International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5200 -5206
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0680

Feature Based Survey on Fake News Detection: Statistical and Semantic

Ms. Itika U. Lakkewar[1], Dr. R. N. Jugele[2]
[1]Research Scholar, Department of Computer Science, Science College, Nagpur, Maharashtra, 440001, India
[2]Professor, Department of Computer Science, Science College, Nagpur, Maharashtra, 440001, India
Emails: *itika_lakk@yahoo.co.in[1], rn_jugele@yahoo.com[2]*

Abstract
The digital news portals and social media are rapidly expanding, which has significantly increased the spread of fake news, which affects public opinion, social harmony, and trust in information sources. Detection of fake news at an early stage is a critical research challenge. In recent years, researchers have applied multiple methods for the classification of news articles as fake using Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL) techniques. This paper discusses a feature-based analysis of text-oriented fake news detection methods published from 2017 to 2025. The analysis demonstrates how different textual features, such as linguistic, stylistic, psychological, statistical, semantic, and syntactic features, are used for the identification of fake or real news. A comparative analysis of existing studies shows that most research primarily depends on statistical and semantic representations like N-grams, TF-IDF, and word embeddings, whereas linguistic, stylistic, and psychological cues are comparatively less explored. In addition, syntactic features have gained very limited attention despite their potential to enhance detection performance. The review emphasizes integrating multiple feature types to develop more reliable and interpretable detection systems. It also identifies research gaps and suggests future directions for developing comprehensive feature-based frameworks for fake news detection.
Keywords: *Fake News Detection, text-based Analysis, NLP Features, Linguistic Features, Psychological Features, Syntactic features, Feature Engineering*

1. Introduction
In past decades, the use of online content, social media and digital platforms has increased rapidly. News is accessed through digital platforms and is available in different formats, such as text, images, and videos. Digital media and online platforms play a vital role in the creation and distribution of news. Unlike traditional news, online news spreads rapidly because it can be shared instantly through digital platforms. Before sharing news, users often do not verify its authenticity or fact-check whether it is real or fake. Fake news refers to false information designed to deceive people. There are different forms of fake news, such as propaganda, hoaxes, and satire. In today's digital world, fake news significantly impacts society, politics and public trust. Fake political news changes public opinion, while health-related fake news creates panic, violence, anger and fear. This low barrier to content creation allows misleading or fabricated information to spread rapidly across large audiences. The digital media platforms are highly effective for encouraging discussions and debates on topics including democracy, education and health [1]. Therefore, an automated system is required to stop the fabrication of news. The text-based fake news identification remains fundamental due to the widespread use of online news through blogs, tweets, news articles, and social media posts. Most news has textual content such as title, description, headline, captions, or full article. Moreover, early-stage misinformation detection is primarily based on content-based features, as social context and user interaction signals require time to emerge. In many cases, multimedia elements such as images or videos are either absent or insufficient for reliable analysis. Therefore, the fake news detection techniques are more efficient and timely than the multimodal methods, especially in the early intervention scenarios. Researchers have investigated a broad spectrum of techniques such as stylistic and linguistic analysis, extraction of sentiment and psychological features, statistical representations, and semantic modeling with the use

e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5200 -5206
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0680

