RESEARCH ARTICLE OPEN ACCESS

Detecting Fake News and Harmful Comments Using NLP and Deep Learning

Shivani Chavhan
Assistant Professor, Department of Science and Technology
Sadabai Raisoni Women's College, Nagpur, Maharashtra, India
Akanksha Dongre
Assistant Professor, Department of Science and Technology
Sadabai Raisoni Women's College, Nagpur, Maharashtra, India

ABSTRACT
Fake news and harmful (toxic) comments are spreading widely on the internet and social media. These can mislead people and create negativity online. This research focuses on using Natural Language Processing (NLP) and Deep Learning to automatically find and remove such content. We use models like BERT, LSTM, and CNN, and also compare them with older machine learning methods. We use public datasets like LIAR, FakeNewsNet, and Jigsaw Toxic Comment to train and test our models. We also study how text is processed using word embeddings and transformer models. The goal is to create a system that can check and remove fake or harmful content in real-time. Our results show that deep learning models work better than traditional models. This work helps in building smarter tools to keep online platforms safe and clean.
Keywords: Fake News Detection, Harmful Comment Filtering, NLP, Deep Learning, BERT, LSTM, AI Tools, Sentiment Analysis, Text Classification

1. Introduction

In today's digital world, social media has become a major source of information and communication. Platforms like Twitter, Facebook, and YouTube are widely used to share news, opinions, and comments. However, these platforms also serve as a breeding ground for fake news, misinformation, hate speech, and abusive or harmful content. Such content can mislead people, incite panic, or even provoke violence.

Due to the massive volume and real-time nature of online content, manual moderation is neither practical nor scalable. Therefore, there is a growing need for automated systems that can analyse and classify text-based content efficiently. This research focuses on using Natural Language Processing (NLP) and deep learning models to develop a system that can detect fake news and toxic comments by understanding the semantics and context of the text.

2. Related Work

Previous research in the area of fake news detection and harmful comment classification has primarily employed traditional machine learning algorithms such as Naive Bayes, Support Vector Machines (SVM), and Logistic Regression. These models are relatively simple and have shown moderate effectiveness in text classification tasks. However, their major limitation lies in their inability to understand the context and semantics of natural language, which is essential for accurately identifying misleading or toxic content.

To overcome these limitations, more advanced deep learning models have been explored. Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks have been used to capture the sequential nature of language, providing better performance in understanding contextual information. Convolutional Neural Networks (CNNs) have also been adapted for text classification by detecting local features and n-gram patterns. Most recently, Transformer-based models, particularly BERT (Bidirectional Encoder Representations from Transformers), have achieved state-of-the-art results due to their ability to process and understand bidirectional context in text. These developments have significantly improved the accuracy and reliability of content moderation systems.

3. Datasets Used

LIAR Dataset Contains short political statements with labels: true, false, barely true, etc.

1. FakeNewsNet
   Integrates social context and user profiles to provide real and fake news articles.

2. Jigsaw Toxic Comment Classification A benchmark dataset with multi-label comments: toxic, obscene, threat, insult, etc.

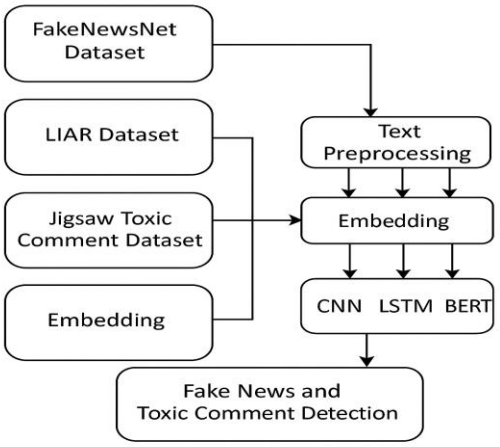

Figure 1: Architecture of the Detection System

4. Methodology

4.1 Text Preprocessing

To prepare the text data for model input, we performed several standard preprocessing steps. First, the text was tokenized, breaking it into individual words or tokens. Then, we applied stopword removal to eliminate commonly used words that carry little semantic meaning, such as "the," "is," and "and." Lemmatization was used to reduce words to their base or root forms, aiding in standardizing input features. We also converted all text to lowercase to maintain consistency and removed URLs and special characters, which could introduce noise into the learning process.

