Distinguishing Arabic GenAI-generated Tweets and Human Tweets utilizing Machine Learning

Noura Saad Alghamdi
Computer Science Department, University of Jeddah, Jeddah, Saudi Arabia
2300333@uj.edu.sa (corresponding author)

Jalal Suliman Alowibdi
Computer Science and AI Department, University of Jeddah, Jeddah, Saudi Arabia
04100427@uj.edu.sa

Received: 27 June 2024 | Revised: 13 July 2024 and 21 July 204 | Accepted: 24 July 2024

 DOI: https://doi.org/10.48084/etasr.8249

ABSTRACT

Generative Artificial Intelligence (GenAI) tools, like ChatGPT, have made it easy to create text, music, images, and other types of media. GenAI, a type of AI technology, has rapidly gained fame and popularity for its ability to generate new content. Notably, its applications allow anyone to produce natural conversations and content, making it increasingly challenging to distinguish between human-written and GenAI-generated material. The current research focuses on Arabic content to differentiate GenAI-generated content from authentic human-written content on the X platform (Twitter). Datasets from both real human-written tweets and GenAI-generated tweets were collected. Then, three Machine Learning models were built to predict whether a tweet source is GenAI-generated or human-written. The highest achieved accuracy was 93%.

Keywords-GenAI; Machine Learning (ML); OpenAI; Twitter; ChatGPT

I. INTRODUCTION

The advancements in GenAI have revolutionized data generation, employing sophisticated methods such as Deep Learning (DL) and Machine Learning (ML). Unlike traditional tasks like regression and classification, GenAI excels in automatically creating new data—whether text, music, or images—that closely mirror the distribution of original datasets. Central to GenAI is the generative model, which captures the potential distribution of data and synthesizes new instances that exhibit similar characteristics to the original. Today, GenAI's applications are vast and diverse, spanning from language processing to music generation, and image creation. One of the most renowned generative models is the Generative Adversarial Network (GAN), celebrated for its ability to produce images almost indistinguishable from real photographs. On the other hand, in the realm of Natural Language Processing (NLP), models such as transformer networks and recurrent neural networks have demonstrated exceptional abilities in generating coherent and contextually relevant textual data. Similarly, in music generation, techniques like the automatic encoder and the variational automatic encoder have shown promise in composing novel musical pieces. In the past decade, GenAI saw substantial advancement, offering new concepts and ways to boost Artificial Intelligence (AI) technology, with successes like AlphaGo defeating top human Go players, using a combination of DL and reinforcement learning. Overall, literature traces the evolution of GenAI from its early language-focused beginnings to its more recent breakthroughs in areas like computer vision and game-playing, driven by key innovations in DL [1].

Since the release of ChatGPT, the concept of GenAI has garnered widespread attention, sparking significant interest in its potential impacts across various domains. In the context of Arabic tweets, distinguishing between those generated by GenAI and those authored by humans presents a unique challenge. The linguistic characteristics, cultural references, and contextual details inherent in human-written Arabic tweets render this task particularly complex. This paper explores the methodologies and techniques employed to differentiate GenAI-generated Arabic tweets from those written by human users. By leveraging advanced NLP models and analyzing various linguistic features, this research aims to explore the distinguishing characteristics of GenAI-generated Arabic tweets, contributing to the broader discourse on the integration of AI in social media and communication.

Large Language Models (LLMs) have advanced significantly, leveraging abundant digital text data and computational power to produce human-like text that is increasingly difficult to distinguish from real human-written content. Models like GPT-3, GPT-4, and LLaMA have acquired remarkable sophistication in language generation and understanding. However, the ability of LLMs to produce original content also introduces risks around the potential creation of false information. Even seemingly fluent and coherent text from LLMs must be carefully fact-checked, as chatbots or AI-enhanced searches could generate fraudulent

claims or explanations on sensitive topics like finance or health, which could then be shared and cited by users. There are also concerns that LLMs could be intentionally misused for malicious purposes, such as powering fake news sites, scams, or bot-generated content. For example, AI-driven fake news sites have already attracted large online followings [2].