of machine learning (ML), deep learning (DL) and natural language processing (NLP). There are different types of linguistic features, such as function words, word count and readability index, which help to classify whether the news data sources from various domains are real or fake. Stylistic features include the text tone, writing style, language formation. Sentiment and psychological feature extraction classify the article and include the positive or negative emotion words. Statistical and semantic features convert the text into numerical vector format. This helps in finding semantic relationships between texts or words. Existing studies suggest that misleading information or fake news is frequently different from authentic news in terms of linguistic structure, emotional tone and semantic coherence. Consequently, researchers have explored a broad set of text-based fake news classification approaches using NLP and ML or DL algorithms. Although several research papers have been published on fake news detection, many of the researchers focus on machine learning and deep learning algorithms without extracting text features. Text feature extraction is important for classification. Moreover, existing reviews often lack a structured taxonomy of text-based features, making it difficult to analyze the contribution of individual feature types. In addition, there is no clear and consistent distinction between linguistic, stylistic, and psychological features, as these categories are frequently used interchangeably across studies. These limitations highlight the need for a feature-centric review that systematically classifies text-oriented fake news classification techniques. This review examines feature-centric approaches for text-based fake news classification methods published from 2017 to 2025. The study systematically categorizes extracted textual features into linguistic, stylistic, psychological, statistical and semantic and compares the corresponding machine learning and deep learning techniques used for classification. The key contributions of this paper are: (1) a unified taxonomy of text-based features in fake news detection including linguistic, stylistic, psychological, statistical, semantic and syntactic features; (2) a comparative analysis of the selected studies according to the types of textual features used; and (3) the recognition of the underexplored feature categories and future research directions for the development of more comprehensive and interpretable fake news detection systems.

2. Review Methodology

Relevant studies published between 2017 and 2025 were identified using academic databases and digital libraries. The search focused on keywords such as "fake news detection," "text-based fake news detection," "feature extraction," "linguistic features," "statistical features," and "semantic features." Studies were selected based on their relevance to textual feature extraction and fake news classification. The selected studies were analyzed according to the types of features used, including linguistic, stylistic, psychological, statistical, semantic, and syntactic features. The studies were then compared to identify commonly used feature categories and underexplored research areas. The researchers have been studying fake news detection (FND) with several different methods to increase the accuracy. Some examples of these approaches are machine learning (ML), deep learning (DL), linguistic features, and multimodal techniques. Additionally, many studies have noted that feature extraction methods are critical for developing a FND system. This section presents a review of the different major contributions to fake news classification systems. Only a limited number of researchers worked on linguistic, stylistic, and psychological features in fake news detection, and these features are often combined with other feature types. Ahmad et al. (2020) used the LIWC tool to extract a comprehensive set of the linguistic, writing style and psychological characteristics of published false news articles [1]. Similarly, Rashkin et al. (2017) applied comprehensive linguistic analyses of fake news and political fact-checking. Using LIWC, subjectivity, hedging and intensifier lexicons, the study demonstrated that misleading and unreliable news differs from trusted news primarily in terms of linguistic style, emotional tone and sensible phrasing rather than outright falsehood [5]. Mashuri et al. (2025) used lexicon-based strategies and Long Short-Term Memory (LSTM), to improve the performance of fake news classification. The lexicon-based methods have been used to detect phrases and words that are commonly found in fake news [10]. In

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5200 -5206
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0680

