Journal of Computer Science

Research Article

A Machine Learning-Based Sentiment Analysis of Article 370 Tweets to Support Government Policy Decisions

Subhasis Mohapatra[1], Sudhir Kumar Mohapatra[1], Sweta Samantaray[1], Aliazar Deneke Deferisha[2] and Prasanta Kumar Bal[3]

[1]*Faculty of Engineering and Technology, Sri Sri University, Cuttack, India*
[2]*Faculty of Computing and Software Engineering, Arba Minch University, Arba Minch, Ethiopia*
[3]*GITA Autonomous College, Bhubaneswar, India*

Article history
Received: 04-08-2025
Revised: 06-03-2026
Accepted: 22-05-2026

Corresponding Author:
Aliazar Deneke Deferisha
Faculty of Computing and Software Engineering, Arba Minch University, Ethiopia
Email: aliazar.deneke@amu.edu.et

Abstract: This study proposes a robust sentiment analysis framework to evaluate public opinion on the abrogation of Article 370 using Twitter data. The methodology begins with tweet collection through the Twitter API, followed by systematic preprocessing. Sentiment labels were generated using the Text Blob lexicon-based polarity scoring approach to facilitate the construction of a large-scale sentiment dataset. Features are extracted, and the dataset is split into training (80%) and testing (20%) sets. A variety of models—including lexicon-based approaches, traditional machine learning algorithms, and ensemble learning techniques are trained and optimized using hyperparameter tuning. Additionally, a hybrid CNN–LSTM deep learning model is employed to capture both spatial and temporal dependencies in the text data. Experimental results reveal that the tuned Voting Ensemble model achieved the highest agreement with lexicon-derived labels, achieving an accuracy of 94.05% and an F1-score of 95.7%. The CNN–LSTM model also demonstrated strong performance. Lexicon-based polarity trends show that the dataset we looked at had mostly positive feelings during the time we chose. The results show how well ensemble and deep learning methods work together for automated sentiment classification. Future work could include validation datasets that have been annotated by people, support for multiple languages, more advanced transformer-based models for detecting sarcasm and emotion, and analysis of temporal sentiment trends.

Keywords: Sentiment Analysis, Article 370, Ensemble Learning, Twitter Data, Public Opinion

Introduction

Social media platforms have made public opinion more accessible in the digital era, offering real-time insights into the attitudes and viewpoints of the public. A branch of Natural Language Processing (NLP) called sentiment analysis has become a potent instrument for assessing public opinion by examining textual data from various platforms. Online forums are also being used by people all around the world to voice their thoughts about different government programs, policies, and missions (Chehal et al., 2021; Gupta et al., 2021; Surabhi and Jain, 2022). Globally, governments are realizing the value of this technology in policymaking and utilizing it to determine public concerns, evaluate the effects of actions, and practice data-driven governance. Policymakers can better address the wants and concerns of citizens when they can glean valuable insights from unstructured text data.

Sentiment analysis helps spot trends in public conversation and gives early warning signs of possible discontent or support. Governments can track disinformation, monitor reactions, and take prompt action, for example, by analyzing emotions expressed on social media sites like Facebook and Twitter during elections, natural disasters, or public health emergencies. Additionally, sentiment patterns may guide communication tactics, assisting leaders in creating messages that speak to the public's concerns and steer clear of misunderstandings (Chehal et al., 2021; Gupta et al., 2021; Kumar and Jha., 2022; Isnain et al., 2021; Alotaibi and Nadeem, 2024). Sentiment analysis fills the

void between the public and decision-makers by methodically recording emotions like fear, pleasure, rage, and trust (Nath, 2024; Hristova and Netov, 2023; Tori et al., 2024; Das et al., 2025).

Understanding the opinions of various demographic segments is essential, especially for democratic administrations. Sentiment analysis helps promote inclusive government by providing a voice to underprivileged or underrepresented groups, who frequently use social media instead of more conventional routes for criticism (Sujatha and Radha, 2023; Verma and Jamwal, 2020). This inclusion guarantees a wider base of citizen interaction and aids in developing more egalitarian policies. Sentiment analysis is, therefore, becoming a vital tool in the current policy toolbox, providing empirical evidence in favor of more responsible and transparent governance.

Sentiment analysis's accuracy and scalability have greatly increased with the use of machine learning. Machine learning algorithms, particularly those that use supervised learning approaches, can learn from context, identify feelings based on large quantities of training data, and get better with each new set of data (Widiastutie et al., 2024; Qi and Shabrina, 2023). These models are capable of automatically identifying linguistic subtleties that are sometimes difficult for rule-based systems to grasp, including sarcasm or ambiguity. Machine learning's versatility and effectiveness make it the perfect tool for tracking public opinion on quickly changing political problems in real time.

In August 2019, the abrogation of Article 370 of the Indian Constitution was one such subject that generated a lot of internet discussion in India. Jammu & Kashmir's unique status under Article 370 was withdrawn by the Narendra Modi-led Union Government of India (Verma and Jamwal, 2024). It caused a great deal of media agitation. Public mood and media discourse were separated into two categories based on their reactions: Positive and negative. Many hailed the action as a step toward progress and cohesion, while others saw it as an overreach of the constitution that may worsen tensions in the area. The revocation of Article 370 was enthusiastically applauded by right-wing supporters, while supporters of Prime Minister Modi referred to the move as a historic and admirable decision. On the other hand, a retired Supreme Court justice and left-leaning parties denounced the decision to repeal Article 370. Former Supreme Court of India judge Justice Gopal Gowda characterized the move to repeal Article 370 as a constitutional breach (Verma and Jamwal,2024; Singh et al., 2023; Patil et al., 2019).

According to media sources, researchers have observed a range of opinions on the repeal of Article 370 among individuals with diverse beliefs and backgrounds. Researchers were therefore inspired by these divergent viewpoints to assess the mixed opinion using sentiment analysis, utilizing machine learning techniques. The amount and variety of viewpoints expressed on Twitter make the public's response to the repeal of Article 370 an excellent case study for sentiment analysis.

Existing research on sentiment analysis for policy evaluation is frequently limited by several factors, despite the rising interest in this field. Much of the research mostly depends on lexicon-based or conventional machine learning techniques, which are inadequate for properly interpreting complex, multilingual, or sardonic information that is frequently present in social media political conversation. The majority of earlier studies also fail to thoroughly investigate ensemble or hybrid deep learning models and rely on tiny or language-biased datasets. This limits the results' precision, generalisability, and practical relevance in situations when policy is at stake. In order to properly capture public opinion during important political events like the repeal of Article 370 in India, a more reliable, scalable, and context-aware sentiment analysis framework is desperately needed.

The main goal of this study is to provide a thorough framework for sentiment analysis that combines sophisticated deep learning architectures like CNN-LSTM, ensemble techniques like voting and stacking, and conventional machine learning classifiers. This work uses a large-scale Twitter dataset focused on the public conversation around the repeal of Article 370 to compare the performance of these models to popular lexicon-based tools such as TextBlob, AFINN, VADER, and SentiWordNet (Loria et al., 2014; Nielsen, 2011; Hutto and Gilbert, 2014; Baccianella et al., 2010). During this politically significant event, the study aims to identify the major emotional tones whether positive or negative expressed by users by examining the polarity and subjectivity of tweets. The ultimate objective is to offer practical insights that can assist data-driven governance, enabling decision-makers to better comprehend public opinion and respond with more inclusive and successful communication tactics.

Motivation

Social media's widespread use has drastically changed how people participate in political debate by providing large-scale, real-time access to public opinion. Significant national events, like the repeal of Article 370 of the Indian Constitution, spark intense internet responses that represent a range of intellectual and emotional viewpoints. Events of great national significance, like the repeal of Article 370 of the Indian Constitution, elicit strong online responses from a wide range of intellectual and emotional viewpoints. For policymakers hoping to promote open, inclusive, and data-driven government, it is crucial to comprehend these responses. Most of the research conducted in policy contexts, despite significant advancements in sentiment analysis, is hampered by small

datasets, monolingual analysis, or dependence on simple machine learning or lexicon-based models, which frequently fall short in capturing the complexities and subtleties of political sentiment.

Considering these drawbacks, this study offers a thorough framework for sentiment analysis that combines supervised learning algorithms, ensemble classifiers, lexicon-based approaches, and a hybrid deep learning CNN–LSTM model. It is then applied to a sizable dataset of more than 2.2 lakh tweets pertaining to the repeal of Article 370. The objective is to build a strong pipeline that can facilitate real-time policy sentiment monitoring and methodically assess the efficacy of various models. By offering a decision-support tool and comparison standard for assessing public opinion in politically delicate situations, our work seeks to close a significant gap.

Novelty of the Study

This study is innovative since it uses a multi-layered sentiment analysis framework to examine public opinion over the repeal of Article 370. It uses a large-scale, real-time Twitter dataset (more than 2.2 lakh tweets), which has been refined through meticulous preprocessing and lexicon-based labeling to enable high-quality supervised learning, in contrast to previous works that used small datasets or only used traditional models. Additionally, this research does one of the most thorough comparison analyses to date, including more than 15 different sentiment classification strategies, such as supervised models, sophisticated ensemble methods, deep learning architectures, and lexicon-based baselines. It highlights the framework's scalability, robustness, and adaptability in real-time public opinion mining for democratic governance. This kind of comprehensive analysis is extremely uncommon in policy-driven sentiment research.

Related Work

In the past decade, sentiment analysis has become a vital tool for digital governance. It allows governments to measure public mood in real time in response to significant policy choices. This section systematically examines the broad strategies, techniques, and conclusions from earlier research on how social media might increase public participation in governmental policy and decision-making.

Chehal et al. (2021) used topic modeling and sentiment analysis to examine the emotional reactions of Indian Twitter users during COVID-19 lockdowns 2.0 and 3.0. They classified tweets into emotions and detected changes in public sentiment using the NRC Emotion Lexicon and LDA. Using TextBlob and VADER for annotation and machine learning classification, (Gupta et al., 2021) performed sentiment analysis on 12,741 tweets posted by Indian users during the COVID-19 shutdown. Following preprocessing and vectorization using CountVectorizer, eight classifiers were evaluated, with a maximum accuracy of 84.4% was obtained using LinearSVC and unigrams. The study shows how well NLP and machine learning work together to detect sentiment in real time and inform public policy during medical emergencies. Kumar et al. (2022) performed sentiment analysis on tweets surrounding the repeal of India's farm laws and the end of the farmers' protest using NVivo 12. The results highlight Twitter's significance in political discourse and imply that examining social media sentiment can help guide public policy and communication. Chandra and Krishna (2021) designed a deep learning system for multi-label sentiment analysis of Indian tweets during the COVID-19 peak using LSTM, BD-LSTM, and BERT. Over 150,000 tweets were analyzed using BERT, which performed better than other models after being trained on the Senwave dataset. To analyze public sentiment from Twitter data in Ostrava, Czechia, (Švaň, 2023) proposed a decision-support framework that combines fuzzy logic, TextBlob for sentiment analysis, and BERTopic for topic modeling. The findings indicated that while subjects like pensions and wages exhibited more negativity, topics like politics and education had significant levels of positive emotion. Through the collection of detailed, topic-specific public feedback, the study demonstrates how social media analytics may assist municipal decision-making. Isnain et al. (2021) applied the Naive Bayes Classifier with TF-IDF and N-Gram (Unigram, Bigram, Trigram) feature extraction to conduct sentiment analysis on Indonesian tweets about the government's COVID-19 "New Normal" policy. Trigram N-Gram produced the greatest results, with an 85% F1-score, 84% accuracy, 84% precision, and 86% recall. The study demonstrates that improving classification accuracy for policy-related tweet sentiment involves integrating Naive Bayes with Trigram features. Using data from Facebook, Obiedat et al. (2021) created a sentiment-based DSS that combines SVM and Whale Optimization Algorithm (WOA) to assess public opinion about Jordan's COVID-19 rules. Their WOA-SVM model achieved an accuracy of 78.78% and an F-measure of 84.64%, outperforming other metaheuristics and traditional classifiers. Alotaibi and Nadeem (2024) employed the AraBERT transformer model to apply sentiment analysis to 216,858 Arabic tweets about the educational reforms in Saudi Arabia. With an F1-score of 0.89, the refined AraBERT outperformed other Arabic transformer models (CAMeLBERT, MARBERT) by 4% and conventional ML models (SVM, RF, LR) by 5%. Dandannavar et al. (2019) examined several sentiment analysis techniques used to assess public opinion toward government activities using data from Twitter. The study showed how SA may direct policy restructuring through real-time public input and underlined the significance of integrating ML and lexicon-based approaches for precise

sentiment categorization. Hristova and Netov (2023) analyzed mood and emotion in more than 13,000 Bulgarian tweets about public policy. For sentiment scoring, they used Python's TextBlob package with the NRC emotion lexicon. The results of sentiment analysis showed an almost equal distribution: 45.2% were positive, 43.1% were negative, and the remaining percentage was neutral. The results highlight the potential of social media analytics to assist public policy and show how well lexicon-based approaches can reveal the emotional reactions of the people to government activities. Patil et al. (2020) investigated public response to the repeal of Article 370 by doing sentiment analysis on 2,200 tweets between August 5 and August 30, 2019. They discovered that many tweets were neutral to slightly positive using TextBlob's polarity score feature. The authors picked prominent people (sometimes known as "thought leaders") whose tweets had the widest reach and utilized data visualization techniques to emphasize sentiment trends. Tori et al. (2024) examined 1,998 multilingual tweets concerning Brussels' mobility strategy using XLM-T and GPT-4. Accuracy was 67% for XLM-T and 66% for GPT-4, which used zero-shot learning and was better at handling nuanced emotion and sarcasm. The study demonstrated how LLMs might aid in urban policymaking by offering insightful information on public opinion. Rahman et al. (2023) used boosting models to examine 300,000 COVID-19 tweets from Bangladesh, the United Kingdom, and the United States. CatBoost has the highest F1-score, 85.8%. Srivastava et al. (2024) compared the effectiveness of k-Nearest Neighbors (kNN) and neural networks in sentiment analysis of COVID-19-related tweets using a Kaggle dataset. Neural networks demonstrated low accuracy and great recall, but kNN demonstrated the converse. The study emphasizes how important model selection and preprocessing are to successfully classifying sentiment in social media data. Alotaibi and Nadeem, (2024) examined Arabic tweets to gauge Saudi Arabian public opinion toward educational changes. With an F1 score of 0.89, their refined AraBERT beat other models using a manually labeled dataset, proving the usefulness of Arabic-specific transformers for sentiment analysis. Firdaus et al. (2024) examined recent research on sentiment analysis for policymaking and concluded that the most accurate models were deep learning models such as LSTM and BERT. To enhance sentiment analysis for morally sound decision-making, they addressed important issues such multilingual data and fraudulent accounts and suggested NLP-based solutions. Widiastutie et al. (2024) used SVM, LSTM, and Bi-LSTM to evaluate tweets from Indonesia on the nickel export embargo; Bi-LSTM achieved 91% accuracy. The majority of tweets expressed support for the strategy, pointing to its economic advantages, but others expressed concern over trade tensions with the EU. The study highlights sentiment analysis as a tool for policy insight.

