FakeTweet Finder: Deep Learning Detection

GROUP 2

ABDUL MANNAN F (CEC21CS003)

ACHAL V V (CEC21CS006)

ADVAITH B (CEC21CS010)

GOPINANDAN P S (CEC21CS048)

Guided By:

ASWATHY PARAPPURAM

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

COLLEGE OF ENGINEERING CHERTHALA

JANUARY 08 , 2024

Overview

1. INTRODUCTION
2. PROBLEM STATEMENT
3. OBJECTIVE
4. LITERATURE SURVEY
5. PRODUCT FUNCTIONS
6. SOFTWARE REQUIREMENTS
7. HARDWARE REQUIREMENTS
8. DESIGN
9. IMPLEMENTATION
10. CONCLUSION
11. REFERENCES

Introduction

- The rise of advanced text generation models has made it increasingly difficult to distinguish between human and machine-generated content.
- Social media platforms are particularly vulnerable to the spread of misinformation through deepfake text.
- Effective detection mechanisms are crucial to maintain the integrity of information shared online.

PROBLEM STATEMENT

- To develop a framework to reliably distinguish between human- and bot-generated tweets, helping to protect the integrity of information on social media."

OBJECTIVE

- Robust Tweet Classification
- Real-Time Detection
- Model Training and Fine-Tuning
- User Feedback System

LITERATURE SURVEY 1

1.TweepFake: About detecting deepfake tweets
Author: Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella,Maurizio Tesconi.
Year of Publication: 2021

- **Methods Used:** The dataset includes various generation techniques like Markov Chains, RNN, LSTM, and GPT-2.
- **Advantages:**
  1.Unique Dataset: First dataset of real deepfake tweets posted on Twitter.
  2.Publicly Available: Accessible dataset and code, encouraging further research.

LITERATURE SURVEY 1(CONTD..)

- **Disadvantages:**
  1.Short Texts: Focus on tweets may limit applicability to longer texts.
  2.Limited Techniques: Not all text generation methods are covered.
- **Future Scope:**
  1.Dataset Expansion: Include more models and languages.
  2.Cross-Platform Analysis: Extend to other social media platforms.

LITERATURE SURVEY 2

2. No Rumours Please! A Multi-Indic-Lingual Approach for COVID Fake-Tweet Detection
Author: Debanjana Kar,Mohit Bhardwaj,Suranjana Samanta,Amar Prakash Azad.
Year of Publication: 2020

- **Methods Used:**
  mBERT (Multilingual BERT) is a transformer-based model for feature extraction, providing contextual embeddings across languages to help the system analyze tweets in their linguistic context.
- This paper aims to tackle misinformation in Indic languages by developing a system to detect fake news in COVID-19-related tweets, supporting a diverse linguistic population.

LITERATURE SURVEY 2(CONTD..)

- **Advantages:**
  1.Multilingual Capability: This approach targets multiple Indic languages, ensuring inclusivity and a broader understanding of misinformation among diverse linguistic groups in India.
  2.Localized Detection: Detecting fake news in local languages enhances accuracy and relevance by considering cultural and regional contexts often overlooked by models trained only on English or major languages.
- **Disadvantages:**
  1.Computational Cost: Training and maintaining a multi-lingual model can be resource-intensive.
  2.Data Scarcity: For some Indic languages, there may be limited labeled data available for training and evaluating fake-tweet detection models.

LITERATURE SURVEY 2(CONTD..)

- **Future Scope:**
  1.Language Expansion: Include more Indic languages, dialects, and potentially other regional languages globally.
  2.Advanced Models: Utilize newer NLP models and cross-lingual transfer learning for improved performance.
  3.Data Improvement: Enhance data collection methods, including crowdsourcing and dynamic updates for better accuracy.
  4.Policy and Impact: Align with regulations, support public education on misinformation, and explore international applications.

LITERATURE SURVEY 3

3. A Framework for Hate Speech Detection Using Deep Convolutional Neural Network
Author: Pradeep Kumar Roy,Asis Roy Tripathy,Tapan Kumar Das.
Year of Publication: 2020

- **Methods Used:** DCNN and Other methods used for comparison LSTM and other traditional machine learning models
- A Framework for Hate Speech Detection Using Deep Convolutional Neural Network," proposes a Deep Convolutional Neural Network (DCNN) framework for detecting hate speech on Twitter. The model automatically extracts features from tweet text using convolutional layers, reducing the need for manual feature engineering.

