makan, klinik, final, mari, melayu, fatihah, al, imam,	n/a	1.6

	baca, fa, piala, baik, roti, sabtumutu, dunia, pilih, bangsa, tumbang, jimat, pokok, kuda, operator, tepat		
7	Puchong, kampung, sungguh, tolak, motif, balas, petang, kecewa, suasana, well, lee, seksyen, seri, chair, wheelchair, kalian, alasan, selipar, ubat, bukan, south, usaha, settle, moga, jual, senja, pangsapuri, kozato, keluli, kilang	n/a	1.5
8	Menang, sunigakura, negeri, liga, sepak, join, bola, sedia, team, lah, depan, peristiwa, sure, muthu, suami, angkut, ceo, prasarana, projek, pie, kondo, changgang, diba, naman, bazaar, Ramadan, pahit, cuti, do, potong	n/a	1.2

Similar to the LDA_English model, the same experiment was conducted for the LDA_Manglish model to determine its detection rate. Table 10 presents the detection rate of the LDA_Manglish model on the service disruption event tweets for unseen LDA_test_Manglish data. Topic 5 to Topic 8 were excluded due to their low prevalence value. At the likelihood measure of 60%, 93% of event tweets were identified under Topic 1, which confirms the prior hypothesis from the visualisation of Topic 1.

Table 10. The Detection Rate of The Lda_manglish Model

Likelihood (percentage)	**Tweets class**	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 4**
80%	Event	182 (80%)	0	0	0
	Non-Event	22 (29%)	6 (7%)	4 (5%)	6 (7%)
70%	Event	200 (88%)	0	0	1 (0.4%)
	Non-Event	24 (31%)	9 (12%)	5 (6%)	10 (13%)
60%	Event	211 (93%)	0	0	2 (1%)
	Non-Event	29 (38%)	12 (16%)	6 (7%)	13 (17%)

V. CONCLUSION

In this study, LDA was applied to detect rapid transit service disruption events in Malaysia based on two types of tweets, namely English and Manglish tweets. The pre-processing of Manglish (or multilingual) tweets were performed and the standardization of words was used to ‘uniform’ the text.

In the experiment, LDA parameters were tuned and the best number of topics were decided for each type of tweets. The LDA model for Manglish tweets in this study provided better performance in detecting event tweets, which may be likely due to the error in identifying English tweets using the MALAYA package.

In the case of topic modeling, the terms of the most common topics were generally related to time and waiting; thus, suggesting that delays are the most common disruption events in the rapid transit service. Service providers can use the data to have a general view of the service disruption and commuters' complaints, helping them maintain and enhance the service.

However, approaches other than LDA can be used for detecting the disruption events such as clustering techniques, in addition, dictionaries and classification algorithms can be used for standardization. Hence, the effectiveness of these methods against the proposed one need to be examined.

ACKNOWLEDGEMENT

The authors received no funding from any party for the research and publication of this article.

REFERENCES

[1] H. Sun, J. Wu, L. Wu, X. Yan, and Z. Gao, "Estimating the influence of common disruptions on urban rail transit networks", Transportation Research Part A: Policy and Practice, vol. 94, pp. 62-75, 2016.

[2] Y. Yuan, S. Li, L. Yang and Z. Gao, "Real-time optimization of train regulation and passenger flow control for urban rail transit network under frequent disturbances", Transportation Research Part E: Logistics and Transportation Review, vol. 168, pp.102942, 2022.

[3] I. Y. Oh and Y. Y. Lim, "Commuter chaos as Kelana Jaya Line breaks down", The Star, 2019. Available: https://www.thestar.com.my/news/nation/2019/02/18/commuter-chaos-as-kelana-jaya-line-breaks-down/.

[4] T. Lee, and M. Tso, "A universal sensor data platform modeled for realtime asset condition surveillance and big data analytics for railway systems: Developing a "Smart Railway" mastermind for the betterment of reliability, availability, maintainability and safety of railway systems and passenger service", Sensors, pp. 1-3, 2016.

