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

Links to data: https://github.com/RichardYao08/PinionAnalysisSocialMedia

DOI: 10.14324/111.444/000154.v1

Preprint first posted online: 16 June 2022

Keywords: Public opinion evaluation, Civil infrastructure projects, Machine learning, Sentiment analysis, Topic modelling, Policy and law, Environmental policy and practice, Transport, Sustainability

Dear Editors of *UCL OPEN ENVIRONMENT*:

I am writing to submit an article entitled "Public Opinion Analysis on Social Media Platforms: A Case Study of High Speed 2 (HS2) Rail Infrastructure Project" for consideration for publication in *UCL OPEN ENVIRONMENT*.

Environmental impact assessment is critical in civil infrastructure planning since civil infrastructure projects significantly impact the environment. In EIA research, public participation schemes are receiving increasing popularity. Thus, we used social media analysis and machine learning to evaluate public opinion on civil infrastructure projects in this manuscript. We adopted sentiment analysis and topic modelling on a case study of public Twitter data around a United Kingdom railway project, High Speed 2 (HS2). The results showed that the public raised environmental concerns over railway planning and construction. These opinions are highly valuable in decision and policymaking. We believe this manuscript is appropriate for publication by *UCL OPEN ENVIRONMENT* because it presents a study exploring a civil infrastructure project's environmental and social sustainability.

Please let me know if you have any questions or concerns regarding this submission, and I will be happy to address them. The contact information given below is the best ways to reach me. I look forward to hearing from you.

We confirm that neither the manuscript nor any parts of its content are currently under consideration or published in another journal.

All authors have approved the manuscript and agree with its submission to *UCL OPEN ENVIRONMENT*.

Sincerely,

Ruiqiu Yao
MPhil/PhD Student
Department of Civil, Environmental, and Geomatic Engineering
University College London
Email: Ruiqiu.yao.19@ucl.ac.uk

Public Opinion Analysis 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
* Correspondence: Ruiqiu.yao.19@ucl.ac.uk

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 public opinion evaluation framework with machine learning algorithms, including sentiment analysis and topic modelling. We selected the United Kingdom railway project, High Speed 2, as the case study. The sentiment analysis showed that around 53% to 63% of tweets expressed a negative sentiment, suggesting the public may have an overall negative perception of the project. Topic modelling with text corpora showed key topics of public opinion. The proposed framework demonstrates the feasibility of using supervised machine learning to evaluate public opinion on infrastructure projects, as the framework can save time and cost. Furthermore, assessment results can aid policymakers and managers in decision-making.

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, as well as 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 delivering 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(4). Checkoway (5) stated some drawbacks of public hearings. For instance, the technical terms are hard to understand for the public, and participants are often not representative of the actual population. As for polling, Heberlein (6) 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 (4) 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 (4).

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 Evaluating public opinion 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 (10,11). Moreover, with machine learning and natural language processing, researchers can perform advanced and automated algorithms, such as sentiment analysis and topic modelling on social media posts. Sentiment analysis can categorise the textual data in social media into different emotional orientations (positive, negative, neutral), providing an indicator of public satisfaction. Topic modelling can uncover the important topics of public opinion from users' social media posts. Recent research in infrastructure project evaluation with social media analysis (i.e. sentiment analysis/topic modelling) revealed the feasibility of using social media analysis as an alternative public opinion evaluation method (4,8,12). In terms of sentiment analysis techniques, the abovementioned research adopted a lexicon-based approach (i.e. using a predefined dictionary). However, the performance of lexicon-based sentiment analysis methods is limited because they cannot consider contextual information, nuanced indicators of sentiment messages, and internet slang (13). Jiang, Qiang and Lin (8) reported that the lexicon-based approach in their study performed poorly in analysing sarcastic text data.

We developed a public opinion evaluation framework using natural language processing (sentiment analysis and topic modelling), focusing on adopting supervised machine learning for sentiment analysis—a case study with the U.K. high Speed 2 (HS2) project was conducted with this framework. There were two reasons HS2 was selected as a case study. Firstly, HS2 is the most significant civil infrastructure investment in the U.K., where the total cost was approximately £65 billion to £88 billion (based on the 2015 price) in December 2019. Secondly, HS2 is highly controversial. Although HS2 aims to improve the U.K. transport system and economy, it is criticised for its financial viability and environmental impact. As a result, it is facing opposition from environmental groups (14) and campaign groups (15,16). All the opposition parties/groups are a strong presence on social media platforms, especially Twitter.

The main contributions of this study include 1) presenting a public opinion evaluation framework with a machine learning algorithm; 2) demonstrating the feasibility of sentiment analysis with supervised machine learning for civil infrastructure projects; 3) Comparing the accuracy of the Multinomial Naïve Bayes (MNB) classifier and Support vector machine (SVM) classifier.

