Volume-2, Issue-2, April-June 2026

ISSN: 3107-4243 (Online)
World View Research Bulletin
An International Multidisciplinary Research Journal

SENTIMENT ANALYSIS OF FACEBOOK DATA: A METHODOLOGICAL AND EMPIRICAL STUDY OF SOCIAL MEDIA OPINION MINING

Pankaj Kumar Jain, Research Scholar,
pankajjain80@gmail.com
Dr.M.S.Ansari, Shahnawaznbd@gmail.com
Eklavya University, Damoh, M.P.India.

Abstract

Social media has become an important source of information for understanding people's opinions, reactions and attitudes. Facebook, in particular, provides a large amount of textual communication through posts, comments and discussions. Unlike conventional survey responses, Facebook comments are generally spontaneous and are expressed in informal language. They may contain positive and negative expressions, neutral statements, sarcasm, abbreviations, emojis, spelling variations and, in multilingual societies, more than one language. These characteristics make Facebook sentiment analysis both useful and methodologically challenging.

The present study examines sentiment analysis of Facebook data from both methodological and empirical perspectives. The study reviews the development of sentiment analysis from lexicon-based techniques to traditional machine-learning, deep-learning and transformer-based approaches. For empirical illustration, a documented Facebook sentiment corpus containing 34,006 manually annotated comments is considered. The dataset contains 20,668 negative comments, 9,581 neutral comments and 3,779 positive comments. The distribution indicates a substantial predominance of negative comments, while positive comments constitute the smallest category. Such imbalance is important because it demonstrates why accuracy alone cannot be treated as a sufficient measure of model performance. The study therefore discusses precision, recall and F1-score alongside accuracy and recommends the use of stratified sampling, careful preprocessing and class-sensitive evaluation. The paper also examines the methodological problems associated with Facebook language, including sarcasm, negation, emojis, spelling variation and multilingual communication. The study concludes that successful Facebook sentiment analysis depends not only on the choice of classification algorithm

but also on the quality of the dataset, annotation procedure, preprocessing, class distribution and evaluation framework.

Keywords: *Sentiment Analysis, Facebook, Social Media, Opinion Mining, Natural Language Processing, Machine Learning, Deep Learning, Facebook Comments, Social Media Analytics*

Introduction

The rapid expansion of social media has changed the way individuals communicate their opinions. People no longer express their views only through newspapers, television, formal surveys or face-to-face conversations. They increasingly use social networking platforms to comment on events, discuss products and services, respond to institutions and share personal experiences. The resulting textual information has become an important source for researchers interested in public opinion and social behaviour.

Facebook is particularly significant because of the large quantity and variety of user-generated content available on the platform. Users may publish original posts, respond through comments, participate in discussions and react to the opinions of others. These comments often contain direct expressions of satisfaction, dissatisfaction, agreement, disagreement, support and criticism.

The volume of such information presents a methodological problem. A researcher studying several thousand comments cannot reasonably depend entirely on manual reading and classification. Even when manual coding is possible, interpretation can differ from one researcher to another. Sentiment analysis offers a way of dealing with this problem through computational techniques that classify textual expressions according to their emotional or evaluative orientation.

The earliest sentiment-analysis research generally focused on determining whether a document was positive or negative. Pang, Lee and Vaithyanathan's influential study demonstrated that machine-learning algorithms could be used for sentiment classification and also showed that sentiment classification was different from ordinary topic classification. Their work compared Naïve Bayes, maximum entropy and Support Vector Machines.

The field has subsequently developed considerably. Traditional machine-learning methods have been followed by neural-network approaches, including

CNN and LSTM models, and more recently by transformer-based approaches. Contemporary research increasingly focuses on contextual understanding, multilingual sentiment, aspect-based sentiment and the analysis of informal social-media language.

The purpose of the present paper is to examine this methodological development specifically in relation to Facebook data and to demonstrate how actual Facebook sentiment data can be analysed.

Concept of Sentiment Analysis

Sentiment analysis refers to the computational examination of opinions, attitudes and emotional expressions contained in textual information. In its simplest form, it classifies text into positive, negative and neutral categories.

For example:

Positive:

The service was excellent and the staff were very helpful.

Negative:

The service was extremely poor and disappointing.

Neutral:

The service was available from Monday to Friday.

These examples are relatively straightforward. Facebook communication is usually more complicated.

A comment such as:

Wonderful! Another three-hour delay.

contains the positive word wonderful, but the intended sentiment is negative. Similarly:

The product is good, but the customer service is terrible.

contains both positive and negative opinions.