[5] F. Zantalis, G. Koulouras, S. Karabetsos and D. Kandris, "A review of machine learning and IoT in smart transportation.", Future Internet, vol. 11, no. 4, p. 94, 2019.

[6] P. Fraga-Lamas, T. M. Fernández-Caramés L. and Castedo, "Towards the Internet of smart trains: A review on industrial IoT-connected railways", Sensors, vol. 17.6, pp. 1457, 2017.

[7] A. Lim, "Train door sensors may cut delays at MRT stations", The Straits Times, 2018. Available: https://www.straitstimes.com/singapore/transport/train-door-sensors-may-cut-delays-at-mrt-stations.

[8] J. X. Chew, "Condition monitoring of train door system 2", Nanyang Technological University, 2020. Available: https://hdl.handle.net/10356/141443.

[9] J. Weng, Y. Zheng, X. Qu and X. Yan, "Development of a maximum likelihood regression tree-based model for predicting subway incident delay", Transportation Research Part C: Emerging Technologies, vol. 57, pp. 30-41, 2015.

[10] J. Weng, Y. Zheng, X. Yan, and Q. Meng, "Development of a subway operation incident delay model using accelerated failure time approaches", Accident Analysis & Prevention, vol. 73, pp. 12-19, 2014.

[11] S. M. Grant-Muller, A. Gal-Tzur, E. Minkov, S. Nocera, T. Kuflik and I. Shoor, "Enhancing transport data collection through social media sources: methods, challenges and opportunities for textual data", IET Intelligent Transport Systems, vol. 9.4, pp.407-417, 2014.

[12] A. M. Ertugrul, B. Velioglu and P. Karagoz, "Word embedding based event detection on social media", International Conference on Hybrid Artificial Intelligence Systems, Springer, 2017, pp. 3-14.

[13] D. Ramachandran and P. Ramasubramanian, "Event detection from Twitter–a survey", International Journal of Web Information Systems, 2018.

[14] L. Zou and W. W. Song, "Lda-tm: A two-step approach to Twitter topic data clustering", IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), 2016, pp. 342-347.

[15] D. Shang, X. Y. Dai, W. Ge, S. Huang and J. Chen, "A Multi-view Clustering Model for Event Detection in Twitter", International Conference on Computational Linguistics and Intelligent Text Processing, Springer, 2017, pp. 366-378.

[16] N. N. Haghighi, X. C. Liu, R. Wei, W. Li and H. Shao, "Using Twitter data for transit performance assessment: a framework for evaluating transit riders' opinions about quality of service", Public Transport, vol. 10.2, pp. 363-377, 2018.

[17] T. Hoang, P. H. Cher, P. K. Prasetyo and E. P. Lim, "Crowdsensing and analyzing micro-event tweets for public transportation insights", IEEE International Conference on Big Data (Big Data), 2016, pp. 2157-2166.

[18] K. Balakrishnan, "Influence of Cultural Dimensions on Intercultural Communication Styles: Ethnicity in a Moderating Role", JCLC, vol. 2.1, pp. 46–62, 2022.

[19] T. Ji, K. Fu, N. Self, C. T. Lu and N. Ramakrishnan, "Multi-task learning for transit service disruption detection", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2018, pp. 634-641.

[20] E. Mogaji and I. Erkan, "Insight into consumer experience on UK train transportation services", Travel Behaviour and Society, vol. 14, pp. 21-33, 2019.

[21] S. L. Lo, R. Chiong and D. Cornforth, "An unsupervised multilingual approach for online social media topic identification", Expert Systems with Applications, vol. 81, pp. 282-298, 2017.

[22] G. Abalı,, E. Karaarslan, A. Hürriyetoğlu and F. Dalkılıç, "Detecting citizen problems and their locations using twitter data", IEEE International Istanbul Smart Grids and Cities Congress and Fair (ICSG), 2018, pp. 30-33.

[23] M. O. Pratama, W. Satyawan, R. Jannati, B. Pamungkas, M. E. Syahputra and I. Neforawati, "The sentiment analysis of Indonesia commuter line using machine learning based on twitter data", Journal of Physics: Conference Series, IOP Publishing, Vol. 1193.1, 2019.