Table 1: Summary of Related Work

Author	Objective/Focus	Method/Model	Data Size	Key Finding	Limitation
Chehal et al. (2021)	Emotional health during lockdowns	NRC, LDA	~80K tweets	Rise in negative emotions in 3.0	English only, short window
Gupta et al. (2021)	Lockdown sentiment in India	TextBlob, VADER, ML classifiers	12,741 tweets	LinearSVC: 84.4% accuracy	Emojis, hashtags excluded
Kumar et al. (2022)	Sentiment on farm laws repeal	NVivo, NCapture	37,857 tweets	Mostly negative sentiments	Limited hashtags, tool-defined labels
Chandra and Krishna (2021)	Sentiments during COVID-19 peak	LSTM, BERT, GloVe	~178K tweets	BERT best; optimism & annoyance common	Cultural humor, data mismatch
Švaňa (2023)	City-level sentiment for decision-making	BERTopic, TextBlob, TFN	~3K tweets	Richer topic-level insight	TFN symmetry assumption
Isnain et al. (2021)	Sentiment on New Normal policy	NB + TF-IDF + N-gram	1,823 tweets	Trigram NB: 84% accuracy	No DL or sarcasm detection
Obiedat et al. (2021)	Jordan policy sentiment via Facebook	WOA-SVM	5,250 comments	F1- score: 84.64%	Arabic only, limited generalizability
Alotaibi and Nadeem (2024)	Education reforms in Saudi Arabia	AraBERT, transformers	216K tweets	F1- score: 0.89	Arabic dialect variation, tweet length
Dandannavar et al. (2019)	Framework for public sentiment	Lexicons + ML + Hybrid	Not specified	Modular SA framework	No experiments or metrics
Hristova and Netov (2023)	Govt sentiment and trust in Bulgaria	VADER, NRC	~4K tweets	Trust is linked to positive sentiment	Lexicons lack nuance
Patil et al. (2019)	Article 370 revocation sentiment	TextBlob	2,200 tweets	Mostly neutral-positive	No emotion detection
Tori et al. (2024)	Mobility policy sentiment (Brussels)	XLM-T, GPT-4	1,998 tweets	GPT-4 handled context best	Needs local knowledge, not generalizable

Table 1: Continued

Rahman et al. (2023)	Best model for COVID-19 sentiment	CatBoost, XGBoost, USE	300K tweets	CatBoost: F1- score: 85.8%	Random split, regional bias
Srivastava et al. (2024)	Model comparison on COVID-19 tweets	kNN, NN, syuzhet, etc.	Not clear	NN better in RMSE, kNN had high recall	Low precision, no DL comparison
Alotaibi and Nadeem (2024)	Repeat of [12] (similar results)	AraBERT and others	216K tweets	AraBERT best performer	Same as [12]
Widiastuti et al. (2024)	Nickel export policy in Indonesia	SVM, LSTM, BD-LSTM	7,070 tweets	BD-LSTM: 91% accuracy	Narrow topic, single year
Qi and Shabrina (2023)	Lexicon vs ML for lockdown tweets (UK)	TextBlob, VADER, SVC	77K tweets	VADER best for SM; SVC best ML	Only 3K labeled, city-specific
Mahrenbach and Pfeffer (2023)	Aadhaar sentiment over time	LIWC, thematic coding	250K tweets	Negative sentiment grew over time	Two time periods, English only
Rulandari (2024)	TAPERA housing policy sentiment	Traditional SA	Not specified	Public reaction was largely negative	No model detail, sarcasm handling missing
Bodaghi and Zhu (2024)	Twitter policy on quoting during election	SARIMAX, VADER	304M tweets	Quoting rose briefly; retweets fell	Limited to Twitter only
Adamu et al. (2021)	Palliatives sentiment in Nigerian Pidgin	SVM, RF, LR, etc.	9,803 tweets	SVM: 88% accuracy	Only Pidgin, no cross-validation

Research Gap and Problem Statement

The previous research based on policy-oriented sentiment analysis exhibits several limitations. Most of the studies rely on very small, context-oriented datasets, lexicon-based approaches, and limited model comparison (Table 1). Comprehensive analysis of traditional machine learning, ensemble and deep learning approaches within a single framework is scarce.

Therefore, the current study addresses the core research problem.

How can a scalable and systematically benchmarked sentiment analysis framework integrating diverse modeling paradigms provide reliable and policy-relevant insights for large-scale political discourse?

Methodology

An extensive methodological framework that included data collection, text preprocessing, polarity labelling, feature extraction, and classification using both conventional and cutting-edge learning models was used to evaluate public opinion on the repeal of Article 370.

In this study, tweets gathered after the policy announcements are analyzed using various models such as lexicon-based, machine learning, and ensemble models to gauge public opinion toward the repeal of Article 370. The approach is intended to guarantee methodical data collection, pre-processing, modeling, and assessment to extract precise and pertinent policy insights from Twitter data. The complete research methodology architecture is illustrated in Fig. 1.

Data Collection

This study uses the Twitter API to collect 2,27,259 tweets, concentrating mostly on the hashtags #Article370, #Kashmir, #370Abrogation, and #IndianGovernment from 8th August 2019 to 17th August 2019, with an emphasis on the period August 15 to 17, capturing public sentiment during key political developments. The dataset was saved as 'Tweets_08-17Aug2019.csv'. This Twitter hashtag collection was done to analyze popular sentiment over the Union Government of India's repeal of Article 370. The timeframe was intentionally selected to avoid the extreme surge in tweet volume observed on the announcement day (August 5), which recorded several lakh tweets per day and consisted largely of reactionary and high-frequency reposted content. By selecting a slightly later window, the study aimed to analyze relatively stabilized public discourse while maintaining dataset balance. A sample of data collected is provided in Fig. 2.

The dataset consists of the following attributes, which are described below in Table 2.

Data Pre-Processing

Preprocessing data is an essential aspect of any machine learning pipeline, but it is especially important for natural language processing applications like sentiment analysis. A thorough preprocessing step was performed on the raw Twitter data to eliminate noise, normalize textual content, and guarantee linguistic consistency before machine learning techniques were applied for sentiment classification. The detailed procedure is outlined in Algorithm 1.

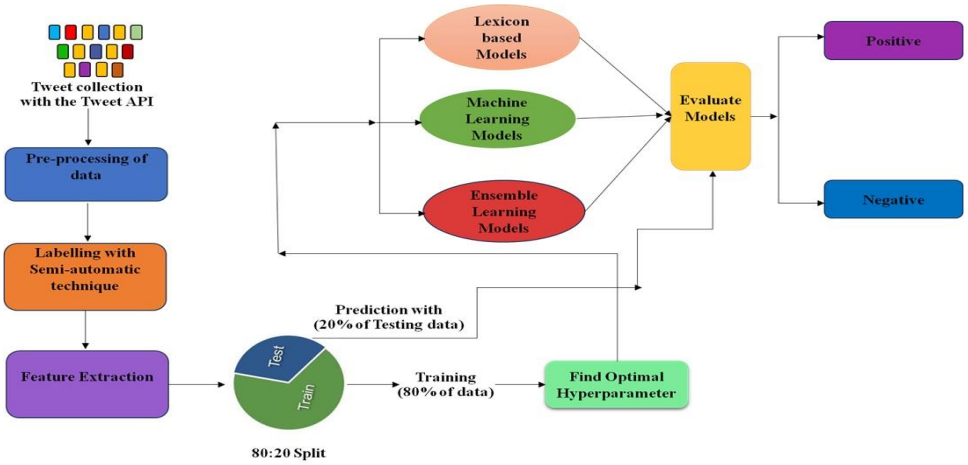

Fig. 1: Architecture of Research Methodology

	ID	datetime	has_media	is_reply	is_retweet	medias	nbr_favori	nbr_reply	nbr_retwe	text	url	user_id	usernameTweet
2	1.16E+18	10-08-2019 13:00		FALSE	FALSE		1	0	0	#	/Hilladventure	1.15E+18	Hilladventure1
3	1.16E+18	10-08-2019 11:19		FALSE	FALSE		0	0	0	Shocking & Shame	/samwhitefyr/	9.91E+17	samwhitefyr
4	1.16E+18	14-08-2019 01:30		TRUE	FALSE		0	0	0	What makes " True	/rashidakhan1	7.11E+17	rashidakhan1977
5	1.16E+18	16-08-2019 00:21		TRUE	FALSE		0	0	0	Just sickening to	/indianbyheart	579614977	indianbyheart76
6	1.16E+18	11-08-2019 20:31		TRUE	FALSE		0	0	0	Don't worry Pyari E	/Shivaprakash	1.13E+18	ShivaprakashYe2
7	1.16E+18	11-08-2019 19:33		FALSE	FALSE		0	0	0	The Kasmiris didn't	/RishiBhargava	8.05E+17	RishiBhargava17
8	1.16E+18	15-08-2019 10:55		FALSE	FALSE		0	0	0	#	/abhinavchanc	1.16E+18	abhinavchand16
9	1.16E+18	11-08-2019 08:08		TRUE	FALSE		0	0	0	When a big tree fa	/sshonti/statu	8.01E+17	sshonti
10	1.16E+18	11-08-2019 11:02		TRUE	FALSE		0	0	0	Indian action regar	/abdulrehman	2911998450	abdulrehman670
11	1.16E+18	16-08-2019 15:06		TRUE	FALSE		1	0	0	Well Article 370	/rahulmanuwa	297279988	rahulmanuwas
12	1.16E+18	10-08-2019 23:39		TRUE	FALSE		0	2	0	Understand but IPC	/ganabhat/sta	17894096	ganabhat
13	1.16E+18	15-08-2019 13:36		TRUE	FALSE		0	0	0	Congress against A	/Jrai_x/status/	1.06E+18	Jrai_x
14	1.16E+18	10-08-2019 21:03	TRUE	FALSE	FALSE	['https://t.	2	0	2	BBC News - Article	/stephen8275	636415876	stephen82754737
15	1.16E+18	16-08-2019 16:34	TRUE	FALSE	FALSE	['https://t.	0	0	0	Madras Bar Associ	/VIJAYraghavS	235529423	VIJAYraghavSING
16	1.16E+18	15-08-2019 00:55		TRUE	FALSE		0	0	0	Please do	/realkulwantk/	7.59E+17	realkulwantk
17	1.16E+18	14-08-2019 15:36		TRUE	FALSE		10	1	2	Again she starting	/sudip_ind/sta	4346490554	sudip_ind
18	1.16E+18	12-08-2019 10:12		TRUE	FALSE		3	2	0	Had thee been a Hi	/vishalkmumb	245264896	vishalkmumbai
19	1.16E+18	11-08-2019 00:21	TRUE	FALSE	FALSE	['https://t.	1	0	1	Article 370 : First	/eye_thinker/s	1599438462	eye_thinker
20	1.16E+18	11-08-2019 03:29	TRUE	FALSE	FALSE	['https://t.	0	1	0	This is one	/sartajanand/s	19639769	sartajanand

Fig. 2: Raw data gathered from Twitter.

