Unveiling Clickbait and Unmasking Fake News: A Multimodal Analysis Leveraging User Influence on Twitter

- Presented by
Anurag (13000120103)
Abhinab Paul (13000120107)
Uttkarsh Ranjan (13000120109)
Priyasu Guin (13000120117)
- Under the guidance of Prof. Poulami Dutta

Contents

Motivation

The motivation behind our project are :

1. Rising threat of misinformation and fake content
   - In the digital age, the swift dissemination of misinformation and fake content poses a substantial threat to public discourse and decision-making. It can lead to widespread panic, fear, and confusion, impacting public order.
2. Technological advancements in content manipulation
   - Advancements in technology, notably in deepfake and AI-generated content, empower malicious individuals to craft incredibly convincing fake multimedia. This poses a significant threat, as such technology also makes it easier to generate a lot of fake media very quickly.
3. Protecting users against phishing
   - Phishing attacks pose a constant threat to individuals, leading to identity theft, financial losses, and unauthorized access to personal information. Businesses incur financial losses due to fraud, loss of customer trust, and expenses associated with remediation.

Introduction

- In the age of information, the pervasive influence of social media has granted unprecedented reach to news and content dissemination. However, this democratization of information comes with the looming threat of fake news and malicious activities such as phishing, jeopardizing the integrity of online discourse.
- To counteract these challenges, our project endeavors to develop an advanced system for detecting and mitigating fake news spread on social media platforms. Beyond the conventional realms of misinformation, our solution extends its vigilance to the identification of phishing links embedded within social media posts.

Problem Statement

- Train a model which can identify and categorize false news that are spread through the social media platform Twitter (now X) using various forms of media (text, images, videos, audios).
- The model must also detect any malicious links that might be embedded in the social media post.

Our Contribution

- We curated the datasets “TweetaVerse” and “URL Guardian”.
- We were able to derive the category of tweet headline using Zero Shot Classification.
- We have used User Influence to figure out whether a tweet is true or false.
- We calculated ‘Exclusivity’ of a tweet which helped in finding out whether a tweet is true or not.
- We were able to classify fake news into further sub-categories – Misinformation, Disinformation, Satire and Spam.
- We have implemented image classification which is a step towards the multimodal nature of our project. So far we are able to classify news, URLs and images as fake or real.

Existing Solutions

1. Natural Language Processing
   - Utilizing NLP techniques, social media posts undergo analysis for language patterns, sentiment, and semantic structures, facilitating the identification of potentially misleading content.
2. Source Credibility Assessment
   - Evaluating the credibility of sources involves an assessment of their reputation, track record for accuracy, and reliability. This process aids in distinguishing between reliable and questionable information.
3. Fact Checking Integration
   - Integration with fact-checking databases enables real-time verification of claims, allowing the system to flag or label content that contradicts established facts.

Existing Solutions

4. Community Reporting and Feedback
   - Incorporating mechanisms for users to report and provide feedback on potentially false content allows for community involvement in the identification process.
5. Cross Referencing Multiple Sources
   - Cross-referencing information from different sources is done to verify or refute claims, utilizing varied perspectives to improve the precision of identifying misinformation.
6. Metadata Analysis
   - Examination of metadata associated with multimedia, such as timestamps, geolocation, and digital fingerprints, aids in assessing the authenticity of content and detecting inconsistencies.

Existing Solutions

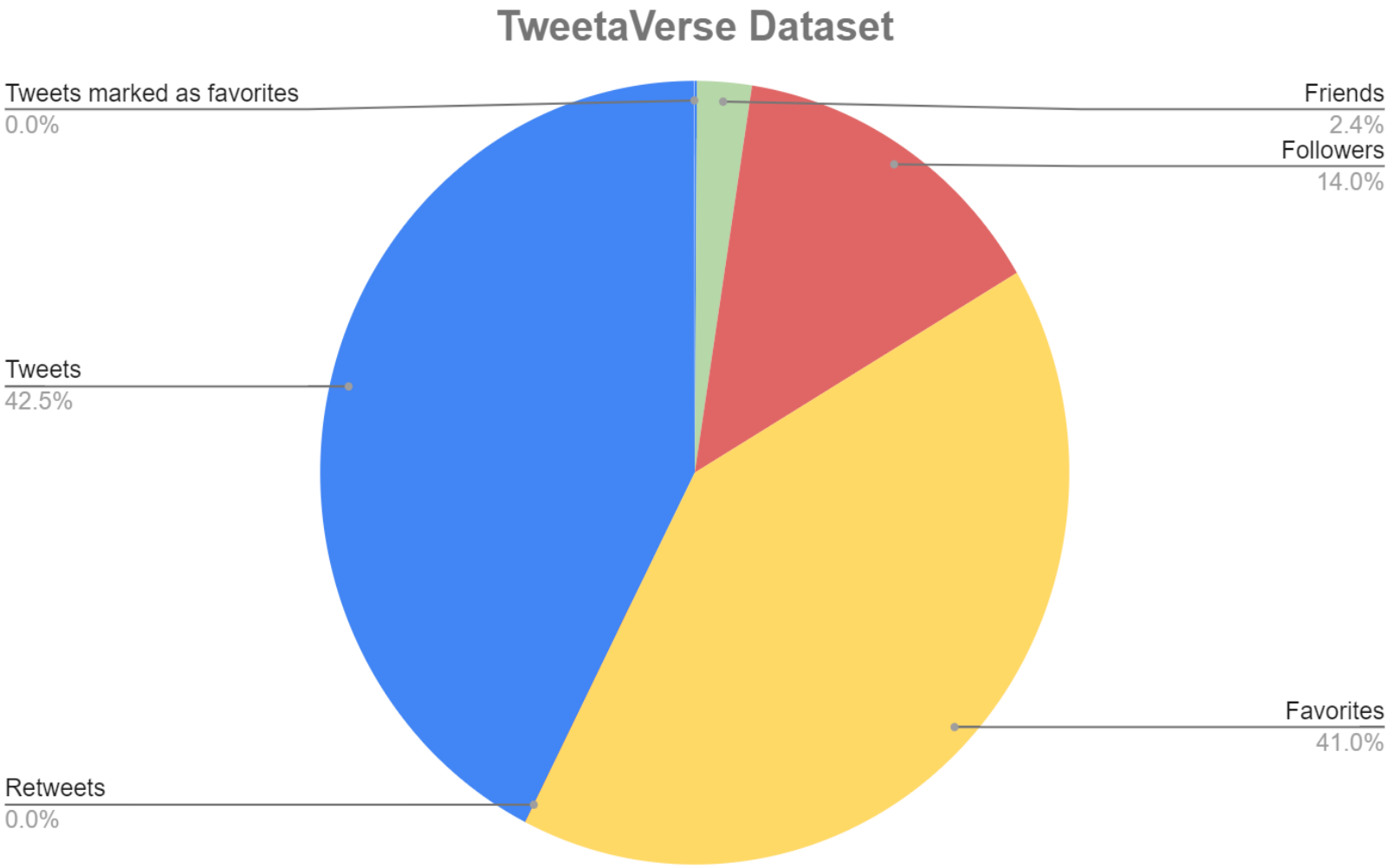
7. Deep Learning and Neural Networks
   - Deep learning methodologies have shown remarkable achievements across diverse tasks in natural language processing (NLP) and computer vision. This success renders them particularly auspicious for detecting fake content, as these techniques leverage neural networks to autonomously acquire complex patterns and representations from large datasets.
   - For textual analysis, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been employed to scrutinize textual content and identify linguistic features indicative of fake news or misinformation.