2. Literature Review

2.1 Sentiment analysis

Sentiment analysis, also known as opinion mining, is an analysis method which adopts automated natural language processing to classify the sentiment polarity of text data (positive, neutral, and negative) (17). The text data can be retrieved from various online platforms, such as news articles, social media platforms, and web forums (18–20). Sentiment analysis has been used extensively in evaluating and predicting business performance and social issues. For example, Rui, Liu, and Whinston (21) found that word of mouth (WOM) has a high valence on movie sales, whereas positive Twitter WOM (positive sentiment) is linked with high movie sales. Smailović et al. (22) used sentiment analysis to monitor the changes in public interest in companies and their products. As a result, sentiment polarity on Twitter can predict stock market price in advance. Budiharto and Meiliana (23) made an algorithm to count maximum word frequency and predict the polarity of tweet sentiment about presidential candidates in Indonesia.

Apart from applications in business and social issues, sentiment analysis can also be used in understanding public opinion on companies and public projects. For example, Aldahawi (24) 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 has also been used on public opinion data for the California High-Speed Rail and Three Gorge Project (4,8).

Lexicon-based sentiment analysis is used to calculate the sentiment of documentation from the polarity of words (25). 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 vale pairs, replacing words in the document with polarity values, and calculating the sentiment polarity for the document.

Ding (4) tailormade the dictionary by removing some words (e.g. like work, etc.) from a positive word list. Jiang, Qiang, and Lin (8) built a dictionary for hydro projects by integrating the Nation Taiwan Sentiment Dictionary (26), Hownet (a Chinese/English bilingual lexicon database) (27), and a hydro project-related word list. Recent research showed the practicality of implementing the lexicon-based sentiment analysis for public opinion studies on civil projects. However, lexicon-based classifiers perform poorly in analysing text data in a sarcastic tone and cannot consider the context of textual data. Therefore, our study utilised a machine learning classifier to perform sentiment analysis.

In supervised sentiment analysis, classifying text data with different labels (positive, neutral, negative) is called classification. Current mainstream classifiers used in sentiment analysis are probabilistic classifier (Naïve Bayes, Bernoulli Bayes, multinomial Bayes, etc.) and Support Vector Machines (SVM).

The Naïve Bayes classifier is a probabilistic classification model which applies the Bayes theorem (28) in probability. The term "Naïve" means the naïve assumption of independence among each pair of features (attributes) and class variable values (29). For example, in natural language processing, the "naïve" assumption implies the classifier will process the text data independently as bag-of-words, ignoring the relationships among words and only considering the word frequency in the document.

The mathematical formula of the Bayes Theorem is given as follows:

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

where
vector *x* = (x_1,⋯,x_n) represent a problem instance with n features
vector *y* is the given class variable

For Multinomial Naïve Bayes (MNB) classifier, it is assumed that each P(xn│y) is a multinomial distribution, which means MNB is a special case of Naïve Bayes classifier (30). The support vector machine (SVM) is a classifier which utilises a separating hyper-plane and is developed based on statistical theory (31). SVM classifies text vectors with a hyperplane (straight line) or a Non-linear decision boundary, such as the kernel method (32).

2.2 Topic Modelling

Topic modelling is a technique to extract topics from text documents. Deerwester et al. (33) 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 (34). Blei, Ng, and Jordan (35) introduced Latent Dirichlet allocation (LDA), 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 the 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. (36) used LDA variant topic modelling to uncover the probabilistic relationship between Adverse Drug Reaction topics. They 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. (37) 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. (38). In this study, we used topic visualisation to determine the number of topics that returned the most consistent results.

3. Methodology

In this study, we analysed public opinion with social media data around High Speed 2 (HS2). The schematic flowchart is shown in Figure 1. Our procedure included: 1) Data collection and pre-processing; 2) Supervised machine learning classifier training and evaluation; 3) Sentiment analysis; and 4) topic modelling.

