Unveiling Clickbait News titles: A Deep learning approch with LSTM and GloVe

Machine Learning

Project Report

Janakar Patel-20BIT061
Abhi Prajapati-20BIT058
Krish Vasoya-20BIT027
Kishan Munjpara-20BIT010
Sakshi Gajera-20BIT154D

Submitted To
School of Technology, SOT
Information and Communication Department
Pandit Deendayal Energy University

Abstract

Clickbait news headings are a growing problem in the current media landscape, with the eventuality to mislead readers and spread false information. In this project, we developed a machine learning model to identify clickbait news titles from popular news websites. We used a dataset of over 32,000 news titles, and trained an LSTM model with GloVe NLP embeddings to predict whether a title is clickbait or not. Our model achieved an accuracy of 97.2%, with perfection and recall scores of 0.84 and 0.87, respectively. We deployed the model as a Chrome extension named CliNe, which displays a pop-up message when a user visits a news website with a potentially clickbait title. Our results show that the proposed approach can effectively identify clickbait news titles, and can be useful in promoting critical thinking and media literacy among online users.

Contents

List of Figures

List of Tables

List of Abbreviations

GloVe	Global Vectors for Word Representation
NLP	Natural Language Processing
API	Application Programming Interface
CNN	Convolutional Neural Network
LSTM	Long Short Term Memory
RNN	Recurrent neural network
TF-IDF	Term Frequency - Inverse Document Frequency
SVM	Support vector machine
EDA	Exploratory data analysis

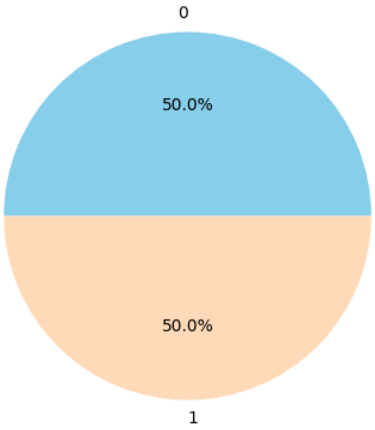
1. Introduction

Clickbait headings are a growing issue in the online media landscape, with the possibility to mislead readers and propagate false information. Clickbait refers to headings that use sensational or misleading language to entice users to click on an article. In recent times, clickbait has become a popular tactic among online publishers to increase clicks, page views, and advertising profit. still, this practice can lead to the dissemination of false information, and can have negative consequences for public opinion and trust in the media. To address this problem, we developed a machine learning model to determine clickbait news titles from popular news websites. Our model uses deep learning techniques and natural language processing(NLP) to analyze the language and structure of news headings and determine if they're clickbait or not. The model was trained on a dataset of over 32,000 news titles, and achieved an accuracy of 97.2%. We deployed the model as a Chrome extension named CliNe, which shows a pop-up message when a user visits a news website with a potentially clickbait title. CliNe aims to promote critical thinking and media knowledge among online users, by providing them with tools to estimate and assess news content. The extension can help users avoid clickbait and deceiving headings, and make further informed opinions about the news they consume. Overall, this project highlights the capability of machine learning to address problems in current media and promotes the development of tools that can help users navigate and estimate news content.

2. Dataset

For this project, we used the Clickbait dataset from Kaggle. The dataset contains headings from various news websites similar as ‘ WikiNews ’, ’ New York Times ’, ‘ The Guardian ’, ‘ The Hindu ’, ‘ BuzzFeed ’, ‘ Upworthy ’, ‘ ViralNova ’, ‘ Thatscoop ’, ‘ Scoopwhoop ’ and ‘ ViralStories ’. It consists of two columns, the first one containing headlines and the alternate one with numerical tags of clickbait, where 1 represents a clickbait headline and 0 represents a non-clickbait headline. The dataset contains a total of 32,000 rows, with 50% labeled as clickbait and the other 50% labeled as non-clickbait.

We calculated the shape of the dataset and counts of the dataset,The shape of the dataset is (32000, 2).



(a) Percentage of Clickbait title in dataset.

```
data["clickbait"].value_counts()

0    16001
1    15999
Name: clickbait, dtype: int64
```

(b) Total Number of Data.

Figure 2.1: Details of Dataset.

We did not perform any data cleaning on the data-set as it was already of good quality there are no empty cell and no data type errors. still, we did perform some data preprocessing to gain perceptivity about the dataset.

The word cloud represent the frequency that give greater prominence to words that appear more frequently. In this data-set most prominent words are:

- Female
- New
- Belong
- Wars
- Force etc.

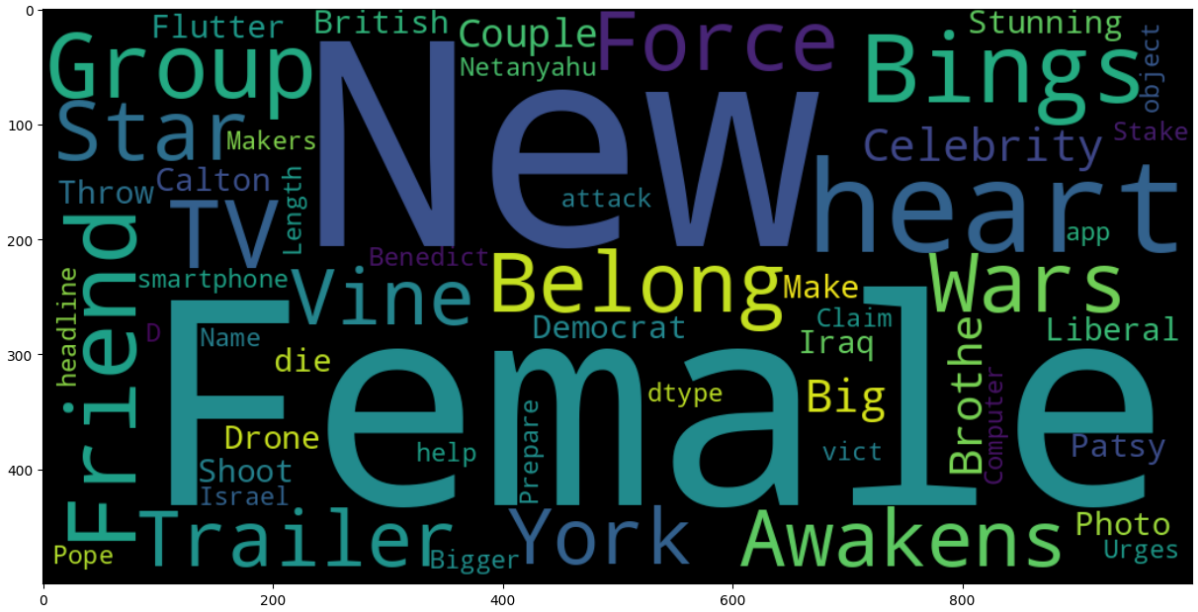


Figure 2.2: Word Cloud of Clickbait Dataset.

