Fake Reviews Detection

Masoomeh Akbari, Hafsa Salad, Ziqi Yang, Anqi Zhu
University of Ottawa, Ottawa, Canada
Email: {makba074, hsala089, zyang040, azhu056}@uottawa.ca

***Abstract*—The rise of fake reviews has become a critical issue in online platforms, misleading consumers and compromising the credibility of products and services. This project reviews current research on fake review detection methods, including traditional machine learning techniques like Support Vector Machines (SVM) and newer deep learning approaches such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). By analyzing the strengths and limitations of existing methods, this study highlights the importance of feature extraction techniques and proposes improvements to enhance detection accuracy. Additionally, the project evaluates recent advancements in deep learning and their effectiveness in handling the complexities of fake review detection, providing recommendations for future research and system development. Our proposed contribution is a detection process combining CNNs, RNNs, Capsule Networks (CapsNets), and explainable AI (XAI) to create a robust, interpretable solution. This novel approach addresses the limitations of existing methods, enhancing adaptability, interpretability, and efficiency in detecting fake reviews.**

***Index Terms*—Fake Reviews, CNN, RNN, CapsNets, XAI**

I Introduction

As the Internet becoming increasingly integrated into our daily lives, people depend heavily on online reviews to guide their decision-making process. Over 90% of customers rely on reviews to choose products or services [1], [2], [3]. While reviews can offer valuable insights, there is a growing concern about the presence of fake reviews designed to mislead users. Fake reviews are often posted by spammers with the intent to deceive by providing inaccurate or biased information. With the growing threat of fake reviews to the credibility of online platforms, machine learning and deep learning techniques have emerged as key tools to address the problem [1], [2], [3], [4], [5]. Traditional machine learning methods, such as SVM and Naive Bayes, perform well in classification on small-scale datasets, but their ability to detect fake reviews in dynamic, multi-domain contexts is limited [4], [5]. In contrast, deep learning techniques, like CNN and RNN, have shown greater adaptability and robustness [1], [3], [5]. This project will analyze existing research and methods proposed in the field of fake review detection and will present a contribution aimed at improving these techniques.

II Problem or Domain Description

A. Problem

As consumers become increasingly reliant on reviews to make purchasing decisions, the presence of fake reviews poses a serious challenge to the credibility and reliability of these platforms. The scale of fake reviews has grown exponentially with the rise of e-commerce and review-driven platforms. Fake reviews not only distort people's perceptions of the products' quality, they also create a competitive disadvantage for companies [1].

B. Challenge

Detecting fake reviews is facing multiple challenges:

1) Similarity to Genuine Reviews

Fake reviews are often crafted so well that they mimic genuine reviews. Fake reviews use realistic language patterns, specific product details, and varying sentiments to avoid detection. Spammers intentionally emulate authentic review styles to deceive detection algorithms, and some fake reviews are generated by sophisticated bots using AI-based language models, further complicating detection [1], [3].

2) Large Datasets

The large scale of reviews posted on online platforms makes real-time detection computationally intensive. Platforms like Amazon and Yelp generate millions of reviews daily, so it is challenging to analyze all content effectively [2]. Detecting fake reviews at this scale requires highly efficient algorithms and computing power.

3) Evolving Spamming Technology

Spammers continuously evolve their techniques to deceive detection systems. Research has highlighted that spammers often create temporal patterns, such as leaving multiple fake reviews within a short timeframe, to manipulate product ratings [4]. Additionally, spammers use fake accounts to distribute reviews across multiple platforms, which increases the complexity of detection.

4) Multilingual Detection

Detection models that are trained on a single platform often fail to generalize across other platforms due to differences in review structures and user behavior. For example, a model trained that is based on an English-language reviews dataset may perform poorly when applied to other languages. So, it's crucial to develop multilingual detection frameworks [1].

5) Imbalanced Datasets

The imbalance between genuine and fake reviews in datasets poses a significant challenge for machine learning models. Most datasets contain a small number of fake reviews compared to genuine ones, which leads to biased models that favor the majority class [3].

C. Problem statement

The primary question we face is: How can we effectively and accurately detect fake reviews?

III Fake Review Overview

Fake reviews are deceptive content created on online platforms to mislead consumers and manipulate business reputations. Understanding their types and detection processes is crucial for developing effective detection models.

