UCLPRESS

Article title: Public Opinion Analysis on Social Media Platforms: A Case Study of High Speed 2 (HS2) Rail Infrastructure Project

Authors: Ruiqiu YAO[1], Andrew Gillen[2]
Affiliations: civil, environmental and geomatic engineering / university college london / london / the united kingdom[1], department of civil and environmental engineering / northeastern university / boston / the united states[2]
Orcid ids: 0000-0002-2596-5031[1]
Contact e-mail: ruiqiuyao@gmail.com

Preprint statement: This article is a preprint and has not been peer-reviewed, under consideration and submitted to UCL Open: Environment Preprint for open peer review.
Links to data: https://github.com/RY7415/OpinionAnalysisSocialMedia
DOI: 10.14324/111.444/000154.v2
Preprint first posted online: 02 June 2023
Keywords: Public opinion evaluation, Civil infrastructure projects, Machine learning, Sentiment analysis, Topic modelling, Transformers, neural network, Policy and law, Environmental policy and practice, Transport, Sustainability, Statistics

Public Opinion Evaluation on Social Media Platforms: A Case Study of High Speed 2 (HS2) Rail Infrastructure Project

Ruiqiu Yao [1*], Andrew Gillen [2]

[1] University College London; Ruiqiu.yao.19@ucl.ac.uk
[2] Northeastern University; a.gillen@northeastern.edu
***** Corresponding author

Abstract: Public opinion evaluation is becoming increasingly significant in infrastructure project assessment. The inefficiencies of conventional evaluation approaches can be improved with social media analysis. Posts about infrastructure projects on social media provide a large amount of data for assessing public opinion. This study proposed a hybrid model which combines pre-trained RoBERTa and gated recurrent units for sentiment analysis. We selected the United Kingdom railway project, HighSpeed 2, as the case study. The sentiment analysis showed the proposed hybrid model has good performance in classifying social media sentiment. Furthermore, the study applies LDA topic modelling to identify key themes within the tweet corpus, providing deeper insights into the prominent topics surrounding the HS2 project. The findings from this case study serve as the basis for a comprehensive public opinion evaluation framework driven by social media data. This framework offers policymakers a valuable tool to effectively assess and analyse public sentiment.

Keywords: Public opinion evaluation; Civil infrastructure projects; Machine learning; Sentiment analysis; Topic modelling

1. Introduction

Infrastructure systems lay the foundation of the economy for a nation by providing primary transportation links, dependable energy systems, and water management systems to the public. In the United Kingdom, the National Infrastructure Strategy 2020 reveals the determination of the U.K. government to deliver new infrastructure and upgrade existing infrastructure across the country to boost growth and productivity and achieve a net-zero objective by 2050 [1]. Although infrastructure projects positively affect the national economy, they can negatively impact the environment and society. For instance, they may disrupt the natural habitat of wildlife by filling up water lands. As a result, the wildlife may have to migrate to other regions, causing problems to regional ecology [2].

Environmental Impact Assessments (EIA) are a critical part of the planning and delivery of large infrastructure projects. In EIA research, public participation schemes are receiving increasing popularity. O'Faircheallaigh [3] emphasised the importance of public participation in EIA decision-making processes. Social media platforms are gaining increasing ubiquity and are emerging methods for the public to participate in decision-making processes and raise environmental concerns. Thus, the research objective of this study is to evaluate the feasibility of using social media data to perform public participation analysis.

1.1 Conventional approaches to public opinion evaluations

Public hearings and public opinion polling are the two most adopted public consultation approaches. Checkoway [4] stated some drawbacks of public hearings. For instance, the technical terms are hard to understand for the public, and participants often do not represent the actual population. As for polling, Heberlein [5] revealed that conducting polling can usually take a month or even years. Since civil infrastructure projects typically have tight project timelines, there is a need for a more efficient public opinion evaluation method.

Moreover, Ding [6] argued that the data collection process is costly for conventional opinion polling. A typical 1,000-participant telephone interview will cost tens of thousands of U.S. dollars to operate [7]. Besides conducting surveys, costs associated with data input and data analysis should also be considered [6].

Public hearings and polling are not ideal for obtaining public opinions for infrastructure projects. They can be costly, invasive, and time-consuming. Therefore, researchers have drawn attention to developing an alternative method for obtaining and assessing public opinion. A new opportunity in bringing and evaluating public opinion has emerged with the growing popularity of various social media platforms [8]. User-generated content on social media platforms provides a tremendous amount of data for text mining. This text data is an alternative resource for opinion evaluation toward civil infrastructure projects.

1.2 Related work on public opinion evaluation with social media analysis

Kaplan and Haenlein [9] defined social media platforms as internet-based applications adopting Web 2.0 (participative Web). Due to the number of active users on Facebook and Twitter, the massive amount of user-generated content provides valuable opportunities for researchers to study various social topics [10]. Moreover, with machine learning and natural language processing, researchers can perform advanced and automated algorithms on social media posts,

such as sentiment analysis and topic modelling. Sentiment analysis can categorise the textual data in social media into different emotional orientations, providing an indicator of public opinion. Recent research in infrastructure project evaluation with social media analysis revealed the feasibility of using social media analysis as an alternative public opinion evaluation method.

Aldahawi [11] investigated social networking and public opinion on controversial oil companies by sentiment analysis of Twitter data. Kim and Kim [12] adopted lexicon-based sentiment analysis for public opinion sensing and trend analysis on nuclear power in Korea. Lexicon-based sentiment analysis with domain-specified dictionaries and topic modelling has also been used on public opinion data for the California High-Speed Rail and Three Gorge Project [6], [8]. Lexicon-based sentiment analysis calculates the sentiment of documentation from the polarity of words [13]. In lexicon-based sentiment analysis, it is assumed that words have inherent sentiment polarity independent of their context. A user must establish dictionaries containing words with sentiment polarity to build a lexicon-based classifier. After building up the classifier, the polarity of a document is calculated in three phases: establishing word-polarity value pairs, replacing words in the document with polarity values, and calculating the sentiment polarity for the document. Ding [6] tailormade the dictionary by removing unrelated words from a positive word list. Jiang, Qiang, and Lin [8] built a dictionary for hydro projects by integrating the Nation Taiwan Sentiment Dictionary [14], Hownet (a Chinese/English bilingual lexicon database) [15], and a hydro project-related word list. Recent research showed the practicality of implementing the lexicon-based sentiment analysis for public opinion evaluation on civil projects. The recent developments in deep learning show a promising future for public opinion evaluation.

1.3 Recent development of natural language processing

In 2014, Bahdanau, Cho, and Bengio [16] introduced a novel neural network architecture named attention mechanisms. Attentional mechanisms are designed to mimic cognitive perception, which computes the attention weight on input sequences so that some parts of the input data obtain more attention than the rest. In 2017, Vaswani et al. [17] published their ground-breaking research paper "Attention is all you need", where Vaswani et al. proposed an influential neural network named transformer. The transformer architecture leverages self-attention and multi-head attention to enable parallel computation. Using multiple attention

heads and self-attention mechanism, the transformer architecture can obtain different aspects of input data through learning different functions. As a result, transformer architecture can handle increased model and data size. Kaplan et al. [18] demonstrated that transformer models have remarkable scaling behaviour as model performance increases with training size and model parameters. Hence, natural language processing can benefit from large-language models, such as GPT [19], [20] and BERT [21].