Proposed Methodology

1. Tweet Dataset (Tweetaverse) Preparation
   - "TruthSeeker: The Largest Social Media Ground-Truth Dataset for Real/Fake Content" was chosen as the dataset upon which we would work. Multiple fields which were of no consequence to out desired output were removed.
   - To the dataset we added the 'tweet_category' field which categorizes the tweet content into categories – world, business, crime, politics, health, sports, social, technology, finance, education and entertainment. This was done using Zero Shot Classification.
   - The fields – 'Bot_Score', 'cred_score', 'UAS', 'TAS', 'UAS_normalized', 'TAS_normalized' and 'Exclusivity' were further derived and added.
   - Total number of features : 27
   - Total number of records : 1,34,194
   - 80 : 20 split for Training and Testing was done.
   - For the Training set, 107355 records and 6 features. For the Test Set, 26839 records and 5 features.

Proposed Methodology

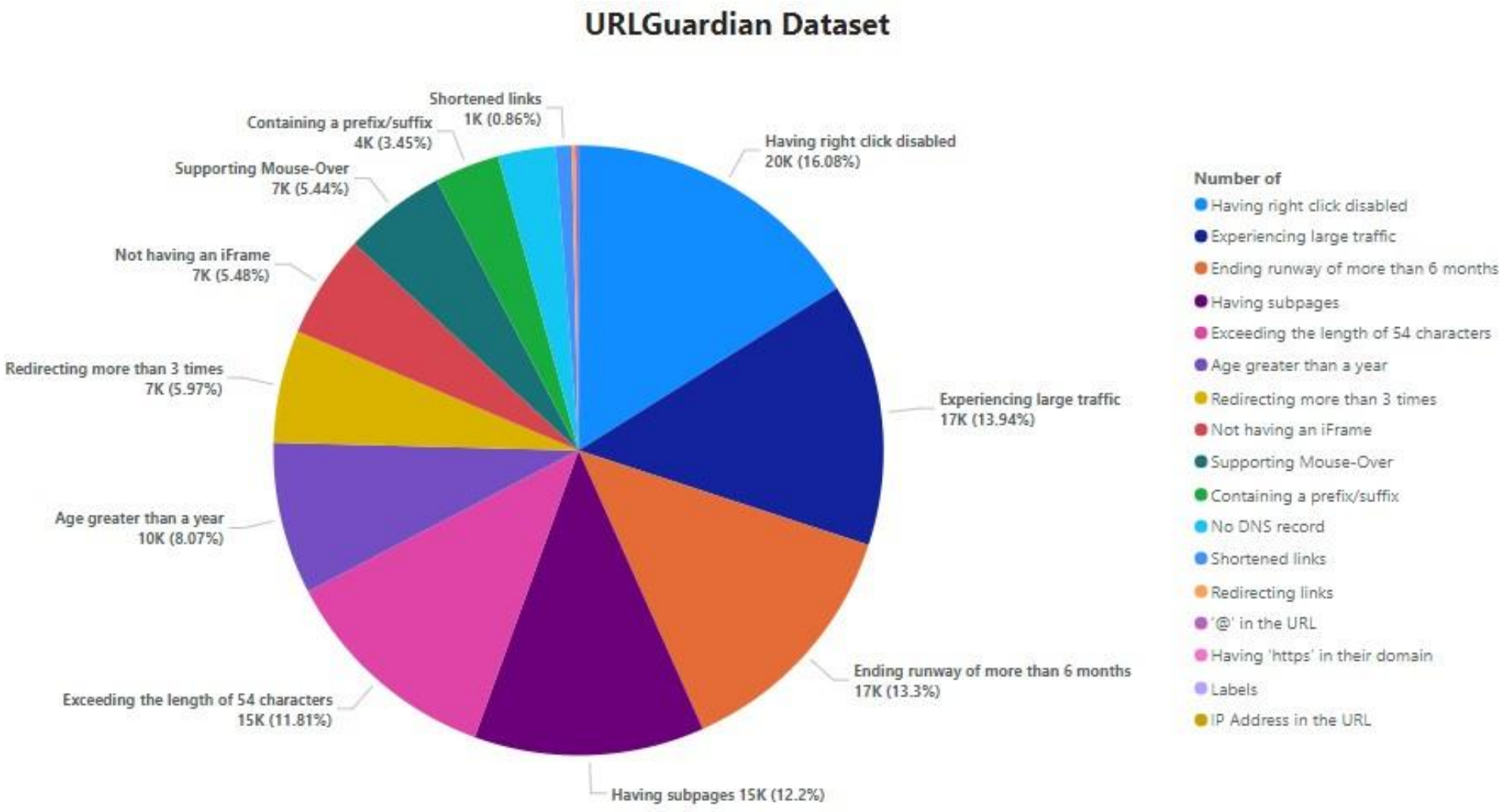
Lists	9835397
Labels	5
Mentions	186384
Tweet categories	11
Friends	254053386
Replies	256882
Tweet URLs	79
Followers	1515446387
Favorites	4425917915
Quotes	76952
Retweets	895678
Tweets	4588528087
Tweets marked as favorites	3700159
Hashtags	14052