Table 2: Attributes of Dataset

Field Name	Datatype	Possible Values	Description	Data Category
ID	Integer	Unique numeric values (e.g., 1, 2, 3, ...)	Unique identifier for each tweet	Identifier
Datetime	Datetime	e.g., 2025-07-18 14:32:00	Timestamp of when the tweet was posted	Temporal
has_media	Boolean / Nullable	True, False, NULL	Indicates presence of media content in tweet	Content Metadata
is_reply	Boolean	True, False	Indicates if the tweet is a reply to another tweet	Interaction Type
is_retweet	Boolean	True, False	Indicates if the tweet is a retweet	Interaction Type
medias	String / Nullable	e.g., photo, video, gif, NULL	Type of media in the tweet, if any	Content Metadata
nbr_favorite	Integer	0, 1, 2, ...	Number of times the tweet was liked	Engagement Metric
nbr_reply	Integer	0, 1, 2, ...	Number of replies to the tweet	Engagement Metric
nbr_retweet	Integer	0, 1, 2, ...	Number of times the tweet was retweeted	Engagement Metric
text	String	Free-form tweet text	Main content of the tweet used for sentiment/text analysis	Textual Content
url	String	URL format (e.g., https://twitter.com/...)	URL of the tweet	Reference/Link
user_id	Integer	Unique user IDs (e.g., 123456)	Unique identifier of the tweet author	User Identifier
usernameTweet	String	e.g., @user123	Twitter handle of the tweet author	User Metadata

Algorithm 1: Text Preprocessing for Twitter Dataset

Input: Raw tweet texts in a dataset

Output: Cleaned dataset

Begin:

1.Read the text in the dataset;

2.While (! End of the text in a dataset):

If the text is not in English, remove the entry;

If the text contains null value, then remove the entry;

If the text contains extra white space, then trim the text;

If the text is empty after stripping, then remove the entry;

If the text contains duplicate text, then remove duplicate text entries;

If the text contains hashtags [e.g., #example…], Remove all hashtags;

If the text contains Twitter handles (e.g., @username), Remove Twitter handles;

If the text contains URLs (e.g., http://... or https://...), then remove all URLs;

If the text contains special characters, then remove all special characters;

If the text contains multiple consecutive spaces, then substitute multiple consecutive spaces with a single space;

If the text contains only a number or is purely numeric, then remove the entry;

3. Return the cleaned dataset;

End:

This thorough preprocessing ensured high-quality input for sentiment analysis by reducing the dataset to 30,559 tweets. The sample output of tweets is shown in Figure 3.

Polarity Calculation and Sentiment Analysis

Polarity information was extracted from Twitter posts about the repeal of Article 370 using sentiment analysis. To generate sentiment scores, a lexicon-based approach was used, which allowed for the systematic and scalable annotation of massive amounts of unstructured textual data. The TextBlob (Loria et al., 2014) lexicon-based polarity scoring mechanism was chosen for automated label generation out of the assessed tools AFINN (Nielsen, 2011), VADER (Hutto and Gilbert, 2014), SentiWordNet (Baccianella et al., 2010), and TextBlob because of its integrated polarity and subjectivity scoring framework.

Polarity scores for this study ranged from -1 to +1, with values near zero indicating neutral orientation, positive values indicating positive sentiment, and negative values indicating negative sentiment. To concentrate on blatantly divisive viewpoints, neutral tweets were not included in the analysis. As a result, a binary sentiment problem with positive and negative categories was used to structure the classification task.

	text
2	KashmirWithModi With the scrapping of Article 370 the dream of Sardar Patel Baba Saheb Ambedkar Dr Syama Prasad Mukherjee Atalji and of crores of patriots has been fulfilled pictwitterc
3	What makes True Islam followers like you conflate Faith with real estate Did India ever throw anyone out on the basis of religion And note it was Article 370 that was allowing child marriage
4	Just sickening to read this thread guess it doesnt matter if if article 370 tramples hindus Dalits rights making them second class citizens unlike Kashmiris who can do whatvr they want in Ind
5	Dont worry Pyari Behna whoever opposing or protesting against the lifting of 370 shortly you will see them in CENTERS in queue giving their biometrics
6	KashmirWithModi Article 370 and 35A gave terrorism separatism dynasty politics corruption to JK Pakistan used it to destabilise the region 42000 people had to give up their lives due to the
7	When big tree falls the earth shakes Which is the big tree Indira or Article 370 For me Indira Gandhi because its cost was thousands of Sikh life
8	Indian action regarding abrogation of article 370 is against freedom of Kashmiri peoples
9	Well Article 370 that Hasan mentions was supposed to be temporary and the alarming situation he is so worried about does not exist Stop making decisions based on biased reports
10	Understand but IPC Indian Penal Code must apply post the Article 370 abolition Case must stand
11	Congress against Article 370 35a Congress is divisive and anti national
12	Please do enlighten yourself on some basic history Article 370 was temporary Im law student so understand these things quite well than you
13	Again she starting kashmir kashmir kashmir Article 370 sub articles revoked from all regions in jammu kashmir ladakh But theye only see Kashmir Kashmir kashmir think people of jammu lad
14	Had thee been Hindu majority there wouldnt HV been need for article 370 in first place
15	Isnt he right Restoration of Article 370 vl free Kashmir Than why even after 70yrs India is still there in Kashmir Pol struggles and these Cons Articles are meaningless You will either have to
16	You may show what want to show but the reality is that Majority Common ppl in Kashmir Jammu and Laddakh and also Whole of India are Solidly standing Supporting Abrogation of Maliciou
17	The fact is its article 370 not section 370 The fact is article 370 doesnt exist anymore
18	narendramodi wish and Congratulation your Successfull Leadership Government of india Article 370 and 35A End in JK
19	KashmirWithModi With the scrapping of Article 370 the dream of Sardar Patel Baba Saheb Ambedkar Dr Syama Prasad Mukherjee Atalji and of crores of patriots has been fulfilled via NaMo
20	PC instead of telling BJP has removed Article 370 because its Muslim state rascal stop spreading non existence issue tell to people of India how 3300 cr is with UNEDUCATED son of you as

Fig. 3: Output after cleaning of data

TextBlob offers a subjectivity score in addition to polarity, with values closer to 0 denoting objective statements and values closer to 1 denoting subjective or opinion-driven content. Subjectivity scoring was added to help prioritise tweets that expressed personal opinions and filter factual statements.

The sentiment classification criteria were defined as follows:

- If the polarity score > 0.05 and the subjectivity score ≥ 0.2, the tweet was labeled as positive
- Otherwise, the tweet was categorized as negative

These automatically produced polarity labels served as training annotations for deep learning and supervised models that followed. Notably, these labels are not human-validated ground truth; rather, they are automated annotations derived from lexicons. Consequently, the evaluation of the model that follows shows predictive agreement with lexicon-based polarity scoring. TextBlob outputs were used for final dataset labelling to maintain methodological consistency across experiments, even though several lexicon-based tools were investigated for comparative purposes.

Figure 4 shows a partial depiction of the data following score assignment.

Supervised Machine Learning Methodology

This study used a supervised learning approach that involved data partitioning, feature extraction, model training, and sentiment prediction in a methodical order. In the supervised learning phase, sentiment classification was carried out by training various machine learning models on labeled tweet data. The process began with a training prediction pipeline, wherein the labeled dataset was divided using an 80:20 ratio for training and testing. Each tweet in the training set was associated with a binary sentiment label positive or negative previously assigned using lexicon-based methods. These labels served as target variables during model learning.

Training Prediction Process

In this step, each machine learning model was trained to identify patterns in the tweet text and link them to the appropriate sentiment labels. Each labelled feature vector, which represented the linguistic structure of a tweet, was taken from the training set and used to train the models. Following training, the models were applied to the test set to predict sentiment labels. To evaluate classification accuracy and dependability, predictions were then contrasted with actual labels.

Feature Extraction

The cleaned tweets were transformed into numerical vectors using the Term Frequency–Inverse Document Frequency (TF-IDF) approach to turn textual data into a machine-readable format. Bigram and unigram configurations were used to capture the relevance of neighbouring word pairs as well as individual words, which are crucial for comprehending context.

	text	polarity_score	subjectivity	sentiment
2	KashmirWithModi With the scrapping of Article 370 the dream of Sardar Patel	0	0	0
3	What makes True Islam followers like you conflate Faith with real estate Did I	-0.029166667	0.529166667	0
4	Just sickening to read this thread guess it doesnt matter if if article 370 trampl	-0.3	0.333333333	0
5	Dont worry Pyari Behna whoever opposing or protesting against the lifting of	0	0.3	0
6	KashmirWithModi Article 370 and 35A gave terrorism separatism dynasty poli	-0.125	0.375	0
7	When big tree falls the earth shakes Which is the big tree Indira or Article 370	0	0.1	0
8	Indian action regarding abrogation of article 370 is against freedom of Kashm	0.1	0.1	0
9	Well Article 370 that Hasan mentions was supposed to be temporary and the	-0.1	0.6	0
10	Understand but IPC Indian Penal Code must apply post the Article 370 abolitic	0	0	0
11	Congress against Article 370 35a Congress is divisive and anti national	0	0	0
12	Please do enlighten yourself on some basic history Article 370 was temporary	0	0.125	0
13	Again she starting kashmir kashmir kashmir Article 370 sub articles revoked fr	0	0.55	0
14	Had thee been Hindu majority there wouldnt HV been need for article 370 in f	0.25	0.333333333	1
15	Isnt he right Restoration of Article 370 vl free Kashmir Than why even after 70	0.061904762	0.778571429	1
16	You may show what want to show but the reality is that Majority Common pp	0.03	0.321428571	0
17	The fact is its article 370 not section 370 The fact is article 370 doesnt exist ar	0	0	0
18	narendramodi wish and Congratulation your Successfull Leadership Governme	0	0	0
19	KashmirWithModi With the scrapping of Article 370 the dream of Sardar Patel	0	0	0
20	PC instead of telling BJP has removed Article 370 because its Muslim state ras	0	0	0
21	KashmirWithModi It is Modi govt which is finally fulfilling the aspirations of th	0.033333333	0.733333333	0
22	Bold of you to say after India revoked Article 370 Seriously mean wonder wha	0.010416667	0.677083333	0
23	What did you do for all Kashmiris Pandits Are you preparing your way to Rajya	0.2	0.35	1
24	Article 370 35A are part of Indian constitution nd dont think revoking those sh	0.214285714	0.428571429	1
25	KashmirWithModi The abolition of article 370 will herald new dawn for the pe	0.118181818	0.402272727	1

Fig. 4: Score of polarity, subjectivity, and sentiment

Classification of Text

Several supervised classifiers were used to learn the sentiment patterns encoded in the TF-IDF vectors following feature extraction. The labelled dataset was used to train models, including logistic regression, support vector machines, random forests, MLP, etc. Grid search and cross-validation were used to adjust hyperparameters to maximize performance. Predictive resilience was further improved by implementing ensemble techniques, such as a stacking classifier with logistic regression as the meta-learner and a soft voting classifier. The preceding pseudocode describes the processes involved in the classification process, from training to assessment on the test set.

Validation Strategy

To ensure the reliability of model evaluation, 3-fold stratified cross-validation was employed. In this current study, large-scale dataset (over 2.2 lakh tweets) and the comprehensive analysis of multiple models, 3-fold cross-validation provided a computationally efficient evaluation framework. Performance metrics were computed for each fold and averaged to obtain the final reported results. The consistency of scores across folds indicates stable model generalization without significant variance.

Pseudo-code Algorithm for Classification:

Begin
Step 1: Load and Preprocess the Dataset
- Clean text (remove stopwords, punctuation, lowercase conversion)
- Tokenize the text data

Step 2: Feature Extraction
- Use TF-IDF vectorization with bigrams
- Generate feature matrix

Step 3: Lexicon-Based Sentiment Analysis
For each tool in [TextBlob, AFINN, VADER, SentiWordNet]:
- Assign sentiment scores
- Evaluate Accuracy, Precision, Recall, F1-Score, AUC

Step 4: Machine Learning Classification
For each model in [Logistic Regression, Decision Tree, Random Forest, Extra Trees, AdaBoost, Gradient Boosting, Bagging, XGBoost, MLP, SVM]:
- Perform train-test split
- Apply hyperparameter tuning
- Train and evaluate using TF-IDF features

Step 5: Ensemble Voting Classifier
- Use base models: Logistic Regression, Random Forest, SVM
- Tune parameters and apply soft voting
- Evaluate ensemble performance

Step 6: Stacking Ensemble Model
- Base models: Multinomial NB, Logistic Regression, XGBoost
- Use a meta-classifier for final prediction
- Train and evaluate performance

Step 7: Deep Learning Classification using CNN + LSTM
- Preprocess and tokenize text; apply padding to fixed length
- Split the dataset into training and testing sets
- Construct a hybrid model: Embedding → Conv1D → MaxPooling → LSTM → Dropout → Dense → Sigmoid
- Compile with binary crossentropy loss and Adam optimizer
- Train with early stopping; evaluate using standard classification metrics

Step 8: Result Comparison
- Compare all models using performance metrics
- Identify the best-performing approach for reporting

End

Supervised Classification Models

These models ranged from linear classifiers and decision-tree-based ensembles to neural networks and support vector machines. All models were trained on TF-IDF vectorized data using unigram and bigram features. Hyperparameter optimization was conducted using RandomizedSearchCV with 3-fold cross-validation to identify the best-performing configurations. Below is an overview of each classifier used in this study.

Logistic Regression

It is a linear classification algorithm that estimates class probabilities using the logistic sigmoid function. It performs well in high-dimensional spaces, especially with sparse data such as text. In this study, the optimal configuration included a regularization strength of C = 1 and the 'liblinear' solver, which allows support for both L1 and L2 regularization.

Support Vector Machine (SVM)

It is a powerful classifier known for its effectiveness in high-dimensional feature spaces. It constructs a decision boundary that maximizes the margin between classes. The best results were obtained using a 'linear' kernel with a regularization parameter C = 1, which ensured a good balance between margin width and classification error.