1.4 Research question and main contributions

The recent developments in deep learning research motivate this study to assess how state-of-art machine learning algorithms can help public opinion evaluation on infrastructure projects. The main contributions of this study include:

1) This study proposed a hybrid transformer-recurrent neural network model for sentiment analysis, which combines pre-trained RoBERTa [22] and bidirectional gated recurrent neural networks [23].

2) This study employed tweets data of HighSpeed 2 as a case study, utilising it to compare the performance of proposed RoBERTa-BiGRU with baseline classifiers. Moreover, this study applied topic modelling with Latent Dirichlet Allocation (LDA) on tweet corpus.

3) Based on the insights from the case study results, the study proposes a public opinion evaluation framework that leverages social media data with RoBERTa-BiGRU and topic modelling. This framework provides a valuable tool for policymakers to evaluate public opinion effectively.

The rest of this manuscript is organised as follows: Section 2 provides a detailed exposition of the machine learning algorithms used in this study. Section 3 presents the case study with HS2, delving into the specific details and findings. In Section 4, the study outlines the limitations of this research and suggests potential avenues for future research. Finally, Section 5 concludes the study, summarising the main findings and contributions.

2. Machine learning models

This section provides a comprehensive overview of implementing machine learning algorithms for public opinion evaluation. In Section 2.1, the formulation of the MNB classifier is presented. Section 2.2 introduces the proposed RoBERTa-BiGRU model, highlighting its essential components and architecture. Finally, Section 2.3 discusses the topic modelling technique using LDA.

2.1 Sentiment analysis with MNB classifier

The Naïve Bayes classifier is a family of probabilistic classification models based on the Bayes theorem [24]. The term "Naïve" means the naïve assumption of independence among each pair of features (attributes) and class variable values [25]. More specifically, the "naïve" assumption means that classifiers process the text data independently as bag-of-words, ignoring the relationships among words, such as sequences, and only considering the word frequency in the document. The mathematical formula of the Bayes theorem Eq.(1) states that given n feature vectors $x_1, \dots, x_n$ and class variable y, the probability distribution of y is:

$$P(y|x_1, \cdots, x_n) = \frac{P\ (y)\ P(x_1, \cdots, x_n|y)}{P(x_1, \cdots, x_n)} \tag{1}$$

Because the probability distribution of feature vectors $P(x_1, \cdots, x_n)$ are given by model input, the following classification rule Eq.(2) and Eq.(3) can be obtained [26]:

$$P(y|x_1, \dots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y) \tag{2}$$

$$\hat{y} = {}^{argmax}_{y} P(y) \prod_{i=1}^{n} P(x_i|y) \tag{3}$$

Where $P(y)$ is the frequency distribution of y in the training dataset and $P(x_i|y)$ is determined by the Naïve Bayes classifier assumptions. For example, the Gaussian Naïve Bayes classifier assumes $P(x_i|y)$ follows Gaussian distribution.

In the case of MNB classifier, the multinomial distribution is parameterised by $(\theta_{y1}, \dots, \theta_{yn})$ vectors for each y with n features. θ_{yi} indicates the probability distribution of x_i under class y in the training set. In other words, $\theta_{yi} = P(x_i|y)$. Then, smoothed maximum likelihood estimation [27] can be used to estimate θ_{yi}:

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n} \tag{4}$$

Where N_{yi} is the number of occurrences feature i for sentiment class y. N_y is the number of occurrences of all features for y. α is smoothing prior, which is a hyperparameter to be tuned.

2.2 Sentiment analysis with RoBERTa-BiGRU

As mentioned in Section 1.3, transformer architectures have remarkable scaling ability to handle large training data sizes and model parameters. As a result, researchers have proposed

fine-tuning a pre-trained large-scale transformer model for specific downstream natural language processing tasks. This approach is referred to as transfer learning which leverages knowledge learned from the large-scale database to other downstream tasks [28]. The Bidirectional Encoder Representations from Transformers (BERT) [21] is a large language model that has state-of-the-art for natural language processing performance. The BERT model encodes text data in a bidirectionally way such that BERT can process text tokens in both left-to-right and right-to-left directions. This study used a variant of the BERT model, named Robustly optimised BERT approach (RoBERTa) [22], because RoBERTa is pre-trained on a much larger scale of text data than BERT.

Details of fine-tuning the RoBERTa model for sentiment analysis are shown in Figure 1. RoBERTa used similar transformer architecture as BERT. The input token sequence is passed to multiple self-attention heads, followed by a layer normalisation [29]. The normalised data is subsequently sent to feed-forward networks and a second layer normalisation. Figure 1 shows the transformer architecture of a single encoder layer. RoBERTa model contains multiple encoders based on model preference. A RoBERTa encoder's hidden states can then be fed into a classifier for classification tasks. Noticeably, the "<cls>" token indicate the global representation of input text [28].

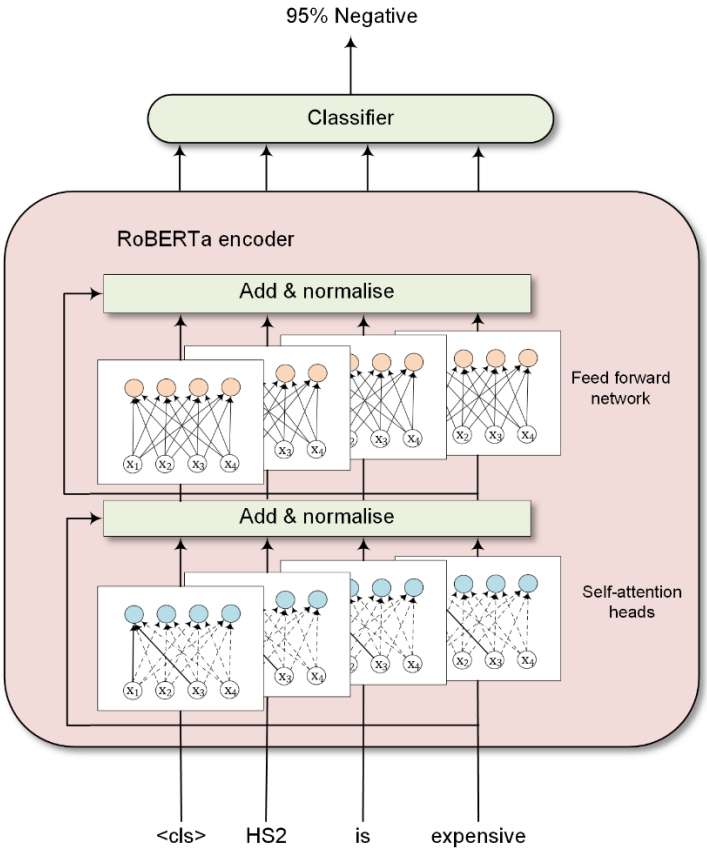

Figure 1 Fine-tuning RoBERTa for sentiment analysis

The classifier can be different neural network architectures, such as FNN or RNN. The Long Short-Term Memory (LSTM) architecture is a prevalent choice as the classifier [30]. The LSTM introduced internal states and gates in addition to RNN to process information in sequenced data [31]. The gated recurrent unit (GRU) architecture, proposed by Cho [23] in 2014, is a streamlined adaptation of LSTM architecture which retains internal states and gating mechanism. This study adopted the GRU architecture as a classifier from RoBERTa outputs because GRU has a faster computation speed than LSTM with comparable performance [32].

