



Fake News Detection System

Aditya Reddy Borra - 862393087
Srikar Kilambi - 862395242
Sree Yashwanth Sai Venkatesh - 862394400
Rahul Sharma - 862395453

The Rise of Fake News in Digital Era

- The digital age has brought an abundance of information, allowing easy access and consumption of data from various sources.
- Fake news refers to intentionally false or misleading information presented as legitimate news stories.
- Fake news poses a significant threat to society:
- Damages reputations and spreads misinformation.
- Manipulates public opinion and influences election outcomes.





Problem Statement

- Escalating the scale of fake news on the internet presents a complex and urgent challenge.
- Traditional methods of news verification and fact-checking struggle to keep up with the volume and speed of false information spread.
- Creators of fake news employ evolving tactics:
- Exploit individual vulnerabilities and manipulate images and videos.
- Mimic the appearance of legitimate news sources and write as objective reporting.
- Individuals seeking accurate information face a significant dilemma and distrust mainstream media.

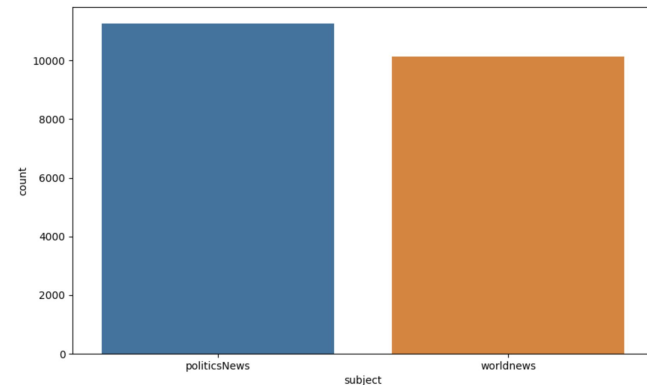
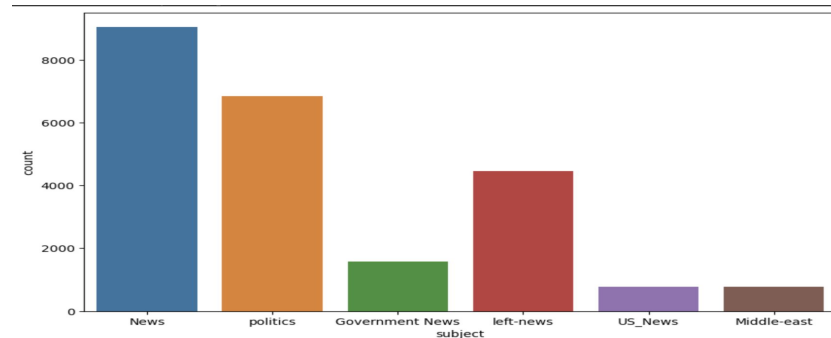


Proposed Solution

- Utilising Long Short-Term Memory (LSTM) networks with dropout regularisation for fake news detection.
- LSTM networks are specialised RNNs capable of processing sequential data, making them suitable for analysing text.
- LSTM networks capture contextual and long-term dependencies in news articles, enabling the detection of patterns indicative of fake news.
- Dropout regularisation is employed to enhance model performance:
- Randomly deactivated neurons during training to prevent overfitting.
- Prevents reliance on specific features, promoting a more robust and generalizable system.
- Extensive experimentation and evaluation yielded an impressive accuracy rate of 99.5%.
- Our proposed solution provides a strong foundation to combat fake news in the digital landscape.
- Our solution serves as a stepping stone for future advancements in fake news detection technology.

DATASET

- The dataset consists of 44,897 news articles, with 23,481 articles classified as fake news and 21,416 articles classified as true news.
- The dataset covers a wide range of topics including politics, and other international news.
- The dataset aims to facilitate the detection and analysis of fake news
- Each news article is labeled as either fake or true, providing ground truth for training and evaluating classification models.
- With an equal distribution of fake and true news, the dataset provides a balanced representation of both classes, enabling reliable model training and evaluation.
- The inclusion of diverse news topics ensures the dataset's relevance and applicability across various domains and research areas.
- The dataset serves as a valuable resource for studying the characteristics, patterns, and linguistic cues of fake and true news articles.



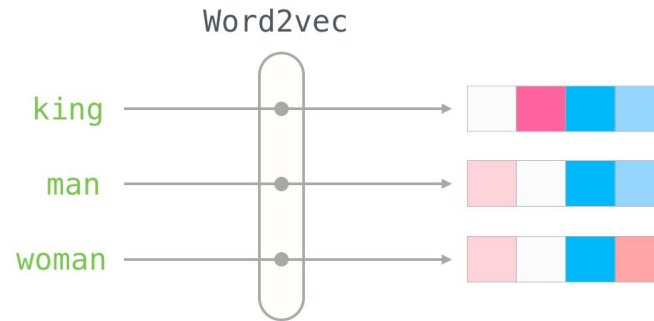


Preprocessing Text

- Dropped an assumed empty row from the 'true' DataFrame.
- Removed empty or blank texts from the 'fake' DataFrame.
- Combined the title and text columns, creating a unified 'text' column for each DataFrame.
- Converted all 'text' values to lowercase for consistency.
- Assigned class labels (1 for true, 0 for fake) to the respective DataFrames.
- Selected only the 'text' and 'class' columns for further analysis.
- Concatenated the 'true' and 'fake' DataFrames into a single DataFrame, 'data'.
- Shuffled the rows of 'data' for randomness and better training/testing.
- Prepared labels by extracting the 'class' column values into a list, 'y'.
- Split the text into individual words, storing them in a list called 'X'.