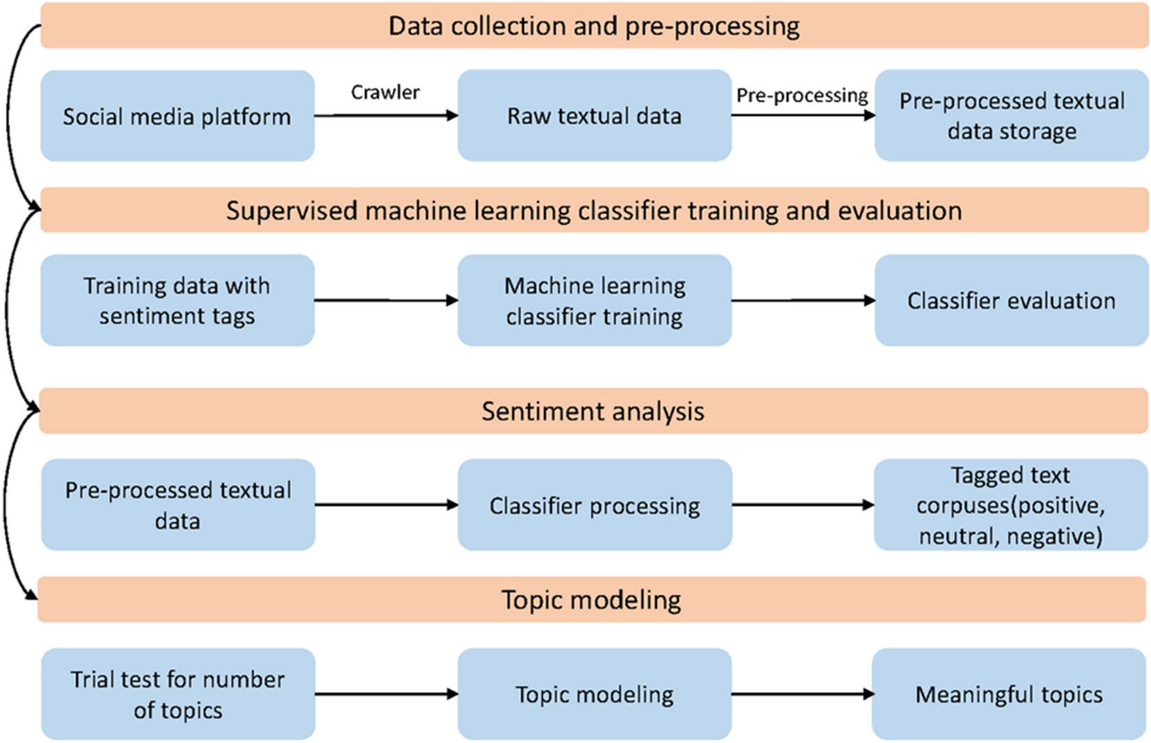
Data collection and pre-processing
Social media platform
Crawler
Raw textual data
Pre-processing
Pre-processed textual data storage
Supervised machine learning classifier training and evaluation
Training data with sentiment tags
Machine learning classifier training
Classifier evaluation
Sentiment analysis
Pre-processed textual data
Classifier processing
Tagged text corpuses(positive, neutral, negative)
Topic modeling
Trial test for number of topics
Topic modeling
Meaningful topics

Figure 1. Public opinion analysis with machine learning for infrastructure projects

3.1 Data collection and pre-processing

The HS2-related social media text data were collected from Twitter. Twitter provides a wide range of Application Programming Interfaces (API) for developers to interact with Twitter data, such as PowerTrack API (streaming real-time tweets with filters) and Search API (search and retrieve tweets with predefined rules). Search API provides seven days, 14 days, 30 days, and full archive (tweets dated back to 2006) search windows for collecting tweets. Our study adopts Search Tweets API: Full archive to collect Tweets in different project stages of HS2. Tweets containing the string "#HS2" were collected by Search Tweets: Full Archive. Due to the availability of Twitter API, the number of tweets collectable is limited to under 10,000 tweets during the time of this research. Thus, the number of tweets was allocated between training and test datasets. Related tweets between June and July 2020 were used for training machine learning classifiers (918 tweets). According to Bailey (39), HS2 phase 1 was approved in December 2016, and construction of phase 1 began in 2017. We were interested in the public opinion on the HS2 project after construction work began, so we chose May as a benchmark. We collected data for the following months: May 2017 (1544 tweets), May 2018 (1130 tweets), May 2019 (2909 tweets), and May 2020 (3040 tweets), which in total returned 8623 tweets. Noticeably, the tweets collected were in extended mode, so the full text was collected (more than 140 characters).

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 usually contains a lot of uninformative text, such as URL links, Twitter usernames, and email. Furthermore, some words that don't have sentiment orientation, such as "me", "you", "is", "our", "him", and "her", are called stop words. 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 (40). Other text pre-processing techniques include text lowercasing and text stemming.

3.2 Supervised machine learning classifier training and evaluation

Since the sentiment analysis focuses on labelling (classifying) the sentiment (negative, neutral, positive) of the tweets, the sentiment analysis is treated as a classification problem. The classification problem can be solved with supervised machine learning by adopting Naïve Bayes Classifier, Multinominal Naïve Bayes Classifier, and Support Vector Machine.

Firstly, a classifier algorithm was determined (Muiltinominal Naïve Bayes or Support Vector Machine). Secondly, the text data was pre-processed for a machine to extract features, such as term frequency from text data. Thirdly, the machine learning algorithm paired sentiment tags with features towards building a classifier model. Finally, the classifier model performed the prediction (sentiment analysis).

3.2.1 Machine Learning Training data set up

In the training procedure, the machine learning algorithm learns to relate the input text to the responding tags based on the training data. One thousand unique tweets were randomly picked from the tweets collection to build up that training data for a classifier. After deleting the duplicated tweets, 918 tweets were fed to the machine learning algorithm.