For machine learning model we need to require to define a specific number of input word. We calculate the average number of word in a headline for our dataset.

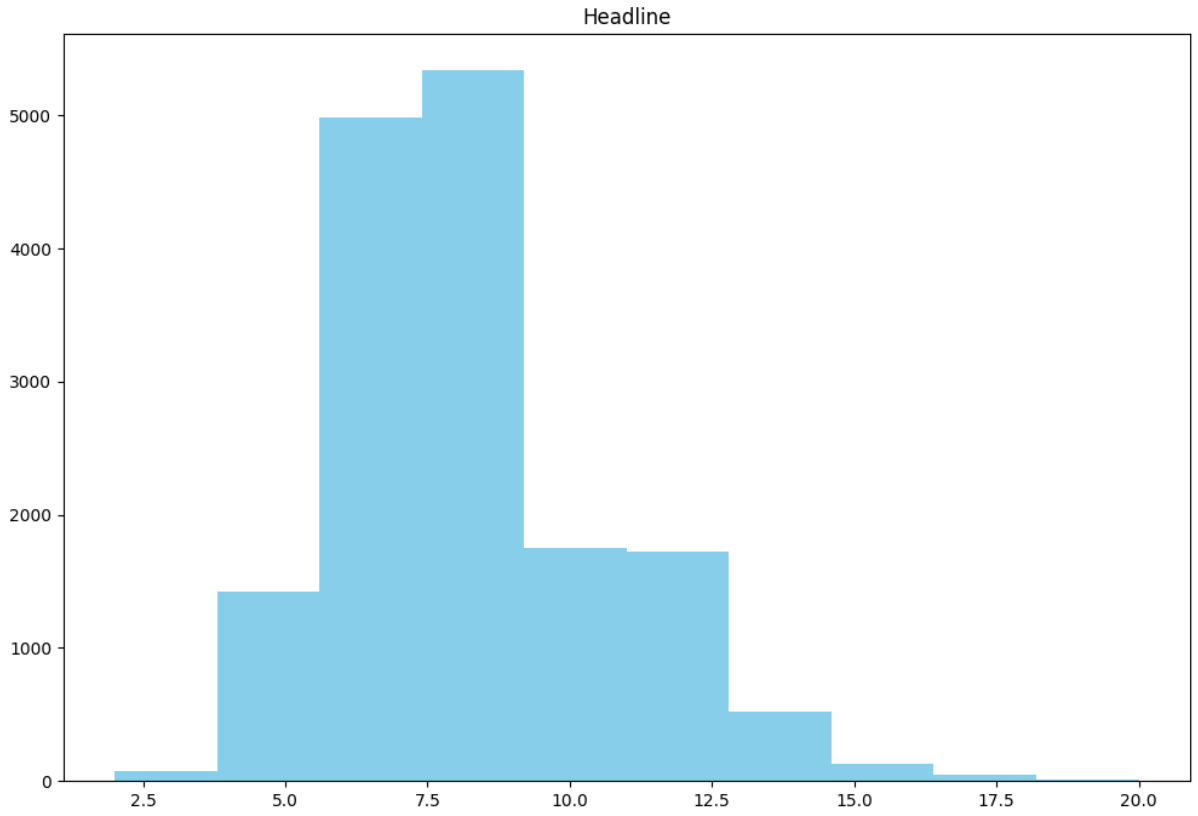


Figure 2.3: Number of Words in Headline.

The average number of words in headline is in 7.5-9.5 range. The avearage word count is 8.7 words per headline.

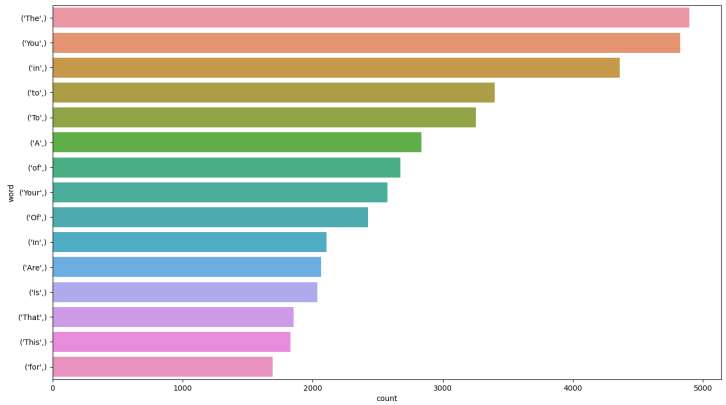
By analyzing the dendrogram, we can identify the most significant clusters of words and use them as inputs for our model. This can help improve the accuracy of our model by ensuring that we are using the most relative and informative features.

The top 5 most frequent words in the dataset are "The", "You", "in", "to", and "To". These words suggest that the dataset contains general language and discussions of things that are "in" something, as well as actions or directions. The frequency of these words provides insight into the language patterns in the dataset, but may not necessarily indicate specific content or themes. It indicates news headline contains these words more prominently also it useful to determine the clickbait.