Word2Vec for Word Embeddings

- Word2Vec is an algorithm used for generating word embeddings, representing words as dense vectors.
- It captures semantic relationships and contextual similarities between words, enhancing natural language processing tasks.
- We utilized the gensim library to convert words to vectors using the Word2Vec model.





Steps for Word2vec

1. Extracted the 'class' column from the 'data' DataFrame into the 'y' variable.
2. Converted the text data into a list of lists, with each inner list representing a document.
3. Created a Word2Vec model using the 'X' list of documents, with a window size of 10 and a minimum word count of 1.
4. Stored the Word2Vec model in the 'w2v_model' variable.
5. Utilized the 'most_similar' function of the Word2Vec model to find the most similar words to 'india' based on learned word embeddings.



Tokenizer and word Embedding

- A tokenizer is a tool or library used to break down text into smaller units like words or sentences.
- It facilitates text preprocessing and enables further analysis by converting text into a manageable format.
- Tokenizers are commonly used in natural language processing tasks.
- Word embedding is the process of representing words or entities as dense vectors in a high-dimensional space.
- These vectors capture semantic relationships and enhance machine learning models' understanding of textual data.
- Word embeddings improve performance in tasks like language modeling, sentiment analysis, and information retrieval.



Steps

1. Create a Tokenizer and fit it on the 'X' data.
2. Convert the 'X' data to sequences using the Tokenizer's `texts_to_sequences()` method.
3. Create a numpy array, 'nos', to store the lengths of each sequence in 'X'.
4. Set the maximum sequence length to 1050 by filtering out longer sequences and using the `maxlen` parameter.
5. Pad the 'X' sequences using the `pad_sequences()` function to ensure uniform sequence length.
6. Determine the vocabulary size based on the length of the tokenizer's word index plus 1.
7. Define a function, 'get_weight_matrix', to create a weight matrix for word embeddings.
8. Iterate over the vocabulary and assign corresponding word vectors from the Word2Vec model to the weight matrix.
9. Obtain the weight matrix using the 'get_weight_matrix' function and store it in 'embedding_vectors'.



Creating model using LSTM in Deep learning

Long Short-Term Memory (LSTM) for Sequential Data

- LSTM is a recurrent neural network (RNN) architecture designed to model long-term dependencies in sequential data.
- It is commonly used in natural language processing tasks to capture context and sequential information.

Model Architecture:

- Our model architecture includes:
 - An embedding layer that takes in the vocabulary size, embedding dimension of 128, and maximum sequence length.
 - An LSTM layer with 128 units to learn long-term dependencies.
 - Dropout regularization layers to prevent overfitting.
 - A dense layer with a sigmoid activation function for binary classification.



Results

- The fake news detection model achieved an impressive overall accuracy of 99%.
- The model demonstrated high precision, recall, and F1-scores for both fake news (class 0) and real news (class 1).
- For class 0 (fake news):
 - Precision: 99%
 - Recall: 99%
 - F1-score: 99%
- For class 1 (real news):
 - Precision: 99%
 - Recall: 99%
 - F1-score: 99%

	precision	recall	f1-score	support
0	0.99	0.99	0.99	5783
1	0.99	0.99	0.99	5442
accuracy			0.99	11225
macro avg	0.99	0.99	0.99	11225
weighted avg	0.99	0.99	0.99	11225



Fake News Detection Web Application

Key Components:

- **Backend Code:** We utilized Flask, a lightweight and flexible Python web framework, to handle HTTP requests and perform computations.
- **Machine Learning Model:** We employed the Keras API for easy and fast prototyping of deep learning models. Our LSTM model was trained using this framework.
- **Model Storage:** The trained model is saved in the Hierarchical Data Format version 5 (HDF5 or .h5) file format, which efficiently stores large amounts of numerical data such as neural network weights and biases.

Web Application Functionality:

- The web application serves as an interactive interface for users to input news articles.
- The inputted articles are processed and evaluated by our LSTM model.
- The model predicts the likelihood of the news being fake or not.
- This showcases the practical application of LSTMs and deep learning in addressing real-world issues like misinformation.



Web app Home Page

News Classifier

Fake News Classification

Enter News below to check wheather it is Real or Fake:

Enter your text here...

Submit



Detection with input being real news

News Classifier

Results : News is Real!