The GRU model consists of two internal gates: a reset gate and an update gate. The reset gate determines the extent to which information from the previous state is retained, while the update gate controls the proportion of the new state that replicates the old state. The mathematical formulate of the reset gate and update gate are:

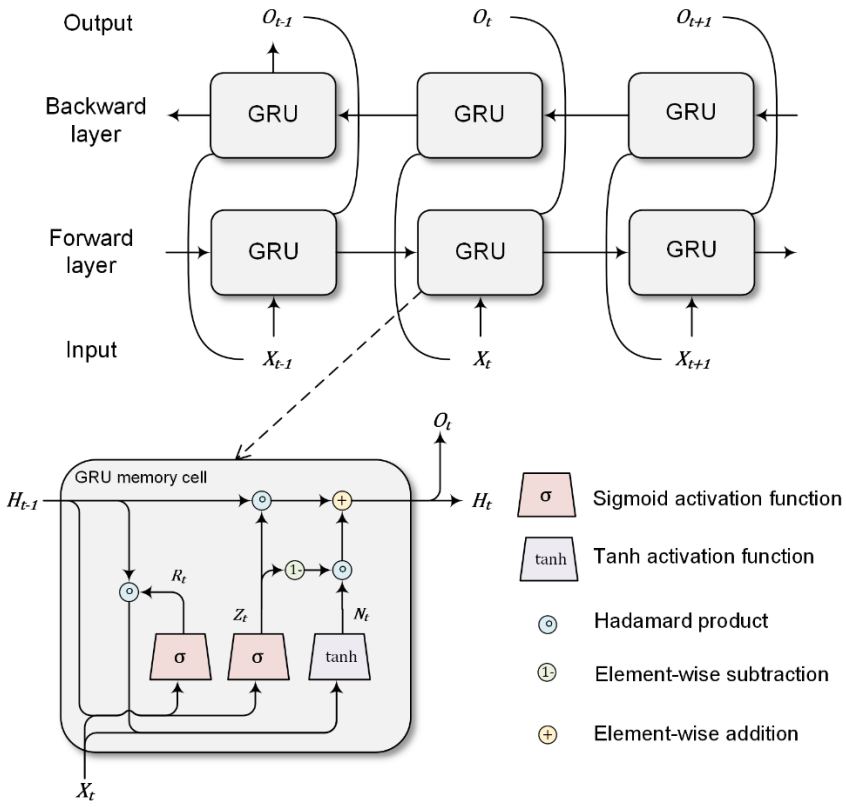

Figure 2 Bidirectional GRU model

$$\boldsymbol{R}_t = \sigma(\boldsymbol{W}_{ir}\boldsymbol{X}_t + b_{ir} + \boldsymbol{W}_{hr}\boldsymbol{H}_{t-1} + b_{hr}) \tag{5}$$

$$\boldsymbol{Z}_t = \sigma(\boldsymbol{W}_{iz}\boldsymbol{X}_t + b_{iz} + \boldsymbol{W}_{hz}\boldsymbol{H}_{t-1} + b_{hz}) \tag{6}$$

$$\sigma(x) = \frac{1}{1 + \exp(-x)} \tag{7}$$

Where $\boldsymbol{X}_t \in \mathbb{R}^{n\times d}$ is a minibatch input of memory cell (n is the number of sample and d is the dimension of features). $\boldsymbol{H}_{t-1} \in \mathbb{R}^{n\times h}$ is the hidden state of previous step (h is the number of hidden units of a GRU memory cell). $\boldsymbol{W}_{ir}, \boldsymbol{W}_{hr} \in \mathbb{R}^{d\times h}$ and $\boldsymbol{W}_{iz}, \boldsymbol{W}_{hz} \in \mathbb{R}^{h\times h}$ are model weights. b_{ir}, b_{hr}, b_{iz}, and b_{hz} are model bias parameters. The reset gate $\boldsymbol{R}_t \in \mathbb{R}^{n\times h}$ and update gate $\boldsymbol{Z}_t \in \mathbb{R}^{n\times h}$ are computed based on Eq.(5) and Eq.(6). In other words, two gates are fully connected layers with sigmoid activation function Eq.(7).

The reset gate is designed to yield candidate hidden state $\boldsymbol{N}_t \in \mathbb{R}^{n\times h}$ with Eq.(8) and $tanh$ activation function Eq.(9). The influences of previous information $\boldsymbol{H}_{t-1}$ in Eq.(8) is reduced by the Hadamard product of $\boldsymbol{R}_t$ and $\boldsymbol{H}_{t-1}$. The candidate hidden state $\boldsymbol{N}_t$ is then passed to Eq.(10) to calculate the new hidden state $\boldsymbol{H}_t$, in which the update gate $\boldsymbol{Z}_t$ controls the degree to which $\boldsymbol{H}_t$ resembles $\boldsymbol{N}_t$.

$$\boldsymbol{N}_t = \tanh\big(\boldsymbol{W}_{in}\boldsymbol{X}_t + b_{in} + \boldsymbol{R}_t \odot (\boldsymbol{W}_{hn}\boldsymbol{H}_{t-1} + b_{hn})\big) \tag{8}$$

$$\tanh(x) = \frac{1 - \exp(-2x)}{1 + \exp(2x)} \tag{9}$$

$$\boldsymbol{H}_t = (1 - \boldsymbol{Z}_t) \odot \boldsymbol{N}_t + \boldsymbol{Z}_t \odot \boldsymbol{H}_{t-1} \tag{10}$$

Where $\boldsymbol{W}_{in} \in \mathbb{R}^{d \times h}$ and $\boldsymbol{W}_{hn} \in \mathbb{R}^{h \times h}$ are model weights. b_{in} and b_{hn} are bias parameters. $\odot$ is Hadamard product, which is also referred to as element-wise product.

Similar to the bidirectional setting of BERT, a two-layer GRU is also able to process the text data bidirectionally with a forward layer and a backward layer, as shown in Figure 2. The hidden state of the forward layer and backward layer is denoted as $\overrightarrow{\boldsymbol{H}_t} \in \mathbb{R}^{n \times h}$ and $\overleftarrow{\boldsymbol{H}_t} \in \mathbb{R}^{n \times h}$. The forward layer hidden states $\overrightarrow{\boldsymbol{H}_t}$ is then multiplied with a dropout rate δ which is a Bernoulli random variable with δ probability of being 0. The output of a GRU is a concatenate of $\overrightarrow{\boldsymbol{H}_{t\_\delta}}$ and $\overleftarrow{\boldsymbol{H}_t}$ with dimension $n \times 2h$.