Decision Tree Classifier

It is a non-parametric model that splits data based on feature thresholds, forming a hierarchical decision structure. It can handle both linear and non-linear relationships but is prone to overfitting if not pruned. The tuned model performed best with max_depth = 20 and min_samples_split = 5.

Random Forest Classifier

It is an ensemble method that aggregates predictions from multiple decision trees, reducing variance and

improving generalization. The model achieved its best performance with n_estimators = 200, max_depth = 20, and min_samples_split = 2.

Extra Trees Classifier

It further increases diversity among trees by selecting split thresholds at random. This technique reduces overfitting and computational cost. The optimal configuration matched that of random forest: n_estimators = 200, max_depth = 20, and min_samples_split = 2.

AdaBoost Classifier

It works by iteratively combining weak learners, typically decision stumps, and focusing more on misclassified instances at each step. The model performed best with n_estimators = 100 and learning_rate = 1.0.

Gradient Boosting Classifier

It builds decision trees in a stage-wise manner, optimizing a differentiable loss function using gradient descent. In this task, the best parameters were n_estimators = 200, learning_rate = 0.1, and max_depth = 5.

Bagging Classifier

It trains multiple instances of a base estimator on different random subsets of the training data, thereby reducing variance. The most suitable configuration used n_estimators = 100.

XGBoost Classifier

It is an optimized gradient boosting framework that includes regularization to prevent overfitting and supports efficient computation. In this implementation, use_label_encoder=False and eval_metric='logloss' were specified. The best configuration was obtained with n_estimators = 200, learning_rate = 0.1, and max_depth = 5.

Multilayer Perceptron (MLP)

It is a feedforward artificial neural network that maps input features to output labels through multiple hidden layers. It can model complex, non-linear relationships in data. The optimal model architecture included hidden_layer_sizes = (100, 50), the 'relu' activation function, and the 'adam' optimizer, with a maximum of 300 iterations (max_iter = 300).

Ensemble Classification Strategies

The study used ensemble learning approaches to improve generalisation and classification performance. A soft voting ensemble and a stacked generalisation model (stacking ensemble) were the two ensemble techniques used. The strengths of many basic classifiers were merged in both methods to create a more reliable decision-making process.

Soft Voting Classifier

The soft voting ensemble aggregated the probability outputs of three distinct base classifiers: Logistic Regression, Random Forest, and Support Vector Machine. This approach considers the class probability predicted by each classifier and outputs the class with the highest average probability, thereby improving robustness in decision boundaries.

Hyperparameter tuning was conducted using a GridSearchCV approach with 3-fold cross-validation. The best configuration obtained was:

- RandomForestClassifier: n_estimators = 200, max_depth = 20
- LogisticRegression: C = 1.0
- SVC (with probability=True): C = 1.0

The final model used soft voting to combine predictions, leveraging both linear and non-linear decision surfaces to handle the complexity of tweet sentiment classification.

Stacking Classifier

The stacking ensemble utilized a two-layer architecture where base classifiers (first level) were trained independently, and their outputs were passed to a meta-learner (second level) for final prediction. In this setup, the base learners were:

- Logistic Regression (max_iter = 1000)
- Multinomial Naïve Bayes

The meta-classifier was XGBoost, selected for its regularization capabilities and ability to capture complex patterns. It was configured with:

- Use_label_encoder=False
- Eval_metric='logloss'

TF-IDF vectorization (with bigrams, min_df=5, and max_df=0.9) was embedded within a pipeline, ensuring consistent preprocessing across the base and meta models. The stacking model employed 5-fold stratified cross-validation (StratifiedKFold) to prevent overfitting during training.

Hybrid Deep Learning Ensemble: A CNN–LSTM Integrated Framework

A deep learning architecture of convolutional and recurrent layers was used to capture both local characteristics and sequential relationships in Twitter data to supplement traditional and ensemble-based machine learning approaches. This hybrid model made use of Long Short-Term Memory (LSTM) networks to imitate temporal dynamics in the input sequences and Convolutional Neural Networks (CNNs) to extract local n-gram features.

The model architecture began with an embedding layer (input_dim=15000, output_dim=128, input_length=120) to project tokens into dense vectors. This was followed by a 1D convolutional layer with 64 filters and a kernel size of 5, using ReLU activation to detect local patterns. A max-pooling layer with pool size 2 reduced the spatial dimensionality. The pooled output was passed to an LSTM layer with 64 units to capture long-range dependencies in the sequence. To prevent overfitting, a dropout layer with a rate of 0.5 was applied after the LSTM output.

The dense layers included a fully connected ReLU-activated layer with 32 neurons, followed by a sigmoid output layer producing a binary prediction. The model was compiled using the 'adam' optimizer and 'binary_crossentropy' loss function, and trained for up to 10 epochs with a batch size of 128. Early stopping was employed to halt training when the validation loss failed to improve for three consecutive epochs, preserving the best weights.

The detailed layer configuration and hyperparameters are provided to ensure methodological transparency and reproducibility.

Results and Discussion

In this section, the classification models used for sentiment analysis of tweets on the repeal of Article 370 are compared. The performance of individual Lexicon-Based Sentiment models, supervised classifiers, ensemble models, and a deep learning-based hybrid architecture was assessed using a consistent evaluation framework. Each model was trained on 80% of the data and assessed on the remaining 20% using the following common performance metrics: ROC-AUC, F1-score, accuracy, precision, and recall. The model performance metrics reported in this section reflect predictive agreement with lexicon-derived automated sentiment labels, rather than independently human-annotated ground truth.

Performance Evaluation of Lexicon-Based Sentiment Models

To categorise the sentiment of tweets about Article 370, four lexicon-based sentiment analysis algorithms were used: TextBlob, AFINN, VADER, and Senti WordNet (Loria et al., 2014; Nielsen, 2011; Hutto and Gilbert, 2014; BacciAnella et al., 2010). Each approach uses specified sentiment dictionaries to apply rule-based polarity scoring. Accuracy, precision, recall, F1-score, and AUC were used to assess their performance, while confusion matrices were used for visualisation. The distribution of public opinion about the repeal of Article 370 is depicted in the following pie chart in Figure 5. In the dataset under analysis, 40.25% of the tweets were classified as negative and 59.75% as positive based on lexicon-derived polarity scores.

Text Blob

When tested against its threshold-based classification criteria, Text Blob, which was utilised for automated polarity label generation, obtained an accuracy of 94.29%. Consistency in the lexicon-derived labelling scheme is reflected in the high recall. The F1-score was 0.9338, supported by a high precision of 0.8758. Its AUC value of 0.99 further reflected its strong capability to distinguish between positive and negative sentiments in the dataset.

A sentiment label of 0 indicates negativity, whereas a label of 1 indicates positivity. 12,301 positive tweets and 16,514 negative tweets were accurately identified by the model. High recall was demonstrated by the fact that no positive tweets were mistakenly categorised as negative. The model's propensity for recall-oriented predictions was demonstrated by the 1,744 negative tweets that were projected to be positive. Overall, the confusion matrix reflects strong alignment with the lexicon-derived polarity classification criteria (Fig. 6).

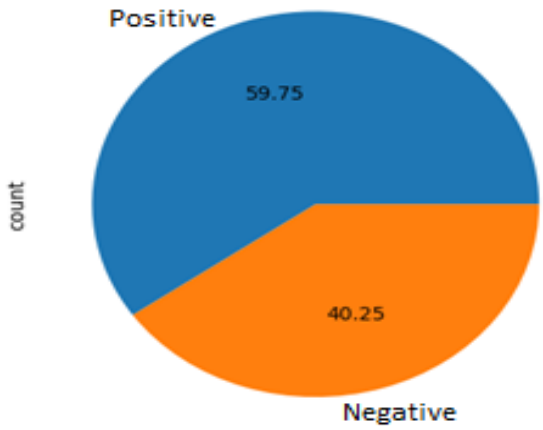

Fig. 5: Sentiment Analysis result of abrogation of Article 370

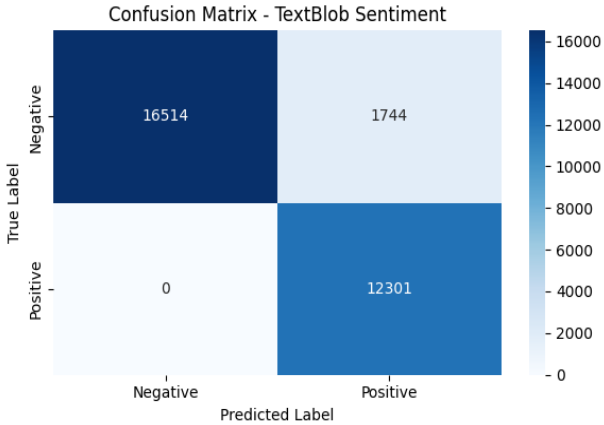

Fig. 6: Confusion matrix of TextBlob sentiment classification

AFINN

AFINN, which calculates sentiment by summing the integer polarity scores of individual words, recorded an accuracy of 67.09%. With a precision of 0.5878 and recall of 0.6107, it demonstrated balanced classification performance across metrics. Its F1-score of 0.599 and AUC of 0.70 confirm its utility as a reliable lexicon-based classifier, especially in scenarios requiring quick and interpretable sentiment scoring.

The model correctly identified 7,512 positive tweets and 12,990 negative tweets. It misclassified 4,789 positive tweets as negative and 5,268 negative tweets as positive. The results indicate that while the model performs moderately well across both classes, there is a relatively balanced distribution of misclassifications affecting both precision and recall (Fig. 7).

VADER

VADER, a sentiment model tailored for social media text, achieved an accuracy of 65.12%. Its strength was reflected in a recall of 0.6936, while the corresponding F1-score and AUC were 0.6155 and 0.70, respectively. These results suggest that VADER is particularly adept at identifying positive sentiment expressions in informal, short-text formats such as tweets.

The model correctly predicted 8,532 positive tweets and 11,369 negative tweets. It misclassified 3,769 positive tweets as negative and 6,889 negative tweets as positive. The distribution indicates that while VADER effectively identifies a substantial portion of each class, some overlap between sentiment boundaries leads to moderate misclassification rates on both sides (Fig. 8).

Senti Word Net

SentiWordNet applies sentiment scores at the synset level using WordNet-based lexical relationships. It achieved an accuracy of 58.20%, a recall of 0.7106, and an F1-score of 0.5778, with an AUC of 0.64. Despite operating at a more granular lexical level, its performance metrics affirm its relevance as a lexicon tool, particularly in linguistic research and comparative evaluations.

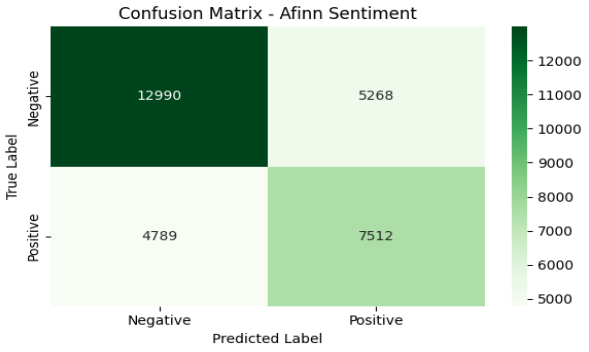

Fig. 7: Confusion matrix for AFINN sentiment classification

The model correctly predicted 8,741 positive tweets and 9,045 negative tweets. It misclassified 3,560 positive tweets as negative and 9,213 negative tweets as positive. The distribution indicates that while SentiWordNet captures a significant number of correct sentiment labels, a notable number of negative tweets are incorrectly classified as positive (Fig. 9).

Table 3 presents a comparative evaluation of four lexicon-based sentiment analysis models: TextBlob, AFINN, VADER, and SentiWordNet. With an accuracy of 94.29%, a perfect recall of 1.000, and an AUC of 0.99, TextBlob outperformed the others overall, indicating strong consistency within the lexicon-derived labelling framework.

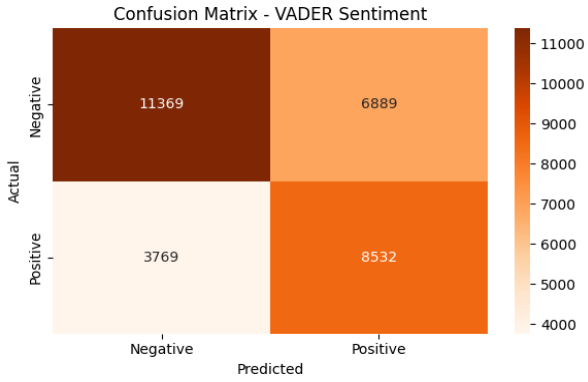

Fig. 8: Confusion matrix for VADER sentiment classification

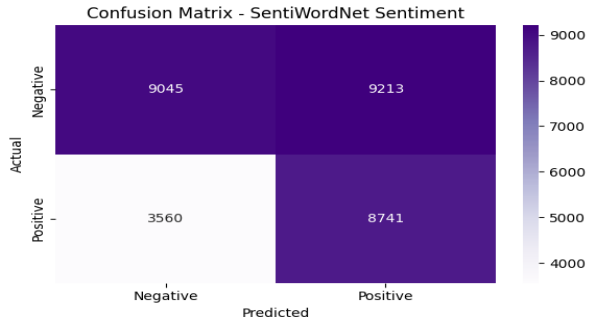

Fig. 9: Confusion matrix for SentiWordNet sentiment classification

Table 3: Performance Evaluation of Lexicon-Based Sentiment Models

Model	Accuracy	Precision	Recall	F1-Score	AUC
TextBlob	0.9429	0.8758	1.0000	0.9338	0.99
AFINN	0.6709	0.5878	0.6107	0.5990	0.70
VADER	0.6512	0.5533	0.6936	0.6155	0.70
SentiWordNet	0.582	0.4869	0.7106	0.5778	0.64

The comparative performance of four lexicon-based sentiment analysis models across five assessment measures is shown in Figure 10. Despite having a reasonably high recall, Senti WordNet exhibits the lowest precision and accuracy, whereas Text Blob consistently surpasses the other models in every metric.