A. Fake review Types

Fake reviews generally fall into three main categories:

1) Untruthful Opinions

These reviews are intentionally written to praise or criticize a product to mislead potential buyers. For example, they may overstate the benefits of a product to increase sales, or falsely criticize a product to damage the reputation of a competitor. Untruthful reviews are difficult to detect because they often mimic real consumer reviews.

2) Brand-only Reviews

This type of review focuses solely on the brand instead of discussing specific products or services. These reviews may be promotional or critical of the brand, which often lacks detailed feedback on specific products.

3) Non-reviews

These include Irrelevant content, like advertisements about another product, questions about the product, or comments about a comment. Non-reviews do not provide useful insights about a specific product or service, it will contribute noise to datasets and complicate the detection process.

In summary, understanding the different types of fake reviews is essential to developing a targeted detection strategy, since each type of fake review poses different challenges that affect the accuracy and reliability of the detection model [1].

IV Fake review process

The fake review detection process is consisting of four main steps: data collection, preprocessing, feature extraction, and detection model [1], [3], [5], as shown in Figure 1. First, data like user IDs, review text, and metadata will be collected from online platforms. These data are pre-processed through tokenization and removal of stop words and stems to prepare data for analysis. Then, feature extraction is performed to identify key attributes such as behavioral patterns and textual features. These features are then fed into detection models, such as traditional machine learning algorithms and advanced deep learning techniques, to categorize reviews as fake or not [1].

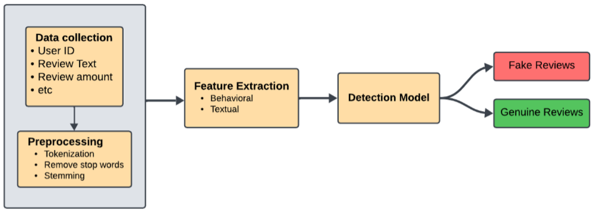


Fig. 1. Fake Review Process

A. Feature Extraction

Feature extraction is a crucial step in the fake review detection process, it converts raw data into meaningful attributes that can be analyzed by detection models. This step involves identifying and quantifying key patterns in user behavior and review content to effectively differentiate between fake and genuine reviews [1], [2], [4], [5]. Feature extraction can be broadly categorized into two types: behavioral features and textural features.

1) Behavioral Features

Behavioral features focus on user activity patterns, which provide valuable insights into potential spam behavior. These features are important because they reveal statistical and behavioral anomalies that distinguish spammers from genuine commenters. Examples of behavioral features include:

- Maximum number of reviews: Spammers typically post an unusually high number of comments in a short period of time. This is an indicator of spam activity since genuine users typically do not engage in such behavior. Research shows that 75% of spammers post more than five reviews in a single day, while 90% of normal users never write more than one review per day [1].
- Percentage of positive reviews: If an individual user has a high percentage of positive remarks, this may suggest promotional intent, especially if these remarks lack critical feedback. A high percentage of overly positive reviews may suggest deceptive behavior, as spammers often try to artificially boost ratings [1], [2].
- Average review length: Fake comments tend to appear abnormally short since spammers prioritize quantity over quality. Studies indicate that 90% of reliable reviewers write longer reviews, with an average length of more than 200 words [1].

2) Textual features

Textual features are extracted from the content of the comments, with a focus on linguistic and semantic attributes that help to differentiate between fake and genuine comments. These characteristics are critical because fake reviews often show unique patterns, such as overly generic language, unnatural wording, or repetitive content.

- Meta data: This involves analyzing metadata such as timestamps, comment lengths, and frequency of specific terms. For example, unusually short comments or comments posted at unusual times may indicate fake reviews [1].
- Bag of words (BOW): The BOW model captures the frequency of individual words or phrases in the review content, it provides a simple but effective representation of textual content. It is useful for identifying common patterns or repeated phrases that are often found in fake reviews [1], [3].
- Word embeddings: Word embeddings capture deep semantic meaning and context but require high computational resources [1], [3]. Advanced technologies like Word2Vec or GloVe capture the semantic relationships

between words, which enables the detection model to understand the contextual meaning of the comment [1].