The RoBERTa model can be fine-tuned by optimising loss-function of the above-mentioned bidirectional GRU and connecting the output of bidirectional GRU with a fully connected layer. The loss function to be optimisation in GRU is cross entropy function [33]. Moreover, the fully connected layer uses the soft-max activation function Eq.(11):

$$s(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \tag{11}$$

Where n is the number of sentiment classes. The fully connected layer converts the hidden states of bidirectional GRU to the probability of each sentiment class.

Figure 3 demonstrates the complete structure of the RoBERTa–BiGRU model. First of all, tweets are tokenised with the RoBERTa tokeniser. Then, the tokens are passed to 12 encoders with multiple self-attention heads to obtain 768 tweets hidden representations. The tweets hidden representations can then be allocated to sentiment classes through bidirectional GRU and fully connected layer.

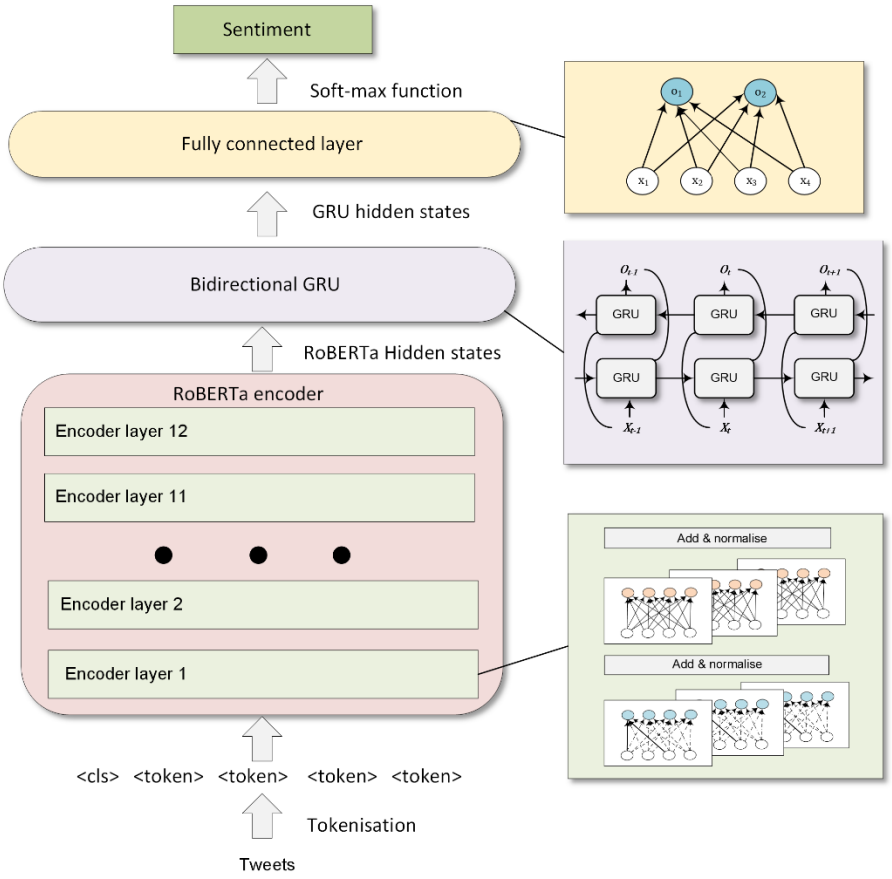

Figure 3 Structure of RoBERTa-BiGRU for sentiment analysis

2.3 Topic modelling with LDA

Deerwester et al. [34] proposed a Latent Semantic Indexing method for topic modelling, applying Singular Value Decomposition (SVD) to drive the latent semantic structure model from the matrix of terms from documents. SVD is a linear algebra technique to decompose an arbitrary matrix to its singular values and singular vectors [35]. Blei, Ng, and Jordan [36] introduced LDA, which is a general probabilistic model of a discrete dataset (text corpus).

LDA is a Bayesian model, which models a document as a finite combination of topics. Each topic is modelled as a combination of topic probabilities. For example, an article that talks about the structural design of a building complex may have various topics, including "structural layout" and "material". The topic "structural layout" may have high-frequency words related to structural design, such as "beam", "column", "slab", and "resistance". Also, the "material" topic may have the words "concrete", "steel", "grade", and "yield". In short, a document has different topics with a probabilistic distribution, and each topic has different words with a

probabilistic distribution. Human supervision is not required in LDA topic modelling, as LDA only needs a number of topics to perform analysis.

Topic modelling with LDA has a wide range of applications in research. Xiao et al. [37] used LDA variant topic modelling to uncover the probabilistic relationship between Adverse Drug Reaction topics. Xiao et al. found that the LDA variant topic modelling has higher accuracy than alternative methods. Jiang, Qiang, and Lin [8] showed the feasibility of LDA topic modelling to extract topics about the Three Gorges Project on a Chinese social media platform. Apart from focusing on extracting terms from textual corpus, topic modelling is another trend-finding tool, as it will reveal the relationship between topics. Chuang et al. [38] proposed a method to visualise topics with circles in a two-dimensional plane, whose centre is determined by the calculated distance between topics. The distance is calculated by Jenson-Shannon divergence, and Principal Components analysis determines the size of the circle [39].

3. Case study with HighSpeed 2 project

This section provides implementation details of sentiment classifiers and topic modelling methods for HighSpeed 2 case study. In Section 3.1, the background of the HS2 project is presented, offering insights into the rail infrastructure project. Section 3.2 explains data collection and processing, detailing the methods employed to gather social media data related to HS2. Section 3.3 presents the evaluation metrics used to assess the performance of sentiment classifiers, enabling a thorough examination of sentiment classification models. Sections 3.4 and 3.5 show the results of sentiment analysis and topic modelling respectively. Finally, Section 3.6 introduces a framework for evaluating public opinion based on social media data.

3.1 Background on HighSpeed 2 project

The transportation demand for the U.K. railway network has steadily grown over the past decades. According to the Department for Transport [40], rail demand has doubled since 1994-95, with a rising rate of 3% every year. Therefore, HS2 programme is proposed to construct a new high-speed and high-capacity railway, aiming to boost the economy in the U.K., improve connectivity by shortening journey time, provide sufficient capacity to meet future railway network demand, and reduce carbon emission by reducing long-distance driving. Figure 4 shows that HS2 will connect London, Leeds, Birmingham, and Manchester, joining existing railway infrastructure to allow passengers to travel to Glasgow, Newcastle, and Liverpool [41].

Figure 4 HS2 infrastructure map [41]

3.2 Data preparation

The collection of HS2-related tweets was carried out by using the Twitter Application Programming Interfaces (API). Specifically, tweets that containing the hashtags "#HS2" and "#HighSpeed2" were collected. However, the number of collectable tweets is constrained by the limitations imposed by the Twitter API, which restricts the collection to under 10,000 tweets. Thus, the total number of tweets collected is 8623 tweets. The tweets were sampled over a five-year period from 2017 to 2020. The tweets distribution across the years is: 2017 (1544 tweets), 2018 (1130 tweets), 2019 (2909 tweets), and 2020 (3040 tweets). Noticeably, the tweets collected were in extended mode, allowing the retrieval of the complete text, surpassing the 140-character limit.