LITERATURE SURVEY 3(CONTD..)

- **Advantages:**
  1.No Manual Feature Engineering: The model automatically extracts features, reducing the need for manual input.
  2.Handles Imbalanced Data: The model effectively deals with imbalanced datasets, which is common in hate speech detection.
- **Disadvantages:**
  1.Complexity: The deep learning model is computationally intensive and requires significant resources.
  2.Language Limitation: The model was trained on English tweets, limiting its applicability to other languages.
- **Future Scope:**
  1.Multilingual and Cross-lingual Models: Expanding the framework to support multiple languages and dialects can enhance its applicability globally.

LITERATURE SURVEY 4

4.A Framework to detect fake tweet images on social media
Author: Shivam B.Parikh,Saurin R.Khedia,Pradeep K.Atrey.
Year of Publication: 2019

- **Methods Used:**
  Text Extraction: Uses Microsoft Azure's Computer Vision to extract text (name, username, content, timestamp) from tweet images.
  Data Verification: Compares extracted text with data from Twitter's API to detect discrepancies.
  Tampering Detection: Flags tweet images as tampered if there are mismatches in content, username, or timestamp.

LITERATURE SURVEY 4(CONTD..)

- The paper proposes a framework for detecting tampered tweet images shared across various social media platforms. The framework focuses on identifying fake or altered screenshots of tweets that may be used to spread misinformation. It validates the proposed framework using a self-collected dataset of real and tampered tweet images. The dataset includes 100 tweet screen captures, comprising 50 real and 50 fake tweets, and covers multiple types of tampering.
- **Advantages:**
  1.High Accuracy: 83.33 percentage success rate in detecting fake tweet images.
  2.Automated: Reduces manual verification through automated tools.

LITERATURE SURVEY 4(CONTD..)

- **Disadvantages:**
  1.Service Dependency: Relies on external services like Azure and Twitter API.
  2.False Positives: Can incorrectly flag tweets due to deletions or changes in user information.
- **Future Scope:**
  1.Broader Research Use: By publishing the dataset, other researchers can use it to test and improve detection methods.
  2.Collaboration Across Platforms: Develop tools that work across different social media platforms to stop the spread of tampered tweets.

LITERATURE SURVEY 5

5.The Detection of Fake Messages using Machine Learning
Author: Maarten S. Looijenga .
Year of Publication: 2019

- **Methods Used:**
  Data Collection: Utilizes a Twitter dataset from the Dutch 2012 election, focusing on tweets with relevant hashtags.
  Machine Learning Classifiers: Trains eight different supervised machine learning classifiers, including Decision Tree, Naïve Bayes, and Random Forest, using a manually labeled dataset of 300 tweets.
  Preprocessing: Applies data cleaning, tokenization, and TF-IDF vectorization to convert tweet text into numerical form for analysis.

LITERATURE SURVEY 5(CONTD..)

- Investigates the use of fake messages on Twitter during the Dutch 2012 election. Develops and compares eight machine learning classifiers to detect fake tweets, finding that the Decision Tree algorithm performs best with an F-Score of 88
- **Advantages:**
  1.Provides a detailed comparison of multiple machine learning algorithms for detecting fake tweets.
  2.Achieves high accuracy with the Decision Tree classifier.
  3.Focuses on a specific real-world event, making the findings contextually relevant.

LITERATURE SURVEY 5(CONTD..)

- **Disadvantages:**
  1.Limited to textual content; does not consider user account data due to the age of the dataset.
  2.Relies on a small sample size (300 tweets) for training, which may not generalize well to other contexts.
  3.Dataset is outdated, which limits access to additional user information for more comprehensive analysis.
- **Future Scope:**
  1.Integration of More Data Sources: Combine text analysis with user account data to create a more accurate detection model.
  2.Exploration of New Features: Include additional tweet characteristics, such as sentiment or how often they are shared, to improve detection.

LITERATURE SURVEY 6

6.Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection
Author: Hajime Watanabe,Mondher Bouazizi,Tomoaki Ohtsuki.
Year of Publication: 2018

- Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," presents a methodology combining sentiment-based, semantic, unigram, and pattern features to detect hate speech on Twitter.
- **Advantages:**
  1.Pragmatic Method: Offers a practical approach to collecting hate speech patterns for future use.