Each method has its own advantages, and usually a combination of these methods produces better results. Combining behavioral and textual features significantly improves the performance of fake review detection models [1], [2].

V Review of Current Work

The field of fake review detection has seen significant advancements. Current research leverages both traditional machine learning and deep learning methods. Traditional ML models, such as Support Vector Machines (SVM) and Naive Bayes, rely on structured features and have been widely studied for their effectiveness in identifying fake reviews. In contrast, deep learning approaches, such as CNN and RNN use the power of neural networks to automatically extract complex patterns from reviews. Both of these approaches create a foundation of ongoing research in detecting fake reviews.

A. *Traditional Statistical Machine Learning In Detecting Fake Reviews*

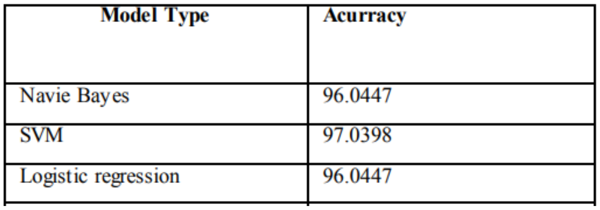
Traditional machine learning methods can effectively detect fake reviews. These models are applied for review classification after relevant features are isolated for analysis. Numerous studies have evaluated the performance of these models in identifying fake reviews, showcasing their strengths and limitations

C. Silpa, et al. [6], proposed a framework utilizing supervised machine learning methods to classify reviews as genuine or fake. The study compared algorithms like Support Vector Machine (SVM), Naive Bayes, and Logistic Regression in their capability to fake detect reviews. In their study, data of reviews was retrieved from various sources, such as Amazon and websites for making airline, hotel, and restaurant reservation [6]. The variety of reviews from different sources was to increase the diversity of the data. After Feature extraction, key metrics were analyzed to assess model performance.

- **Accuracy:** measures how often the model's predictions are correct [6].
- **Precision:** evaluates how many reviews classified as "fake" are actually fake [6].
- **Recall:** assesses how many of the actual "fake" reviews the model correctly identified [6].
- **F1 score:** mean of precision and recall, balancing the trade-off between these two metrics [6].

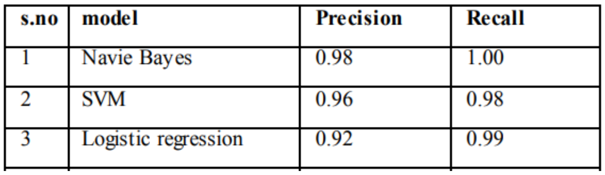
To ensure fairness in model training, the dataset was sampled to include an equal proportion of fake and genuine reviews. This strategy prevented the models from becoming biased toward one class of review. The study compared algorithms like Support Vector Machine (SVM), Naive Bayes, and Logistic Regression. SVM is a supervised machine learning algorithm that maps data into a feature space for categorization. It works by finding a decision boundary (hyperplane) that separates different classes in the feature space. An advantage of SVM is its effectiveness in handling huge datasets and high accuracy [6]. Naive Bayes is a probabilistic machine learning algorithm based on Bayes' Theorem. It calculates the conditional probability of a class (such as "fake" or "genuine") given the features. Logistic Regression is a linear model used for binary classification tasks. It uses a standard sigmoid function, which maps the input features to a probability value between 0 and 1.

When comparing the models performances, the study found SVM to be the most accurate with a precision of 97% [6]. Naive Bayes demonstrated the highest recall at 100%, along with a precision of 98% and an F1 score of 98% [6]. Logistic Regression achieved a similar accuracy to Naive Bayes, at approximately 96% [6]. What can be extracted from this data is SVM is the most balanced and effective model, while Naive Bayes excels in thoroughly identifying fake reviews.