addition, Aslam et al. (2025) extracted linguistic, stylistic, psychological, statistical and semantic features from the text. The readability scores, Named Entity Recognition (NER) and Part-of-Speech (POS) tagging were used to extract linguistic features. The results for the evaluation of the models provide important information regarding their ability to differentiate between real and fake news articles [4]. Collectively, these studies indicate that linguistic, stylistic, and psychological features provide valuable information for identifying deceptive content. However, compared with statistical and semantic representations, these three features categories are less frequently explored in recent studies and are often integrated with other feature categories rather than being investigated independently. This suggests the need for further research on comprehensive linguistic, stylistic, and psychological feature engineering and its integration with complementary textual features. The most popular representation techniques are statistical and semantic methods. Statistical features convert text data into numerical vectors using BoW, CV, TF-IDF, N-grams, and frequency-based representations. Almarashy et al. (2023) implemented an innovative model for fake news identification based on fusion of a deep neural network with CNN-BiLSTM. This study extracted and combined global, temporal and spatial features of text. Global features of the text were extracted from TF-IDF, temporal features were captured from BiLSTM and CNN was used to extract local features. This feature fusion achieved a high level of accuracy [3]. Mahmoud et al. (2025) explored ML techniques and NLP-based statistical approaches for fake news classification. The study extracted statistical features using N-grams, Count Vector, and TF-IDF and employed Logistic Regression, Support Vector Machines, Naïve Bayes, and Decision Tree algorithms for classification. Traditional machine learning models such as Naïve Bayes, Logistic Regression and Support Vector Machines have been effective in classifying deceptive content based on linguistic and statistical features [7]. Mahmud et al. (2024) adopted a deep learning based approach integrated with a word embedding technique for fake news identification. The authors extracted semantic features from news content using the Word2Vec technique and implemented various deep learning models including Convolutional Neural Networks, Artificial Neural Networks, LSTM and BiLSTM. The approach achieved competitive performance. However, stylistic, syntactic, sentiment, and statistical features were not explored in the study [6]. Abduljaleel et al. (2025) implemented deep learning-based techniques for fake news detection using contextual semantic representations. The study extracted semantic features through BERT word embedding and applies a hybrid CNN–BiLSTM model with an attention mechanism for classification. Experimental results showed improved detection accuracy across multiple datasets. However, this technique mainly emphasizes semantic embeddings and does not explicitly incorporate linguistic, stylistic or psychological features, reflecting a model-centric rather than feature-centric design [11]. Another approach presented by Khan et al. (2023) a hybrid feature engineering approach that combined a Word2Vec embeddings with TF-IDF weighting to capture both contextual meaning and term importance. The ensemble-based framework demonstrates that hybrid semantic representations outperform standalone word embedding models in textual fake news detection [2]. Zhang et al. (2025) primarily focused on statistical and semantic features for text-driven fake news classification. The authors extracted statistical features from N-gram and TF-IDF, semantic features were derived from BERT and word2vec and applied both ML and DL models for classification. This approach achieved competitive performance, indicating the effectiveness of the selected features. However, the approach mainly emphasizes statistical and semantic embeddings and does not explore stylistic, syntactic or psychological features [8]. Jadhav et al. (2025) used NLP for automated fake news identification and to categorize news as authentic or fraudulent using both machine learning and deep learning models. TF-IDF, word embeddings (Word2Vec, GloVe) are used for statistical feature extraction and BERT was used for contextual embeddings. However, this study does not explore linguistic, stylistic, syntactic or psychological features, limiting comparative analysis across feature types [9].

3. Taxonomy Of Textual Features

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5200 -5206
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0680

In fake news identification, the news articles rely heavily on textual features. Therefore, the text-based feature extraction is important. Over the years, researchers have explored a diverse set of textual cues to differentiate deceptive news from trustworthy information. Based on an extensive review of current literature, this section proposes a systematic taxonomy of text-based features, categorizing them into linguistic, stylistic, psychological, statistical, semantic and syntactic features. This categorization helps to understand the contribution of each feature type and identify trends in feature engineering techniques for fake news detection.

3.1.Linguistic Features

Linguistic features capture the structural and grammatical properties of language used in news articles. Parts-of-Speech (POS) tagging and readability analysis are NLP techniques that help to extract these types of features. Linguistic features include Part-of-Speech (POS) tagging distributions (nouns, verbs, adjectives, and adverbs), pronoun usage (first-person and second-person), sentence length and word length, vocabulary richness and complexity, readability index including Flesch Reading Ease and Gunning Fog Index. Several studies have revealed that linguistic patterns of fake news are different from real news. According to Rashkin et al., untrustworthy news typically uses superlatives and modal verbs [5]. Similarly, Ahmad et al. implemented an ensemble-based technique using LIWC features to show that linguistic markers considerably improve fake news classification performance when combined with traditional machine learning models [1].

3.2.Stylistic Features

Stylistic features reflect the author's writing style and rhetorical strategies used to attract or manipulate readers. Stylistic features include excessive use of punctuation, capitalization, superlatives, intensifiers, informal language, headline–body inconsistency, sentence variation and formatting patterns such as number of long sentences, number of short sentences and number of special characters. Previous research demonstrated that fake news frequently uses a sensational tone, implementing exaggerated expressions and dramatic phrasing to elicit emotional responses. Linguistic analyses show that stylistic cues are particularly effective in differentiating satire, hoaxes and propaganda from trustworthy news articles.