LITERATURE SURVEY 6(CONTD..)

2.Comprehensive Features: Combines sentiment, semantic, unigram, and pattern features for robust hate speech detection.

Disavantages:

1.Classification Challenges: Struggles with distinguishing between hateful and merely offensive content

2.Surface-level Analysis: Relies on basic features, potentially missing deeper context.

Future Scope:

1.User Behavior Analysis: Study the impact of hate speech on users to inform better intervention strategies.

LITERATURE SURVEY 7

7.Weakly Supervised Learning for Fake News Detection on Twitter
Author: Stefan Helmstetter,Heiko Paulheim.
Year of Publication: 2018

- **Methods Used:**
  1.BOW: BOW is used to convert tweets into vectors, which can then be analyzed by machine learning algorithms to determine if the content is fake or real.
  2. Doc2Vec Model:Doc2Vec modelis an advanced technique for text representation. Unlike BOW, Doc2Vec captures the semantic meaning of documents by considering the context of words within the document.

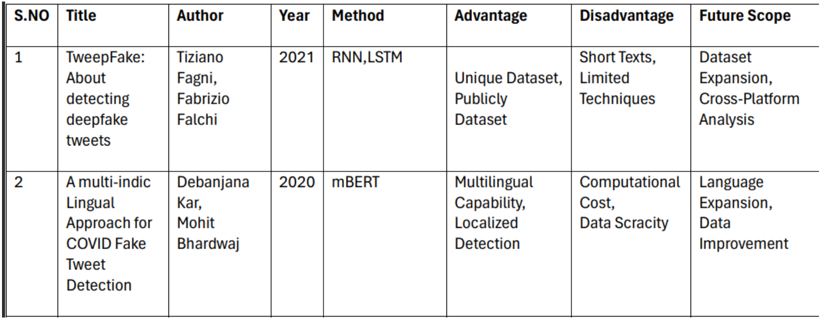
LITERATURE SURVEY 7(CONTD..)

- The paper suggests that while traditional models like BOW are still in use, more advanced techniques like Doc2Vec are being adopted to improve the detection of fake news on platforms like Twitter. These advanced models help in better understanding the context and semantics of the text, which is crucial for accurately identifying misleading or false information.
- **Advantages:**
  1.Scalability: Automatically collects a large dataset from Twitter, enabling effective training without extensive manual labeling.
  2.Adaptability: Easily updates to reflect new trends and changes in Twitter content.

LITERATURE SURVEY 7(CONTD..)

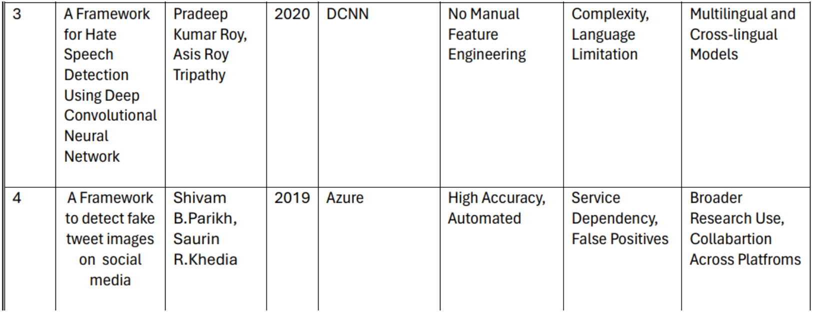
- **Disadvantages:**
  1.Feature Dependency: Performance drops if user-level features are unavailable, such as with new accounts.
  2.Real-World Uncertainty: May not perform as well in diverse, real-world scenarios.
- **Future Scope:**
  1.Real-Time Detection: Developing methods to apply this approach in real-time, especially for emerging news sources and accounts, would make it more practical for live fake news detection.
  2.Cross-Platform Application: Extending the approach to detect fake news across multiple social media platforms (e.g., Facebook, Instagram) could increase its utility and robustness.