Figure 11 illustrates that TextBlob achieves the highest performance with an AUC value of 0.99. AFINN and VADER follow, each showing an AUC of 0.70. In contrast, SentiWordNet records the lowest performance with an AUC of 0.64. The diagonal line in the figure indicates the baseline for random guessing, which corresponds to an AUC of 0.5.

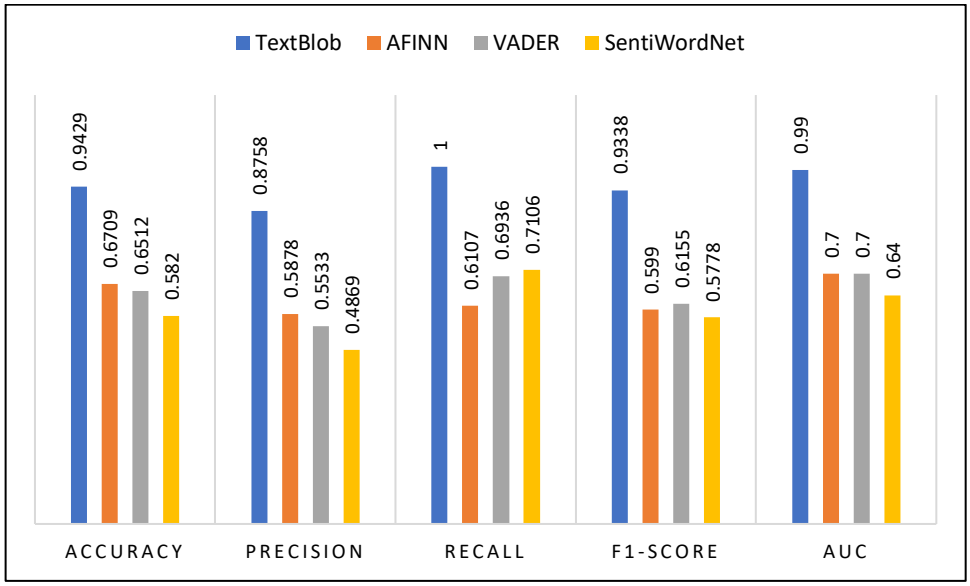

Fig. 10: Performance Comparison of Lexicon-Based Sentiment Models

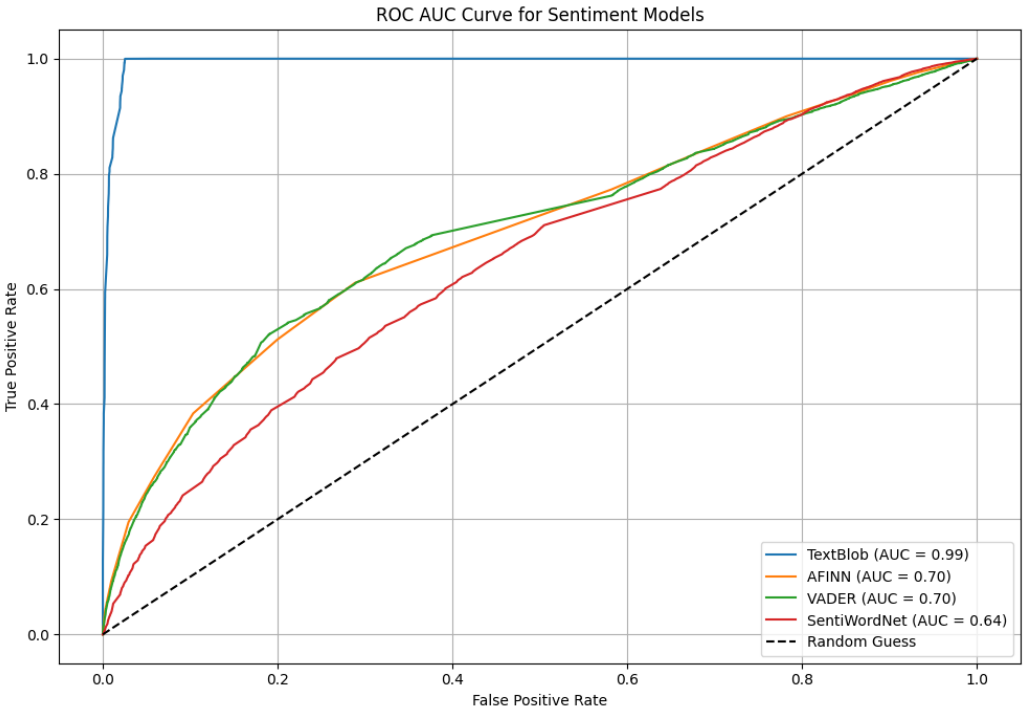

Fig. 11: ROC AUC Curve comparison of four Lexicon-Based Sentiment Models

Performance Evaluation of Supervised Machine Learning Models

Machine learning models were built in this phase to better categorize tweet emotions by utilizing bigrams and TF-IDF features, which capture context at both the single-word and phrase levels. The models included Logistic Regression, Decision Tree, Random Forest, Extra Trees, AdaBoost, Gradient Boosting, Bagging, XGBoost, MLP, and SVM. Hyperparameter tuning was performed using RandomizedSearchCV to optimize model configurations. The models' efficacy in binary sentiment categorization of Article 370 tweets was assessed using accuracy, precision, recall, and F1-score. Ten supervised machine learning models that were evaluated on the binary sentiment classification task using TF-IDF bigram features are compared in Table 4.

With an accuracy of 92.75% and an F1-score of 0.9477, Logistic Regression was the best-performing model, showing a high degree of consistency across precision 0.9315 and recall 0.9645. This indicates that it was highly effective in classifying both positive and negative sentiments with minimal error. SVM (Support Vector Machine) closely followed, achieving an accuracy of 92.46% and an F1-score of 0.9461. The high recall of 0.9715 highlights its ability to correctly identify a large proportion of positive tweets. Random Forest, Extra Trees, MLP, and Bagging classifiers also performed competitively, each achieving an accuracy above 90% and maintaining F1-scores above 0.93. These ensemble and neural network-based models benefited from their ability to model complex patterns and reduce overfitting. Gradient Boosting and XGBoost achieved moderately high performance, with F1-scores of 0.9203 and 0.9096, respectively, although they were slightly lower in precision compared to other ensembles. On the lower end, Decision Tree and AdaBoost showed reduced F1-scores of 0.9012 and 0.8278, respectively. While AdaBoost achieved the highest recall of 0.9935, its low precision of 0.7095 suggests a high number of false positives, which reduces its overall reliability. The performance comparison of Supervised Machine Learning Models is illustrated in Fig. 12.

Table 4: Performance Evaluation of supervised machine learning models

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.927583	0.931524	0.964548	0.947749
Decision Tree	0.867236	0.913043	0.889754	0.901248
Random Forest	0.911098	0.919133	0.953307	0.935908
Extra Trees	0.910509	0.933535	0.935149	0.934341
AdaBoost	0.718575	0.709478	0.993515	0.82781
Gradient Boosting	0.885487	0.874319	0.971466	0.920336
Bagging	0.906094	0.923534	0.939905	0.931648
XGBoost	0.867825	0.851433	0.976221	0.909567
MLP Classifier	0.908154	0.929953	0.935581	0.932759
SVM	0.924639	0.922035	0.971466	0.946105

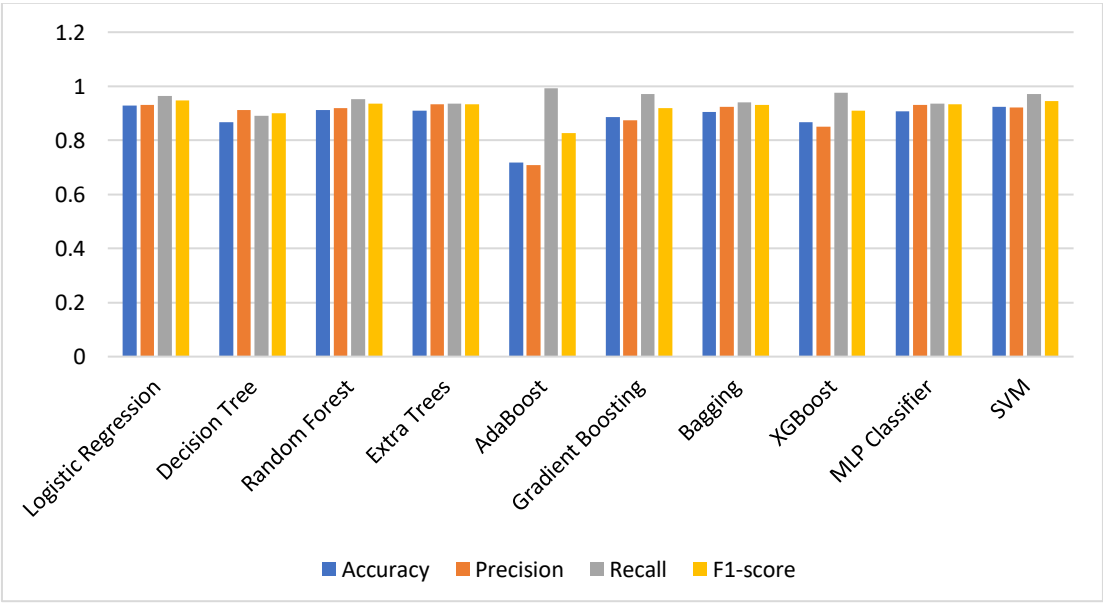

Fig. 12: Performance Comparison of Supervised Machine Learning Models.

The performance of the Logistic Regression model was assessed using a confusion matrix. As shown in the confusion matrix (Figure 13), the model correctly classified 780 negative samples and 2274 positive samples, while it misclassified 297 negative samples as positive (false positives) and 46 positive samples as negative (false negatives). This indicates strong predictive performance, particularly in identifying positive instances.

The Random Forest model's performance was assessed using a confusion matrix. As shown in the confusion matrix (Figure 14), the model correctly classified 882 negative samples and 2194 positive samples, while it misclassified 195 negative samples as positive (false positives) and 126 positive samples as negative (false negatives).

A confusion matrix was used to evaluate the performance of the Extra Trees model. The model accurately identified 938 negative samples and 2166 positive samples, as seen in the confusion matrix (Figure 15). However, it incorrectly identified 139 negative samples as positive (false positives) and 154 positive samples as negative (false negatives). With a balanced capacity to identify both positive and negative instances this suggests dependable performance.

The Bagging model's performance was assessed using a confusion matrix. As shown in the confusion matrix (Figure 16), the model correctly classified 903 negative samples and 2100 positive samples, while it misclassified 174 negative samples as positive (false positives) and 220 positive samples as negative (false negatives). This reflects strong overall performance, with a solid ability to identify both classes, though slightly more errors were observed in predicting positive instances compared to the previous models.

The MLP Classifier model's performance was assessed using a confusion matrix. As shown in the confusion matrix (Figure 17), the model correctly classified 922 negative samples and 2154 positive samples, while it misclassified 155 negative samples as positive (false positives) and 166 positive samples as negative (false negatives). This demonstrates robust classification capability with relatively balanced error distribution across both classes.