Proposed Methodology

2. URL Dataset (URL Guardian) Preparation
   - We used 10,000 phishing URLs from “PhishTank”, a service that provides a set of phishing URLs in multiple formats like csv, json etc. that gets updated hourly.
   - We used 10,000 legitimate URLs from “ISCX-URL2016”, which is a source that has a collection of legitimate, benign, spam, phishing, malware and defacement URLs. The number of legitimate URLs in this collection are 35,300.
   - We ended up with 10,000 each of Phishing and Legitimate URLs. We only needed the URL attribute from both “PhishTank” and “ISCX-URL2016” datasets. From this attribute we extracted 17 more features and hence URL Guardian dataset was created containing 20,000 records and 18 features.
   - 80 : 20 split for Training and Testing was done. URL feature was dropped from both Training and Testing sets as it inconsequential to the output.
   - For the Training set, 16,000 records and 17 features. For the Test Set, 4000 records and 16 features.

Proposed Methodology



Proposed Methodology

3. UAS, TAS and Exclusivity
   - UAS (User Activity Score) measures the average engagement the user's usual tweets receive. To calculate it we have used the following factors – UFS, UFRC, UFVC, UTC, ULC, UMC, UQC. UAS was calculated using the below formula, where the w_i are the weights that were assigned to the respective features based on their variance (higher the variance, higher the weight).

     $$\text{UAS} = w1 * UFS + w2 * UFRC + w3 * UFVC + w4 * UTC + w5 * ULS + w6 * UMC + w7 * UQC$$

   - TAS (Tweet Activity Score) measures the average engagement a particular tweet is receiving. To calculate it we have used the following factors –TRTC, TFVC, TRC and THC. TAS was calculated using the below formula, where the w_i are the weights that were assigned to the respective features based on their variance (higher the variance, higher the weight).

     $$TAS = w1 * TRTC + w2 * TFVC + w3 * TRC + w4 * THC$$

   - Exclusivity is a binary metric which signifies whether a particular tweet is exclusive (not a part of) a user's previous other tweets based on their activity scores namely, UAS and TAS". If TAS is greater than k*UAS then Exclusivity is 1 or else it is 0.
4. Bot Score and Cred Score
   - The Bot Score in the TruthSeeker dataset is a numerical value ranging from 0 to 1, signifying the likelihood that a given user is a potential bot account. A score equal to or less than 0.5 suggests that the user is not a bot, while a score exceeding 0.5 identifies them as a potential bot.
   - Cred Score metric is also based on the Cred Score in the TruthSeeker Dataset, which is also a value between 0 and 1 showing the credibility of the user.

Proposed Methodology

5. Semantics Classifier

- The Semantic Classifier categorizes false tweets into one of the following types: misinformation, disinformation, spam, or satire.
- Disinformation: If a bot user shares false news deliberately. This is determined by the bot score of the tweet.
- Misinformation: If a non-bot user shares false news by mistake.
- Spam detection:
  - Considers factors like repetitive words, presence of links, use of special characters, text length, and frequency of special words.
  - Uses majority voting with higher weightage for Repetitive Words Score (RWS), Special Words Score (SWS), and Phishing Links Present (PLP). Other scores are Special Characters Score (SCS) and Text Length Score (TLS).
- Satire detection:
  - Utilizes a pre-trained sentiment analysis model (Twitter-roberta-base-sentiment-latest) to detect positive and negative sentiments in sub-sentences.
  - If both positive and negative sentiments are present in a tweet, it is classified as Satire.
- If a tweet has both Spam and Satire elements, it is classified as Misinformation.