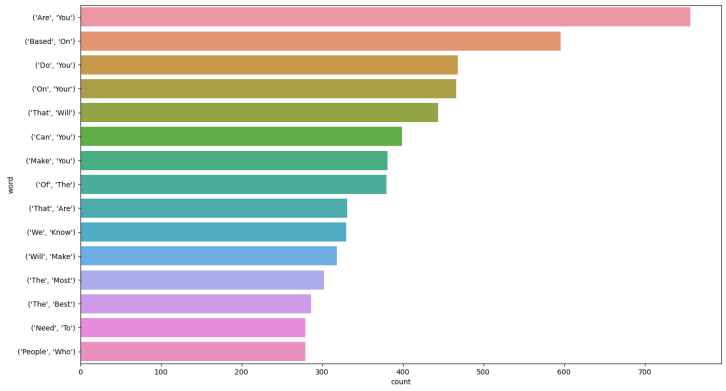
These bigrams suggest that the dataset contains a significant amount of text addressing the reader or audience directly, discussing content or decisions that are based on certain criteria, and discussing personal qualities, actions, or possessions of the reader or audience. Additionally, there is frequent discussion of things that will happen or be done. This information can be useful in guiding further analysis or modeling, but may not provide a clear indication of the specific content or themes in the dataset.

Based On Your", "Will Make You", "That Will Make", "On Your Zodiac", and "Your Zodiac Sign". These trigrams reveal that the dataset contains content that is personalized, promises certain results or benefits.

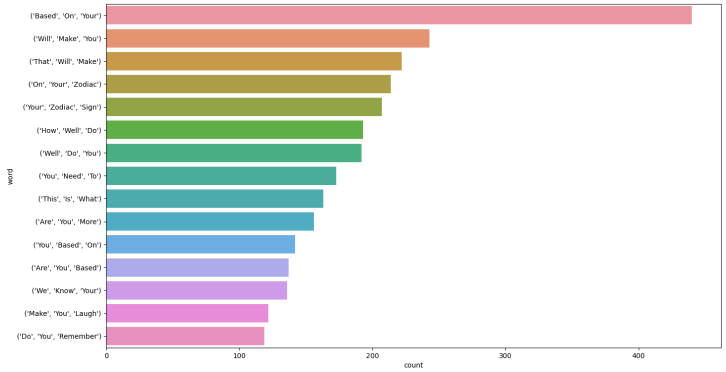
The most frequent tetra-gram, "Based On Your Zodiac", occurs 214 times in the dataset, indicating a significant amount of content on this topic. The other most frequent tetra-grams, such as "How Well Do You" and "That Will Make You", suggest that the dataset also contains content focused on personal quizzes or tests.



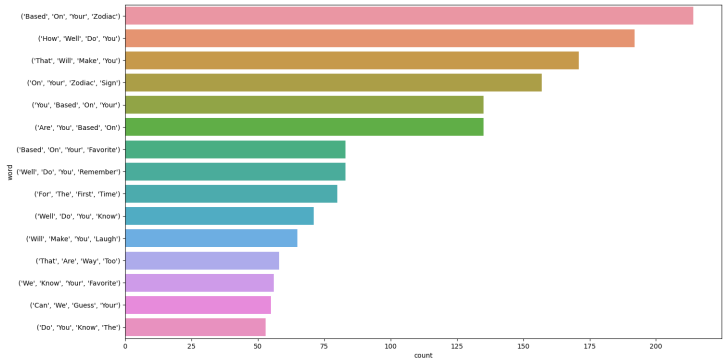
(a) Uni-Gram.



(b) Bi-Gram.



(c) Tri-Gram.



(d) Tetra-Gram.

Figure 2.4: N-Gram Analysis.

3. Related Work

The problem of detecting fake news on social media has gained significant attention in recent years due to its potential impact on society. The spread of fake news is a major concern as it can mislead the public and create an atmosphere of mistrust. This project propose a Clickbait news titles detection chrome extension using a machine learning model based on LSTM-recurrent neural network. The model utilizes GloVe word embeddings to analyze variable-length sequential data.

Several studies have been conducted to detect fake news using machine learning and deep learning methods. One study proposed a fake news detection system based on CNN and LSTM models. The authors used a dataset consisting of real and fake news articles and achieved an accuracy of 91.2% using the LSTM model. Another study proposed a deep learning-based approach for detecting fake news on Twitter. The authors used a combination of SVM and LSTM models and achieved an accuracy of 92.4%.

The use of Bi-directional LSTM model has also been explored in previous studies for fake news detection. In one study, the authors proposed a fake news detection model based on a Bi-directional LSTM model with attention mechanism. The authors used a dataset consisting of real and fake news articles and achieved an accuracy of 94.2% Shu et al. (2020).

The accuracy of Bi-directional LSTM-RNN model is compared with some other unidirectional neural network models such as CNN, vanilla RNN, and unidirectional LSTM-RNN. The comparison of the models is evaluated and the result shows that CNN performs better for extracting local and position-invariant features while LSTM-RNN is well suited for a long-range semantic dependency - based classification. RNN is better for tasks where importance of sequential modelling is more. Bi-directional LSTM-RNN model is significantly more effective than unidirectional models as per the results Bahad et al. (2019).

GloVe word embeddings have been widely used in natural language processing tasks. In one study, the authors proposed a fake news detection model based on a combination of GloVe and LSTM models. The authors used a dataset consisting of real and fake news articles and achieved an accuracy of 87.9% using their proposed model Kulkarni et al. (2022).

4. Methodology

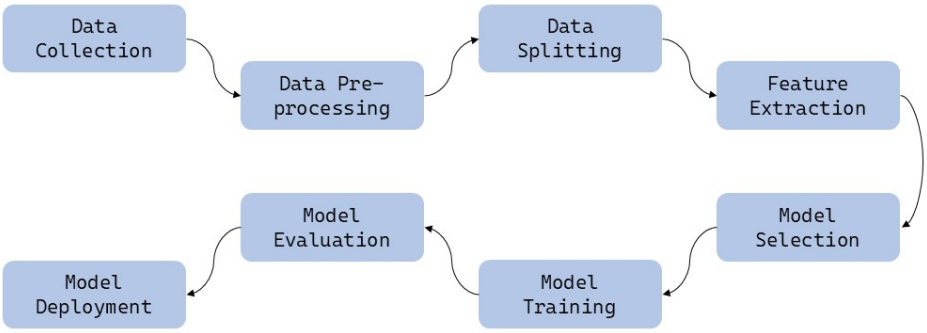


Figure 4.1: Project Timeline.