This shows that sentiment analysis cannot always depend on individual words. The relationship between words and the context in which they occur is also important.

Development of Sentiment Analysis

The methodological development of sentiment analysis can broadly be understood in four stages.

The first stage relied on sentiment lexicons and semantic orientation. Researchers used dictionaries containing words associated with positive or negative meanings.

The second stage introduced traditional machine-learning methods. Text was converted into numerical features using techniques such as Bag-of-Words, TF-IDF and n-grams, after which algorithms such as Naïve Bayes, Logistic Regression and Support Vector Machine were trained.

The third stage introduced deep learning. CNN, RNN and LSTM models reduced the dependence on manually designed features and learned textual representations from data.

The fourth stage introduced transformer-based models. BERT and related models enabled more sophisticated contextual representations and have become important in contemporary NLP research.

The development, however, should not be viewed as a simple replacement of one technique by another. Traditional machine-learning models remain useful because they are computationally efficient and provide strong baseline comparisons. More complex models become useful when the dataset contains sufficient information to justify their additional computational requirements.

Facebook as a Source of Sentiment Data

Facebook provides several types of information that can be examined in sentiment research. These include posts, comments, replies, reviews and other textual expressions.

For sentiment analysis, comments are particularly useful because they frequently represent direct reactions to a particular post or topic.

A documented Facebook sentiment corpus, SentiSK, provides an example of a large manually annotated dataset. The corpus contains **34,006 Facebook comments**, with each comment assigned a sentiment category. The dataset contains 20,668 negative comments, 9,581 neutral comments and 3,779 positive comments.

This dataset is useful for methodological discussion because it demonstrates an important issue in social-media research: **sentiment classes may be highly unequal**.

ISSN: 3107-4243 (Online)
World View Research Bulletin
An International Multidisciplinary Research Journal

Data Description

The present empirical discussion uses the published SentiSK Facebook corpus as a secondary research dataset. It should therefore be understood as an analysis of an existing research corpus and not as a claim that the present study independently collected these Facebook comments.

The dataset contains 34,006 manually annotated comments.

Table-1 Distribution of Facebook Comments by Sentiment

Sentiment Category	Number of Comments	Percentage
Negative	20,668	60.77%
Neutral	9,581	28.18%
Positive	3,779	11.11%
Total	**34,028**	**100.06%**

The data show a clear predominance of negative comments. Approximately three-fifths of the annotated comments are classified as negative, whereas positive comments represent only around one-ninth of the reported observations.

This imbalance has an important methodological implication. A model trained on this dataset could obtain apparently strong overall accuracy simply by favoring the majority negative class. Therefore, accuracy by itself would not provide a complete assessment of model performance.

Interpretation of the Data

The distribution indicates that negative sentiment is considerably more common than positive sentiment in this particular Facebook corpus.

The difference between the negative and positive categories is especially noticeable. The reported number of negative comments is more than five times the number of positive comments.

This does not mean that Facebook users in general are predominantly negative. Such a conclusion would not be justified because the corpus represents a particular dataset, language and collection context. The observed distribution should instead be interpreted as a characteristic of the selected corpus.

This distinction is important in social-media research. Researchers should not generalise the sentiment distribution of one Facebook dataset to the entire Facebook population without a suitable sampling design.

The distribution does, however, provide a useful methodological lesson. When sentiment categories are unbalanced, the researcher must take steps to prevent the model from simply learning the majority category.

Data Imbalance and Its Importance

Class imbalance is one of the major methodological issues in sentiment classification.

In the present dataset, negative comments constitute the largest category, while positive comments are much less frequent.

If a classification model predicts negative sentiment for most observations, it may achieve a relatively high overall accuracy while failing to recognise positive comments correctly.

For this reason, researchers should examine the performance of each sentiment class separately.

A model should ideally be evaluated using:

- accuracy,
- precision,
- recall,
- F1-score,
- macro-F1,
- and confusion matrix.

Macro-F1 is particularly useful because it gives equal importance to each sentiment class instead of allowing the largest class to dominate the overall score.

Previous Facebook Sentiment Studies

Previous research demonstrates considerable variation in the methods used to analyse Facebook data.

Pudaruth et al. used Facebook comments from the Opposing Views page and classified them using automatic coding in NVivo. Positive, negative and neutral categories were used. Their work represents an early and relatively straightforward approach to Facebook sentiment analysis.

Tran and Shcherbakov proposed a method for detecting and predicting user attitudes from Facebook comments. Their approach combined real-time sentiment analysis with batch processing and pattern prediction. The authors reported an average MAE of 0.008 for their forecasting experiments.