LITERATURE SURVEY



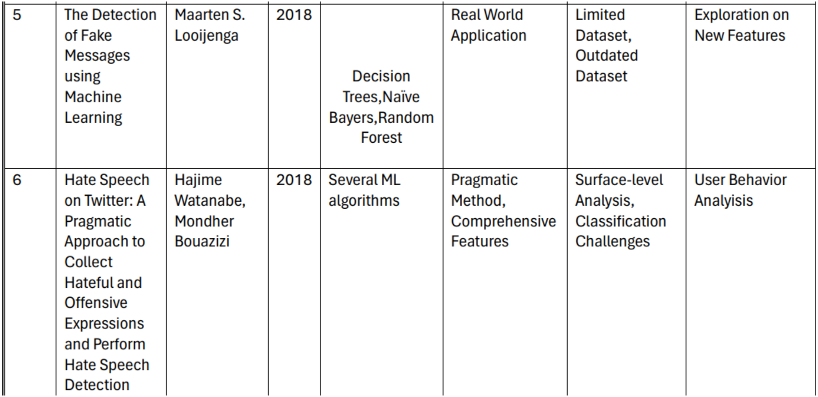
S.NO	Title	Author	Year	Method	Advantage	Disadvantage	Future Scope
1	TweepFake: About detecting deepfake tweets	Tiziano Fagni, Fabrizio Falchi	2021	RNN,LSTM	Unique Dataset, Publicly Dataset	Short Texts, Limited Techniques	Dataset Expansion, Cross-Platform Analysis
2	A multi-indic Lingual Approach for COVID Fake Tweet Detection	Debanjana Kar, Mohit Bhardwaj	2020	mBERT	Multilingual Capability, Localized Detection	Computational Cost, Data Scracity	Language Expansion, Data Improvement

LITERATURE SURVEY



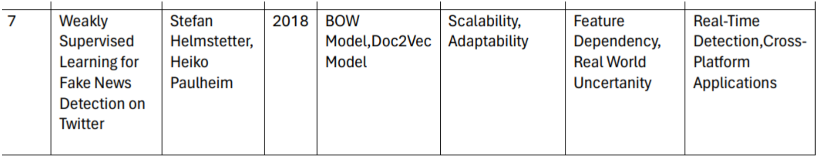
3	A Framework for Hate Speech Detection Using Deep Convolutional Neural Network	Pradeep Kumar Roy, Asis Roy Tripathy	2020	DCNN	No Manual Feature Engineering	Complexity, Language Limitation	Multilingual and Cross-lingual Models
4	A Framework to detect fake tweet images on social media	Shivam B.Parikh, Saurin R.Khedia	2019	Azure	High Accuracy, Automated	Service Dependency, False Positives	Broader Research Use, Collabartion Across Platfroms

LITERATURE SURVEY



5	The Detection of Fake Messages using Machine Learning	Maarten S. Looijenga	2018	Decision Trees,Naïve Bayers,Random Forest	Real World Application	Limited Dataset, Outdated Dataset	Exploration on New Features
6	Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection	Hajime Watanabe, Mondher Bouazizi	2018	Several ML algorithms	Pragmatic Method, Comprehensive Features	Surface-level Analysis, Classification Challenges	User Behavior Analyisis

LITERATURE SURVEY



7	Weakly Supervised Learning for Fake News Detection on Twitter	Stefan Helmstetter, Heiko Paulheim	2018	BOW Model,Doc2Vec Model	Scalability, Adaptability	Feature Dependency, Real World Uncertanity	Real-Time Detection,Cross-Platform Applications

Conclusion from Literature Survey

The above paper mostly consist of :

- Language diversity, especially with multilingual datasets, remains a critical area. Models such as mBERT address this by enabling fake content detection across multiple languages, but resource intensiveness and data scarcity in some languages pose limitations.
- Some models are designed to scale efficiently, with potential to be extended across various social media platforms, enhancing their utility in combating misinformation.
- Some frameworks rely on external services, like APIs and cloud services, which can introduce dependencies and limitations, especially if these services experience downtime or changes in availability.

PRODUCT FUNCTIONS

- FakeTweet Finder will accurately classify tweets, distinguishing between human- and machine-generated content
- It will support batch processing to analyze large volumes of tweets in formats like CSV or JSON, making it adaptable for high-frequency social media data.
- Users will receive notifications when potential misinformation or suspicious tweet activity is detected, with seamless integration for continuous, automated analysis using social media monitoring tools.
- Enable users to retrain or fine-tune the model on new datasets, making the tool adaptable to the latest advancements in AI-generated text and evolving social media trends.