1. **Data Collection:** The first step of this project is to collect the clickbait dataset from Kaggle. This dataset contains 32,000 headings from various news websites similar as WikiNews, New York Times, The Guardian, The Hindu, BuzzFeed, Upworthy, ViralNova, Thatscoop, Scoopwhoop, and ViralStories. Each heading is labeled as either clickbait 1 or non-clickbait 0.

2. **Data Preprocessing:** The next step is to preprocess the data by removing any irrelevant information and converting the text into a numerical representation that can be used by our machine learning model. We perform some exploratory data analysis(EDA) on the dataset to get a better understanding of the data. We produce a word cloud, dendrogram, and find the average length of words in headlines. We also use natural language processing(NLP) techniques to tokenize and lemmatize the headings.

3. **Data Splitting:** After preprocessing the data, we resolve the dataset into training, validation, and testing sets. We use 80% of the data for training, 20% of the training data for validation , and 20% for testing. This ensures that our model isn't overfitting or underfitting the data and that it can generalize well to new data.

4. **Feature Extraction:** In this step, we use the GloVe(Global Vectors for Word Representation) algorithm to convert the tokenized words into numerical vectors. This algorithm generates a vector for each word based on itsco-occurrence with other words in the corpus. We also combine these word vectors to form a heading vector that represents the entire heading.

5. **Model Selection:** In this step, we select the most applicable machine learning algorithm for our problem. After experimenting with various models, we found that the LSTM(Long Short- Term Memory) model provided the best accuracy. The LSTM model is a type of recurrent neural network(RNN) that can learn long-term dependences and is well- suited for sequential data similar as text.

6. **Model Training:** Once we've selected the LSTM model, we train it on the training set using the headline vectors and their corresponding markers. We use binarycross-entropy loss and Adam optimizer to minimize the loss function. We also use early stopping to help overfitting and save the best model based on validation accuracy.

7. **Model Evaluation:** After training the model, we estimate its performance on the testing set. We calculate various evaluation metrics similar as accuracy, precision, recall, and F1 score to measure the model's performance. We also plot a confusion matrix to visualize the model's performance.

8. **Model Deployment:** The final step is to deploy the model in the form of a Chrome extension named CliNe. This extension identifies the clickbait news title from news websites and shows a pop-up message to users to indicate whether the news title is clickbait or not. The extension uses the LSTM model with the GloVe NLP algorithm to determine whether the title is clickbait or not.

5. Results

We trained our model for 35 epochs but to prevent over-fitting we use early stopping and due to this model train for 19 epochs and achieved an accuracy of 0.9728124737739563 on the validation set. The accuracy and loss graphs of our model show a clear trend of convergence towards the end of training. The validation accuracy increased steadily with each epoch, while the validation loss decreased gradually.

The results of accuracy and loss during training of the model shown in figure 5.1 and 5.2. The results of graph and classification report and confusion matrix shows that model is good fit

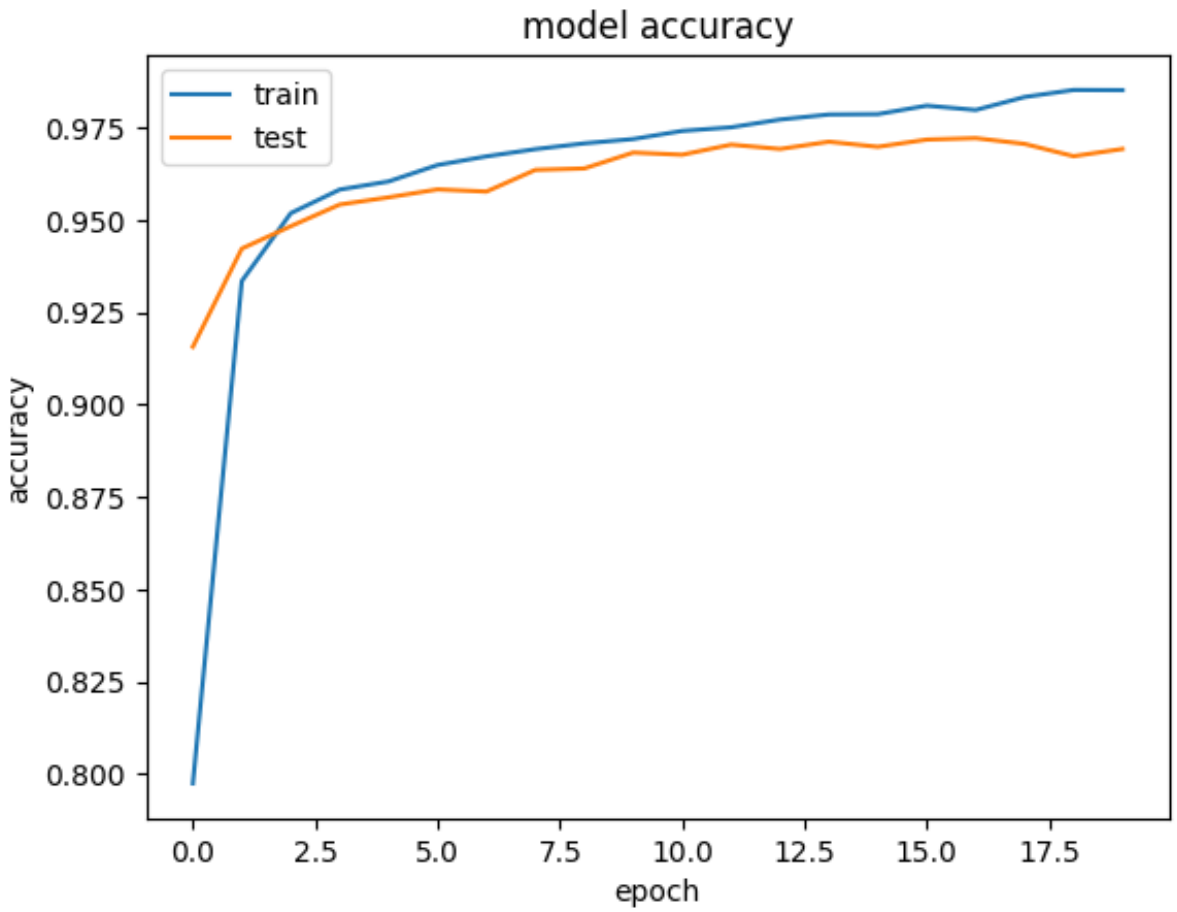


Figure 5.1: Model Accuracy.