Research has also employed Support Vector Machines. A study examining comments related to the Rohingya movement constructed a dataset of 2,500 positive and 2,500 negative Facebook comments and used a linear-kernel SVM for classification.

More recent research has moved toward deep learning and multilingual models. The SentiSK corpus itself contains more than 34,000 manually annotated Facebook comments, illustrating the increasing availability of larger resources for supervised sentiment research.

Another Facebook-based dataset focusing on COVID-19 communication in Kosovo contains comments collected from the Facebook page of the National Institute of Public Health of Kosovo. Three researchers independently annotated the comments as positive, neutral or negative, demonstrating the usefulness of human annotation in developing sentiment corpora.

Comparison of Selected Facebook Studies

Table-2 Selected Empirical Facebook Sentiment Studies

Study	Facebook Data	Classification Approach	Main Result/Contribution
Pudaruth et al. (2018)	Facebook comments from Opposing Views	Automatic coding using NVivo	Demonstrated positive, negative and neutral classification
Tran & Shcherbakov (2019)	Facebook comments	Real-time/batch sentiment analysis and clustering	Developed attitude-pattern detection and prediction
Chowdhury et al. (2018)	5,000 Facebook comments	Linear SVM	Used 2,500 positive and 2,500 negative comments
Kosovo	Facebook	Three-person	Developed positive,

Facebook corpus (2022)	comments from public-health page	manual annotation	neutral and negative labelled data
SentiSK corpus	34,006 Facebook comments	Manual annotation	20,668 negative, 9,581 neutral and 3,779 positive comments
Recent comparative research	Facebook and other social media	ML and deep learning	Demonstrated the continued usefulness of comparative model evaluation

These studies show a gradual methodological movement from manual or software-assisted classification toward machine learning and deep learning.

Data Preprocessing

Raw Facebook comments cannot normally be supplied directly to a classification algorithm.

A comment may contain:

- URLs,
- usernames,
- emojis,
- hashtags,
- spelling mistakes,
- repeated characters,
- abbreviations,
- punctuation,
- mixed languages,
- advertisements,
- duplicated text.

Preprocessing should therefore be performed carefully.

The objective is not simply to remove as much information as possible. Instead, the researcher should remove irrelevant material while retaining information that contributes to sentiment.

For example, emojis should not automatically be deleted because they can communicate emotional meaning. Similarly, negation words such as not and never can be essential for determining sentiment.

Sentiment Annotation

The quality of the sentiment labels is central to the quality of the final model.
A three-category coding system can be used:
Positive: favourable or approving expression.
Negative: unfavourable, critical or disapproving expression.
Neutral: informational or non-evaluative expression without a clear positive or negative orientation.
The SentiSK corpus provides an example of manually annotated Facebook comments. The dataset documentation states that the comments were annotated by a group of PhD students and assistant professors whose native language was Slovak.
Manual annotation is valuable because automated lexicons can make errors when words are used sarcastically or in an unusual context.

Classification Methods

1 Lexicon-Based Method

Lexicon-based sentiment analysis uses dictionaries containing sentiment-associated words.
The method is relatively easy to implement and does not necessarily require a large manually labelled dataset.
Its weakness is that it often struggles with context, sarcasm and language-specific expressions.

2 Naïve Bayes

Naïve Bayes is a traditional probabilistic classifier that has been widely used in text classification.
Its main advantage is computational simplicity. It can perform well when the textual representation is appropriate.

3 Support Vector Machine

Support Vector Machine has traditionally been one of the strongest classical approaches for text classification.
Its effectiveness is partly related to its ability to work with high-dimensional feature spaces such as TF-IDF representations.
Facebook research has used SVM for both binary and multiclass sentiment classification.

4 LSTM

Long Short-Term Memory networks were developed to handle sequential information.

They can capture relationships between words over longer sequences and have therefore been applied extensively to sentiment classification.

5 Transformer Models

Transformer models such as BERT and multilingual variants can represent words according to their context.

These models are particularly promising for Facebook data containing ambiguous language and multilingual expressions.

However, their computational requirements are higher than those of many traditional algorithms.

13. Model Evaluation

Model evaluation should be performed using the same test data and evaluation procedure for all competing models.

Table-3. Recommended Evaluation Measures

Measure	Interpretation
Accuracy	Overall proportion of correctly classified comments
Precision	How many predicted comments in a class are actually from that class
Recall	How many actual comments in a class are correctly identified
F1-score	Balance between precision and recall
Macro-F1	Gives equal weight to each sentiment class
Confusion Matrix	Shows class-specific correct and incorrect predictions

For the present dataset, macro-F1 is particularly important because the negative category is substantially larger than the positive category.