There are numerous concerns regarding the use of LLMs in education, as their ability to provide instant solutions could threaten the development of critical thinking and problem-solving skills. LLMs also raise issues around credibility, as they can be used to generate "deepfake" news that could mislead audiences. Their potential misuse in legal and cybersecurity domains is also worrying. Importantly, even fluent text produced by LLMs requires careful fact-checking, as the technology's rapid expansion necessitates swift societal responses in the form of regulations and awareness-building. Several studies have been conducted with promising results in the field of detecting and classifying text generated by AI language models like ChatGPT. Authors in [3] proposed an XGBoost-based classification model that exhibited excellence in distinguishing between ChatGPT-generated text, with a high accuracy of 96%. Authors in [4] presented TSA-LSTMRNN, an algorithm that incorporates the Tunicate Swarm Algorithm with the Long Short-Term Memory Recurrent Neural Network, achieving decent accuracy rates of 93.17% and 93.83% over the human-generated and ChatGPT-generated datasets, respectively. Authors in [5] introduced a state-of-the art overview of detecting large LLM-generated text and also put great emphasis on the necessity of complete metrics for evaluation. Authors in [6] presented an ML-based solution to differentiate between human-generated and ChatGPT-generated text, achieving an accuracy of 77%. Authors in [7] distinguished between human-generated text and that produced by ChatGPT, using the T5 and RoBERTa language models and obtaining over 97% accuracy. Authors in [8] focused on the discrimination of medical texts written by human specialists from those created by ChatGPT, obtaining accuracy over 95%. Authors in [9] looked at the detection of homework assignments generated by AI through the HowkGPT program, which would ensure fair grading and the maintenance of academic integrity. Authors in [10] noted that, on ethical grounds, complex detection procedures about LLMs are called for and showed that the detection of AI-generated content is possible. Authors in [11] compared the identification of texts created by AI and humans, providing a new dataset and evaluating the performance of various ML models. Authors in [12] assessed the effectiveness of generative AI text detectors and emphasized the challenges posed by manipulated content. They recommended a critical approach to the implementation of AI text detectors in higher education. There are also many studies interested in analyzing tweets related to humans, e.g. [13, 14].

The present study developed models that can discern authentic human-written Arabic tweets from the GenAI-generated Arabic tweets in the social network, X, using GenAI literacy. It also looked for the answers to the following two research questions: Q1) What unique characteristics or patterns distinguish human-generated content on Twitter from GenAI-generated content? Q2) What is the most effective ML model that can quickly recognize and highlight content generated by GenAI on X with high accuracy?

II. MOTIVATION

Recently, the advent of GenAI technologies, such as ChatGPT, has made it remarkably easy to generate text, audio, images, and other media content. GenAI has surged in popularity due to its ability to produce novel content that closely mimics human creativity in authored texts. This technological advancement allows users to generate natural-sounding conversations and content with minimal effort and easy-to-distribute text content. However, this convenience comes with a challenge: distinguishing whether the content was created by GenAI or actual humans is becoming increasingly difficult. Thus, the current research is driven by this challenge, specifically focusing on Arabic text content on the X platform, (Twitter). The primary aim is to differentiate between GenAI-generated tweets and legitimate tweets that were written by humans [15], which is a matter of paramount importance for several key reasons:

- GenAI can produce highly convincing content that can be easily mistaken for genuine human-created material. This poses the significant risk of spreading misinformation.
- Ensuring content authenticity is crucial for maintaining the integrity of information.
- Human-generated content typically implies a level of accountability and responsibility that may not be present with GenAI-generated content.
- There are substantial ethical and legal considerations surrounding the use of GenAI.
- Users often prefer engaging with authentic, human generated content.
- Effective content moderation is critical for maintaining a healthy digital environment.
- For researchers and developers working on AI, understanding the differences between AI-generated and human-generated content is essential for improving GenAI systems.
- Businesses and marketers can tailor their strategies based on the source of content.

In summary, the ability to distinguish between GenAI-generated tweets and human-created tweets is critical for maintaining the integrity, trustworthiness, and overall health of digital communication platforms. It has wide-ranging implications for misinformation prevention, ethical considerations, user experience, and the future development of AI technologies. Therefore, to achieve this, a dataset comprising real human-created tweets from the X platform and GenAI-generated tweets from ChatGPT was collected. Certain models were then developed to predict whether a tweet was produced by GenAI or a human.