Data pre-processing involves cleaning and preparing data to increase the accuracy and performance of text-mining tasks, such as sentiment analysis and topic modelling. Tweet text

data tend to contain uninformative text, such as URL links, Twitter usernames, and email. For MNB and lexicon-based classifier, the stop words need to be removed. To be more specific, stop words are words that don't have sentiment orientation, such as "me", "you", "is", "our", "him", and "her". Since each word in text data is treated as a dimension, keeping stop words and uninformative text will complicate the text mining by making text mining a high dimension problem [42]. Other text pre-processing techniques for MNB and lexicon-based classifier include text lowercasing and text stemming. Noticeably, the transformer architectures do not require removing stop words, lowercasing, and text stemming, as transformers are able to handle the implied information in stop words.

Upon conducting a manual inspection of collected tweets, the number of tweets expressing positive sentiment was significantly lower than those with negative or neutral sentiment. The sentiment classification task is set to binary to address the imbalance issue. The task was designed to classify tweets as either having negative sentiment or non-negative sentiment (include neutral and positive sentiment). A set of 1,400 tweets was carefully annotated to train classifiers in this case study. Within this annotated dataset, 700 tweets were labelled negative sentiment, while the remaining 700 tweets were labelled non-negative sentiment. To access the annotated training tweets, a GitHub link is provided in the *Open data and materials availability statement*, facilitating transparency and reproducibility of this study. The annotated tweets were split into 70% training dataset (980 tweets) and 30% validation dataset (420 tweets).

3.4 Sentiment analysis results

Three sentiment classifiers were used in this case study: 1) VADER [43], a rule-based lexicon sentiment classifier. 2) an MNB classifier which is built following details in Section 3.1. 3) a RoBERTa-BiGRU model that is developed based on the architecture presented in Section 3.2. The model details of each classifier are shown in Table 1. The hyperparameters in MNB and RoBERTa-BiGRU, such as smoothing priors α, batch size, hidden units, and dropout rate, were tuned by grid search. The RoBERTa-BiGRU model is trained on a Tesla T4 GPU on Google Colab with a total training time of 2421.23 seconds for 100 epochs.

Table 1 Model details of each classifier

Name	Model parameters
VADER	Rules specified in [43]

MNB	Smoothing priors: $\alpha = 0.1$
RoBERTa-BiGRU	Batch size: 16
	Hidden units: 256
	Dropout rate: 0.5
	Optimiser: AdamW
	Learning rate: $2 \times e^{-6}$
	Epoch:100

The performances of three classifiers were evaluated with accuracy and ROC curve. Accuracy, as shown in Eq.(12), measures the accuracy of the classifier with all correctly identified cases overall identified cases. A Receiver Operating Characteristic(ROC) curve plots the true positive rate, as shown Eq.(13), along Y axis and false positive rate, as shown in Eq.(14), along X axis. A ROC curve shows the graphical interpretation of gain (true positive rate) and loss (false positive rate) [44]. Area Under the Curve (AUC) score calculates the total area under the ROC curve. The AUC score quantitatively evaluates the performance of a classifier, which represents the possibility of a random positive datapoint ranks higher than a random negative datapoint [45].

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$True\ positive\ rate(recall) = \frac{TP}{TP + FN} \tag{13}$$

$$False\ \ positive\ rate = \frac{FP}{FP + TN} \tag{14}$$

Where $TP = True\ Positive$, $TN = True\ Negative$, $FP = False\ Positive$, and $FN = False\ Negative$

Table 2 demonstrates the accuracy of each sentiment classifier. The lexicon-based VADER have the lowest accuracy (70.24%) among the three classifiers. MNB and RoBERTa-BiGRU show better accuracy performance than VADER, where MNB and RoBERTa-BiGRU increased accuracy 12.38% and 19.28% respectively. MNB and RoBERTa-BiGRU are then compared with respect to AUC scores. MNB has an AUC score of 0.9023, while RoBERTa-BiGRU has an slightly lower AUC score of 0.8904. Both MNB and RoBERTa-BiGRU have around 0.9 AUC score, which indicates both models have a high level of classification ability

to classify tweets sentiment. Noticeably, Figure 5 (b) has a much steeper curve. The steeper curve means that RoBERTa-BiGRU can achieve higher recall with a low false positive rate, which is desirable behaviour in sentiment analysis. As a result, RoBERTa-BiGRU has the best performance in terms of both accuracy and ROC curve. Then, the RoBERTa-BiGRU is used for sentiment analysis with all collected tweets.

Table 2 Model accuracy performance

Name	Accuracy
VADER	70.24%
MNB	82.62%
RoBERTa-BiGRU	89.52%

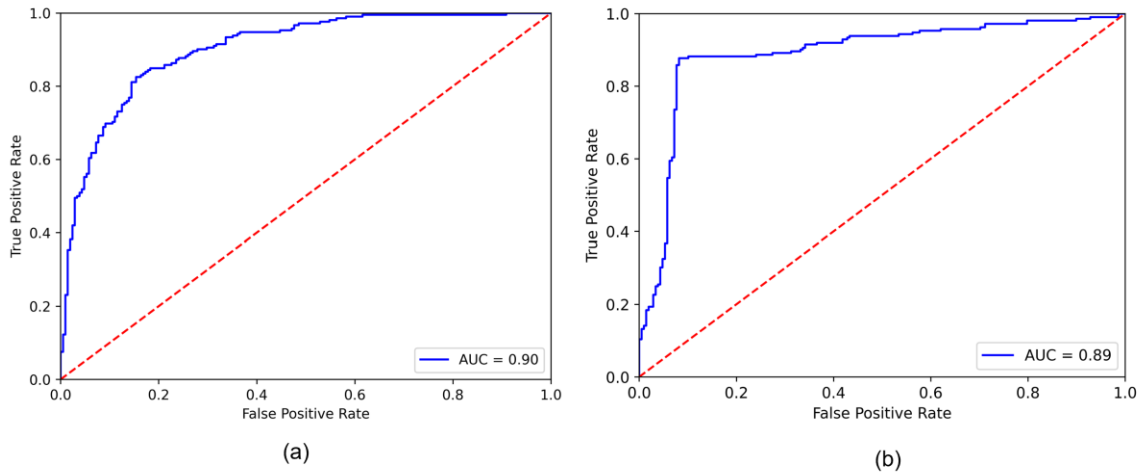

Figure 5 (a) ROC curve for MNB classifier. (b) ROC curve for RoBERTa-BiGRU.

Figure 6 shows the sentiment distribution of HS2-related tweets from 2017 to 2020. Notably, there was a substantial increase in the number of tweets in 2019, indicating a heightened presence of the HS2 project in social media discussions during and after that year. Moreover, it is worth mentioning that the majority of collected tweets across all time periods exhibited a negative sentiment. Specifically, negative tweets accounted for 57.77% in 2017, 53.32% in 2018, 60.64% in 2019, and 65.19% in 2020.