Software Requirements

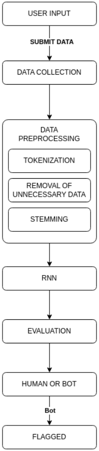
- **Programming Language:** Python, due to its rich ecosystem for machine learning and deep learning.
- **Framework:** TensorFlow, Keras, scikit-learn, FastText
- **Deployment Tools:** Flask

Hardware Requirements

- **Processor:** Minimum Intel Core i5 or AMD Ryzen 5
- **GPU:** NVIDIA RTX 3060 or higher (e.g., RTX 3090, A100, or Tesla GPUs for larger datasets)
- **Memory(RAM):** Minimum 8 GB
- **Storage:** 256 GB SSD

DESIGN

Architecture Diagram



Activity Diagram

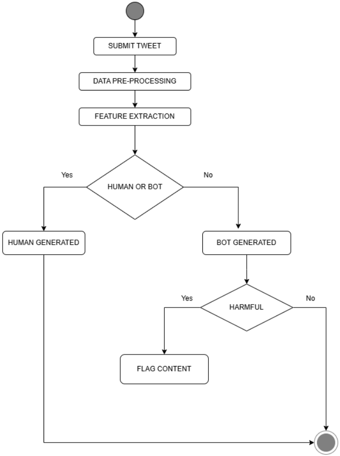


Fig : Activity Diagram

DFD Level 0 Diagram

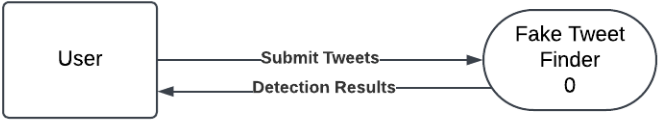


Fig : DFD Level 0 Diagram

DFD Level 1 Diagram

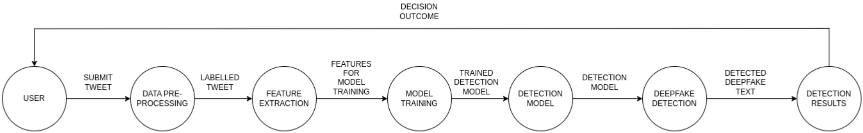


Fig : DFD Level 1 Diagram

Gantt Chart

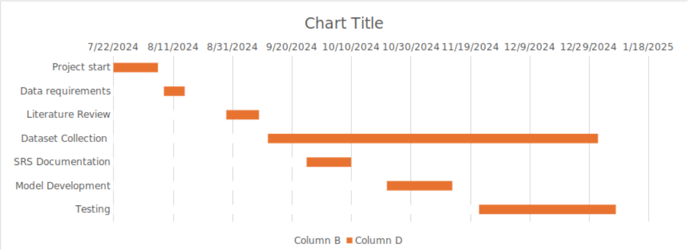


Fig : Gantt Chart

COST ESTIMATION

Overview

- Cost Estimation is performed using the COCOMO Model, which helps predict:
  - **Effort** required to complete the project.
  - **Time (Months)** needed for development.
  - **Cost** based on estimated effort and developer salary.
- The project is classified as **Organic** due to relatively simple project complexity, involving basic development tasks.

COST ESTIMATION

Input Parameters

- Lines Of Code (LOC) = ˜3000
- Project Type = Organic
- Salary per month = 50,000

COCOMO Model Formulas

$$\text{Effort} = a \cdot (\text{KLOC})^b \quad \text{PM}$$

$$\text{Time} = c \cdot (E)^d$$

$$\text{Cost} = \text{Effort} \cdot \text{Salary}$$

where $a = 2.4$, $b = 1.05$, $c = 2.5$, $d = 0.38$ for Organic projects.

COST ESTIMATION

KLOC (Kilo Lines of Code)

$$\text{KLOC} = \frac{3000}{1000} = 3.0$$

Effort (E) in Person-Months (PM)

$$\text{Effort} \approx 2.4 \times (3.0)^{1.05} \approx 7.59 \text{ person-months}$$

Time (T) in Months

$$\text{Time} = 2.5 \times (7.59)^{0.38} \approx 2.5 \times 1.86 \approx 4.65 \text{ months}$$

Cost (C)