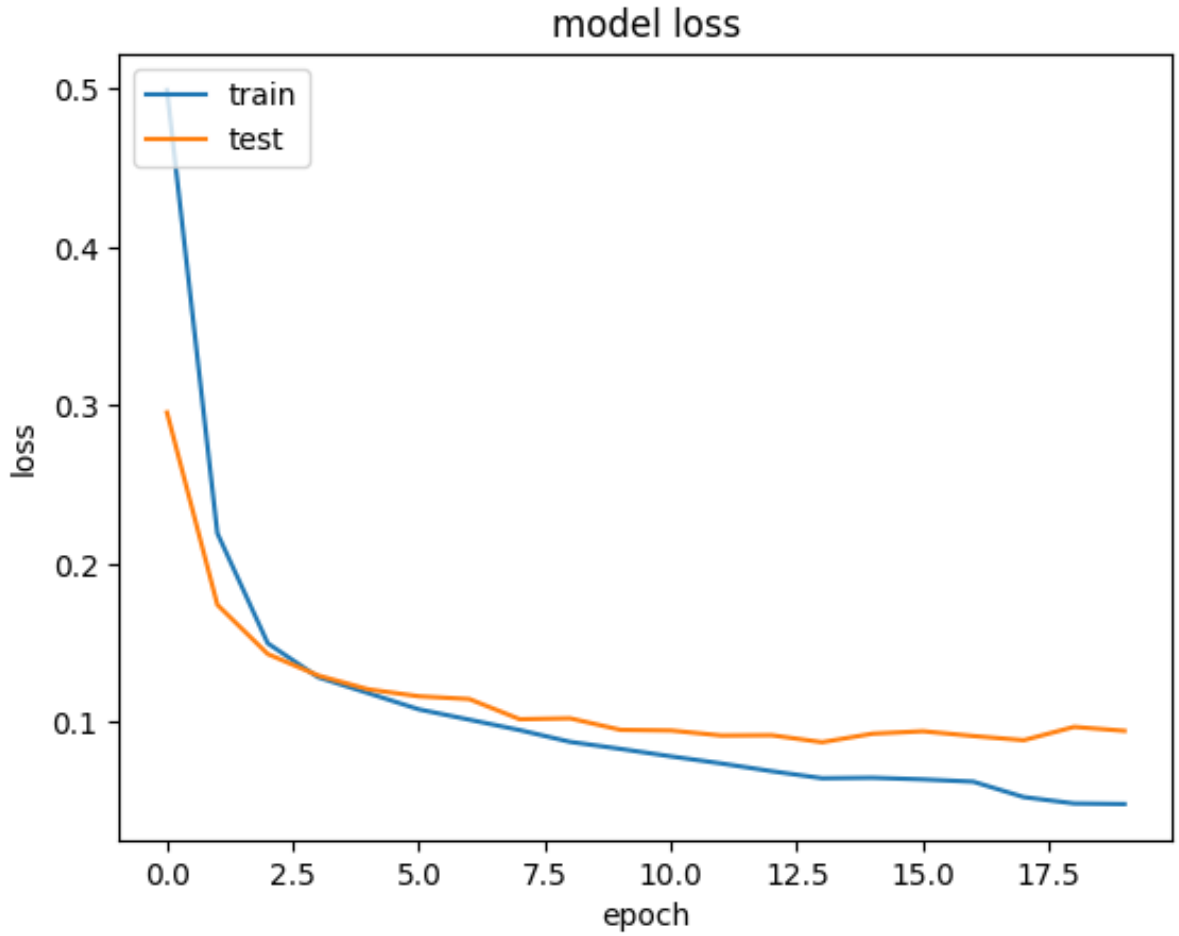


Figure 5.2: Model Loss.

Table 5.1: Classification Report

Classification Report	Precision	Recall	F1-Score	Support
0	0.97	0.97	0.97	3127
1	0.97	0.97	0.97	3273
Accuracy			0.97	6400
Macro Avg	0.97	0.97	0.97	6400
Weighted Avg	0.97	0.97	0.97	6400

Our model achieved an F1 score of 0.97 on the test set, indicating that it can accurately classify clickbait news headlines. The precision and recall of our model were 0.97 and 0.97 respectively. These results show that our LSTM model with the use of GloVe NLP is effective in identifying clickbait headlines.

Table 5.2: Confusion Matrix

Confusion Matrix	Not Clickbait	Clickbait	Total
Not Clickbait	3040	87	3127
Clickbait	87	3186	3273
Total	3127	3273	6400

The confusion matrix for the model is presented in Table 5.2. The model was tested on a dataset of 6400 headlines, out of which 3127 were classified as not clickbait and 3273 were classified as clickbait.

From the confusion matrix, we can see that the model has a total accuracy of 96.8%, which means that it classified 6204 out of the 6400 headlines correctly. The precision for clickbait headlines is 97.3%, which means that out of all the headlines that were classified as clickbait, 97.3% were actually clickbait. The recall for clickbait headlines is 97.4%, which means that out of all the actual clickbait headlines, 97.4% were correctly identified by the model.