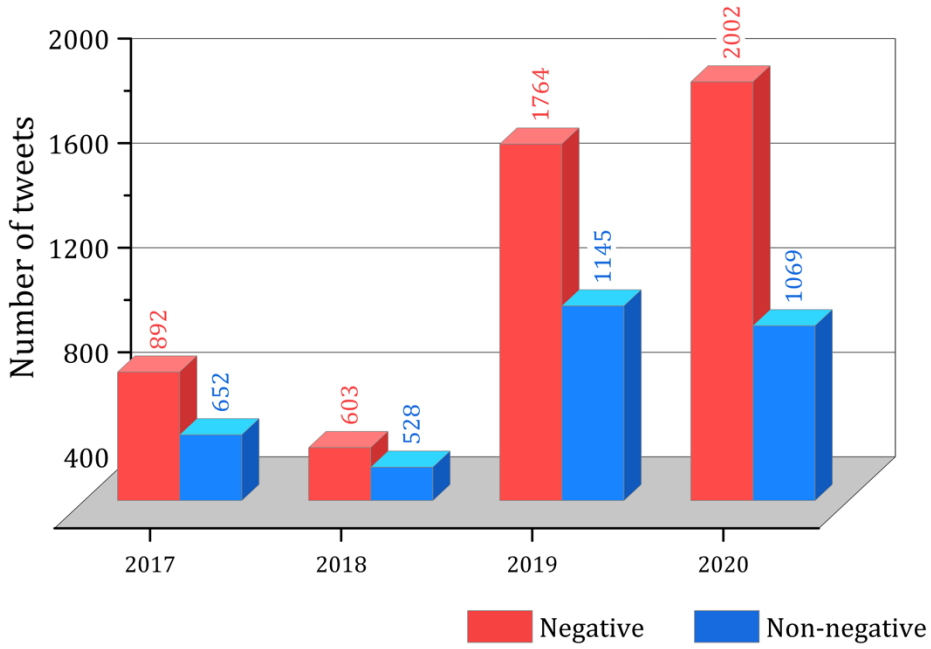

Figure 6 Sentiment analysis results for 2017 to 2020

The substantial proportion of negative tweets in all periods indicates a prevailing negative sentiment among the public regarding HS2, highlighting the importance for policymakers and decision-makers to take this sentiment into consideration. However, it is essential to approach these findings with caution. While the high percentage of negative tweets may raise concerns, it is crucial to note that this alone does not necessarily imply a public relationship emergency for HS2. It is worth acknowledging that certain Twitter users might repeatedly express their negative sentiment towards HS2 [46], potentially influencing the overall sentiment distribution. Given the sentiment analysis results, it is important to uncover the key topics within the tweets discussions, necessitating the application of topic modelling.

3.5 Topic modelling results

The tweets dataset is then classified by the RoBERTa-BiGRU model into two collections: negative corpus and non-negative corpus. Each collection was performed with topic modelling and visualisation individually. Topic modelling with Latent Dirichlet Allocation (LDA) was performed with genism, a collection of Python scripts developed by Rehurek and Sojka [47]. We used pyLDAvis for visualising topics such that we could determine the most suitable number of topics. Several models were constructed with a number of topics ranging from 3 to 20. We selected five as the number of topics through manual inspection of term distribution and topic relevance.

3.5.1 Negative tweets corpus

Table 3 shows major topics in the negative corpus. Topic 1 is the largest topic which accounts for 35.3% of the negative corpus. Topic 1 contains words like "need", “money”, “nhs”, “badly”, “billion”. These words express the negative sentiment on HS2 budget spending. These tweets criticise the over-spending of HS2 and argue that the money should be invested in National Health Service (NHS) rather than HS2. Topic 2 and Topic 4 have a similar focus. Topic 2 has words like “government”, “protester”, “social”, and Topic 4 include words like “stophs2”, “petition”, “media”, “political”. Both topic 2 and topic 4 discuss the campaign to stop HS2 project by petition. Topic 3 and Topic 5 show some relevance. Topic 3 contains “stop”, “please”, “trees”, “contractors”, “changed”, “essential”, which raises environmental concerns about construction work on woodland. Topic 5 also discusses the environmental issues with the words “construction”, “damage”, and “destroy”.

Table 3 Topics in negative corpus

Topic number	Terms	Topic percentage
1	Borisjohnson, hs2, work, time, need, money, say, nhs, use, uk, course, amp, transport, nt, cancel, even local, badly, billion, ancient, public, needed, boris, way, think, country, rishisunak, trains, know	35.3%
2	Rail, government, going, still, protesters, like, news, case, go, social, could, economic, train, people, home, London, times, business, ltd, working, travel, back, road, north, sense, says, dont	24.2%
3	Stop, post, mps, please, another, anti, away, seems, trees, make, already, without, contractors, may, changed, control, steeple, long, big, bill, sign, essential, protest, claydon, likely, means, yet, billions, station, caught	13.9%
4	Sopths2, workers, petition, sites, via, take, destruction, ever, change, media, track, year, ukparliament, least, investment, everyone, account, despite, find, continue, political, wants, white, along, british, longer, evidence, called, massive, elephant	13.6%
5	Report, scrap, construction, costs, last, end, law, latest, true, tax, first, damage, full, job, trident, nesting, figures, wonder, share, read, unnecessary, questions, destroy, failed, coming, vital	13.1%

3.5.2 Non-negative tweets corpus

Table 4 shows topics in a non-negative corpus. Topic 1 includes words like “new”, “railway”, “good”, “midlands”, “important”, where tweets express positive sentiment on HS2 by mentioning the positive effect on the Midland area. A similar result can be found in Topic 3, which includes words like “planning”, “Manchester”, “airport”, “benefit”, “better”. Topic 3 highlights the transportation infrastructure in Manchester could benefit from HS2 project. Topic 2 discusses the business case of HS2 with words “project”, “business”, “build”, “network”, and “industry”. Topic 4 and 5 both discuss on potential improvements on accessibility to the airport with words “heathrow”, “airports”, “opportunities”. Overall, the LDA topic modelling showed good execution on obtaining key topics from the tweet corpus.

Table 4 Topics in non-negative corpus

Topic number	Terms	Topic percentage
1	Work, new, project, one, railway, station, first, time, may, ever, people, plans, common, good, midlands, find, watch, still, well, way , may, could, largest, part, back, important, day	35.4%
2	Construction, hs2ltd, rail, post, projects, train, business, build, track, road, read, network, phase, industry, latest, leaders, think, green, big, please, works, air, know, local, year, along	24.4%
3	High, speed, need, old, north, planning. would, capacity, built, engineering, course, Manchester, building, another, plan, recent, airport, must, benefit, needs, evidence, better, needed, chief, funding	15.9%
4	Government, news, trains, us, would, home, two, heathrow, cost, start, railways, service, suppliers, roads, update, every, keep, seems, question, longer, join, money	13.3%
5	Stations, use, lake, community, following, scheme, economic, really, opportunities, spending, committee, supply, benefits, due, chain, role, ealy, daily, fund, freight, article, essential, airports	11.1%

3.6 Proposed public opinion evaluation framework using social media data

The case study results showed that the RoBERTa-BiGRU and LDA topic modelling has a good performance in evaluating public opinion on HS2 with tweet data. Hence, a public opinion evaluation framework using social media data is proposed to facilitate the decision-making of policymakers.