3.2.2 Supervised Machine Learning Classifier Training and Evaluation

The label training dataset was transferred to MonkeyLearn API, a third-party toolkit for building and accessing supervised machine learning classifiers (41). Multinomial Naïve Bayes (MNB) and Support Vector Machine (SVM) classifiers were trained individually with training datasets. We set the N-gram Range to include Unigram, Bigram, and Trigrams. N-grams define how sentences are divided into N number of consecutive words. Unigram (Bigram, Trigram) means the classifier treats a sentence as a combination of consecutive one (two, three) words. Moreover, the number of features was set to the maximum value, 10,000. The performance of classifiers was evaluated with *Accuracy* and *F1 score*, including *precision* and *recall.*

$$precision = \frac{TP}{TP + FP} \tag{2}$$

$$precision = \frac{TP}{FN + TP} \tag{3}$$

$$accuracy = \frac{TP + TG}{TP + TN + FP + FG} \tag{4}$$

$$F1\, score \; = 2 \times \frac{precision \; \times recall}{precision + recall} \tag{5}$$

In which

TP: True Positive
TN: True Negative
FP: False Positive
FN: False Negative

Accuracy and *F1 score* are widely adopted quality metrics. *Accuracy* measures the accuracy of the classifier with all correctly identified cases overall identified cases, and the *F1 Score* also represents the accuracy of the classifier with the harmonic mean of precision and recall. Notably, the *F1 Score* takes false positive and false negative into account more than *accuracy*, and the *F1 Score* is more accurate with the uneven class distribution. Since the number of tweets with negative sentiment was significantly larger than neutral and positive sentiment, the *F1 Score* was more representative in evaluating classifier accuracy for this study. The machine learning classifier with a higher *F1 score* was chosen to perform sentiment analysis.

Due to the data size limitation, the classifiers were tested with the training data set. The test results are shown in Table 1 and Table 2.

Table 1. MNB classifier test results

	true positive	true negative	false positive	false negative	precision	recall
negative tweets	460	165	258	37	64%	93%
neutral tweets	107	532	65	216	62%	33%
positive tweets	22	812	8	78	73%	22%
accuracy						64%
F1 Score						60%

Table 2. SVM classifier test results

58	true positive	true negative	false positive	false negative	precision	recall
negative tweets	409	283	140	88	75%	82%
neutral tweets	181	494	103	142	64%	56%
positive tweets	58	791	29	42	67%	58%
accuracy						70%
F1 Score						70%

The *accuracy* and *F1 Score* for SVM were 70% and 70% respective, while the MNB performed poorly with only 64% *accuracy* and 60% in the *F1 Score*. Although MNB achieves 93% recall in negative sentiment tweets, MNB struggled in predicting tweets with 33% neutral and 22% positive sentiment. Zimbra et al. (13) evaluated 29 sentiment analysis commercial packages, in which the domain-specific have accuracies ranging from 63.83% to 72.09%. Considering the training data has only 918 tweets, the performance of the SVM classifier for HS2-related tweets is satisfactory. Therefore, this study used SVM for the sentiment analysis task.

3.3 Sentiment Analysis and Topic modelling

Tweets labelled with sentiment orientation were then divided into three collections (positive, neutral, and negative). 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 (42). 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 reverence.

4. Results and Discussion

4.1 Sentiment Analysis Results

Figure 2 shows the sentiment analysis results in which tweets are marked with a different colour scheme. Although there was a slight dip in 2018, the overall trend shows the number of HS2-related tweets increased. Notably, the number of tweets increased considerably between May 2018 (1130 tweets) and May 2019 (2909 tweets), with a 157% increase in HS2-related tweets. The trend may imply that the public was more interested in HS2 and willing to discuss HS2 on social media platforms.