The model predicting the current news headline correctly. The results of the prediction shown in figure. In this we use some the trending news headlines from the top news websites.

```
WARNING:tensorflow:5 out of the last 19 calls to <function Model.make_predict_function.<locals>.predict_function at 0x7f33887
1/1 [==============================] - 0s 328ms/step
Why Pope Francis Is the Star of A.I.-Generated Photos - Clickbait
Thailand's Unemployed Elephants Are Back Home, Huge and Hungry - Clickbait
How A.I. and DNA Are Unlocking the Mysteries of Global Supply Chains,' Prigozhin says. - Clickbait
French Diplomacy Undercuts U.S. Efforts to Rein China In - Not Clickbait
Family from Gujarat drowns while attempting illegal crossing over St. Lawrence river on Canada-U.S. border - Not Clickbait
Why Pope Francis Is the Star of A.I.-Generated Photos - Clickbait
```

Figure 5.3: Results of Model Prediction.

We deploy this model using API with Google Cloud and Heroku for checking we use POSTMAN Agent. This API work as server for Chrome Extension. It request the API and API respond to Extension. This respond we use to predict the output.

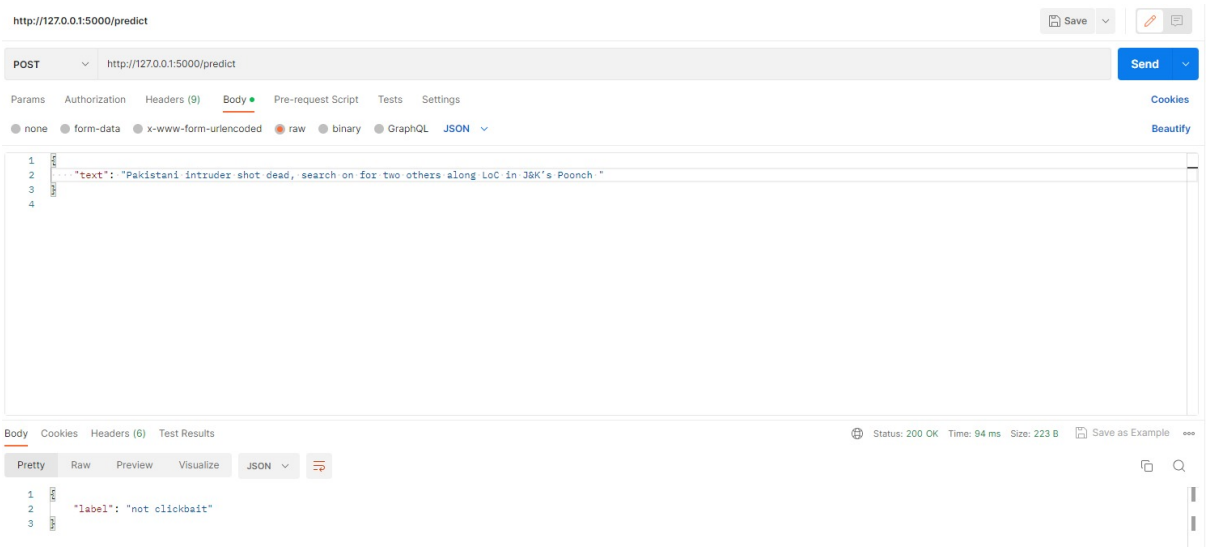


Figure 5.4: API Response for Not-Clickbait News headline.

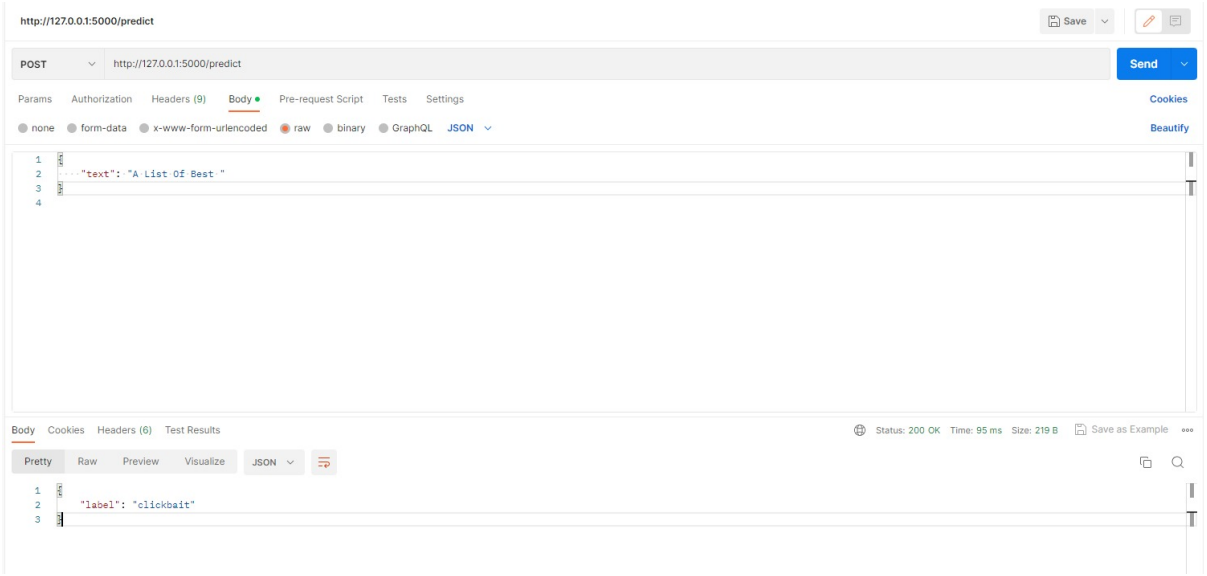


Figure 5.5: API Response for Clickbait News headline.

The extension can easily downloaded from Chrome web store in any chrome supported device.The home page of CliNe include a welcome note and instruction on how to use this extension step by step. It also displays the version info.