III. THE PROPOSED APPROACH

This work focuses on finding the features relevant to human- or AI-generated content and develop a supervised model to distinguish between the two. Thus, initially, the required data were collected and accounts were created on the X platform. The tweets written by ChatGPT were shared and the trending hashtags were used to see how ChatGPT writes about different topics. Then, these tweets and real-human tweets from the same hashtag and topic were assembled into an Excel file. Afterwards, some preprocessing was performed to make the data clearer and more useful. Then, the selected models' building was initiated and the data were split into a training set and a test set. To evaluate the results, accuracy, recall, precision, and F1-score were deployed.

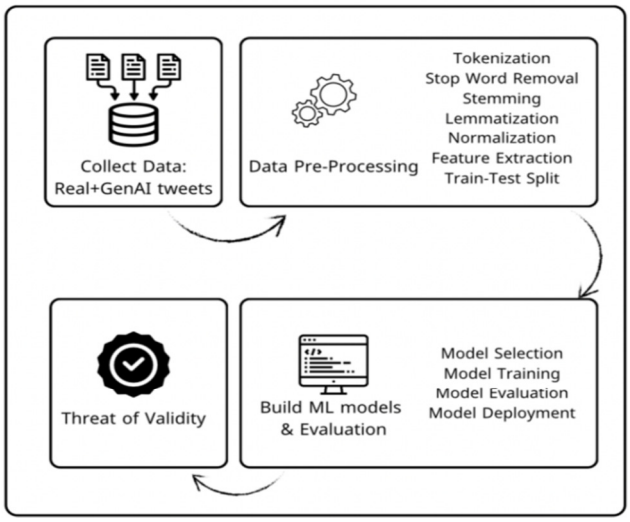

Fig. 1. The proposed methodology.

IV. EXPERIMENTAL RESULTS

A. Dataset

The two types of tweets that were collected for the current research were those created by real human users and those that GenAI systems generated. This study started with choosing the most popular hashtags at the moment. Then, authentic human-written tweets for the corresponded hashtags were collected. The current study searched for and scraped tweets using those hashtags, closely investigating each account to ensure that it was not a spam or an automated bot account. Thus, for the proposed dataset, 375 human-written tweets were scraped.

To collect a dataset of GenAI-generated tweets, three new Twitter accounts were created. Each account had a unique name, bio, and a profile photo to resemble real profiles of actual people. The next step involved creating tweets by employing the GenAI of the ChatGPT language model. The most popular hashtags used in human-written tweets were monitored for specific hashtags, while GenAI was instructed to generate tweets that appeared to be from human users communicating over the same hashtags. To enhance the naturalness of the tweets, GenAI was also instructed to produce tweets on subjects similar to the human-written tweet topics utilizing hashtags. This process resulted in the collection of 375 GenAI-generated tweets to be used for model testing and training. The entire dataset comprises 750 tweets, with an equal number of tweets from humans and AI. This balanced dataset provides a robust foundation for training and testing models to distinguish between human-written and GenAI-generated tweets.