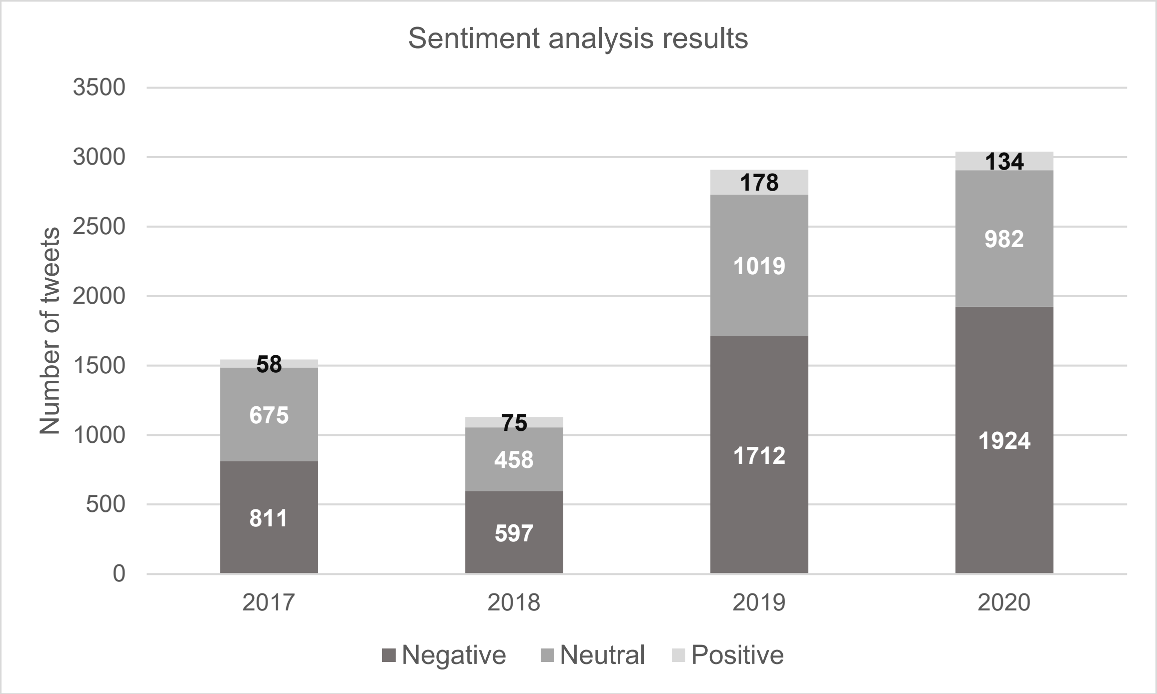

Figure 2. Sentiment analysis results

As shown in Figure 2, the negative sentiment tweets have the largest percentage for all the time windows, increasing from 53% in 2017 to 63% in 2020. Compared with the other two sentiment tags, positive sentiment tweets only account for around 4% to 7% of total tweets. Neutral sentiment tweets account for 32% to 44% of total tweets. Neutral sentiment tweets include tweets without apparent sentiment polarity and tweets with irrelevant information.

The sentiment analysis results show that a significant proportion of online tweets have a negative opinion of HS2. Considering the neutral sentiment tweets also include irrelevant tweets, the percentages of tweets with negative sentiment may be even higher. The high percentage of negative tweets shows that the public had a generally negative sentiment towards HS2, which should be considered for policymakers and managers. However, it should be noted that the high proportions of negative tweets do not necessarily mean HS2 is under a public relationship emergency since some Twitter users repeatedly tweet anti-HS2 tweets, such as @stophs2 and @Hs2Rebellion. These online influencers tend to tweet about HS2 more often than the general public, and most of their tweets have a negative sentiment. For example, @stophs2 tweeted 61 times in May 2019.

Additionally, since supervised machine learning is used for sentiment analysis, the classifier's accuracy should be considered. The *F1 Score* of the SVM classifier is 70%, which means neutral and positive tweets could be misclassified as negative, and negative tweets could be classified with other sentiment orientations. Apart from the abovementioned issues, there are challenging problems to be addressed for adopting sentiment analysis as a reliable method for public opinion evaluation.

4.2 Topic Modelling

4.2.1 Positive Tweets Corpus

The topics in the positive corpus are shown in Table 3. Topic 1 is the largest topic and accounts for 35.6% of the positive corpus. Topic 1 includes words such as "work", “job”, “need”, “north”, “benefit”, and “business” and suggests the topic centres around how HS2 can bring more jobs and work to northern England. Topic 2 is the second-largest topic in the positive corpus, which contains “network”, “new”, “rail”, “station”, “capacity” and “freight”. Hence, topic 2 would appear to focus on how HS2 will upgrade the national rail network with new rail stations, will free up the local rail network, and will increase capacities of local freight lines. Interpreting the meaning of topic four is challenging, as topic 4 includes some words from topic 5. “Environment” in topic 4 implies it may be related to environmental discussions. On the other hand, the meaning of topic 5 is quite explicit. “Flight”, “delivery”, “rail”, “significant”, and “co2” suggest topic 5 is about the transportation mode shift (from flight to rail), and the CO2 reduction from the mode shift.

Table 3. Major topics in positive corpus

Topic number	Terms	Topic percentage
1	need, work, job, transport, infrastructure, benefit, investment, north, capacity, new, deliver, midland, well, create, economy, say, future, more, northern, connectivity, boost, business, build, rail, report, home, mean, opportunity, stationn	35.6%
2	network, new, project, rail, great, capacity, region, community, create, freight, level, include, use, set, important, many, emission, talk, railway, infrastructure, support, station, investment, wide, fast, entire, also, job, process, plan	22.4%
3	increase, line, way, exist, long, travel, take, city, support, improve, become, passenger, station, happen, airport, distance, covid, would, open, pandemic, now, make, could, capacity, question, catalyst, fmajor, opportunity, add, letter	15.9%
4	rail, freight, world, local, train, much, railway, leader, economy, plan, stop, get, receive, look, global, give, really, take, money,	13.7%

	first, industry, go, love, design, additional, company, covid, happy, connect	
5	rail, ever, flight, even, role, play, invest, railway, can, help, delivery, pivotal, project, example, significant, short, report, good, fast, construction, government, however, then, forward, lot, contract, essential, co2, much	12.5%