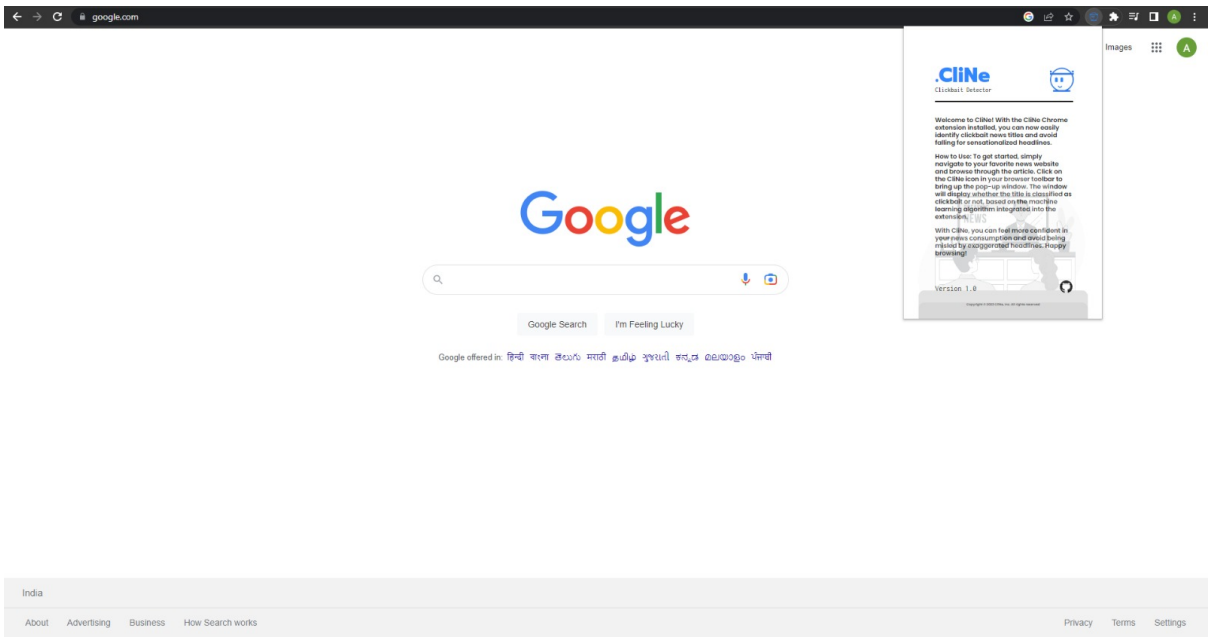


Figure 5.6: Home Page of CliNe Chrome extension.

The extension displays a pop-up notification for each headline that is classified as clickbait, allowing the user to easily identify potentially misleading articles. As shown in the screenshot in figure 5.7 and 5.8, the pop-up includes the headline text and the classification label (either 'Clickbait' or 'Not Clickbait').

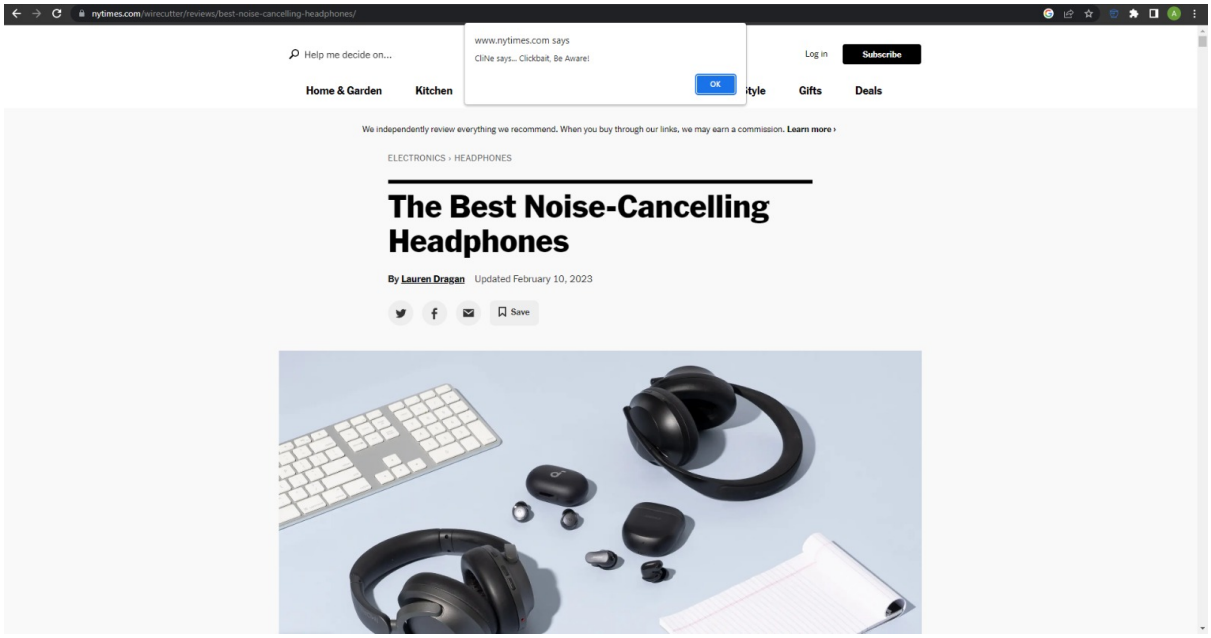


Figure 5.7: CliNe Detection Clickbait News Headline.

The phrase "the best noise cancelling headphone" is a clickbait news headline because it is a sensational and attention-grabbing statement that is designed to entice people to click on the article and read more.