التغريدات المولدة	التغريدات الحقيقية
الحياة مليئة بالأمل والفرص. ابتسم واستمتع بكل لحظة. أنت تستحق السعادة والنجاح. #غرد_بكلمة_جميلة	اللهم عافية وراحة بال ويوم جميل اللهم مساء لايمسّنا فيه سؤء… #غرد_بكلمه_جميله
يا ليل، الوحدة تخنقني وأنا بين جدراني. أبوح لك بأسراري وأحلامي. أنت الصديق الوفي اللي يسمعني بصمته. استقبل بوحي وحكاياتي في أعماقك. #بوح_اخر_الليل	#بوح_اخر_الليل .. لن تنجح في مداواة روحك إن كانت خطتك الوحيدة هي الهرب …
صباح الجمعة الجميلة، يوم السكينة والبركة! يا رب يكون يوم مليء بالطمأنينة والسعادة. استغلوا هذا اليوم للتواصل مع الأحبة وقراءة القرآن والاستمتاع بالأوقات الجميلة. جمعة مباركة على الجميع! #يوم_الجمعة	إنّ الصّلاة على الحبيبِ لَنِعمَةٌ فيها الأجورُ من الإله فأكثِرُوا ". #يوم_الجمعة
الخميس الونيس ما بين الحياء والفرحة! يومنا المفضل اللي نستناه بشوق. نتمنى يكون يوم مليء بالسعادة والإيجابية. ابتسامتك تنور الدنيا، فلا تحرمنا منها يا خميس! #الخميس_الونيس	#الخميس_الونيس التفاؤل شمعة مضيئة داخل كل إنسان ورسالة من رسائل الأمل للقلب والروح والعقل تدعونا للتفاؤل بأن القادم أجمل بمشيئة الله .. مسائكم فرح وسعادة.. وابتسامة تفتح أبواب الأمل.. وتنثر عبير التفاؤل، وتملأ قلوبكم بالمحبة…
مباراة قطر وإيران راح تكون صراع حامي الوطيس! قلوبنا مع الفريقين ونتمنى النجاح للأحسن. يا رب تكون المباراة مليانة إثارة وأهداف حلوة. هيا نشجع فريقنا ونستمتع بالكورة! #قطر_ايران	العنابي الى النهائي … ♥-💚 الف مبروك لكل الخليجيين فوز المنتخب القطري … #قطر_أيران
من الشمال إلى الجنوب، ومن الشرق إلى الغرب، السعودية تعلو وتتسع! بلدنا الغالي يستحق أن يكون في مقدمة الدول العظمى. #السعودية_العظمى	اليوم وكل يوم:"أنتِ ما مثلك بهالدنيا بلد". #يوم_التاسيس #يوم_بدينا #السعودية_العظمى

Fig. 2. Tweets sample.

B. Pre-Processing

To certify that the dataset was not only clean but also primed for advanced analysis, this study employed a comprehensive set of preprocessing techniques, tailored specifically for the complexity of the Arabic language. This advanced preprocessing involved multiple steps to enhance the dataset's quality and relevance.

An initial inspection was conducted to identify any anomalies or irregularities in the dataset. Then, the text was standardized by normalizing characters, such as converting different forms of alef آ, إ, أ to a standard form ا, and normalizing the ta marbuta ة to ha ه. This step is crucial for Arabic texts due to the various forms of characters that need to be normalized. Unnecessary characters and elements that could introduce noise into the dataset were removed. This included removing all punctuation marks, stripping out special characters and symbols, and eliminating URLs and Twitter handles. After that, the tweets were tokenized using advanced tokenization techniques for the text to be split into meaningful tokens. The Farasa segmenter [17], which is specifically designed to handle the tokenization in Arabic texts more accurately, was employed. Then, a comprehensive list of Arabic stop words was deployed and common words that do not contribute significantly to the text meaning, such as prepositions, conjunctions, and pronouns, were removed. Then, Chi-square feature selection was utilized, where the following

input of Dataset D with a tweet, each labeled as GenAI-generated or human-written, and the corresponding classification labels can be found. The output was then selected based on the features that best distinguish between categories. The Chi-square statistic is calculated by:

$$\chi 2 = \sum_{i=1}^{n} \frac{(Oi - Ei)^2}{Ei}$$

where *Oi* is the observed frequency and *Ei* is the expected frequency. Later, to reduce words to their root forms, stemming and lemmatization were applied using the ISRI stemmer, and the Farasa Lemmatizer, respectively. This step helped in reducing the dimensionality of the text data and in grouping similar words together. Principal Component Analysis (PCA) was implemented, where the input is the Matrix *X* of the tweets after TF-IDF transformation and the output is the reduced dimension matrix *X'*. Three types of the PCA formula were also utilized. Firstly, the mean was subtracted and divided by the standard deviation:

$$Z = \frac{X - \mu}{\sigma}$$

Then, the covariance matrix of the standardized data was computed:

$$C = \frac{1}{n-1} Z^T Z$$

Finally, eigen decomposition was performed on the covariance matrix:

$$C.v = \lambda .v$$

where λ is an eigenvalue and v an eigenvector. After that, the top k eigenvectors corresponding to the largest k eigenvalues were chosen for the selection of the principal components.

The original data were then projected onto the selected principal components:

$$X' = Z.V_\kappa$$

where V_κ is the matrix of the top k eigenvectors.