4.2.2 Negative Tweets Corpus

Table 4 shows major topics in the negative corpus. Topic 1 contains “cost”, “stophs2”, “scrap” and “bad”. It is likely that topic 1 covers the overbudget issues about HS2. In addition, “scrap: and “stophs2” are two strong negative sentiment words with high frequency, which means some people tweeted with a fairly strong negative sentiment. Topic 2 includes the words “money”, “people”, “stop”, “pay”, “spend”, “nhsnoths2”, “lie”, and “waste”. These are words mainly criticising spending too much money on HS2. HS2 opposition groups claim the money for HS2 should be invested in other infrastructure projects (existing roads and railways) and the healthcare system. Topic 3 contains “destroy”, “cut”, “protestor”, “wildlife”, “future” and “woodland” and is likely about protesting HS2 against its destruction of woodland and wildlife habitat. HS2 is criticised for destruction of woodlands and parks, including parks around Euston station. Topic 4 and topic 5 also discuss environmental issues as indicated by “environmental”, “destruction”, and “ecocide”.

Table 4. Major topics in negative corpus

Topic number	**Terms**	**Topic percentage**
1	cost, stophs2, government, nhsnoths2, train, spend, well, use, rail, transport, ever, new, come, continue, show, call, support, think, line, covid19, put, must, travel, start, plan, pandemic, will, back, economy	24.7%
2	need, people, go, scrap, borisjohnson, lie, year, take, site, even, make, never, keep, way, many, can, protestor, now, thing, road, tell, find, help, enough, day, include, job, grantshapps, total, least	23.6%
3	project, money, would, still, public, could, see, country, cancel, build, good, much, cut, right, save, future, railway, also, fund, trident, seem, track, hide, try, station, already, benefit, happen, tory, afford	19.6%
4	say, get, tree, destroy, report, know, want, look, woodland, law, local, bill, change, construction, far, budget, long, world, environment, protect, business, parliament, course, environmental, instead, leave, oppose, eviction, kill, ask	17.8%
5	time, stop, work, pay, give, worker, wildlife, really, destruction, great, big, waste, bosses_blindside, impact, ecocide, massive,	14.3%

	infrastructure, do, mean, week, tax, taxis, police, become, thug, end, illegal, life, yet	

4.2.3 Neutral Tweets Corpus

Compared with the positive and negative corpus, the topic modelling result in the neutral corpus is considerably harder to interpret, as shown in Table 5. The words in topic 1 show little consistency. This is because the neutral sentiment corpus includes not only tweets with no evident sentiment but also irrelevant tweets. As a result, the noise in the text data makes the learning process very difficult for machines.

In summary, the topic modelling results are satisfactory for positive and negative corpus, as the topic modelling could provide valuable insights about what is public opinion on HS2.

Table 5. Major topics in neutral corpus

Topic number	**Terms**	**Topic percentage**
1	cost, need, say, stophs2, work, scrap, would, get, train, public, well, could, country, make, rail, think, save, continue, great, long, include, big, massive, parliament, economy, plan, bad, seem, fund, may	24.7%
2	money, people, stop, use, pay, spend, nhsnoths2, lie, year, still, give waste, cancel, borisjohnson, even, site, nhs, call, new, now, many, support, economic, travel, covid19, change, back, face, job, case	24.6%
3	go, government, destroy, see, build, report, good, know, way, keep, want, cut, show, protester, wildlife, reality, tell, look, future, road, put, local, pandemic, start, woodland, day, hide, will, protect	21.4%
4	project, time, tree, take, worker, can, transport, never, thing, destruction, bill, right, find, budget, help, environmental, world, track, live, railway, construction, billion, kill, true, become, cost_potential, protest, station, especially, carbon	17%
5	much, ever, come, taxpayer, line, law, must, far, also, trident, total, course, ecocide, happen, scheme, lot, late, oppose, sense, week, control, thug, forget, furlough, corrupt, expensive, mile, debt, property, ignore	12.3%

4.3 Challenges with Machine Learning in Public Opinion Analysis

4.3.1 Limitation in Training Dataset

For standard practice in training a supervised machine learning model, 500 training data per tag is recommended for satisfactory performance. Although a great effort was made to tag tweets with as much variety as possible, the scarcity of positive sentiment tweets makes the human labelling considerably challenging. Therefore, in the training data set, only 102 tweets out of 918 tweets were tagged with positive sentiment. It is the reason why both Multinomial Naïve Bayes and Support Vector Machine did not perform well with a recall of positive tweets.