Figure 7 presents the comprehensive public opinion evaluation framework that utilises social media data. The process begins by collecting social media data, such as tweets, and storing them in a database. Subsequently, the social media data is processed through sentiment annotation, which involves labelling the data to create training sets. These training sets are then utilised for training a sentiment classifier called RoBERTa-BiGRU. Once the RoBERTa-BiGRU sentiment classifier is trained, it is employed to categorise social media tweets into predefined sentiment labels. Additionally, leveraging LAD topic modelling, the framework extracts key topics from the social media data. Policymakers can subsequently utilise the sentiment analysis results and key topics to evaluate public opinion regarding infrastructure projects.

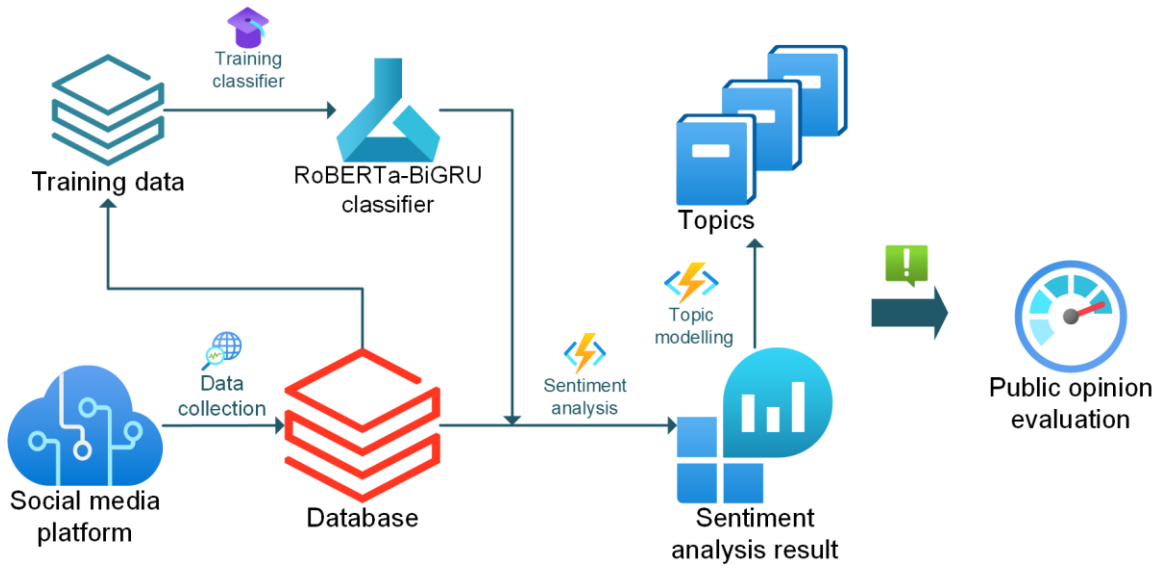

Figure 7 Public opinion evaluation framework

4. Limitation and future research direction.

4.1 Human factors in annotating tweets sentiments

Researchers usually assign multiple annotators (3 to 5) to tag the sentiment orientation to minimise the influence of human annotators [48]. However, in our study, all the training data was tagged by one annotator. As a result, the human factor may have affected the accuracy of the sentiment classifier. The future application on fine-tuning sentiment classifiers could benefit from multiple annotators.

Another impact of human factors could be different sentiment interpretations. For example, the following tweet may be tagged with different sentiment orientations. “#HS2 is a £100bn scheme to have slightly shorter journey times from Manchester and Birmingham to London,

thereby solving Britain's biggest ever problem." One annotator can argue that there are positive sentiment signs (shorter journey time and solving problems). In contrast, another annotator could also argue that the tweet used a sarcastic tone to express a negative sentiment towards over budget issue of HS2.

4.2 Topic modelling challenges

Text documents are combinations of probabilistic distributions of topics, and each topic is a probabilistic distribution of words. However, tweets are short microblogs with character limitations (280 characters), which usually contain one topic. Therefore, LDA may have problems in calculating the probabilistic distribution of topics of tweets data. The performance of tweets topic modelling could be improved with the neural optic models, leveraging deep generative models [49]. Future research on public opinion evaluation with social media data could use the Bayesian networks. In particular, gamma-belied networks showed promising results in yielding structure topics [50].

5. Conclusion

This study utilised tweets data from the HS2 project as a case study. The tweets data were used to compare the performance of the proposed RoBERTa-BiGRU model with MNB and VADER. RoBERTa-BiGRU showed the best performance in terms of accuracy and ROC curves. Additionally, the study employs LDA to uncover key topics within the tweet corpus. This analysis enhances understanding of the prominent themes surrounding the HighSpeed 2 project. The insights derived from the HS2 case study results lay the foundation for a public opinion evaluation framework. This framework, driven by social media data, is an invaluable tool for policymakers to evaluate public sentiment effectively. Overall, this study contributes to the field of public opinion evaluation by introducing a hybrid model, presenting a comprehensive case study analysis, and proposing a practical framework for public opinion evaluation.

Open data and materials availability statement

To ensure the transparency and reproducibility of this study. The collected data (anonymised) and the Python source code in this study are available on the GitHub repository: https://github.com/RY7415/OpinionAnalysisSocialMedia

Authorship Contribution

This research was initially conducted as part of the requirements for the MSc in Civil Engineering at University College London. Mr. Ruiqiu Yao was supervised by Dr. Andrew Gillen for his MSc dissertation. The general topic and use of social media data were proposed by Dr. Gillen, and they met regularly to discuss the research process. Mr. Yao conducted the literature review as well as the data collection and analysis, identifying relevant sources of data and analytical tools. Mr. Yao drafted the manuscript and Dr. Gillen provided feedback on drafts.

Funding Statement

This study did not receive funding.

Declarations and Conflict of Interests statement

The authors declare no conflict of interest with this work

Reference

[1] HM Treasury, "National Infrastructure Strategy ," 2020.

[2] D. J. Hayes, "Addressing the environmental impacts of large infrastructure projects: making 'mitigation' matter," 2014.

[3] C. O'Faircheallaigh, "Public participation and environmental impact assessment: Purposes, implications, and lessons for public policy making," *Environ Impact Assess Rev*, vol. 30, no. 1, pp. 19–27, Jan. 2010, doi: 10.1016/j.eiar.2009.05.001.

[4] B. Checkoway, "The Politics of Public Hearings," *J Appl Behav Sci*, vol. 17, no. 4, pp. 566–582, Jul. 1981, doi: 10.1177/002188638101700411.

[5] T. Heberlein, "Some Observations on Alternative Mechanisms for Public Involvement: The Hearing, Public Opinion Poll, the Workshop and the Quasi-Experiment," *Nat Resour J*, vol. 16, no. 1, 1976.

[6] Q. Ding, "Using Social Media to Evaluate Public Acceptance of Infrastructure Projects," University of Maryland, 2018. doi: doi.org/10.13016/M27M0437D.

[7] B. O'connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series," in *Proceedings of the Fourth International AAAI conference on Webblogs and Social Media*, 2010.