Then, part-of-speech tagging was applied to identify and label the grammatical categories of the words, namely nouns, verbs, adjectives, etc. This was useful for further syntactic and semantic analysis. Yet, named entities, such as names of people, organizations, and locations were identified and tagged. This step was crucial for understanding the context and relevance of the tweets. Also, to ensure the uniqueness and relevance of the dataset, any duplicate tweets and redundant data points that could skew the analysis were removed. The frequency of each word across the dataset was then calculated, while bigrams and trigrams were extracted to capture common phrases and contextual information. After that, Term Frequency (TF) and Inverse Document Frequency (ITF) were were simultaneously applied (TF-IDF) to evaluate the importance of terms in the tweets relative to the entire dataset. The frequency of term *t* in document *d* is defined by:

$$\mathrm{TF}(t,d) = \frac{\text{Number of times term t appears in document } d}{\text{Total number of terms in document } d}$$

where *N* is the total number of documents and $|\{d \in \mathrm{D} : t \in d\}|$ is the number of documents containing the term *t*. After that, TF and IDF were combined:

$$\mathrm{TF-IDF}(t,d) = \mathrm{TF}(t,d) \times \mathrm{IDF}(t)$$

Also, a preliminary sentiment analysis was conducted to categorize tweets into GenAI-generated and human-written sentiments. This involved using pre-trained sentiment analysis models for Arabic texts. After that, the preprocessed text was converted into numerical vectors deploying word embedding techniques such as Word2Vec or BERT embeddings. This transformation enabled the feeding of the text data into ML models to be effectively performed.

Finally, techniques like PCA and t-Distributed Stochastic Neighbor Embedding (tSNE) were deployed to reduce the dimensionality of the feature space while retaining the most informative features, as shown above. Indeed, by employing these advanced preprocessing techniques, it was ensured that the dataset was not only clean and standardized, but also rich in features that are essential for accurately distinguishing between gender-specific language in GenAI-generated and human-written tweets. This meticulous preprocessing laid a solid foundation for the subsequent analysis and model training.

(a) (b)

Fig. 3. Word clouds of most used words. (a) Real-human tweets, (b) AI-generated tweets.

C. Observations

There were several points observed when analyzing the dataset and outputs that could distinguish each type of tweet.

1) GenAI-generated Tweets

- The use of colloquial dialect is inaccurate.

"متحمس لزيارة المدينة اليوم! سيكون الأمر ممتعاً جداً. #أوقات_سعيدة"

- Inconsistency in sentences when using colloquial dialect.

"أحب المشي في الحديقة، الهواء النقي رائع. العصافير تغني، إنه مذهل. #عاشق_الطبيعة"

- The ability to express easily and in a short time about any topic.

"تعلمت كيفية خبز كعكة اليوم، سهل وممتع! #حب_الخبز"

- Less use of emojis.

"سأذهب إلى الشاطئ غداً، لا أستطيع الانتظار لرؤية الشمس والرمال. #يوم_الشاطئ"

- Less spontaneous in expression.

"أقرأ كتاباً عن التاريخ القديم، إنه مثير جداً وغني بالمعلومات. #دودة_الكتب"

- There are fewer spelling errors when using colloquial dialect.

"أشاهد غروب الشمس فوق الجبال، منظر جميل جداً. #مناظر_الغروب"

- Good use of punctuation and spaces.

"تناولت عشاءً رائعاً مع العائلة الليلة. طعام ممتاز ورفقة رائعة. #وقت_العائلة"

- Effort to classify tweets with a hashtag.

"أمارس اليوغا هذا الصباح، أشعر بالراحة والهدوء. #حياة_اليوغا"

- Hashtags are usually placed at end of the tweet.

"أستمع إلى موسيقاي المفضلة، طريقة رائعة للاسترخاء. #عاشق_الموسيقى"

- There is a lack of correct understanding of some of the meanings of hashtags due to their use of colloquial dialect.

"انتهيت من جلسة تمرين، أشعر بالنشاط! #عمل_جيد_مرح"

- More organization of words and spaces.

"أحضر مؤتمر عن الذكاء الاصطناعي، أتعلم الكثير. #مؤتمر_التقنية"

- Less use of photos and video.

"أستمتع بفنجان من القهوة في صباح هادئ. #روتين_الصباح"

2) Human-Authored Tweets

- Better use of colloquial dialect.