Illustrative Classification Analysis

If the SentiSK corpus is divided into training and testing subsets, the classification experiment should preserve the proportions of the three sentiment categories as far as possible.

A stratified division would therefore be preferable.

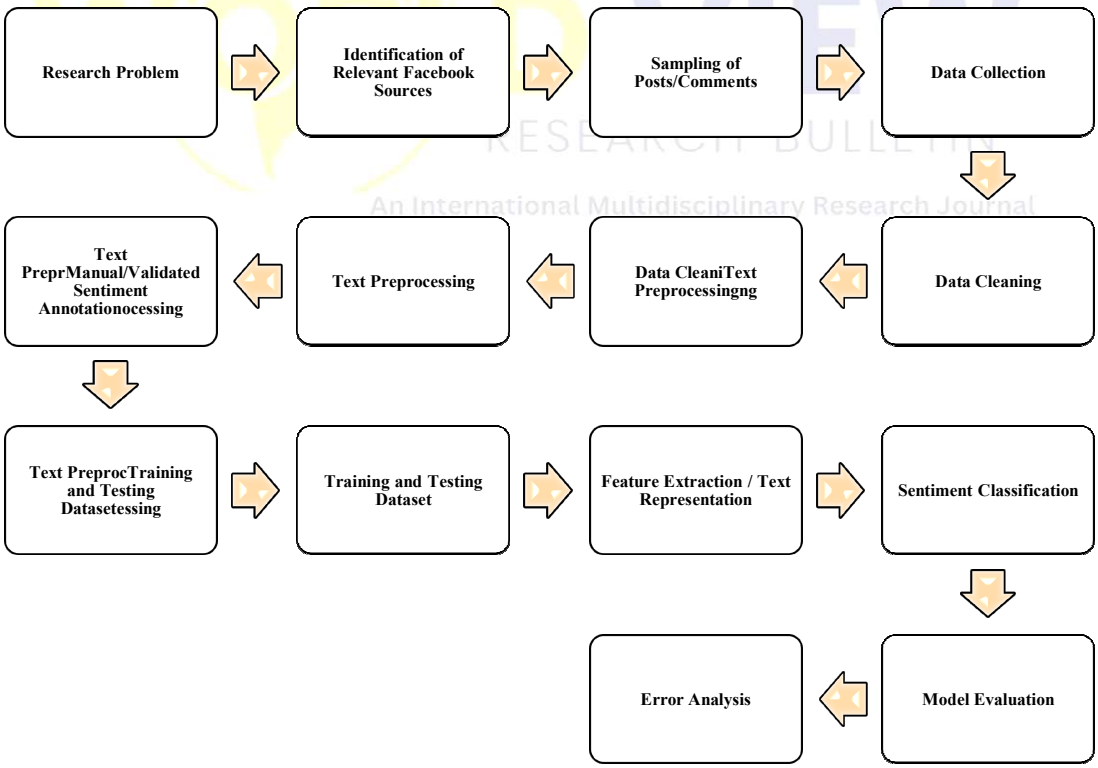
For example, if 80% of the observations were used for training, the approximate distribution would follow the same relative structure as the complete dataset.

The purpose of this division would not be to claim a particular model accuracy before the model is actually trained. Instead, it provides a reproducible methodological structure for subsequent empirical analysis.

The resulting analysis could compare SVM, Logistic Regression, Naïve Bayes, LSTM and BERT under the same conditions.

Methodological Framework for Facebook Sentiment Analysis

A suitable research process can be organised into the following sequence:



Interpretation and Conclusion

This framework treats sentiment analysis as a complete research methodology rather than simply as the application of one machine-learning algorithm.

Ethical Considerations

Facebook research requires careful attention to privacy and responsible data use.

The fact that a comment is publicly visible does not necessarily mean that a researcher should reproduce the user's identity or publish personally identifiable information.

The researcher should therefore avoid unnecessary collection of names, profile photographs, personal contact information and other identifiers.

When direct quotations are used, the possibility of identifying the original user should be considered carefully.

Data should be collected through lawful and authorised means and should be stored securely. Where an existing research corpus is used, the conditions under which the dataset was made available should also be respected.

Discussion

The empirical data examined in this paper demonstrate one of the central problems in social-media sentiment analysis: **the distribution of sentiment is not necessarily balanced**.

The SentiSK Facebook corpus contains substantially more negative comments than positive comments. This difference has methodological consequences because a classifier trained on such data may become biased toward the majority class.