4.3.2 Human Factor in Training Data Tagging

Researchers usually assign multiple annotators (3 to 5) to tag the sentiment orientation to minimise the influence of human annotators (43). 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. That being said, regular debriefing meetings between co-authors mitigated the impact of this limitation.

For example, the following tweet may be tagged with different sentiment orientations. 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 the negative sentiment towards over budget issue of HS2.

"#HS2 is a £100bn scheme to have slightly shorter journey times from Manchester and Birmingham to London, thereby solving Britain's biggest ever problem."

Although the supervised sentiment analysis method faces several problems, it still demonstrates its potential to be an alternative to the traditional public opinion evaluation methods due to its low cost, speed, and utilisation of user-generated content on social media.

4.3.3 Topic Modelling Challenges

Text documents are a probabilistic distribution 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, the assumption in the probabilistic distribution of topics does not fit well in tweet topic modelling. Therefore, adapting LDA topic modelling for tweets can have a more accurate result.

Since the tweets with neutral sentiment in nature will include tweets with irrelevant information, analysing neutral tweets will not give as reliable a result as positive and negative. Therefore, in civil infrastructure project evaluation, the priority of neutral tweets could be lower, so the computing resources could be allocated to positive and negative tweets.

5. Conclusions

The sentiment analysis showed the public's overall negative sentiment toward the HS2 project. Topic modelling on the negative corpus shows that the public is mainly concerned about the overbudget issue and environmental impact. On the other hand, the positive tweets only account for a small portion of total tweet data, mentioning employment and transport capacity improvements. Although the neutral tweet corpus cannot provide a meaningful result, the proposed framework is a useful alternative to existing methods.

In our study, we first performed sentiment analysis with supervised machine learning classifiers (MNB and SVM) in the civil engineering domain. Analysis results show that the SVM is more accurate than the MNB classifier, whose accuracies are 70% and 64%, respectively. Therefore, the SVM classifier is more suitable for sentiment classification tasks in civil engineering projects.

Future research on public opinion assessment with social media data could explore the following areas: 1) exploring methods to remove irrelevant posts effectively when collecting textual data; 2) comparing more machine learning classifiers for infrastructure project assessment, and 3) considering multiple social media platforms.

Funding and Acknowledgement statements

This research did not receive funding.

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.

Data Availability statement

The collected data and the Python source code in this article are available on the GitHub repository:

https://github.com/RichardYao08/PinionAnalysisSocialMedia

To follow the FAIR Data Principles, the data is anonymised.

Declarations and Conflict of Interests statement

There is no conflict of interest.

Ethics Approval

This research used publicly available data and did not require ethics approval.

6. Reference

1. H.M. Treasury. National Infrastructure Strategy [Internet]. 2020. Available from: www.gov.uk/official-documents
2. Hayes DJ. Addressing the environmental impacts of large infrastructure projects: making "mitigation" matter. 2014.
3. O'Faircheallaigh C. Public participation and environmental impact assessment: Purposes, implications, and lessons for public policy making. Environmental Impact Assessment Review. 2010 Jan 1;30(1):19–27.
4. Ding Q. Using Social Media to Evaluate Public Acceptance of Infrastructure Projects. Unvisersity of Maryland; 2018.
5. Checkoway B. The Politics of Public Hearings. The Journal of Applied Behavioral Science. 1981 Jul 26;17(4):566–82.
6. Heberlein T. Some Observations on Alternative Mechanisms for Public Involvement: The Hearing, Public Opinion Poll, the Workshop and the Quasi-Experiment. Natural Resources Journal. 1976;16(1).