[24] G. Currie and C. Muir, "Understanding passenger perceptions and behaviors during unplanned rail disruptions", Transportation research procedia, vol. 25, pp. 4392-4402, 2017.

[25] I. Casas and E. C. Delmelle, "Tweeting about public transit—Gleaning public perceptions from a social media microblog", Case Studies on Transport Policy, vol. 5.4, pp. 634-642, 2017.

[26] R. I. Sarker, S. Kaplan, M. Mailer and H. J. Timmermans, "Applying affective event theory to explain transit users' reactions to service disruptions", Transportation Research Part A: Policy and Practice, vol. 130, pp. 593-605, 2019.

[27] N. Keane, C. Yee and L. Zhou, "Using topic modeling and similarity thresholds to detect events", Proceedings of the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation, 2015, pp. 34-42.

[28] Z. S. Qian, "Real-time incident detection using social media data", Pennsylvania. Dept. of Transportation, (No. FHWA-PA-2016-004-CMU WO 03), 2016.

[29] M. S. Mredula, N. Dey, M. S. Rahman, I. Mahmud and Y. Z. Cho, "A Review on the Trends in Event Detection by Analyzing Social Media Platforms' Data", Sensors, vol. 22.12, pp. 4531, 2022.

[30] Z. Zhang, M. Ni, J. Gao and Q. He, “Mining transportation information from social media for planned and unplanned events”, 2016.

[31] G. Paltoglou, “Sentiment-based event detection in Twitter”, Journal of the Association for Information Science and Technology, vol. 67.7, pp. 1576-1587, 2016.

[32] C. Vicient and A. Moreno, “Unsupervised topic discovery in micro-blogging networks”, Expert Systems with Applications, vol. 42, pp. 17-18, 2015.

[33] M. Hajjem and C. Latiri, “Combining IR and LDA topic modeling for filtering microblogs”, Procedia Computer Science, vol. 112, pp. 761-770, 2017.

[34] K. Rein, R. Coote, L. Sikorski and U. Schade, “Standardization to Deal with Multilingual Information in Social Media During Large-Scale Crisis Situations Using Crisis Management Language”, Application of Social Media in Crisis Management, Springer, pp. 115-131, 2017.

[35] A. M. Bucur, A. Cosma, and L. P. Dinu, "Sequence-to-sequence lexical normalization with multilingual transformers", arXiv preprint, 2021.

[36] R. van der Goot, A. Ramponi, A. Zubiaga, B. Plank, B. Muller, I. S. V. Roncal and W. Sidorenko, "MultiLexNorm: A shared task on multilingual lexical normalization", Seventh Workshop on Noisy User-generated Text, Association for Computational Linguistics, 2021.

[37] K. Dashtipour, S. Poria, A. Hussain, E. Cambria, A. Y. Hawalah, A. Gelbukh and Q. Zhou, “Multilingual sentiment analysis: state of the art and independent comparison of techniques”, Cognitive computation, vol. 8.4, pp. 757-771, 2016.

[38] H. Saadany, C. Orasan, R. C. Quintana, F. D. Carmo and L. Zilio, “Challenges in Translation of Emotions in Multilingual User-Generated Content: Twitter as a Case Study.", arXiv preprint, 2021.

[39] E. D. Gutiérrez, E. Shutova, P. Lichtenstein, G. de Melo and L. Gilardi, “Detecting cross-cultural differences using a multilingual topic model”, Transactions of the Association for Computational Linguistics, vol. 4, pp. 47-60, 2016.

[40] A. Balahur and J. M. Perea-Ortega, “Sentiment analysis system adaptation for multilingual processing: The case of tweets.”, Information Processing & Management, vol. 51.4, pp. 547-556, 2015.

[41] S. L. Lo, R. Chiong and D. Cornforth, “An unsupervised multilingual approach for online social media topic identification”, Expert Systems with Applications, vol. 81, pp. 282-298, 2017.