Proposed Methodology

5. Semantics Classifier (Contd)

- Repetitive Words [Repetitive Words Score (RWS)]
  - Spam messages often contain many repeating words. Calculated frequency of each unique word, sorted in descending order. Found point where frequency drops significantly to determine noteworthy words. Assigned score of 1 if repeating words exist, 0 otherwise
- Presence of Links [Phishing Links Present (PLP)]
  - If phishing links are present, tweet is marked as potential spam. Score of 1 for presence of phishing links, 0 otherwise.
- Use of Special Characters [Special Characters Score (SCS)]
  - Excessive use of special characters may indicate spam. Considered continuous sequences of same or different special characters. Score of 1 for presence of excessive special characters, 0 otherwise
- Text Length [Text Length Score (TLS)]
  - Spam tweets are often too short or too long compared to average. Considered average range of 30-140 characters. Score of 1 if length is outside this range, 0 otherwise
- Frequency of Special Words [Special Words Score (SWS)]
  - Certain words like "discount", "urgent", etc. are common in spam. Used a bag of such special words from online sources. Score of 1 if special words are present, 0 otherwise

Proposed Methodology

6. Image Classifier

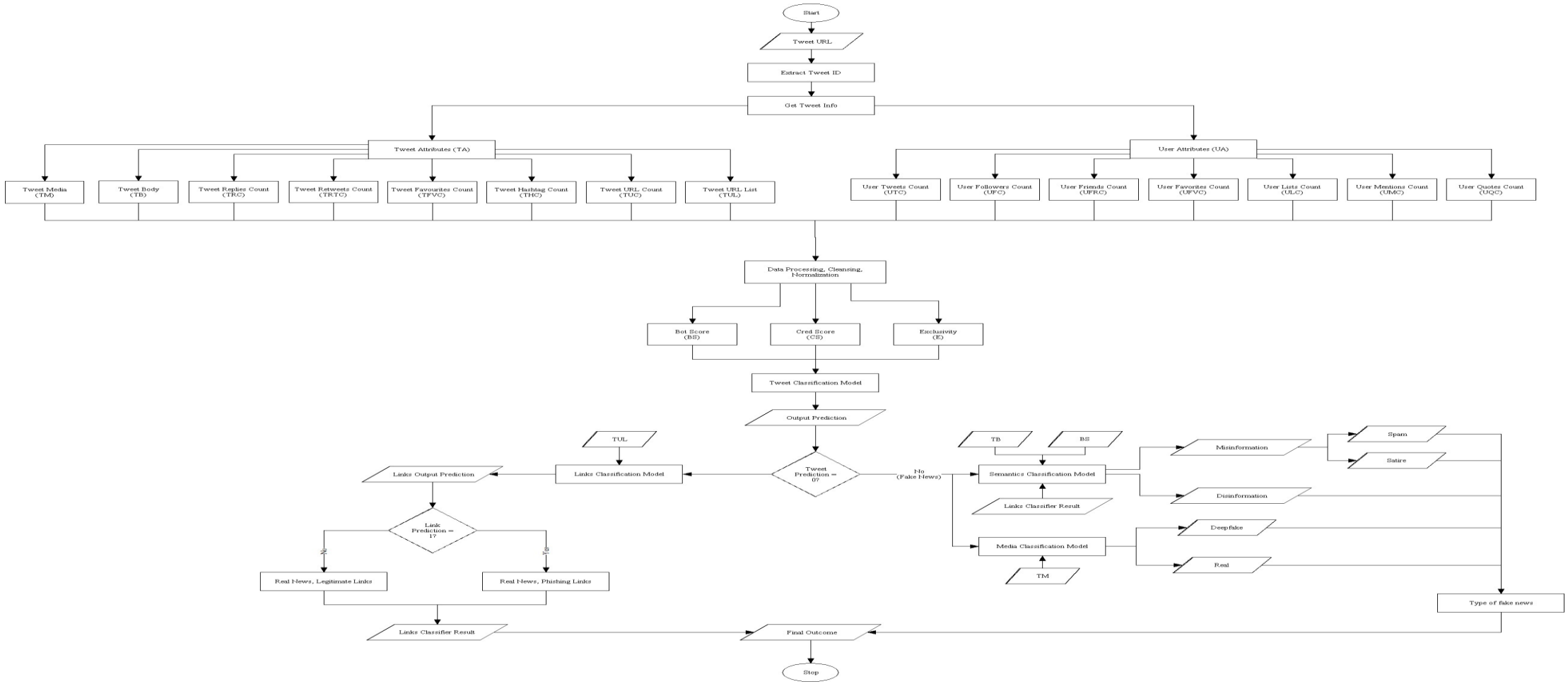
- The Image Classifier is responsible for classifying images as real or fake based on the CIFAKE dataset.
- The CIFAKE dataset contains:
  - Training: 100,000 images (50,000 real, 50,000 fake)
  - Testing: 20,000 images (10,000 real, 10,000 fake)
- Multiple CNN models were compared on the same dataset:
  - ResNet-50 : 50-layer CNN with residual connections to ease training of deep networks.
  - VGG-16 : 16-layer CNN known for its simplicity and effectiveness.
  - EfficientNetV2 : State-of-the-art model using compound scaling and efficient blocks like MBConv.

Proposed Methodology

The overall working of the model is as follows :