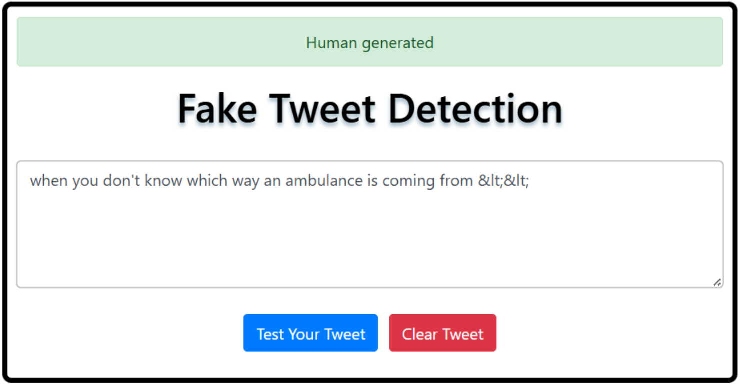
$$\text{Cost} = 7.59 \times 50,000 = 379,500$$

COST ESTIMATION

Final Estimates:

- Effort: 7.59 Person-Months
- Time: ˜4.65 Months
- Cost: 379,500

Implementation

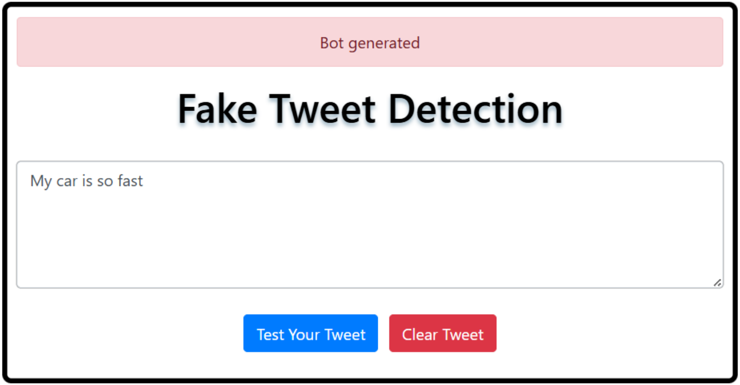


Human generated

Fake Tweet Detection

when you don't know which way an ambulance is coming from &lt;&lt;

Implementation



Bot generated

Fake Tweet Detection

My car is so fast

Fig :BOT GENERATED

Conclusion

- The proposed model demonstrates high accuracy in identifying machine-generated tweets.
- This method shows promise for real-world applications, offering a robust solution to the problem of detecting deepfake texts on social media.
- Future work includes further enhancing the model's performance and adapting it to detect deepfake texts in other languages and formats.

REFERENCES

[1] S. Sadiq, T. Aljrees and S. Ullah, ”Deepfake Detection on Social Media: Leveraging Deep Learning and FastText Embeddings for Identifying Machine-Generated Tweets,” in IEEE Access, vol. 11, pp. 95008-95021, 2023, doi: 10.1109/ACCESS.2023.3308515.

[2] Fagni T, Falchi F, Gambini M, Martella A, Tesconi M (2021) TweepFake: About detecting deepfake tweets. PLoS ONE 16(5): e0251415. https://doi.org/10.1371/journal.pone.0251415

[3] D. Kar, M. Bhardwaj, S. Samanta and A. P. Azad, ”No Rumours Please! A Multi-Indic-Lingual Approach for COVID Fake-Tweet Detection,” 2021 Grace Hopper Celebration India (GHCI), Bangalore, India, 2021, pp. 1-5, doi: 10.1109/GHCI50508.2021.9514012.

REFERENCES

[4] P. K. Roy, A. K. Tripathy, T. K. Das and X. -Z. Gao, "A Framework for Hate Speech Detection Using Deep Convolutional Neural Network," in IEEE Access, vol. 8, pp. 204951-204962, 2020, doi: 10.1109/ACCESS.2020.3037073.

[5] S. B. Parikh, S. R. Khedia and P. K. Atrey, "A Framework to Detect Fake Tweet Images on Social Media," 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), Singapore, 2019, pp. 104-110, doi: 10.1109/BigMM.2019.00-37.

[6] Looijenga, M.S. (2018) The Detection of Fake Messages using Machine Learning.

REFERENCES

[7] H. Watanabe, M. Bouazizi and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," in IEEE Access, vol. 6, pp. 13825-13835, 2018, doi: 10.1109/ACCESS.2018.2806394.

[8] S. Helmstetter and H. Paulheim, "Weakly Supervised Learning for Fake News Detection on Twitter," 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, Spain, 2018, pp. 274-277, doi: 10.1109/ASONAM.2018.8508520.

[9] https://chatgpt.com/

THANK YOU