[8] H. Jiang, M. Qiang, and P. Lin, "Assessment of online public opinions on large infrastructure projects: A case study of the Three Gorges Project in China," *Environ Impact Assess Rev*, vol. 61, pp. 38–51, Nov. 2016, doi: 10.1016/j.eiar.2016.06.004.

[9] A. M. Kaplan and M. Haenlein, “Users of the world, unite! The challenges and opportunities of Social Media,” *Bus Horiz*, vol. 53, no. 1, pp. 59–68, Jan. 2010, doi: 10.1016/j.bushor.2009.09.003.

[10] S. B. Park, C. M. Ok, and B. K. Chae, “Using Twitter Data for Cruise Tourism Marketing and Research,” *Journal of Travel and Tourism Marketing*, vol. 33, no. 6, pp. 885–898, Jul. 2016, doi: 10.1080/10548408.2015.1071688.

[11] H. A. Aldahawi, “Mining and Analysing Social Network in the Oil Business : Twitter Sentiment Analysis and Prediction Approaches,” Cardiff University, 2015.

[12] D. S. Kim and J. W. Kim, “Public Opinion Sensing and Trend Analysis on Social Media: A Study on Nuclear Power on Twitter 1,” *International Journal of Multimedia and Ubiquitous Engineering*, vol. 9, no. 11, pp. 373–384, 2014, doi: 10.14257/ijmue.2014.9.11.36.

[13] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, “Lexicon-basedmethods for sentiment analysis,” *Computational Linguistics*, vol. 37, no. 2, pp. 267–307, Jun. 2011, doi: 10.1162/COLI_a_00049.

[14] L.-W. Ku, Y.-T. Liang, and H.-H. Chen, “Opinion Extraction, Summarization and Tracking in News and Blog Corpora,” 2006.

[15] Z. Dong and Q. Dong, “HowNet - A hybrid language and knowledge resource,” in *NLP-KE 2003 - 2003 International Conference on Natural Language Processing and Knowledge Engineering, Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2003, pp. 820–824. doi: 10.1109/NLPKE.2003.1276017.

[16] D. Bahdanau, K. Cho, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.0473

[17] A. Vaswani *et al.*, “Attention Is All You Need,” Jun. 2017, [Online]. Available: http://arxiv.org/abs/1706.03762

[18] J. Kaplan *et al.*, “Scaling Laws for Neural Language Models,” Jan. 2020, [Online]. Available: http://arxiv.org/abs/2001.08361

[19] T. B. Brown *et al.*, “Language Models are Few-Shot Learners,” May 2020, [Online]. Available: http://arxiv.org/abs/2005.14165

[20] OpenAI, “GPT-4 Technical Report,” Mar. 2023, [Online]. Available: http://arxiv.org/abs/2303.08774

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.04805

[22] Y. Liu *et al.*, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.11692

[23] K. Cho *et al.*, “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation,” Jun. 2014, [Online]. Available: http://arxiv.org/abs/1406.1078

[24] T. Bayes, “LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S,” *Philos Trans R Soc Lond*, vol. 53, pp. 370–418, Dec. 1763, doi: 10.1098/rstl.1763.0053.

[25] L. Jiang, Z. Cai, H. Zhang, and D. Wang, “Naive Bayes text classifiers: A locally weighted learning approach,” *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 25, no. 2, pp. 273–286, Jun. 2013, doi: 10.1080/0952813X.2012.721010.

[26] H. Zhang, “The Optimality of Naive Bayes,” in *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*, AAAI Press, 2004, pp. 562–567. [Online]. Available: www.aaai.org

[27] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008. doi: 10.1017/CBO9780511809071.

[28] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, *Dive into Deep Learning*. 2021.

[29] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer Normalization,” Jul. 2016, [Online]. Available: http://arxiv.org/abs/1607.06450

[30] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, “RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network,” *IEEE Access*, vol. 10, pp. 21517–21525, 2022, doi: 10.1109/ACCESS.2022.3152828.

[31] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[32] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling,” Dec. 2014, [Online]. Available: http://arxiv.org/abs/1412.3555

[33] T. Gneiting and A. E. Raftery, “Strictly Proper Scoring Rules, Prediction, and Estimation,” *J Am Stat Assoc*, vol. 102, no. 477, pp. 359–378, Mar. 2007, doi: 10.1198/016214506000001437.

[34] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” *Journal of the American Society for Information*

Science, vol. 41, no. 6, pp. 391–407, 1990, doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

[35] K. Kanatani, *Linear Algebra for Pattern Processing Projection, Singular Value Decomposition, and Pseudoinverse*. Morgan & Claypool, 2021.

[36] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[37] C. Xiao, P. Zhang, W. Art Chaowalitwongse, J. Hu, and F. Wang, "Adverse Drug Reaction Prediction with Symbolic Latent Dirichlet Allocation," Feb. 2017.

[38] J. Chuang, D. Ramage, C. D. Manning, and J. Heer, "Interpretation and trust: Designing model-driven visualizations for text analysis," in *Conference on Human Factors in Computing Systems - Proceedings*, 2012, pp. 443–452. doi: 10.1145/2207676.2207738.

[39] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," in *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, Association for Computational Linguistics (ACL), Jun. 2014, pp. 63–70. doi: 10.3115/v1/w14-3110.

[40] Department for Transport, "Rail Factsheet 2019," 2019.

[41] HS2 Ltd, "High-speed network map," *https://www.hs2.org.uk/the-route/high-speed-network-map/*, 2023.

[42] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," in *Procedia Computer Science*, Elsevier B.V., Jan. 2013, pp. 26–32. doi: 10.1016/j.procs.2013.05.005.

[43] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," in *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 2014, pp. 216–225. [Online]. Available: http://sentic.net/

[44] J. Davis and M. Goadrich, "The Relationship Between Precision-Recall and ROC Curves," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[45] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit Lett*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.

[46] J. G. Rozema and A. J. Bond, "Framing effectiveness in impact assessment: Discourse accommodation in controversial infrastructure development," *Environ Impact Assess Rev*, vol. 50, pp. 66–73, 2015, doi: 10.1016/j.eiar.2014.08.001.

[47] R. Rehurek and P. Sojka, “Software Framework for Topic Modelling with Large Corpora,” in *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010, pp. 46--50.

[48] C. Callison-Burch, “Fast, cheap, and creative: Evaluating translation quality using Amazon’s Mechanical Turk,” in *EMNLP 2009 - Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: A Meeting of SIGDAT, a Special Interest Group of ACL, Held in Conjunction with ACL-IJCNLP 2009*, Singapore, Aug. 2009, pp. 286–295.

[49] H. Zhao, D. Phung, V. Huynh, Y. Jin, L. Du, and W. Buntine, “Topic Modelling Meets Deep Neural Networks: A Survey,” Feb. 2021, [Online]. Available: http://arxiv.org/abs/2103.00498

[50] H. Zhang, B. Chen, D. Guo, and M. Zhou, “WHAI: Weibull Hybrid Autoencoding Inference for Deep Topic Modeling,” Mar. 2018, [Online]. Available: http://arxiv.org/abs/1803.01328