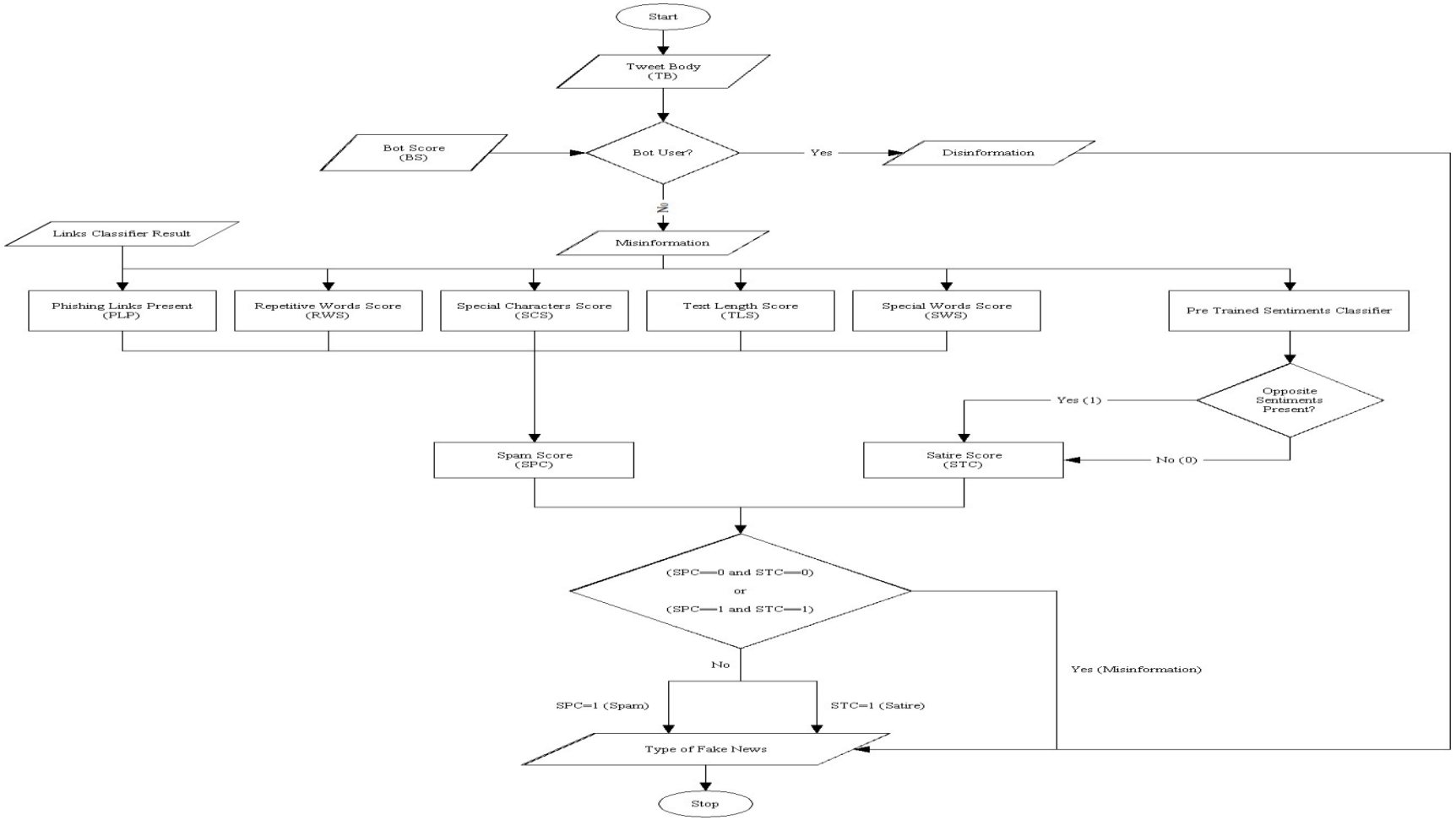
1. A tweet's URL is input by the user.
2. The Tweet ID is extracted using an API. From this we extract the tweet attributes and the user attributes
3. The tweet attributes and the user attributes are used to calculate the Bot Score, Cred Score and Exclusivity. Using these we classify the tweet body as true or false.
4. If the tweet body was false then we pass it into the semantics classifier which further categorizes the fake news into one of four sub-categories.
5. If the tweet body was true, then we check the tweet for any embedded links. If the any such links were found, then we checked the if the link was a phishing link or not.
6. An image can uploaded to verify whether it is real or fake.

Overall Workflow



Link Classifier Workflow

Semantics Classifier Workflow



Uniqueness

1. We have opted for Multimodal Analysis. Comprehensive analysis of both textual and visual elements in multimedia content ensures a more holistic approach to identifying misinformation, especially in contexts where fake news involves multimedia manipulation.
2. We have incorporated Ensemble Analysis to tackle the challenge introduced by the Multimodal Analysis. The outcomes of multiple ML/DL models will be combined into the final outcome.
3. We have included Phishing Link Detection. This adds another layer of security as it has become common for phishers to embed their malicious links in a social media posts which share legitimate news.
4. We have included further classification of false news into sub-categories : Misinformation, Disinformation, Spam and Satire.

Result Analysis

- The tweet classifier's role is to determine if the text in a tweet is false or not, utilizing various metrics discussed in previous sections.
- The training process begins by loading the complete dataset and conducting preprocessing steps.
- During preprocessing, null-valued records were eliminated. Given the textual nature of the data, non-textual elements such as URLs, mentions, hashtags, special characters, and punctuations were removed.
- TV5 unique values were mapped to either true (0) or false (1) in the dataset.
- An 80:20 train-test split was performed on the preprocessed dataset.
- Handling both numerical and textual data, the approach involved combining all numerical data and vectorizing textual data into a numerical format.
- The final step involved merging both sets of numerical data into one, creating a comprehensive input for training the model.