3.3.3.3 Statistical Features

Statistical features are one of the most common ways to represent text in classification tasks. These features look at word occurrence and the importance of words in a sentence. Bag-of-Words (BoW), Term Frequency–Inverse Document Frequency (TF-IDF) and word and character N-grams are all examples of common lexical features. Lexical features are simple but work well for finding fake news, especially when used with machine learning classifiers like Support Vector Machines, Random Forests, and ensemble models. However, these features are limited in capturing contextual meaning and semantic relationships between words.

3.4.3.4 Semantic Features

The aim of semantic features is to understand the meaning and contextual relationships within text. These features represent an important transition from superficial analysis to meaning-driven representations. Word embeddings such as GloVe, FastText, and Word2Vec, Document embeddings (Doc2Vec) and Topic modeling techniques such as Latent Dirichlet Allocation (LDA) and Semantic similarity measures are common examples of semantic representations. Hybrid methods that combine TF-IDF weighting and word embeddings have demonstrated better performance because they integrate both term importance and semantic information. Recent studies further emphasize semantic similarity modeling for short texts such as tweets, demonstrating its importance in social media–based fake news detection.

3.5.3.5 Syntactic Features

Within a sentence, syntactic features focus on the word order and grammatical structure. Unlike lexical features, syntactic cues provide information about the sentence construction and organization. Examples of syntactic features include POS-aware similarity measures, dependency relations, word order and sentence structure patterns, clause complexity. However, syntactic features are less explored compared to semantic features. Recent studies focus on their effectiveness, particularly for short and noisy texts. Syntactic structure has been shown to enhance

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5200 -5206
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0680

similarity-based fake news classification models.

3.6.Psychological and Sentiment Features

Psychological features emphasize the emotional and behavioral aspects of language. These features are especially useful for detecting manipulation and persuasion in fake news. Typical psychological features include sentiment polarity (positive, negative and neutral), emotion categories such as fear, anger, joy and sadness, subjectivity indicators, cognitive and affective process words using Linguistic Inquiry and Word Count (LIWC). Research consistently shows that fraudulent news tends to use emotionally charged language and higher subjectivity compared to factual news. These features have been found to significantly improve the performance of traditional machine learning models and offer better interpretability than deep semantic representations. As shown in Table 1 Psychological sentiment

4. Comparative Analysis of Existing Studies

Table 1 Psychological sentiment

Sr. No.	References	Linguistic	Stylistic	Psychological / Sentiment	Statistical	Semantic	Syntactic
1	Ahmad et al. (2020)	Yes	Yes	Yes	Implicitly	No	No
2	Khan et al. (2023)	Implicitly	No	No	Yes	Yes	No
3	Almarashy et al. (2023)	Implicitly	No	No	Yes	No	No
4	Aslam et al. (2025)	Yes	Yes	Yes	Yes	Yes	Implicitly
5	Rashkin et al. (2017)	Yes	Yes	Yes	Implicitly	Implicitly	No
6	Mahmud et al. (2024)	No	No	No	No	Yes	No
7	Mahmoud et al. *(2025)*	Implicitly	No	No	Yes	Implicitly	No
8	Zhang et al. (2025)	Implicitly	No	No	Yes	Yes	No
9	Jadhav et al. (2025)	No	No	No	Yes	Yes	No
10	Mashuri et al. (2025)	Yes	No	Yes	No	No	No
11	Abduljaleel et al. (2025)	No	No	No	No	Yes	No

5. Discussion And Research Gaps

The analysis has demonstrated that research on misinformation detection frequently employs statistical and semantic attributes of documents. TF-IDF, n-grams and word embeddings are all examples of preferred features. All three have been shown to work well in measuring the importance of terms and meanings. This reliance on these common features