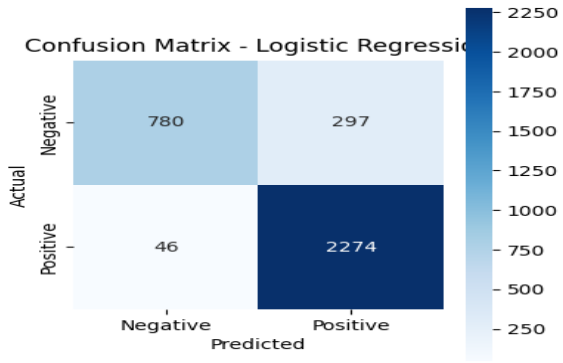

Fig. 13: Confusion matrix of LR

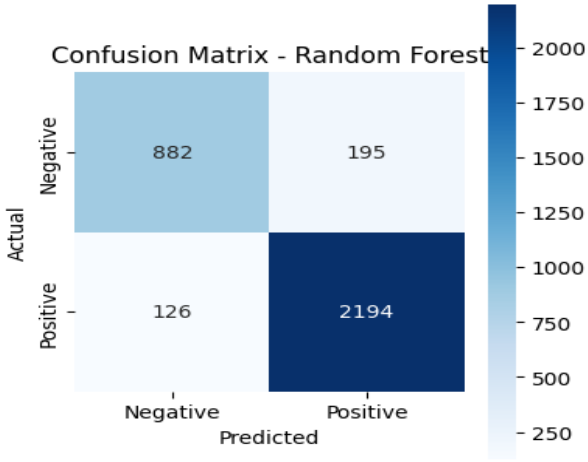

Fig. 14: Confusion matrix of RF

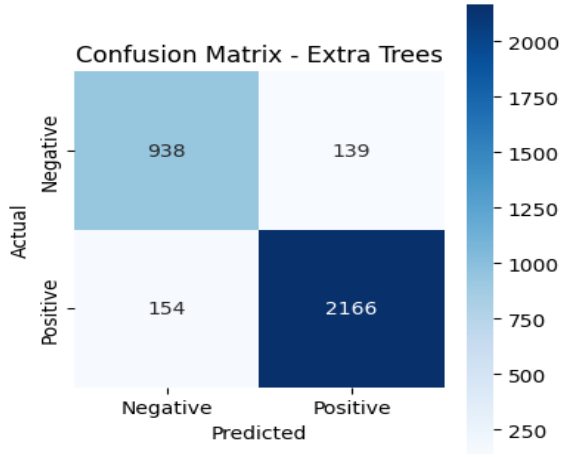

Fig. 15: Confusion matrix of Extra Trees

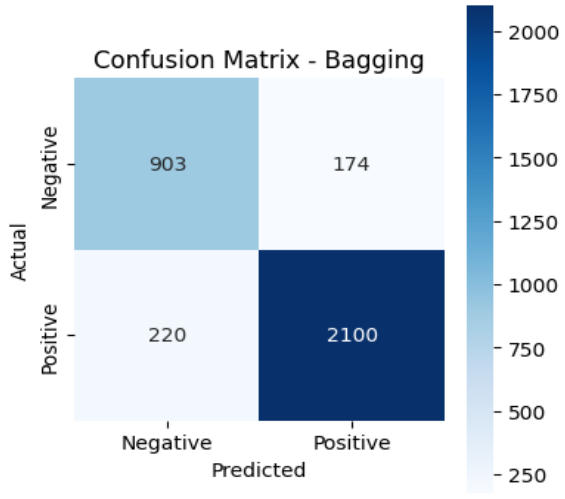

Fig. 16: Confusion matrix of Bagging

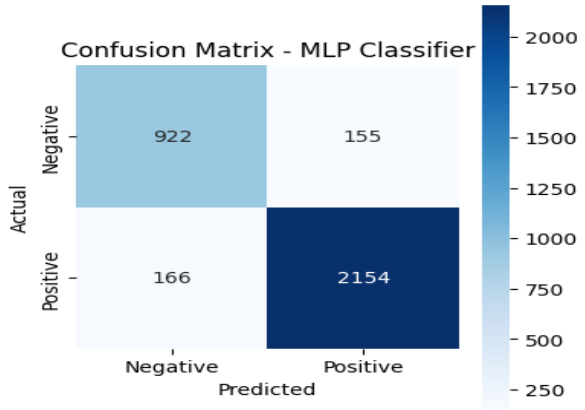

Fig. 17: Confusion matrix of MLP

The Gradient Boosting model's performance was assessed using a confusion matrix. As shown in the confusion matrix (Figure 18), the model correctly classified 437 negative samples and 2282 positive samples, while it misclassified 640 negative samples as positive (false positives) and 38 positive samples as negative (false negatives). This indicates excellent ability in identifying positive instances, though the model struggled with accurately classifying negative cases.

The XGBoost model's performance was assessed using a confusion matrix. As shown in the confusion matrix (Figure 19), the model correctly classified 806 negative samples and 2262 positive samples, while it misclassified 271 negative samples as positive (false positives) and 58 positive samples as negative (false negatives). This reflects high predictive accuracy, particularly in identifying positive cases, with relatively low false negative errors.

The Decision Tree model's performance was assessed using a confusion matrix. As shown in the confusion matrix (Figure 20), the model correctly classified 872 negative samples and 2078 positive samples, while it misclassified 205 negative samples as positive (false positives) and 242 positive samples as negative (false negatives). This indicates moderate performance, with a slightly higher tendency to misclassify positive samples compared to some ensemble models.

The AdaBoost model's performance was assessed using a confusion matrix. As shown in the confusion matrix (Figure 21), the model correctly classified 108 negative samples and 2309 positive samples, while it misclassified 969 negative samples as positive (false positives) and 11 positive samples as negative (false negatives). This highlights the model's exceptional ability to identify positive cases, though it shows a significant weakness in accurately classifying negative instances.

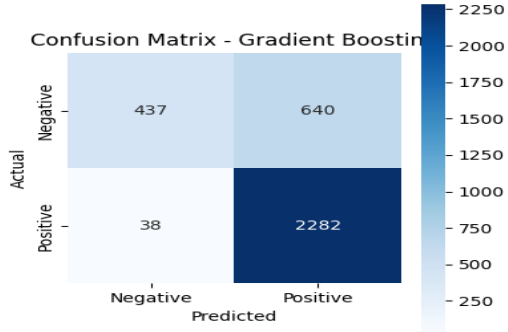

Fig. 18: Confusion matrix of Gradient Boosting

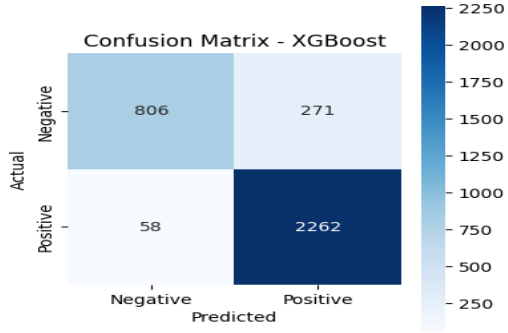

Fig. 19: Confusion matrix of XGBoost

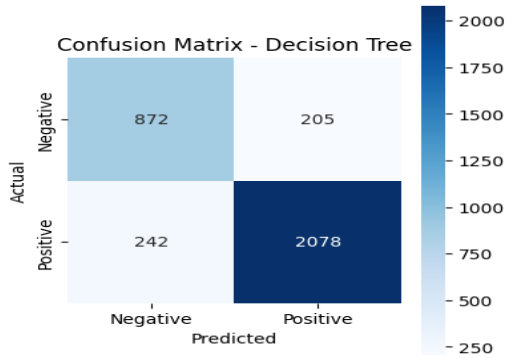

Fig. 20: Confusion matrix of Decision Tree

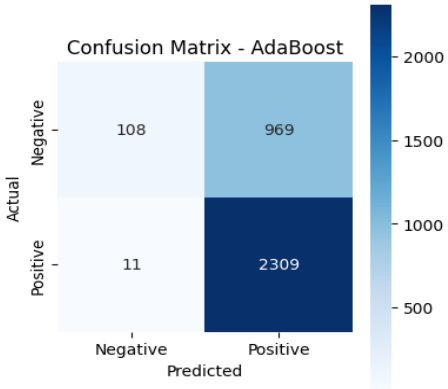

Fig. 21: Confusion matrix of AdaBoost

The ROC AUC curves for each assessed classification model are shown in Figure 22. The best capacity to discriminate between positive and negative classes was demonstrated by SVM, XGBoost, and Logistic Regression, which had the highest AUC values among them (0.960, 0.960, and 0.959, respectively). With AUC values of 0.953 and 0.945, the MLP Classifier and Extra Trees both demonstrated high performance. AdaBoost, on the other hand, performed the lowest, with an AUC of 0.755. Simpler models like Decision Tree and AdaBoost exhibited somewhat lesser discriminative power, but ensemble techniques like Random Forest, Extra Trees, and XGBoost showed strong classification capabilities overall.

Performance Evaluation of Ensemble Learning Models

To robustly classify sentiments expressed in tweets concerning the revocation of Article 370, three ensemble learning approaches were evaluated: Voting Classifier, Stacking Classifier, and a Deep Learning-based CNN-LSTM Hybrid Model. Each model was trained and validated on the same pre-processed dataset and assessed using standard evaluation metrics.

Voting Classifier (With Hyperparameter Tuning)

We used an ensemble voting technique that combined SVM, Random Forest, and Logistic Regression, and probability-based soft voting was used to improve classification performance even more. The optimized model achieved an accuracy of 94.05%, precision of 94.60%, recall of 96.81%, and an F1-score of 95.69%. These results reflect one of the highest overall performances across all models evaluated. The voting mechanism helped mitigate the individual limitations of base models, resulting in a more robust and generalizable classifier. This approach is particularly well-suited for real-world applications due to its high accuracy and interpretability.

The model's excellent performance is demonstrated in Figure 23 by the adjusted Voting Classifier's confusion matrix, which accurately identified 2,246 positive and 949 negative tweets. With only 74 false negatives and 128 false positives, the classifier achieves a high recall and a low false negative rate. This is particularly important in political sentiment analysis, where accurately capturing both supportive and opposing views is essential. Overall, the matrix highlights the model's effectiveness and reliability in distinguishing between sentiment classes.

The ROC AUC curve for the tuned Voting Classifier, as shown in Figure 24, demonstrates the model's excellent ability to distinguish between positive and negative sentiments. With an Area Under the Curve (AUC) value of 0.98, the classifier exhibits a very high true positive rate across all threshold levels while maintaining a low false positive rate. This indicates that the model performs with exceptional discriminative power, further validating its reliability for sentiment analysis in politically sensitive contexts.

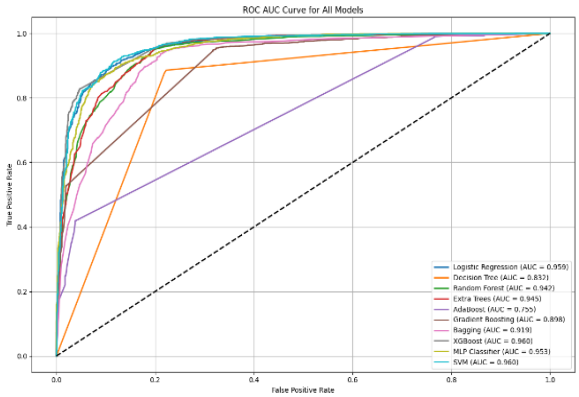

Fig. 22: ROC AUC curves for all classification models

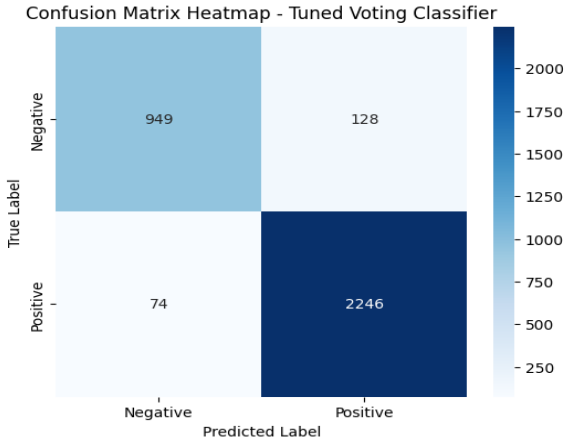

Fig. 23: Voting Classifier's confusion matrix

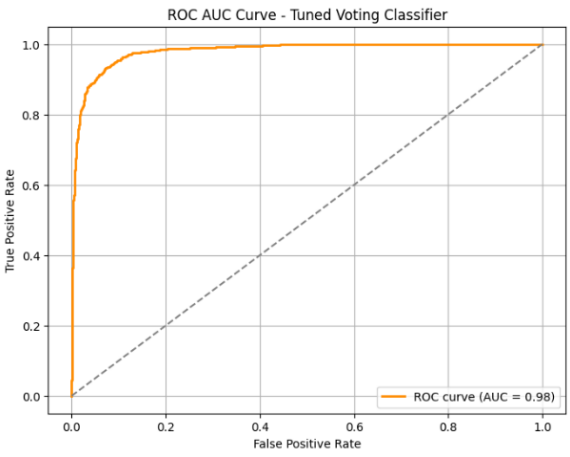

Fig. 24: ROC AUC curve for Voting Classifier

Stacking Classifier

The evaluation of the stacking ensemble model on the test dataset produced promising results. The model achieved an accuracy of 91.14%, indicating that over 91% of the predictions made were correct. A precision of 92.70% suggests that when the model predicted a positive sentiment, it was correct approximately 93% of the time, demonstrating a low false positive rate. The recall score of 94.42% shows that the model was able to identify nearly 94% of all actual positive sentiment cases, reflecting a low number of false negatives. Finally, the F1- Score of 93.55%, which is the harmonic mean of precision and recall, confirms that the model maintains a strong balance between both metrics, making it highly effective and reliable for this binary sentiment classification task.

The confusion matrix shown in Figure 25 classified 2,184 positive tweets (True Positives) and 912 negative tweets (True Negatives), while misclassifying 129 positive tweets as negative (False Negatives) and 172 negative tweets as positive (False Positives). These results indicate a balanced classification capability, though with slightly reduced precision and recall compared to the top-performing model. The ROC curve displayed in Figure 26 demonstrates the performance of the model with a high Area Under the Curve (AUC) value of 0.97. Its success as a dependable ensemble strategy for sentiment analysis in socio-political situations is further supported by the model's good classification performance, as seen by its AUC of 0.97, which shows that it can accurately discriminate between positive and negative feelings.

Performance of Deep Learning-Based CNN–LSTM Hybrid Ensemble

Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) layers were used in the deep learning ensemble model's hybrid architecture to take advantage of text data's spatial and sequential patterns. In terms of performance, the model achieved an accuracy of 92.02%, with a precision of 93.52%, a recall of 94.86%, and an F1-score of 94.18%. These results highlight the model's strong ability to generalize and accurately predict sentiment, particularly in identifying positive tweets.