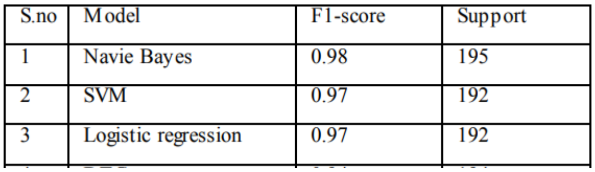
Model Type	Acurracy
Navie Bayes	96.0447
SVM	97.0398
Logistic regression	96.0447

TABLE I
ACCURACY OF 3 ALGORITHMS [6]



s.no	model	Precision	Recall
1	Navie Bayes	0.98	1.00
2	SVM	0.96	0.98
3	Logistic regression	0.92	0.99

TABLE II
CLASSIFICATION REPORT OF PRECISION AND RECALL [6]



S.no	Model	F1-score	Support
1	Navie Bayes	0.98	195
2	SVM	0.97	192
3	Logistic regression	0.97	192

TABLE III
CLASSIFICATION REPORT OF F1-SCORE AND SUPPORT [6]

B. *Neural Network In Detecting Fake Reviews*

Neural networks, particularly deep learning methods, have become essential tools for data classification and analysis, especially in natural language processing (NLP). Unlike traditional machine learning, deep learning automates feature extraction, significantly reducing manual effort. Techniques like word embeddings allow neural networks to capture complex semantic relationships in text. These capabilities make neural networks powerful tools for challenges such as fake review detection. In the next sections, we will discuss the prominent architectures: Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Generative Adversarial Networks (GAN), highlighting their unique strengths and applications [1].

1) Convolutional Neural Networks (CNNs)

CNNs are highly effective for analyzing data with local patterns, such as text or images. They excel at identifying key features by focusing on word combinations and linguistic cues. As shown in Figure 2, a CNN processes text by converting reviews into numerical vectors to form an input matrix. This matrix is then passed through convolutional layers, where filters detect meaningful patterns in small word groups, such as “What is” or “fake review,” creating feature maps that capture essential information. To simplify the data, the output is passed through a pooling layer, which reduces dimensionality while retaining critical features. Finally, the pooled features are concatenated into a single vector and passed through a classifier to determine whether the review is fake or genuine [1].

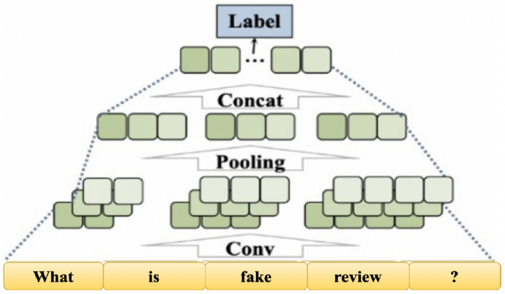


Fig. 2. The architecture of a Convolutional Neural Network (CNN) [1].

Deep Convolutional Neural Networks (DCNNs) extend the standard CNN framework by incorporating additional layers, allowing the network to learn more abstract and complex features from the data. These models have proven to be especially effective in applications like fake review detection, where identifying nuanced patterns in textual data is crucial. By employing a deeper architecture, DCNNs can extract hierarchical representations of text, improving the network’s ability to distinguish between genuine and fake reviews.

While CNNs and DCNNs excel at identifying local patterns in data, they may not capture long-term dependencies and sequential relationships present in textual data. This limitation makes it necessary to complement these models with architectures that can process sequential data. Bidirectional Long Short-Term Memory (BiLSTM), a type of recurrent neural network, is particularly well-suited for this task. BiLSTM processes data in both forward and backward directions, capturing long-range dependencies and enabling a deeper understanding of the text’s context.

In a novel framework discussed in [3], CNN and BiLSTM are combined to achieve significant results in detecting fake reviewers. CNN is used as a behavior-sensitive feature extractor, capturing local dependencies in behavioral data, while BiLSTM processes textual data in both forward and backward directions, capturing long-term dependencies and contextual patterns. Integrated with a context-aware attention mechanism and advanced pre-trained models like Longformer, this approach effectively combines local feature detection and sequential understanding. The framework achieves state-of-the-art accuracy (85.05%) as demonstrated in Table IV, showcasing the superiority of the integrated design.

2) Recurrent Neural Networks (RNNs)

RNNs are effective for processing sequential data, making them suitable for tasks that involve understanding word order and capturing contextual meaning in extended text.

RNNs represent each word in a review as a numerical vector using techniques like Word2Vec or GloVe. The network processes the sequence one word at a time, generating a hidden state that combines information from the current word with the context of previous words, as shown in Figure 3. This allows RNNs to retain and use relevant information from earlier parts of the sequence.