7. O'connor B, Balasubramanyan R, Routledge BR, Smith NA. 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. Jiang H, Qiang M, Lin P. Assessment of online public opinions on large infrastructure projects: A case study of the Three Gorges Project in China. Environmental Impact Assessment Review. 2016 Nov 1;61:38–51.
9. Kaplan AM, Haenlein M. Users of the world, unite! The challenges and opportunities of Social Media. Business Horizons. 2010 Jan 1;53(1):59–68.
10. Stellefson M, Paige S, Apperson A, Spratt S. Social Media Content Analysis of Public Diabetes Facebook Groups. Journal of Diabetes Science and Technology. 2019 May 1;13(3):428–38.
11. Park SB, Ok CM, Chae BK. Using Twitter Data for Cruise Tourism Marketing and Research. Journal of Travel and Tourism Marketing. 2016 Jul 23;33(6):885–98.
12. Kim DS, Kim JW. Public Opinion Sensing and Trend Analysis on Social Media: A Study on Nuclear Power on Twitter 1. International Journal of Multimedia and Ubiquitous Engineering. 2014;9(11):373–84.
13. Zimbra D, Abbasi A, Zeng D, Chen H. The state-of-the-art in twitter sentiment analysis: A review and benchmark evaluation. ACM Transactions on Management Information Systems. 2018;9(2).
14. Woodland Trust. HS2 Rail Link: Ancient Woods Under Threat [Internet]. Woodland Trust. 2020 [cited 2021 Jun 14]. Available from: https://www.woodlandtrust.org.uk/protecting-trees-and-woods/campaign-with-us/hs2-rail-link/
15. Alliance H.A. HS2 Timeline - High Speed 2 Action Alliance [Internet]. [cited 2021 Jun 14]. Available from: http://www.hs2actionalliance.org/
16. Stop HS2. STOP HS2 - The national campaign against High Speed Rail 2 [Internet]. Stop HS2. 2021 [cited 2021 Jun 14]. Available from: http://stophs2.org/
17. Go A, Huang L, Bhayani R. Twitter Sentiment Analysis. 2009.
18. Tetlock P.C. Giving content to investor sentiment: The role of media in the stock market. Journal of Finance. 2007 Jun 1;62(3):1139–68.
19. Dave K, Lawrence S, Pennock DM. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: Proceedings of the 12th International Conference on World Wide Web, WWW 2003. 2003. p. 519–28.
20. Das SR, Chen MY. Yahoo! for amazon: Sentiment extraction from small talk on the Web. Management Science. 2007 Sep;53(9):1375–88.
21. Rui H, Liu Y, Whinston A. Whose and what chatter matters? the effect of tweets on movie sales. Decision Support Systems. 2013 Nov 1;55(4):863–70.
22. Smailović J, Grčar M, Lavrač N, Žnidaršič M. Predictive sentiment analysis of tweets: A stock market application. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, Berlin, Heidelberg; 2013. p. 77–88.
23. Budiharto W, Meiliana M. Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis. Journal of Big Data. 2018 Dec 19;5(1):51.
24. Aldahawi HA. Mining and Analysing Social Network in the Oil Business : Twitter Sentiment Analysis and Prediction Approaches. Cardiff University; 2015.
25. Taboada M, Brooke J, Tofiloski M, Voll K, Stede M. Lexicon-basedmethods for sentiment analysis. Computational Linguistics. 2011 Jun 1;37(2):267–307.
26. Ku LW, Liang YT, Chen HH. Opinion Extraction, Summarisation and Tracking in News and Blog Corpora. 2006.
27. Dong Z, Dong Q. 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. p. 820–4.

28. Bayes T. 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. Philosophical Transactions of the Royal Society of London. 1763 Dec 31;53:370–418.
29. Jiang L, Cai Z, Zhang H, Wang D. Naive Bayes text classifiers: A locally weighted learning approach. Journal of Experimental and Theoretical Artificial Intelligence. 2013 Jun 1;25(2):273–86.
30. Russell S, Norvig P. Artificial Intelligence: A Modern Approach. 4th U.S. edition. 2021.
31. Vapnik V.N. An overview of statistical learning theory. Vol. 10, IEEE Transactions on Neural Networks. 1999. p. 988–99.
32. Sharma A, Dey S. A comparative study of selection and machine learning techniques for sentiment analysis. In: Proceeding of the 2012 ACM Research in Applied Computation Symposium, RACS 2012. 2012. p. 1–7.
33. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. Journal of the American Society for Information Science. 1990;41(6):391–407.
34. Kanatani K. Linear Algebra for Pattern Processing Projection, Singular Value Decomposition, and Pseudoinverse. Morgan & Claypool; 2021. 27–32 p.
35. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003;3:993–1022.
36. Xiao C, Zhang P, Art Chaowalitwongse W, Hu J, Wang F. Adverse Drug Reaction Prediction with Symbolic Latent Dirichlet Allocation. Vol. 31, Proceedings of the AAAI Conference on Artificial Intelligence. 2017 Feb.
37. Chuang J, Ramage D, Manning CD, Heer J. Interpretation and trust: Designing model-driven visualisations for text analysis. In: Conference on Human Factors in Computing Systems - Proceedings. 2012. p. 443–52.
38. Sievert C, Shirley K. LDAvis: A method for visualising and interpreting topics. In: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. Association for Computational Linguistics (ACL); 2014. p. 63–70.
39. Bailey C. Savills U.K. | The Impact Of HS2 On Development [Internet]. Savills. [cited 2021 Jun 13]. Available from: https://www.savills.co.uk/research_articles/229130/197066-0
40. Haddi E, Liu X, Shi Y. The role of text pre-processing in sentiment analysis. In: Procedia Computer Science. Elsevier B.V.; 2013. p. 26–32.
41. MonkeyLearn. Everything There Is to Know about Sentiment Analysis [Internet]. [cited 2021 Jun 14]. Available from: https://monkeylearn.com/sentiment-analysis/
42. Rehurek R, Rehurek R, Sojka P. Software Framework for Topic Modelling with Large Corpora. IN PROCEEDINGS OF THE LREC 2010 WORKSHOP ON NEW CHALLENGES FOR NLP FRAMEWORKS. 2010;45--50.
43. Callison-Burch C. 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; 2009. p. 286–95.