News given:

(In second paragraph, corrects name of Strong's employer to Mental Health Department, not Public Health Department.) By Bernie Woodall (Reuters) - A man claiming to be the person who delivered a gift-wrapped package of horse manure at the Los Angeles home of U.S. Treasury Secretary Steven Mnuchin said on Monday he did it to protest the federal tax overhaul signed into law last week by President Donald Trump. Robert Strong, 45, a psychologist for the Los Angeles County Mental Health Department, said by telephone he left the poop-filled parcel addressed to Mnuchin and Trump in the driveway outside Mnuchin's home in the posh Bel Air community. KNBC-TV, an NBC television affiliate in Los Angeles, reported Mnuchin was not home at the time. The package was found by Mnuchin's neighbor. "Protest really should be funny," Strong told Reuters. "People's eyes glaze over when they just see angry people in the streets." He believes the new tax law will hurt poor people. Neither the U.S. Secret Service nor the Los Angeles Police Department, both of which investigated the incident, would confirm Strong was responsible. The Secret Service interviewed an individual who admitted delivering the package, but no charges had been filed against him as of Monday afternoon. LAPD Lieutenant Rob Weise said it was possible whoever left the package did not break any criminal laws. While he is not assigned to investigate the incident, Weise said if the box did not present any danger, it would not be illegal. The LAPD bomb squad X-rayed the box before opening it on Saturday. In a photo of the card Strong posted on Twitter, he wrote "Misters Mnuchin & Trump, We're returning the 'gift' of the Christmas tax bill" and signed it "Warmest wishes, The American people." Strong said a Secret Service agent, accompanied by six police officers, showed up at his house to question him on Sunday night, and the agent chided him, asking, "Are you ashamed of your behavior?" The White House declined to comment on Monday and officials with the Treasury Department could not be reached.



Detection with input being fake news

News Classifier

Results : News is Fake! Dont Trust it!!!!!!

News given:

Christmas Day message to rebuke Donald Trump without even mentioning his name. The Pope delivered his message just days after members of the United Nations condemned Trump's move to recognize Jerusalem as the capital of Israel. The Pontiff prayed on Monday for the peaceful coexistence of two states within mutually agreed and internationally recognized borders. We see Jesus in the children of the Middle East who continue to suffer because of growing tensions between Israelis and Palestinians, Francis said. On this festive day, let us ask the Lord for peace for Jerusalem and for all the Holy Land. Let us pray that the will to resume dialogue may prevail between the parties and that a negotiated solution can finally be reached. The Pope went on to plead for acceptance of refugees who have been forced from their homes, and that is an issue Trump continues to fight against. Francis used Jesus for which there was no place in the inn as an analogy. Today, as the winds of war are blowing in our world and an outdated model of development continues to produce human, societal and environmental decline, Christmas invites us to focus on the sign of the Child and to recognize him in the faces of little children, especially those for whom, like Jesus, there is no place in the inn, he said. Jesus knows well the pain of not being welcomed and how hard it is not to have a place to lay one's head, he added. May our hearts not be closed as they were in the homes of Bethlehem. The Pope said that Mary and Joseph were immigrants who struggled to find a safe place to stay in Bethlehem. They had to leave their people, their home, and their land, Francis said. This was no comfortable or easy journey for a young couple about to have a child. At heart, they were full of hope and expectation because of the child about to be born; yet their steps were weighed down by the uncertainties and dangers that attend those who have to leave their home behind. So many other footsteps are hidden in the footsteps of Joseph and Mary, Francis said Sunday. We see the tracks of entire families forced to set out in our own day. We see the tracks of millions of persons who do not choose to go away, but driven from their land, leave behind their dear ones. Amen to that. Photo by Christopher Furlong/Getty Images.



Limitations

Requirement for Large Labelled Dataset:

The LSTM-based approach requires a sufficiently large labelled dataset for effective model training.

Influence of Dataset Quality:

The quality of the dataset used for training the LSTM model plays a crucial role.

Impact of Hyperparameter Selection:

The performance of the LSTM model is influenced by the choice of hyperparameters.

Computational Resource Requirements:

Training an LSTM model with dropout regularization can be computationally intensive.



Improvements and Future Work

Improvements for LSTM-Based Fake News Detection

Explore LSTM Architecture Variations:

Experiment with different variations of the LSTM architecture, such as bidirectional LSTMs or stacked LSTMs.

Fine-Tune Hyperparameters:

Fine-tune hyperparameters like the learning rate, batch size, or dropout rate to optimize model performance.

Incorporate Additional Features:

Consider incorporating additional features alongside textual content, such as source credibility or social media engagement.

Explore Ensemble Methods:

Explore ensemble methods by combining multiple models or techniques.



References

- [1] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22-36.
- [2] Zhou, Y., Zafarani, R., Shu, K. (2018). Fake News Detection: A Deep Learning Approach. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp.797-804).
- [3] Gupta, S., Kumaraguru, P., Castillo, C., Meier, P. (2013). A User-Centric Approach to Fake News Detection on Twitter. In International Conference on Social Informatics (pp. 356-369).
- [4] Graves, A., Mohamed, A. R., Hinton, G. (2013). Speech Recognition with Deep Recurrent Neural Networks. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 6645-6649).
- [5] Sutskever, I., Vinyals, O., Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. In Advances in Neural Information Processing Systems (pp. 3104-3112).
- [6] Zhang, X., Zhao, J., LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. In Advances in Neural Information Processing Systems (pp. 649-657)



Thank You