One key feature of RNNs is their use of shared parameters, which ensures consistent processing across all words in a sequence. After processing the entire sequence, the RNN produces a prediction, such as classifying the review as fake or genuine. While RNNs are effective for sequential tasks, they often struggle to capture long-term dependencies in the data [1].

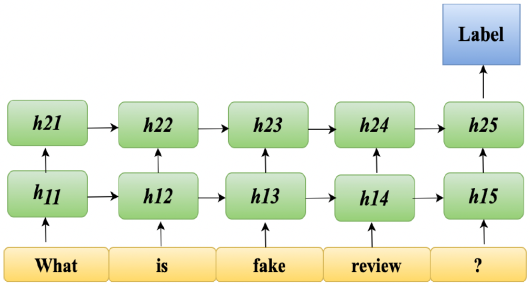


Fig. 3. The architecture of a Recurrent Neural Network (RNN) [1].

To address these challenges, the Bidirectional GRU layer, an enhancement of the traditional Gated Recurrent Unit (GRU), is used to process textual data in both forward and backward directions. By doing so, it captures contextual information more effectively compared to standard RNNs. This bidirectional structure is highly effective in fake review detection, as it enables the model to consider dependencies from both past and future words in the sequence [1].

In addition, [7] introduces a novel framework that utilizes BiLSTM in combination with a feature-rich architecture to improve fake review detection. The model integrates part-of-speech tags, first-person pronoun features, and document embeddings (e.g., GloVe) to create a comprehensive representation of reviews. These features, processed through BiLSTM or Bidirectional GRU layers, allow the framework to effectively capture complex linguistic and contextual patterns.

No.	Algorithm for BF	Algorithm for text	Acc	Precision	Recall	F1-Score	AUC
1	Conv1d	CNN + Attention	77.84%**	0.7139*	0.7255*	0.7197*	0.7778*
2	Conv1d	BiLSTM+Attention	78.48%**	0.7050**	0.6987*	0.7018*	0.7781*
3	Conv1d	CNN + BiLSTM	77.24%***	0.6884***	0.6720**	0.6801***	0.7641***
4	Our model		**85.05%**	**0.7689**	**0.7643**	**0.7664**	**0.8358**

TABLE IV
PERFORMANCE OF "OUR MODEL", THE MODEL PROPOSED IN [3], COMPARED TO ITS VARIATIONS.

By combining these elements, the framework overcomes the limitations of traditional RNN-based methods and achieves significant improvements in accuracy, setting a new benchmark for fake review detection.

3) Generative Adversarial Networks (GANs)

GANs address the challenge of data scarcity in deep learning, particularly in applications like fake review detection, where obtaining labeled data can be labor-intensive. They are also effective in tackling the cold start problem—a situation where there is insufficient or no data available to train models or make predictions, such as when dealing with new users or items in a system.

GANs consist of two competing networks: a generator and a discriminator, as shown in Figure 4. The generator starts with a random noise vector and transforms it into synthetic data resembling genuine reviews. Meanwhile, the discriminator evaluates both real and synthetic data, predicting their authenticity. This adversarial process continues iteratively, with the generator improving its ability to create realistic data and the discriminator enhancing its skill in distinguishing genuine from fake [1].

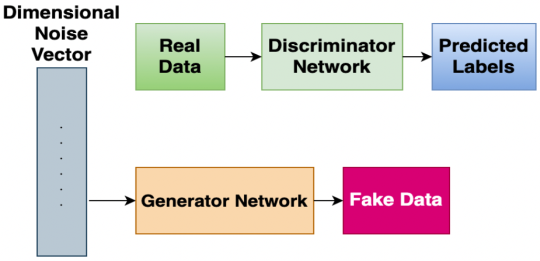


Fig. 4. The architecture of a Generative Adversarial Network (GAN) [1].

In [8], the authors introduce a novel application of GANs to tackle the cold-start problem in fake review detection. Their method generates synthetic behavior features (SBFs) for new users by leveraging easily accessible features (EAFs) such as text, ratings, and attributes. By designing a unique generator and discriminator structure within the GAN framework, they effectively transform EAFs into SBFs, enabling improved detection performance. Their approach demonstrates significant accuracy improvements on Yelp datasets.