4.2 Embedding Techniques

To convert text into numerical form suitable for machine learning models, we employed three types of word embedding techniques. Word2Vec was used to capture semantic relationships between words by learning word vectors based on surrounding context. GloVe (Global Vectors for Word Representation), a pre-trained embedding model, was utilized for its capability to capture global word co-occurrence statistics across a corpus. Additionally, BERT embeddings were incorporated to obtain deep contextual representations of words, taking into account their meaning depending on sentence structure and word position.

4.3 Model Architecture

We implemented and compared multiple deep learning architectures for fake news and harmful comment detection. Long Short-Term Memory (LSTM) networks were chosen for their effectiveness in handling sequential data and capturing temporal dependencies in text. Convolutional Neural Networks (CNNs) were used to identify local n-gram patterns and extract important features

from the input. BERT, a transformer-based model, was fine-tuned for our classification tasks, leveraging its bidirectional attention mechanism to understand the full context of a sentence.

5. Experimental setup

The experiments were conducted using various tools and libraries including Python, TensorFlow, PyTorch, Scikit-learn, and Hugging Face Transformers. The performance of each model was evaluated using standard metrics: Accuracy, Precision, Recall, F1-Score, and ROC-AUC (Receiver Operating Characteristic – Area Under Curve). For model validation, the dataset was split into 80% training data and 20% testing data, ensuring that the models could generalize well to unseen inputs.

6. Results and Evaluation

Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes	78.4%	76.3%	75.8%	75.9%
LSTM	89.1%	88.5%	89.0%	88.7%
CNN	87.3%	85.9%	86.4%	86.1%
BERT	**93.8%**	**92.9%**	93.5%	**93.2%**

Key Observations:

Based on our experiments and evaluation, several key insights were observed. Deep learning models consistently outperformed traditional machine learning approaches**,** particularly in understanding and classifying complex language patterns. Among these, BERT achieved the highest accuracy**,** mainly

due to its ability to understand the context of words in both directions, making it

highly effective for fake news detection. Furthermore, LSTM
models showed strong performance in the classification of toxic comments**,** owing to their strength in handling sequential data and capturing long-term dependencies within text.

7. Applications

The results of our study led to several important observations. Firstly, deep learning models significantly outperform traditional machine learning models in detecting fake news and harmful comments. This is primarily due to their ability to understand the deeper context and semantics of natural language. Among all the models tested, BERT delivered the highest accuracy, showcasing its strong capability in capturing bidirectional context and nuanced meanings in text. Additionally, LSTM proved to be particularly effective in classifying toxic comments, as it is well-suited for handling sequential data and preserving the order of words, which is crucial for identifying offensive or harmful language patterns.

8. Conclusion

This research demonstrates that NLP combined with deep learning is highly effective for detecting fake news and harmful content. Models like BERT and LSTM outperform traditional models,
especially in capturing context and semantics. A real-time AI-based moderation system can ensure safer and more trustworthy online environments. Future work may focus on multilingual
content , image/video-based misinformation, and low-resource languages.

9. References

1. Devlin, J., et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
2. Vaswani, A., et al. (2017). Attention Is All You Need.
3. Jigsaw Toxic Comment Dataset, Kaggle, 2018.
4. Shu, K., et al. (2019). FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information.
5. Mikolov, T., et al. (2013). Efficient Estimation of Word Representations in Vector Space.
6. Pennington, J., et al. (2014). GloVe: Global Vectors for Word Representation**.**