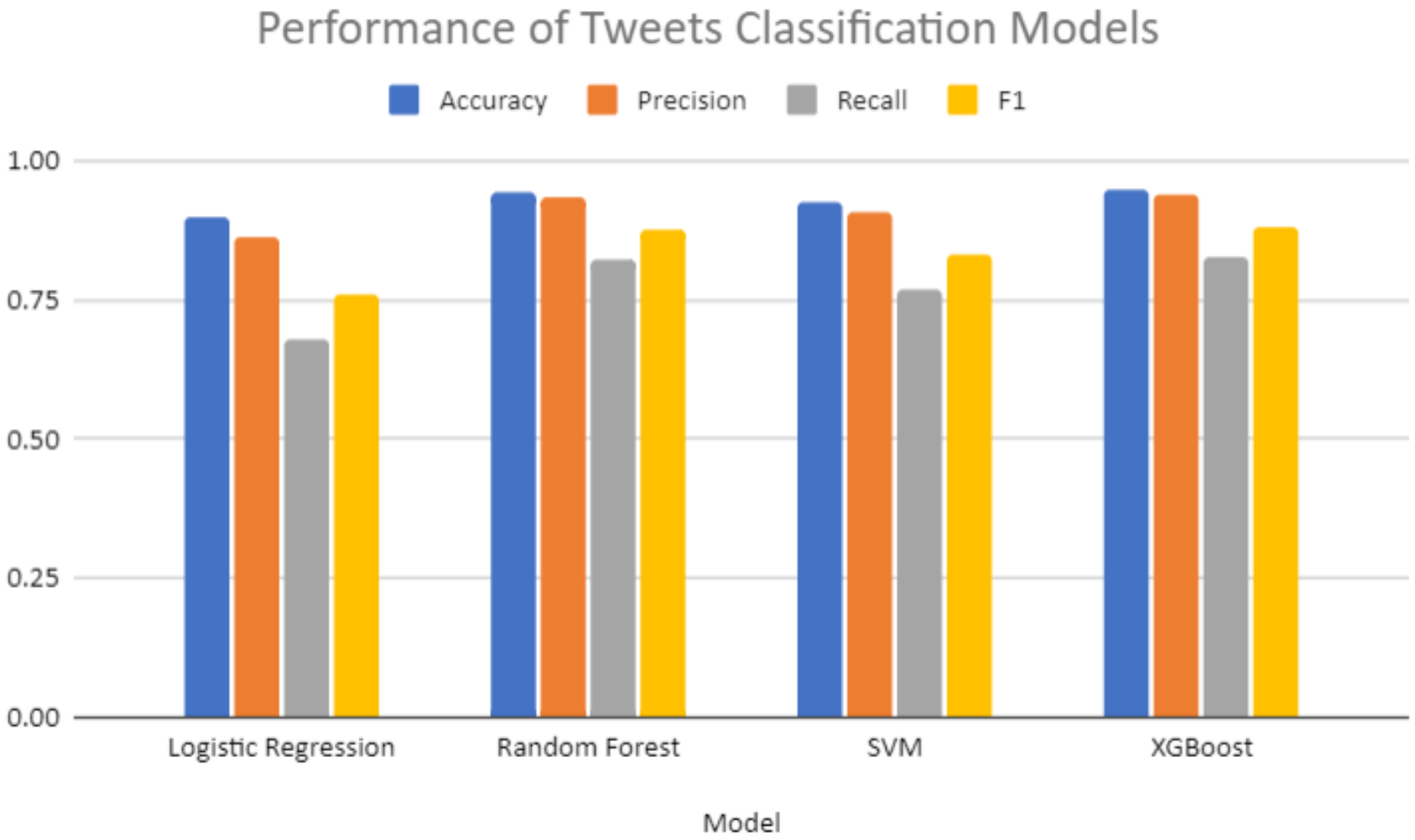
Result Analysis

- As this was a classification stage, we used the following models: Logistic Regression, Random Forest Classifier, SVM, and XG Boost.
- The following were the results from the models :

MODEL	ACCURACY	PRECISION	RECALL	F1
Logistic Regression	0.897	0.862	0.68	0.76
Random Forest	0.943	0.933	0.823	0.875
SVM	0.926	0.908	0.769	0.833
XG Boost	0.946	0.94	0.827	0.88

- XG Boost model has the highest accuracy of 94.6% and was chosen as the model to be used for the project.

Result Analysis



Result Analysis

- Links Classifier focuses on classifying URLs as legitimate or phishing based on predefined features.
- The classifier deals exclusively with numerical data, simplifying the training process compared to the Tweets Classifier.
- The workflow, outlined in Slide 15, involves loading and preprocessing the complete dataset.
- During preprocessing, the 'domain' feature was dropped as it doesn't contribute to the training process.
- The dataset was shuffled to eliminate biases, as the initial 10,000 records are legitimate, followed by 10,000 phishing records.
- A standard 80:20 train-test split was applied, with the goal of classifying input URLs as either legitimate (0) or phishing (1).
- The classifier was trained on numerical data, with the goal of classifying input URLs as either legitimate (0) or phishing (1).