In summary, Traditional machine learning (ML) and deep learning have distinct differences in their approaches and capabilities. Traditional ML relies on handcrafted features, which limits its performance on large, unstructured data such as text or images. On the other hand, neural networks automatically learn features from raw data. They excel when handling complex and unstructured datasets. Traditional ML models are simpler and faster to train. However, they struggle to scale with very large datasets. Neural networks require significant computational resources, but they perform better as the dataset size increases. This makes them more suitable for handling the growing complexity of modern data.

VI Proposed Idea

Fake review detection is a multi-dimensional problem that requires the integration of advanced techniques to effectively capture textual, behavioral, and contextual patterns. While traditional and deep learning methods can address some aspects of this problem, they often fail to achieve the desired results due to their black-box nature, which limits their interpretability, adaptability to changing patterns, and efficiency when dealing with large datasets or multilingual reviews. To overcome these challenges, we proposed a detection process that combines CNNs, RNNs, capsule networks (CapsNets), and explainable AI (XAI) to create a robust and interpretable system. This novel combination leverages the strengths of each model while addressing their individual limitations. Therefore it can create a robust and interpretable solution for fake review detection.

A. Description of the Proposed Idea

1) Description of Capsule Networks

Capsule Networks (CapsNets) have been explored in various domains, such as fake news detection and toxic comment classification, due to their ability to model hierarchical and contextual relationships. However, our research did not find studies applying CapsNets to fake review detection. This gap motivates our proposal to use CapsNets for this task. CapsNets can effectively identify patterns in text data by preserving spatial hierarchies and capturing dependencies between words and phrases, making them a strong candidate for addressing the challenges posed by fake reviews.

A simple architecture of CapsNet is described as follows [9]:

1) **Input Layer:** The network receives raw input, such as text embeddings, which encode semantic information about the input text.
2) **Primary Capsules:** This layer extracts local features and outputs vectors representing the presence and pose of parts, typically using convolutional operations. These

capsule vectors encode not only the presence of a feature but also its spatial properties.

3) **Higher-Level Capsules:** These capsules represent more abstract entities composed of the parts detected earlier. They vote on the pose and likelihood of these entities using learned transformation matrices, preserving the relationships between different text components.
4) **Routing-by-Agreement:** A dynamic mechanism adjusts the connections between capsule layers based on agreement among predictions, clustering them for aligned votes. This mechanism helps the model focus on relevant features while filtering out irrelevant ones.
5) **Output Layer:** The final capsules represent class-specific entities or abstract concepts, which are used to classify the input data into the appropriate category.

CapsNets are particularly effective for tasks requiring a nuanced understanding of hierarchical and contextual relationships [9], such as fake review detection. By capturing dependencies between words and phrases, they can identify subtle patterns indicative of deceptive content, even in complex or varied review structures.

2) Description of Explainable AI

When using a model to determine whether a review is fake or not, it is critical to understand why the model made a certain decision. This understanding not only helps improve credibility, but also identifies potential problems, like data bias or model errors. However, due to the black-box nature of most machine learning models (especially deep learning models), their internal decision-making processes are difficult to explain to humans, which poses a challenge. To address this issue, XAI methods can be utilized to help explain the process of model operation and reveal its decision logic [10], [11], [12].

SHapley Additive Explanations (SHAP) is a powerful XAI technique that provides detailed insights into model predictions by assigning importance scores to individual features. Shap is inspired by game theory, which evaluates the contribution of each feature (e.g., metadata such as specific words, comment length, or comment frequency) to the final classification [13], [14], [10]. For example, in fake review detection, SHAP can highlight that repetitive keywords like "amazing" or unusual behavioral patterns, such as a high number of reviews in a short period, significantly influenced the model's decision to classify a review as fake.

In addition, SHAP's ability to generate intuitive visual interpretations, such as feature importance maps, makes it accessible to non-technical stakeholders [10], [13]. It can also help model developers debug problems like over-reliance on specific features and retrain the model accordingly.