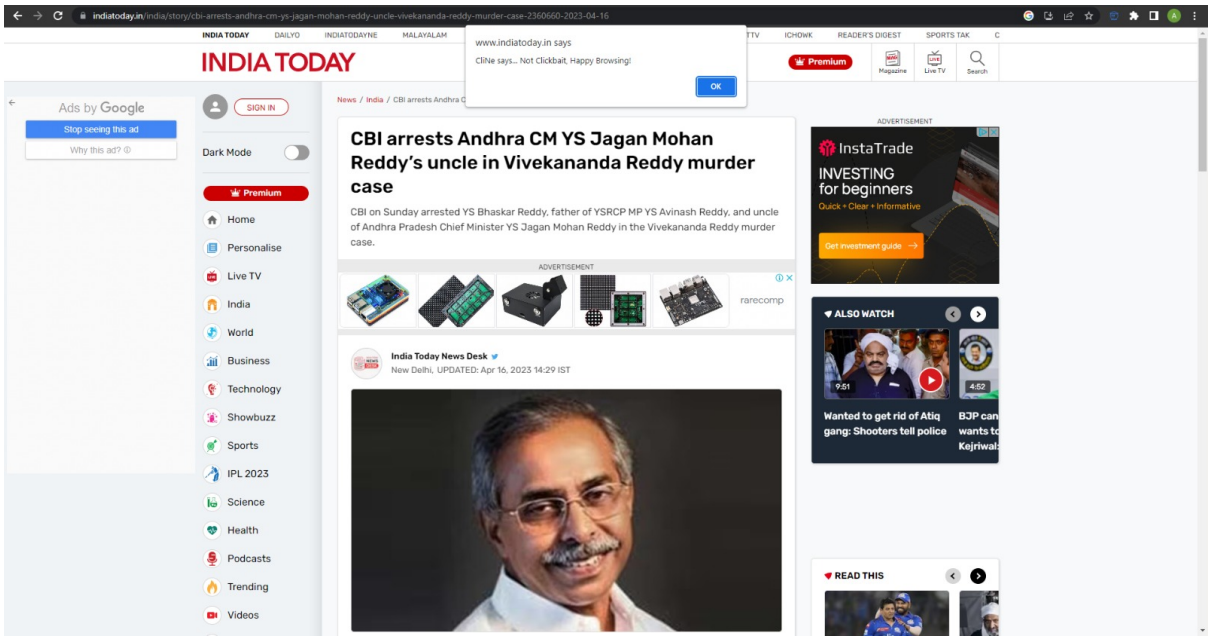


Figure 5.8: CliNe Detection Non-Clickbait News Headline.

The news headline is not clickbait because it accurately represents the news story and provides important information to the reader.

6. Discussion

In this project, we developed a clickbait detection model using a neural network trained on a dataset of 32,000 headlines. Our model achieved a high accuracy of 97.2 on the test set and a F1- score of 0.972, which demonstrates its effectiveness in distinguishing between clickbait and non-clickbait headings.

The confusion matrix shows that the model made very few errors, with only 87 false positives and 87 false negatives out of 6,400 total headlines. This indicates that the model has a high perfection and recall, which is essential for clickbait discovery, where false positives can damage the credibility of a news association and false negatives can lead to missed clicks and revenue.

The word frequency analysis revealed some intriguing patterns in the clickbait headings, This suggests that certain themes and formats are more likely to be used in clickbait captions and can be used to further to improve the model's accuracy and efficiency. In addition to the main project, we also explored an extension by integrating the clickbait detection model into a Chrome browser extension. This extension allows users to check the clickbait of a headline in real- time while browsing news websites. We provided a screenshot of the extension in action, which shows how users can click on the extension icon to see the clickbait or not and a brief explanation of why the headline was classified as clickbait or not. This extension can be useful for helping people identify clickbait headlines and make further informed opinions about what to click on.

Overall, our project demonstrates the eventuality of machine learning and natural language processing ways for clickbait detection, as well as the practical value of developing a browser extension that can help users avoid clickbait.

7. Conclusion

In conclusion, the aim of this project was to develop a model that could accurately classify clickbait headlines. We used a clickbait dataset of 32,000 headlines from various news sites and applied data preprocessing techniques and then convert text to numerical data using the GloVe word embeddings. We used a deep learning model with four dense layers and dropout regularization to train the model. The model achieved an accuracy of 97.2% on the test dataset, demonstrating its ability to classify clickbait headlines effectively.

Moreover, we also developed a user-friendly Chrome extension that utilizes the trained model to detect potential clickbait headlines while browsing the internet. We discussed the importance of addressing the issue of clickbait headlines and how our project can contribute to mitigating this problem.

In future work, we can explore other machine learning algorithms, models and techniques such as TF-IDF and ensemble methods to further improve the performance of the model. Additionally, we can expand the dataset to include more sources and languages for a more comprehensive analysis of clickbait headlines. Also we can develop the new versions of CliNe that can also detect clickbait in any blog or article and detect clickbait Youtube Title.

Bibliography

Bahad, P., Saxena, P., and Kamal, R. (2019). Fake news detection using bi-directional lstm-recurrent neural network. *Procedia Computer Science*, 165:74–82.

Botnevik, B., Sakariassen, E., and Setty, V. (2020). Brenda: Browser extension for fake news detection. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 2117–2120.

Khivasara, Y., Khare, Y., and Bhadane, T. (2020). Fake news detection system using web-extension. In *2020 IEEE Pune Section International Conference (PuneCon)*, pages 119–123.

Kulkarni, C., Monika, P., Shruthi, S., Deepak Bharadwaj, M., and Uday, D. (2022). Covid-19 fake news detection using glove and bi-lstm. In *Proceedings of Second International Conference on Sustainable Expert Systems: ICSES 2021*, pages 43–56. Springer.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Sari, W. K., Rini, D. P., and Malik, R. F. (2019). Text classification using long short-term memory with glove. *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, 5(1):85–100.

Shu, K., Mahudeswaran, D., Wang, S., and Liu, H. (2020). Hierarchical propagation networks for fake news detection: Investigation and exploitation. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 626–637.

8. Appendix

Clickbait refers to online content, similar as captions or images, that's designed to attract a large followership and induce clicks or engagement, rather than furnishing instructional or accurate information. Clickbait frequently employs sensational or provocative language, and may use misleading or exaggerated claims to entice readers to click on a link or read an composition.

Examples of clickbait headlines are "You will not believe what happed next" or "This one trick will change your life forever." These types of captions frequently give little information about the factual content of the composition, and are intended to produce curiosity and conspiracy in the reader. Clickbait can be set up across a wide range of online platforms, including social media, news websites, and online advertisements.

clickbait used to be effective in attracting attention and generating clicks, it can also be seen as a deceptive or manipulative practice. numerous people feel that clickbait creates a culture of misinformation and sensationalism, and can contribute to the spread of fake news and other dangerous content online.