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5200 -5206
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0680

indicates a narrow representation of how deceptive language can be used while ignoring clues from writing style and psychological manipulation found in fake news. The comparative analysis of existing studies shows that linguistic, stylistic, and psychological features are used less frequently. Research indicates that deceptive material has distinct emotional tones, writing styles, and linguistic characteristics. These different types of features remain unexplored, creating a research gap in the area of detecting fake news. One notable observation made in the comparative analysis is that syntactic features are absent in reviewed studies. Syntactic structure such as dependency relationships and parse tree patterns can provide insight into how sentences are constructed and the complexity of the structure. Therefore, the absence of syntactic features in existing studies indicates an area of future research that could yield further structural insight into fraudulent text. Only a few published studies combine multiple types of features into one framework, but none of them integrate all possible feature types identified earlier. The absence of heterogeneous feature fusion limits the generalization and robustness of current detection systems.

Conclusion And Future Work

This study presents a comparative analysis of feature-based approaches used for fake news classification. The analysis of existing research shows a strong preference for statistical and semantic features, while linguistic, stylistic, psychological and syntactic features are less frequently considered. The lack of syntactic features as well as the lack of comprehensive feature integration indicate many areas in which more research needs to be done. These results show that there is an immense demand for fake news detection models that focus on features and are easy to understand in order to make the systems more stable and consistent overall.

References

[1]. Ahmad, I., Yousaf, M., Yousaf, S., & Ahmad, M. O. (2020). Fake news detection using machine learning ensemble methods. Complexity, 2020(1), Article 8885861. https://doi.org/10.1155/2020/8885861

[2]. Khan, Z. A., & Rekha, V. (2023). Fake news detection using TF-IDF weighted with Word2Vec: An ensemble approach. International Journal of Intelligent Systems and Applications in Engineering, 11(3), 1065–1076.

[3]. Almarashy, A. H. J., Feizi-Derakhshi, M. R., & Salehpour, P. (2023). Enhancing fake news detection by multi-feature classification. IEEE Access, 11, 139601–139613. https://doi.org/10.1109/ACCESS.2023.3342458

[4]. Aslam, Z., Missen, M. M. S., Ghaffar, A. A., Mehmood, A., Villar, M. G., Alvarado, E. S., & Ashraf, I. (2025). Advancing fake news combating using machine learning: A hybrid model approach. Knowledge and Information Systems, 67(12), 12137–12177.

[5]. Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 2931–2937). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1317

[6]. Mahmud, T., Akter, T., Aziz, M. T., Uddin, M. K., Hossain, M. S., & Andersson, K. (2024). Integration of NLP and deep learning for automated fake news detection. In 2024 Second International Conference on Inventive Computing and Informatics (ICICI) (pp. 398–404). IEEE. https://doi.org/10.1109/ICICI63465.2024.10730510

[7]. Mahmoud, H. A., & Ibrahim, I. M. (2025). Fake news detection using natural language processing (NLP) and machine learning. Engineering and Technology Journal, 10(5), 5177–5185. https://doi.org/10.47191/etj/v10i05.46

[8]. Zhang, L. (2025). Feature extraction based on Naive Bayes algorithm and TF-IDF for news classification. PLoS ONE, 20(7), e0327347. https://doi.org/10.1371/journal.pone.0327347

[9]. Jadhav, T., Adlinge, S., Dande, N., & Jadhav, S. (2025). Fake news detection on social

media using NLP. International Journal of Scientific Research in Science and Technology, 12(6), 175–181.

[10]. Mashuri, C., Prastyo, E. H. A., & Hariri, F. R. (2025). Improving fake news detection accuracy with a lexicon-based approach and LSTM through text preprocessing. JSINBIS: Jurnal Sistem Informasi Bisnis, 15(2), 285–292.

[11]. Abduljaleel, I. Q., & Ali, I. H. (2025). Detecting fake news using BERT word embedding, attention mechanism, partition and overlapping text techniques. TEM Journal, 14(2), 1152–1165. https://doi.org/10.18421/TEM142-16