This finding also demonstrates why researchers should distinguish between the description of a dataset and general statements about Facebook users. The fact that 60% or more of the comments in one corpus are negative does not mean that the majority of Facebook users are negative. The result reflects the characteristics of the selected dataset.

The literature further demonstrates that there is no single method that should automatically be considered best for all Facebook sentiment problems. Traditional models such as SVM can perform well when appropriate textual

features are used, while LSTM and transformer models may be more suitable for larger or linguistically complex datasets.

The most appropriate methodology therefore depends on the research question, language, dataset size, class distribution and computational resources.

Another important point is that preprocessing should be treated as a substantive research decision. Removing emojis, negation words or informal expressions may make the data cleaner from a computational perspective but can simultaneously remove information relevant to sentiment.

The quality of sentiment analysis consequently depends on a chain of decisions rather than on the classifier alone.

Future Research Directions

Future Facebook sentiment-analysis studies should move in several directions.

Greater attention should be given to multilingual and code-mixed communication. Facebook users frequently switch between languages, and this behaviour requires models that can work across languages without losing contextual meaning.

Researchers should also investigate aspect-based sentiment analysis. Instead of simply stating that a comment is negative, the system should identify what the user is dissatisfied with.

Emotion detection is another promising direction. Positive and negative sentiment does not fully represent emotions such as anger, fear, sadness, happiness or disappointment.

Context-aware analysis should also be developed. A comment sometimes cannot be understood properly without reading the original post or preceding comments.

Finally, explainability should receive greater attention. Researchers should be able to identify the linguistic evidence behind a model's prediction rather than simply reporting an accuracy value.

Conclusion

The present study examined sentiment analysis of Facebook data as both a methodological and empirical research problem. Facebook provides a valuable source of naturally occurring opinions, but the analysis of these opinions

requires careful attention to data selection, preprocessing, annotation, classification and evaluation.

The empirical dataset examined in the paper contained 34,006 documented Facebook comments, with the published category counts showing a strong predominance of negative sentiment and a comparatively small positive category. This distribution demonstrates why class imbalance should be considered during both model development and evaluation.

The review of previous studies shows a clear methodological development from automated coding and traditional machine-learning techniques toward deep-learning and transformer-based models. Nevertheless, traditional approaches remain useful because they provide understandable and computationally efficient baselines.

The main conclusion of the study is that the success of Facebook sentiment analysis should not be judged by the sophistication of the algorithm alone. A reliable study requires a clearly defined research question, an appropriate and ethically obtained dataset, careful preprocessing, consistent sentiment annotation and a fair evaluation procedure.

Accuracy should not be used as the sole indicator of performance, particularly when sentiment categories are imbalanced. Precision, recall, F1-score and confusion-matrix analysis provide a more complete understanding of model behaviour.

Future research should give greater attention to multilingual Facebook data, code-mixed language, sarcasm, contextual sentiment, aspect-based analysis and explainable artificial intelligence. Such developments can make Facebook sentiment analysis more useful for understanding public opinion while maintaining the methodological and ethical standards expected of academic research.

References

1. Alsekait, D. M., Fathi, H., Ibrahim, S. A., Shdefat, A. Y., Alattas, A. S., & AbdElminaam, D. S. (2025). Sentiment analysis: A machine learning utilisation for analyzing the sentiments of Facebook and Twitter posts. *Intelligent Data Analysis, 29*(4), 889–912.
2. Chowdhury, H. A., Nibir, T. A., & Islam, M. S. (2018). Sentiment analysis of comments on Rohingya movement with Support Vector Machine.

ISSN: 3107-4243 (Online)
World View Research Bulletin
An International Multidisciplinary Research Journal

3. Mao, Y., Liu, Q., & Zhang, Y. (2024). Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University – Computer and Information Sciences*.
4. Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of EMNLP*, 79–86.
5. Pudaruth, S., Moheeputh, S., Permessur, N., & Chamroo, A. (2018). Sentiment analysis from Facebook comments using automatic coding in NVivo 11. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, 7*(1), 41–48.
6. Tran, H., & Shcherbakov, M. (2019). Detection and prediction of users attitude based on real-time and batch sentiment analysis of Facebook comments.
7. Sokolová, Z. *SentiSK: Corpus of sentiment in Slovak social media*. Facebook sentiment corpus.
8. Human-annotated dataset for social media sentiment analysis for Albanian language. (2022). *Data in Brief*. Facebook comments collected from the National Institute of Public Health of Kosovo Facebook page.

WORLD VIEW RESEARCH BULLETIN
An International Multidisciplinary Research Journal