B. Proposed Architectures

To further enhance CapsNet performance, we propose hybrid architectures combining CapsNets with CNNs RNNs, and XAI. These combinations allow us to leverage the strengths of each architecture to address different aspects of the fake review detection problem:

1) **CapsNets with CNNs:** CNN layers are used to extract local features, such as n-grams or suspicious patterns. These features are then processed by Capsule layers to model hierarchical dependencies. This architecture is particularly effective for short reviews where local patterns play a critical role in identifying fake content.
2) **CapsNets with BiGRU:** Bidirectional GRU layers capture sequential dependencies, which are then modeled hierarchically by Capsule layers. This architecture is well-suited for longer reviews, where the sequence and flow of information are important for identifying deceptive patterns.
3) **CapsNets with CNN and BiLSTM:** This architecture combines CNN layers, BiLSTM layers, and Capsule layers to integrate local features, sequential patterns, and hierarchical dependencies. CNNs extract local features, BiLSTMs capture temporal patterns, and Capsule layers combine these representations for comprehensive analysis. This hybrid model ensures robust performance across different types of reviews.
4) **CapsNets with DCNN**: Deep Convolutional Neural Networks (DCNN) are used to extract multi-scale local patterns. These patterns refer to features identified at different granularities, such as n-grams, key phrases, and broader semantic structures. The patterns are captured using convolutional filters of varying sizes and passed to Capsule layers, which preserve the spatial and hierarchical relationships between features. This architecture is particularly suited for long reviews, where understanding detailed patterns and their interconnections is critical to identifying if the review is fake.
5) **XAI Integration:** XAI integration in the proposed architectures ensures transparency and trust by providing interpretable explanations for the model's predictions. SHAP highlights the importance of features such as specific words, review patterns, or metadata. It can also help to debug and refine the model.

Pre-trained embeddings such as BERT, Word2Vec, or GloVe can be employed in the input layer to provide rich semantic representations of the text. Additional enhancements, such as soft-routing mechanisms and attention layers, can improve computational efficiency and help the model focus on the most relevant parts of the input data.

C. Justification

CapsNets are well-suited for fake review detection due to their ability to capture hierarchical relationships and contextual information. Their use in related domains, such as fake news detection and toxic comment classification, has demonstrated their effectiveness in identifying subtle patterns in textual data. For example:

- [15] demonstrated that Soft-Capsule Networks outperformed traditional methods in identifying fake news patterns, highlighting the strength of CapsNets in modeling dependencies.

- [16] proposed a Multi-Modal Co-Attention Capsule Network that combined textual and visual features, achieving significant improvements in fake news detection accuracy.
- [17] showed that integrating BERT embeddings with CapsNets enhanced accuracy in toxic comment classification, leveraging the strengths of both contextual embeddings and Capsule layers.
- [18] used Multi-Dimension Capsule Networks for multilingual text classification, demonstrating their versatility across languages and contexts.
- [19] proposed using Capsule Networks in text classification, demonstrating their ability to capture hierarchical and contextual relationships within text. The study also showed how CapsNets enhance feature extraction and classification accuracy, particularly in identifying subtle patterns indicative of fake content.

These findings provide strong evidence for the effectiveness of CapsNets in capturing both local and sequential dependencies, which are critical for detecting fake reviews. Our proposed hybrid architectures aim to build on these strengths, creating a robust and adaptable framework for real-world applications. By addressing the unique challenges of fake reviews, such as varying lengths and structures, these models can significantly improve detection accuracy and reliability in domains where review authenticity is essential.

VII Summary and Conclusion

The detection of fake reviews is a critical challenge in ensuring the reliability of online information. This paper explored the strengths and limitations of traditional machine learning and deep learning approaches, highlighting their contributions to the current research in detecting fake reviews. Additionally, we proposed a detection process integrating CNNs, RNNs, Capsule Networks (CapsNets), and Explainable AI (XAI) to address the multidimensional nature of the problem. By leveraging the strengths of each component, the proposed approach provides a robust and interpretable solution for identifying fake reviews.

There are a few limitations to acknowledge. The computational complexity of the proposed architecture can limit scalability when applied to large datasets. Additionally, the architecture can increase implementation difficulty and requires expertise for fine-tuning. Furthermore, handling of multilingual reviews still remains a challenge. This would potentially require additional preprocessing and embedding techniques for multilingual compatibility.