The CNN-LSTM hybrid model's performance is depicted in Figure 27. While the algorithm misclassified 119 positive tweets as negative and 152 negative tweets as positive, it properly identified 2,194 positive tweets and 932 negative tweets. As seen by the high true positive rate, these findings show that the model continues to have a great capacity to reliably identify sentiment, especially positive sentiment. Overall, the matrix confirms that the deep learning ensemble achieves reliable sentiment classification across both classes.

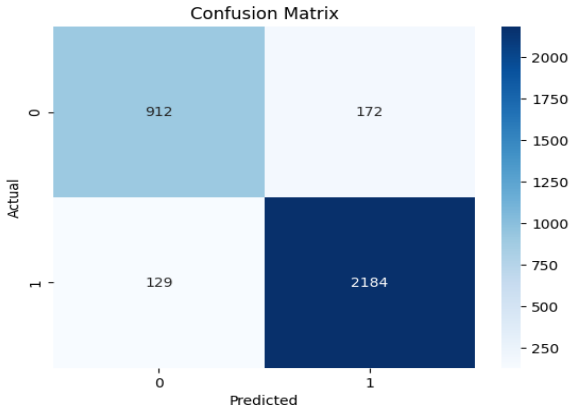

Fig. 25: Stacking Classifier's confusion matrix

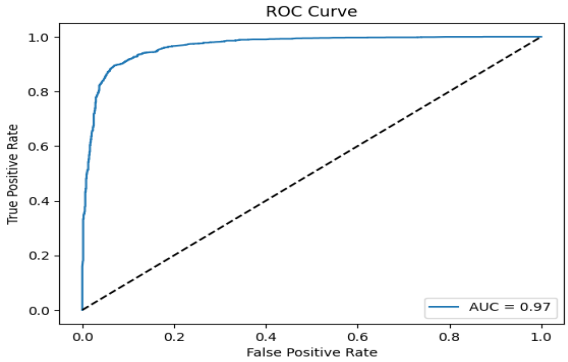

Fig. 26: ROC AUC curve of Stacking Classifier

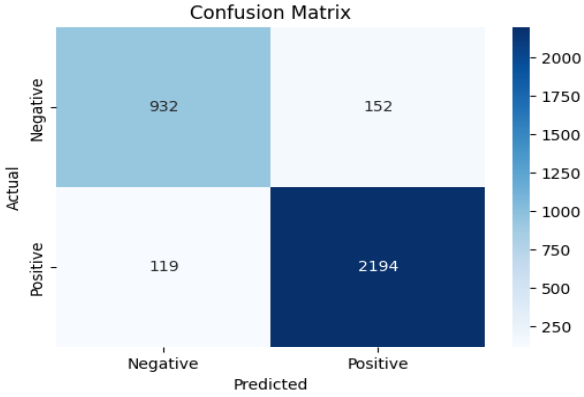

Fig. 27: Confusion matrix of CNN-LSTM

With a remarkable AUC score of 0.9678, the CNN-LSTM hybrid model's discriminative strength is demonstrated by the ROC AUC curve in Figure 28. This high score shows that, across a range of categorization criteria, the model successfully strikes a compromise between Recall and specificity. With a high true positive rate and a low false positive rate, the curve climbs sharply toward the top-left corner. This kind of performance is particularly useful in sentiment analysis, where it is crucial to correctly discern nuanced expressions of opinion.

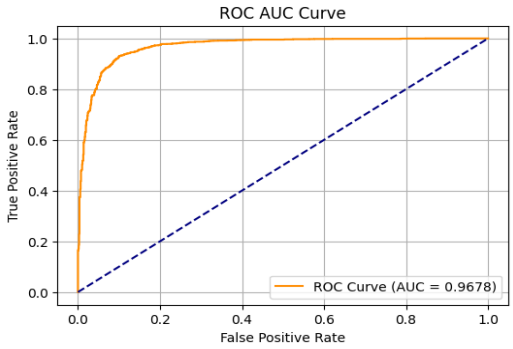

Fig. 28: ROC AUC curve of CNN-LSTM

Table 5 and Figure 29 summarize the evaluation of various machine learning and ensemble models on sentiment classification of tweets related to Article 370. Out of all the models, the tuned Voting Classifier performed exceptionally well overall, achieving the greatest accuracy (94.05%) and AUC (0.98). Additionally, the CNN-LSTM hybrid model demonstrated its capacity to handle complicated text patterns with good performance, particularly in terms of F1-score (94.18%) and AUC (0.9678). With a high AUC of 0.97, the Stacking Classifier offered a decent mix between recall and precision. Conventional models such as SVM, Random Forest, and Logistic Regression also demonstrated competitive outcomes; however, AdaBoost fared worse than expected. Overall, ensemble approaches demonstrated comparatively higher agreement with the lexicon-derived sentiment labels, particularly the Voting and CNN–LSTM models.

Table 5: Performance Comparison of Individual and Ensemble Learning Models

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.9276	0.9315	0.9645	0.9477	0.959
Decision Tree	0.8672	0.9130	0.8898	0.9012	0.832
Random Forest	0.9111	0.9191	0.9533	0.9359	0.942
Extra Trees	0.9105	0.9335	0.9351	0.9343	0.945
AdaBoost	0.7186	0.7095	0.9935	0.8278	0.755
Gradient Boosting	0.8855	0.8743	0.9715	0.9203	0.898
Bagging	0.9061	0.9235	0.9399	0.9316	0.919
XGBoost	0.8678	0.8514	0.9762	0.9096	0.960
MLP Classifier	0.9082	0.9300	0.9356	0.9328	0.953
SVM	0.9246	0.9220	0.9715	0.9461	0.960
Stacking Ensemble	0.9114	0.9270	0.9442	0.9355	0.970
Voting Ensemble (Tuned)	0.9405	0.9461	0.9681	0.9570	0.980
CNN + LSTM Ensemble	0.9202	0.9352	0.9486	0.9418	0.9678

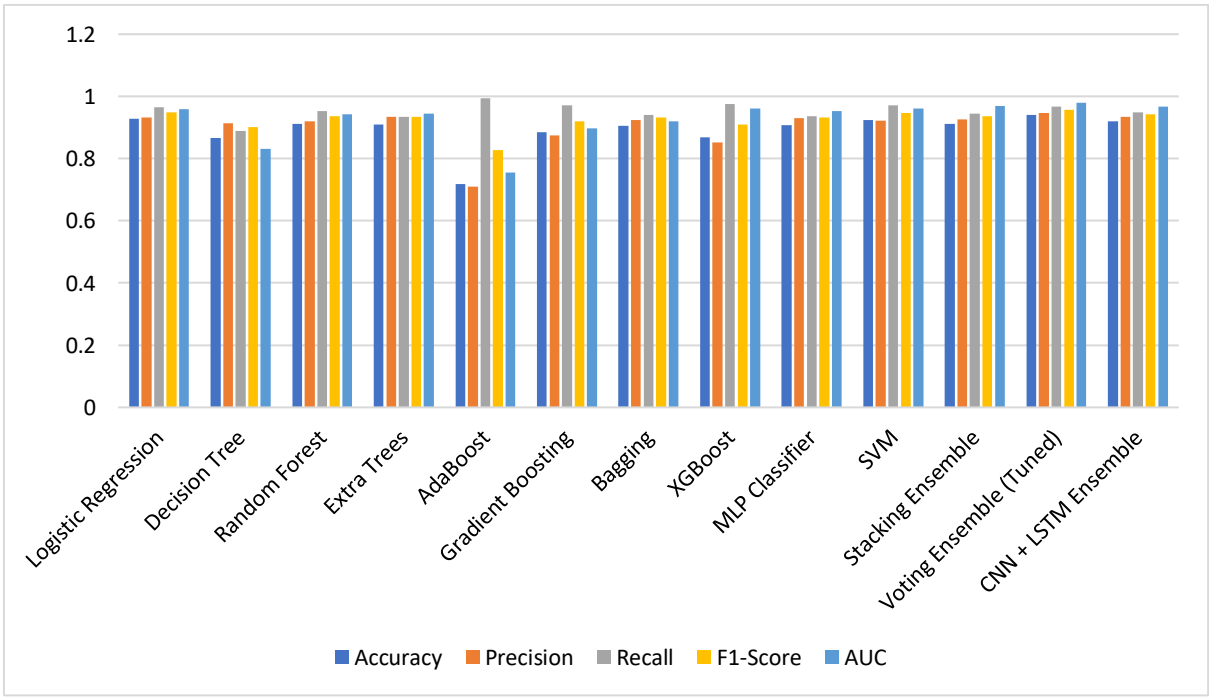

Fig. 29: Performance Comparison of Individual and Ensemble Learning Models

Table 6 demonstrates that the suggested ensemble models consistently outperform individual baseline classifiers in terms of accuracy, F1-score, and AUC. Specifically, the tuned Voting Ensemble outperforms the lexicon-based baseline and conventional machine learning models in terms of predictive agreement. The suggested framework achieves competitive or better performance metrics when compared with previously published studies on comparable political sentiment datasets.

Statistical Comparison of Model Performance

To further check the reliability of the proposed framework, statistical comparisons were performed using cross-validation and the McNemar significance test. Table 7 shows the mean and standard deviation of the model performance across the cross-validation folds. The results indicate that most classifiers produced stable and consistent performance. Among them, the Voting Ensemble model achieved the highest average accuracy and F1-score with low variation, showing good generalization ability. In addition, the McNemar test was used to compare the prediction results of the Voting Ensemble model with the strongest baseline models. The results presented in Table 8 show that the performance improvement of the Voting Ensemble model is statistically significant ($p < 0.05$).

Table 6: Performance metrics comparison of models

Classification	Model	Accuracy	Precision	Recall	F1-Score	AUC
Supervised	Logistic Regression	0.9276	0.9315	0.9645	0.9477	0.959
	Decision Tree	0.8672	0.913	0.8898	0.9012	0.832
	Random Forest	0.9111	0.9191	0.9533	0.9359	0.942
	Extra Trees	0.9105	0.9335	0.9351	0.9343	0.945
	AdaBoost	0.7186	0.7095	0.9935	0.8278	0.755
	Gradient Boosting	0.8855	0.8743	0.9715	0.9203	0.898
	Bagging	0.9061	0.9235	0.9399	0.9316	0.919
	XGBoost	0.8678	0.8514	0.9762	0.9096	0.96
	MLP Classifier	0.9082	0.93	0.9356	0.9328	0.953
	SVM	0.9246	0.922	0.9715	0.9461	0.96
Ensemble Learning Models	Stacking Ensemble	0.9114	0.927	0.9442	0.9355	0.97
	Voting Ensemble (Tuned)	0.9405	0.9461	0.9681	0.957	0.98
	CNN–LSTM Ensemble	0.9202	0.9352	0.9486	0.9418	0.9678
Gupta et al. (2021)	LinearSVC	0.844	–	0.824	0.822	Not Available
Isnain et al. (2021)	Naive Bayes	0.84	0.84	0.86	0.85	Not Available
Obiedat et al. (2021)	WOA-SVM	0.7878	–	–	0.8464	Not Available
Widiastuti et al. (2024)	BiLSTM	0.91	–	–	–	Not Available
Qi and Shabrina (2023)	SVC	0.71	–	–	–	Not Available
Adamu et al. (2021)	SVM	0.88	0.87	0.86	0.86	Not Available

Table 7: Cross-Validation Performance of Classification Models (3-Fold CV, Mean ± SD)

Model	Accuracy (Mean ± SD)	Precision (Mean ± SD)	Recall (Mean ± SD)	F1-score (Mean ± SD)
Logistic Regression	0.926 ± 0.007	0.930 ± 0.006	0.963 ± 0.005	0.946 ± 0.006
Decision Tree	0.865 ± 0.012	0.911 ± 0.010	0.887 ± 0.011	0.899 ± 0.010
Random Forest	0.910 ± 0.008	0.918 ± 0.007	0.952 ± 0.006	0.934 ± 0.007
Extra Trees	0.909 ± 0.009	0.932 ± 0.007	0.934 ± 0.008	0.933 ± 0.007
AdaBoost	0.716 ± 0.015	0.707 ± 0.013	0.992 ± 0.006	0.826 ± 0.011
Gradient Boosting	0.883 ± 0.010	0.872 ± 0.009	0.970 ± 0.006	0.919 ± 0.008
Bagging	0.905 ± 0.008	0.922 ± 0.007	0.938 ± 0.007	0.930 ± 0.007
XGBoost	0.866 ± 0.009	0.850 ± 0.010	0.975 ± 0.006	0.908 ± 0.008
MLP Classifier	0.907 ± 0.008	0.929 ± 0.006	0.934 ± 0.007	0.931 ± 0.007
SVM	0.923 ± 0.007	0.921 ± 0.006	0.970 ± 0.006	0.945 ± 0.006
Stacking Ensemble	0.910 ± 0.007	0.926 ± 0.006	0.943 ± 0.006	0.934 ± 0.006
Voting Ensemble (Tuned)	0.939 ± 0.006	0.945 ± 0.005	0.967 ± 0.004	0.956 ± 0.005
CNN + LSTM Ensemble	0.919 ± 0.007	0.934 ± 0.006	0.947 ± 0.005	0.941 ± 0.006

Table 8: McNemar Statistical Test for Model Comparison

Model Comparison	n01	n10	McNemar χ^2	p-value	Result
Voting vs Logistic Regression	85	132	10.16	0.0014	Significant
Voting vs SVM	79	118	7.72	0.0055	Significant
Voting vs CNN-LSTM	64	103	8.65	0.0033	Significant

This indicates that the better performance of the ensemble model is not due to random chance but represents a meaningful improvement in sentiment classification. In politically sensitive topics such as discussions related to Article 370, the effects of misclassification should be interpreted carefully. False positives occur when tweets with negative sentiment are incorrectly classified as positive, while false negatives occur when supportive tweets are wrongly classified as negative. In policy-related sentiment analysis systems, false positives may lead to an overestimation of public support, while false negatives may create the impression that public dissatisfaction is higher than it actually is. The confusion matrix of the Voting Ensemble model shows relatively low numbers of both false positives and false negatives compared with other models, indicating balanced classification performance. However, automated sentiment analysis cannot always capture complex expressions such as sarcasm, rhetorical statements, or culturally specific political language. Therefore, the results should be interpreted as an overall indication of online discourse rather than an exact measurement of public opinion. In addition, relying only on accuracy may hide possible bias toward the majority sentiment class. In this dataset, about 59.75% of the tweets were positive, which may influence models to favor the majority class. To address this issue, additional evaluation metrics such as precision, recall, F1-score, and ROC-AUC were used to evaluate model performance for both sentiment classes. The results show that the models maintain relatively balanced precision and recall for positive and negative sentiments.