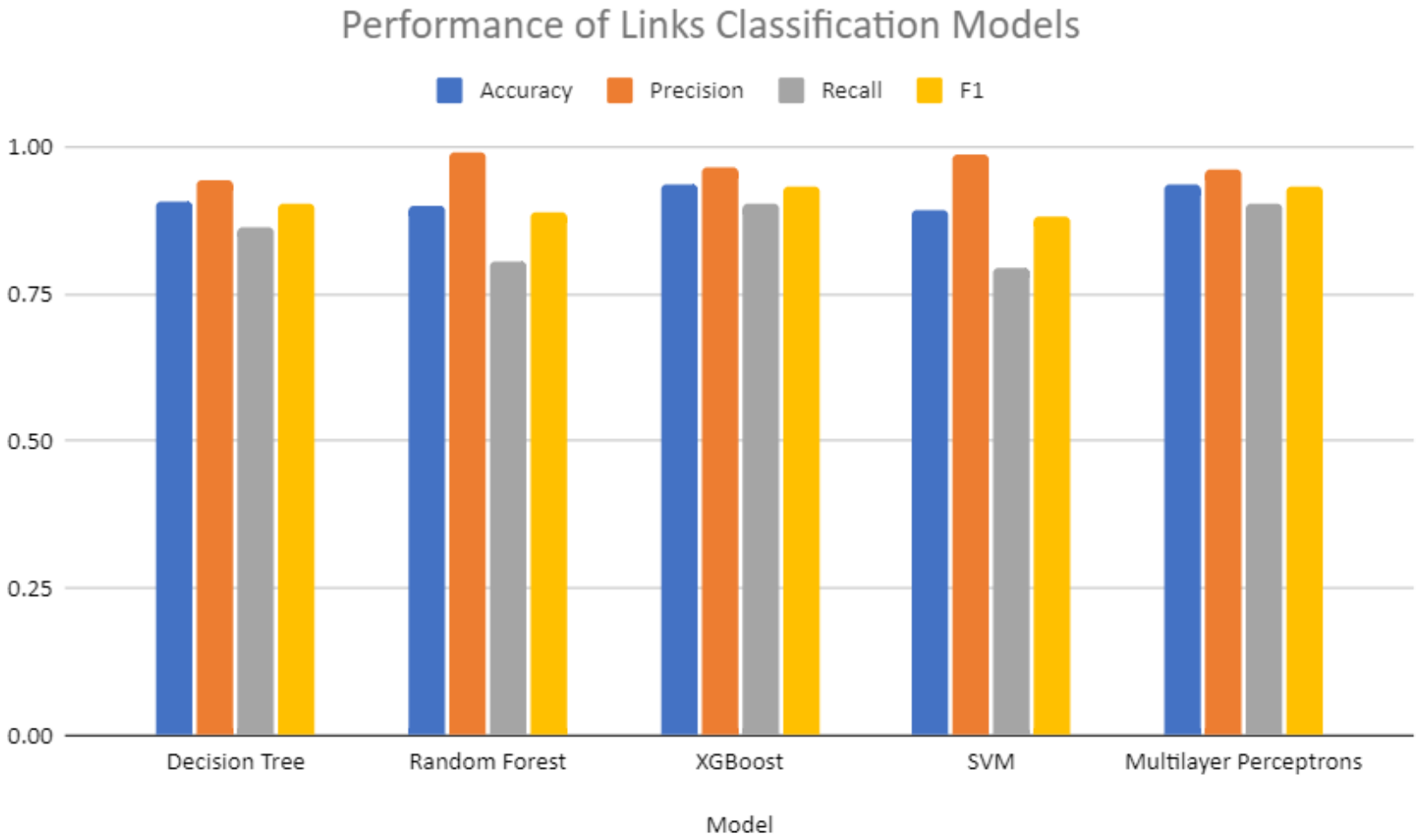
Result Analysis

- As this was a classification stage, we used the following models: Decision Tree, Random Forest, SVM, XG Boost and Multilayer Perceptron.
- The following were the results from the models :

MODEL	ACCURACY	PRECISION	RECALL	F1
Decision Tree	0.907	0.943	0.863	0.902
Random Forest	0.9	0.991	0.804	0.888
XG Boost	0.936	0.965	0.902	0.933
SVM	0.894	0.987	0.796	0.881
Multilayer Perceptron	0.936	0.963	0.904	0.933

- XG Boost Model has the highest accuracy of 93.6% and was chosen as the model to be used for the project, over Multilayer Perceptron which also has the same accuracy as XG Boost Model is computationally lighter.

Result Analysis



Result Analysis

- Image Classifier has the job of classifying images as either fake or real.
- The images were sufficient inputs for the models. Features were extracted by the model itself.
- Training: 100,000 images (50,000 real, 50,000 fake)
- Testing: 20,000 images (10,000 real, 10,000 fake)
- Types of CNNs used : ResNet-50, VGG-16, EfficientNetV2