References

[1] R. Mohawesh, S. Xu, S. N. Tran, R. Ollington, M. Springer, Y. Jararweh, and S. Maqsood, "Fake reviews detection: A survey," *Ieee Access*, vol. 9, pp. 65 771–65 802, 2021.

[2] H. Paul and A. Nikolaev, "Fake review detection on online e-commerce platforms: a systematic literature review," *Data Mining and Knowledge Discovery*, vol. 35, June 2021.

[3] D. Zhang, W. Li, B. Niu, and C. Wu, "A deep learning approach for detecting fake reviewers: Exploiting reviewing behavior and textual information," *Decision Support Systems*, vol. 166, p. 113911, 2023.

[4] W. H. Asaad, R. Allami, and Y. H. Ali, "Fake review detection using machine learning," *Revue d'intelligence artificielle*, vol. 37, 10 2023.

[5] A. M. Elmogy, U. Tariq, and A. Ibrahim, "Fake reviews detection using supervised machine learning," *International Journal of Advanced Computer Science and Applications*, vol. 12, 2021.

[6] C. Silpa, P. Prasanth, S. Sowmya, Y. Bhumika, C. H. S. Pavan, and M. Naveed, "Detection of fake online reviews by using machine learning," *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, vol. N/A, pp. 71–77, 2023.

[7] W. Liu, W. Jing, and Y. Li, "Incorporating feature representation into bilstm for deceptive review detection," *Computing*, vol. 102, pp. 701–715, November 2019.

[8] X. Tang, T. Qian, and Z. You, "Generating behavior features for cold-start spam review detection with adversarial learning," *Information Sciences*, vol. 526, pp. 274–288, July 2020.

[9] F. De Sousa Ribeiro, K. Duarte, M. Everett, G. Leontidis, and M. Shah, "Learning with capsules: A survey," *arXiv preprint arXiv:2206.02664*, 2022.

[10] N. Nobel, S. Sultana, S. P. Singha, S. Chaki, N. Mahi, T. Jan, A. Barros, and M. Whaiduzzaman, "Unmasking banking fraud: Unleashing the power of machine learning and explainable ai (xai) on imbalanced data," *Information*, vol. 15, pp. 298–298, 05 2024.

[11] D. Cirqueira, M. Helfert, and M. Bezbradica, "Towards design principles for user-centric explainable ai in fraud detection," *Artificial Intelligence in HCI*, pp. 21–40, 2021.

[12] L. Valina, B. Teixeira, A. Reis, Z. Vale, and T. Pinto, "Explainable artificial intelligence for deep synthetic data generation models," *2024 IEEE Conference on Artificial Intelligence*, pp. 555–556, 06 2024.

[13] T. Awosika, R. M. Shukla, and B. Pranggono, "Transparency and privacy: The role of explainable ai and federated learning in financial fraud detection," *IEEE access*, vol. 12, pp. 1–1, 01 2024.

[14] I. Psychoula, A. Gutmann, P. Mainali, S. H. Lee, P. Dunphy, and F. Petitcolas, "Explainable machine learning for fraud detection," *Computer*, vol. 54, pp. 49–59, 10 2021.

[15] H. G. Shah and H. Joshi, "Detection of twitter fake news using efficient soft-capsule and improved bigru architecture," *Journal of Artificial Intelligence and Capsule Networks*, vol. 6, no. 4, pp. 393–414, 2024.

[16] C. Yin and Y. Chen, "Multi-modal co-attention capsule network for fake news detection," *Optical Memory and Neural Networks*, vol. 33, pp. 13–27, 2024.

[17] H. Rahman Sifat, N. H. Nuri Sabab, and T. Ahmed, "Evaluating the effectiveness of capsule neural network in toxic comment classification using pre-trained bert embeddings," in *TENCON 2023 - 2023 IEEE Region 10 Conference (TENCON)*, 2023, pp. 42–46.

[18] S. Srivastava and P. Khurana, "Detecting aggression and toxicity using a multi dimension capsule network," in *Proceedings of the Third Workshop on Abusive Language Online*, S. T. Roberts, J. Tetreault, V. Prabhakaran, and Z. Waseem, Eds. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 157–162.

[19] M. H. Goldani, R. Safabakhsh, and S. Momtazi, "X-capsnet for fake news detection," *arXiv preprint*, 07 2023.