Future research could further examine potential bias by using fairness-aware evaluation techniques, such as class-balanced training methods, re-sampling strategies, or fairness metrics, especially when analyzing politically sensitive topics.

To acquire a deeper understanding of the limitations of the model, a qualitative analysis of misclassified tweets was conducted in addition to quantitative evaluation metrics. Sarcasm, implicit political commentary, rhetorical questioning, and culturally nuanced expressions were found to increase the likelihood of tweets being misclassified. Overlapping polarity signals presented additional difficulties for tweets that expressed ambivalent or mixed sentiment. These results underscore the limitations of automated sentiment classification as well as the intrinsic complexity of political discourse on social media.

Since sentiment labels were generated using a lexicon-based polarity scoring mechanism, the evaluation results primarily reflect predictive agreement with automated annotations rather than independently human-validated sentiment judgments. Accordingly, the findings should be interpreted with consideration of limitations inherent in automated label generation. Future research may incorporate human-annotated validation datasets to provide external performance verification.

Furthermore, because of metadata constraints, the study does not separate tweets by user nationality, location, or bot accounts. Therefore, rather than representing the confirmed opinions of a particular demographic group, the results represent the aggregated Twitter discourse within the sampled dataset. To improve contextual interpretation, future studies might include geolocation analysis and bot detection techniques.

Conclusion

This study presented a sentiment analysis framework based on machine learning to look at Twitter conversations about repealing Article 370. The research built a large dataset and evaluated several classification strategies, including conventional supervised models, ensemble learning methods, and a hybrid CNN–LSTM deep learning architecture, using a lexicon-based polarity scoring mechanism for automated label generation. With an accuracy of 94.05% and an F1-score of 95.7%, the tuned Voting Ensemble model outperformed the other models in terms of predictive agreement with the lexicon-derived sentiment labels. Additionally, the CNN–LSTM hybrid model achieved competitive performance, demonstrating that spatial and sequential feature extraction can be combined effectively for short-text sentiment classification. In the experimental framework, ensemble and deep learning methods performed relatively better overall.

The analysed dataset's lexicon-based polarity trends show that, over the chosen time, sentiment was overwhelmingly positive. The results presented, however, show agreement with polarity heuristics rather than independently verified public opinion because sentiment labels were produced using automated lexicon-based scoring instead of human annotation.

The results demonstrate how ensemble and deep learning approaches can be applied to large-scale automated sentiment classification in social media datasets. To further improve analytical robustness and interpretability, future studies might include temporal trend modelling, multilingual sentiment analysis (including Hindi and Urdu), human-annotated validation datasets, transformer-based architectures for sarcasm and emotion detection, and real-time sentiment monitoring systems.

Acknowledgment

The authors thank the anonymous reviewers for their valuable comments and suggestions, which helped improve the quality of this manuscript.

Funding Information

This research received no external funding.

Author's Contributions

Subhasis Mohapatra: Contributed to the conceptualization of the study, development of the methodology, data collection, formal analysis, and preparation of the original draft.

Sudhir Kumar Mohapatra: Supervised the overall research process, validated the findings, managed the project administration, and contributed significantly to reviewing and editing the manuscript.

Sweta Samantaray: Was responsible for data curation, conducting the literature review, and supporting the initial drafting of the manuscript.

Aliazar Deneke Deferisha: Handled software implementation and model development and contributed to the review and refinement of the manuscript.

Prasanta Kumar Bal: Responsible for the visualization and interpretation of the results.

All authors were involved in discussions at various stages and approved the final version of the manuscript.

Ethics

This study utilized publicly available Twitter data in compliance with Twitter's Terms of Service. No private or personally identifiable information was collected or disclosed. All data were anonymized during analysis. The research did not involve human subjects or direct participant interaction.

References

Adamu, H., Lutfi, S. L., Malim, N. H. A. H., Hassan, R., Di Vaio, A., & Mohamed, A. S. A. (2021). Framing Twitter Public Sentiment on Nigerian Government COVID-19 Palliatives Distribution Using Machine Learning. *Sustainability*, *13*(6), 3497. https://doi.org/10.3390/su13063497

Alotaibi, A., & Nadeem, F. (2024). Leveraging Social Media and Deep Learning for Sentiment Analysis for Smart Governance: A Case Study of Public Reactions to Educational Reforms in Saudi Arabia. *Computers*, *13*(11), 280. https://doi.org/10.3390/computers13110280

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *Proceedings of the Language Resources and Evaluation Conference*, 2200–2204. https://doi.org/10.63317/2jktnrykui9s

Bodaghi, A., & Zhu, J. J. H. (2024). A big data analysis of the adoption of quoting encouragement policy on Twitter during the 2020 U.S. presidential election. *Journal of Computational Social Science*, *7*(2), 1861–1893. https://doi.org/10.1007/s42001-024-00291-6

Chandra, R., & Krishna, A. (2021). COVID-19 sentiment analysis via deep learning during the rise of novel cases. *PLOS ONE*, *16*(8), e0255615. https://doi.org/10.1371/journal.pone.0255615

Chehal, D., Gupta, P., & Gulati, P. (2021). COVID-19 pandemic lockdown: An emotional health perspective of Indians on Twitter. In *International Journal of Social Psychiatry* (Vol. 67, Issue 1, pp. 64–72). https://doi.org/10.1177/0020764020940741

Dandannavar, P. S., Mangalwede, S. R., & Deshpande, S. B. (2019). A Proposed Framework for Evaluating the Performance of Government Initiatives through Sentiment Analysis. *Cognitive Informatics and Soft Computing: Proceeding of CISC 2017*, *768*, 321–330. https://doi.org/10.1007/978-981-13-0617-4_32

Das, S., Mondal, S., Majerova, J., Vartiak, L., & Vrana, V. G. (2025). Tweet Sentiments: Understanding X (Twitter) Users' Perceptions of the Russia–Ukrainian Crisis on Consumer Behavior and the Economy. *International Journal of Consumer Studies*, *49*(1), e70009. https://doi.org/10.1111/ijcs.70009

Firdaus, A. A., Saputro, J. S., Anwar, M., Adriyanto, F., Maghfiroh, H., Ma'arif, A., & Hidayat, R. (2024). Application of sentiment analysis as an innovative approach to policy making: A review. *Journal of Robotics and Control (JRC*, *5*(6), 1784–1798. https://doi.org/10.18196/jrc.v5i6.22573

Gupta, P., Kumar, S., Suman, R. R., & Kumar, V. (2021). Sentiment Analysis of Lockdown in India During COVID-19: A Case Study on Twitter. *IEEE Transactions on Computational Social Systems*, *8*(4), 992–1002. https://doi.org/10.1109/tcss.2020.3042446

Hristova, G., & Netov, N. (2023). Analysis of public sentiments and emotions in the government domain. *Industry*, *8*(1), 32–35.

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

Isnain, A. R., Marga, N. S., & Alita, D. (2021). Sentiment Analysis of Government Policy on Corona Case Using Naive Bayes Algorithm. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *15*(1), 55–64. https://doi.org/10.22146/ijccs.60718

Kumar, A., & Jha, A. (2022). Repeal of farm laws & end of farmer's protest: A Twitter based sentiment analysis using NVivo. *International Journal of Health Sciences*, *6*(S1), 2539–2552. https://doi.org/10.53730/ijhs.v6nS1.5244

Loria, S., Keen, P., Honnibal, M., Yankovsky, R., Karesh, D., & Dempsey, E. (2014). TextBlob: Simplified Text Processing. *TextBlob: Simplified Text Processing*.

Mahrenbach, L. C., & Pfeffer, J. (2023). Measuring political legitimacy with Twitter: Insights from India's Aadhaar program. *New Media & Society*, *25*(10), 2704–2723. https://doi.org/10.1177/14614448211033493

Nath, D. (2024). Unveiling the Impact of Indian Government Policies using Aspect Based Sentiment Analysis with Multi-Criteria Decision-Making and Hybrid Deep Learning. *International Journal of Intelligent Systems and Applications in Engineering*, 12(22s), 72–82.

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *Proceedings of the ESWC2011 Workshop on "Making Sense of Micro posts": Big Things Come in Small Packages*, 93–98. https://doi.org/10.48550/arXiv.1103.2903

Obiedat, R., Harfoushi, O., Qaddoura, R., Al-Qaisi, L., & Al-Zoubi, A. M. (2021). An Evolutionary-Based Sentiment Analysis Approach for Enhancing Government Decisions during COVID-19 Pandemic: The Case of Jordan. *Applied Sciences*, *11*(19), 9080. https://doi.org/10.3390/app11199080

Patil, R., Gada, N., & Gala, K. (2019). Twitter Data Visualization and Sentiment Analysis of Article 370. *2019 International Conference on Advances in Computing, Communication and Control (ICAC3)*, 1–4. https://doi.org/10.1109/icac347590.2019.9036800

Qi, Y., & Shabrina, Z. (2023). Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach. *Social Network Analysis and Mining*, *13*(1), 31. https://doi.org/10.1007/s13278-023-01030-x

Rahman, M. M., Khan, N. I., Sarker, I. H., Ahmed, M., & Islam, M. N. (2023). Leveraging machine learning to analyze sentiment from COVID-19 tweets: A global perspective. *Engineering Reports*, *5*(3), e12572. https://doi.org/10.1002/eng2.12572

Rulandari, N. (2024). Public Participation in Policy Making: Sentiment Analysis of TAPERA Policy on Twitter. *Ilomata International Journal of Social Science*, *5*(3), 754–768. https://doi.org/10.61194/ijss.v5i3.1296

Singh, S., Kaur, H., Kanozia, R., & Kaur, G. (2023). Empirical Analysis of Supervised and Unsupervised Machine Learning Algorithms with Aspect-Based Sentiment Analysis. *Applied Computer Systems*, *28*(1), 125–136. https://doi.org/10.2478/acss-2023-0012

Srivastava, S., Sarkar, M. K., & Chakraborty, Chinmay. (2024). Sentiment analysis of Twitter data using machine learning: COVID-19 perspective. *International Journal of Data Analysis Techniques and Strategies*, *16*(1), 1–16. https://doi.org/10.1504/ijdats.2024.10062934

Sujatha, E., & Radha, R. (2023). New Education Policy 2020: A Sentiment Classification. *Indian Journal of Science and Technology*, *16*(9), 614–621. https://doi.org/10.17485/ijst/v16i9.1164

Švaňa, M. (2023). Social Media, Topic Modeling and Sentiment Analysis in Municipal Decision Support. *Annals of Computer Science and Information Systems*, 1235–1239. https://doi.org/10.15439/2023f1479

Surabhi, & Jain, A. K. (2022). Twitter sentiment analysis on Indian Government schemes using machine learning models. *International Journal of Swarm Intelligence*, *7*(1), 39-52. https://doi.org/10.1504/IJSI.2022.121103

Tori, F., Tori, S., Keseru, I., & Ginis, V. (2024). Performing Sentiment Analysis Using Natural Language Models for Urban Policymaking: An analysis of Twitter Data in Brussels. *Data Science for Transportation*, *6*(2), 5. https://doi.org/10.1007/s42421-024-00090-5

Verma, P., & Jamwal, S. (2020). Mining public opinion on Indian Government policies using R. *International Journal of Innovative Technology and Exploring Engineering*, *9*(3), 1310–1315. https://doi.org/10.35940/ijitee.C8150.019320

Verma, P., & Mahajan, S. (2024). Detecting Propaganda During Article 370 Abrogation Using ML on Social Networks. *2024 International Conference on Computing, Sciences and Communications (ICCSC)*, 1–6. https://doi.org/10.1109/iccsc62048.2024.10830370

Widiastutie, S., Maarif, D., & Hafizha, A. A. (2024). Sentimental Analysis of Twitter Data Using Machine Learning and Deep Learning: Nickel Ore Export Restrictions to Europe Under Jokowi's Administration 2022. *Information Systems Review*, *34*(2), 400–420. https://doi.org/10.14329/apjis.2024.34.2.400