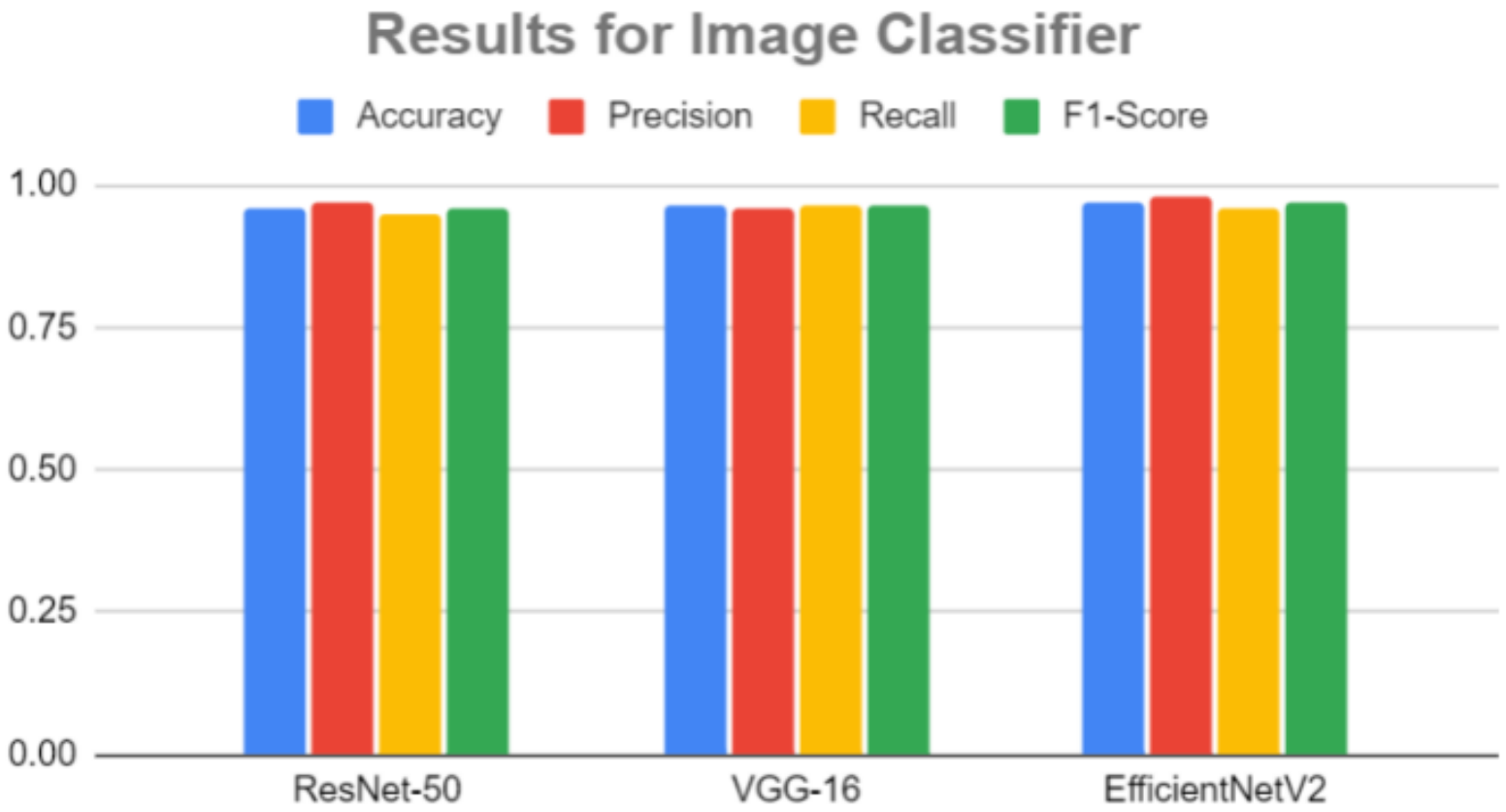
Result Analysis

- Performance of the various CNNs models are as follows :

MODEL	ACCURACY	PRECISION	RECALL	F1
ResNet-50	0.9604	0.9692	0.951	0.96
VGG-16	0.9629	0.9611	0.9649	0.963
EfficientNetV2	0.9715	0.9811	0.9615	0.9712

- EfficientNetV2 has the best accuracy, precision and F1 score and hence has the best performance out of all three CNN types.

Result Analysis



Conclusion

- Textual fake news detection, fake news classification, phishing link detection and fake image detection were implemented into the project.
- XGBoost model was chosen for textual fake news classifier as it gave the highest accuracy of 94.6% and highest precision of 94%.
- XGBoost model was chosen for phishing link classifier as it gave the highest accuracy if 93.6% and precision of 96.5%.
- EfficientNetV2 model was chosen for fake image detection as it gave the highest accuracy of 97% and highest precision of 98.11%.

Future Scope

1. Expansion to other social media platforms
   - As of now, the project uses a dataset which is gathered from only from the social media platform Twitter (now X). The project will expand its scope to include as many social media platforms as possible, in the future.
2. Multimodal Analysis
   - As of now, the project has implemented phishing link detection, fake news detection and classification for textual and image data only. The project will expand its scope to detect fake news using various other forms of media, like audio, videos, etc in the future.
3. Easier use by creating browser extension
   - The project will be made into a browser extension so that the end user can safeguard themselves on the social media platforms while scrolling through their feeds.

References

- Scott Counts Aditya Pal. Identifying topical authorities in microblogs. 2011.
- Jorge Bendahan Rami Puzis Aviad Elyashar. Detecting clickbait in online social media: You won’t believe how we did it. 2022.
- Gwangbin Bae, Martin de La Gorce, Tadas Baltruˇsaitis, Charlie Hewitt, Dong Chen, Julien Valentin, Roberto Cipolla, and Jingjing Shen. Digiface-1m: 1 million digital face images for face recognition. In 2023 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2023.
- Jordan J. Bird and Ahmad Lotfi. Cifake: Image classification and explainable identification of ai-generated synthetic images, 2023.
- Sajjad Dadkhah, Xichen Zhang, Alexander Gerald Weismann, Amir Firouzi, and Ali A Ghorbani. Truthseeker: The largest social media ground-truth dataset for real/fake content. 2023.
- David Hans Rodolfo Villarroel Roberto Munoz Fabi´an Riquelme, Pablo Gonzalez-Cantergiani. Identifying opinion leaders on social networks through milestones definition. 2019.
- Pablo Gonz´alez-Cantergiani Fabi´an Riquelme. Measuring user influence on twitter: A survey. 2016.

References

- Ahmad Lotfi Jordan J. Bird. Cifake: Image classification and explainable identification of ai-generated synthetic images. 2024.
- Talayeh Riahi Asahi Ushio Daniel Loureiro Dimosthenis Antypas Joanne Boisson Luis Espinosa-Anke Fangyu Liu Eugenio Mart´ınez-C´amara Gonzalo Medina Thomas Buhrmann Leonardo Neves Francesco Barbieri Jose Camacho-Collados, Kiamehr Rezaee1. Tweetnlp: Cutting-edge natural language processing for social media. 2022.
- Alex Krizhevsky. Learning multiple layers of features from tiny images. pages 32–33, 2009.
- Mohammad Saiful Islam Mamun, Mohammad Ahmad Rathore, Arash Habibi Lashkari, Natalia Stakhanova, and Ali A Ghorbani. Detecting malicious urls using lexical analysis. In Network and System Security: 10th International Conference, NSS 2016, Taipei, Taiwan, September 28-30, 2016, Proceedings 10, pages 467–482. Springer, 2016.
- James Zou Yiqun T. Chen. Twigma: A dataset of ai-generated images with metadata from twitter. 2023.