"ما أقدر أنتظر لحد ما أروح الشاطئ في الويكند، بيكون حماس! 😍🏖 #شاطئ_الشباب"

- Difficulty expressing a topic if there is not enough knowledge.

"قاعد أحاول أفهم البلوك تشين الجديد، شوي صعب. #مشاكل_تقنية"

- More use of emojis.

"شفت فيلم كوميدي أمس، ما قدرت أوقف ضحك 😂 #ليلة_كوميدية"

- More spelling errors.

"أمضي وقت ممتع مع الأصدقاء الليلة، مررة حلو! #ويكند_الشباب"

- Little use of excessive punctuation marks.

"رايح أتسلق الجبال بكرة، تمرين جيد. #طبيعة"

- Hashtags are not used excessively.

"خلصت قراءة كتاب رائع، أنصحكم تقرأوه."

- Random use of spaces, extra lines, and some repeated characters.

"أناا أحبب هالأغنية الجدييدة!!! 🔥🎶🎶 #موسيقى "

- Sometimes punctuation and diacritic are used for decoration.

"يلا نروح نشرب قهوة ☕ وندردش! #حياة_المقاهي"

- More spontaneous in interaction and writing.

"أخيراً حصلت على هاتف جديد 📱 مررة حلو! #ترقية "

After that, the dataset was separated into training and test sets, the feature extraction processes were performed, and then, before model building initiation, the feature normalization, scaling, and dimensionality reduction were carried out to enhance the models' results.

D. ML Models and Techniques

To ensure the reliability of the considered ML models, we employed a 10-fold cross-validation strategy. The dataset was divided into 10 mutually exclusive subsets or "folds". We then repeated the training and evaluation process 10 times, using 9 folds for training and the remaining 1 fold for testing in each iteration. This approach allowed us to leverage the entire dataset without overlap between training and test sets, and ensured that each sample was used exactly once for evaluation. By averaging the performance across the 10 folds, we obtained a robust estimate of the models' generalization capabilities. This study chose three supervised ML to run through the data and make a comparison. The considered models were:

1) Decision Tree (DT)

DTs are a popular form of supervised learning used to predict and model outcomes for given inputs. This tree-like structure tests attributes at internal nodes, with branches representing attribute values and leaf nodes holding the conclusions. A DT can perform both classification and regression tasks [14].

2) Naive Bayes (NB)

NB is a supervised learning method applied to problems of classification based on Bayes' theorem. It is a simple but efficient probabilistic classifier that offers fast prediction power, especially in high-dimensional text categorization tasks [15].

3) Support Vector Machine (SVM)

SVM is one of the most widely utilized methods applied for either classification or regression problems. Working with SVMs means finding the best hyperplane that separates n-dimensional data into classes and getting the extreme data points—that is, support vectors—to define this separating boundary [15].

To prepare the data, feature extraction was performed for this study's models, following the TF-IDF technique. Numerically, TF-IDF represents the importance of every term in a document and, at the same time, in the whole corpus. Terms with a high TF-IDF value are important, appearing more often in a document but less in the general corpus. Combining TF-IDF in the introduced models allowed them to capture the relative importance of different terms in the Arabic tweet dataset and learn patterns, which can improve the precision of the analysis or the classification of Arabic language content.

E. Results and Discussion

This section presents and discusses the results of the models applied to the proposed dataset. For the assessment of the model's performance, the evaluation metrics used are accuracy, F1-score, recall, and precision. In the following subsections, a detailed analysis of each model's performance for every evaluation metric is provided.

1) SVM Model

The SVM model can distinguish a good number of instances in the dataset, given an accuracy of 92% it obtained. From the F1-score of 0.91, the model seems to have captured the positive cases with reduced cases of false positives and false negatives. In the detection of positive instances, the SVM model proved to have a decent recall of 0.899. The precision of 0.925 is a further evidence of that. It means that the model was likely to be correct when it predicted a positive event.

2) Naive Bayes Model

The accuracy rate of the NB model was 93%, which was relatively smaller in magnitude compared to that of the SVM model. The F1-score of 0.92 shows that this model is unable to establish a good balance between precision and recall. With a value of 0.899 for recall, the model was relatively poor at correctly identifying the positive instances, causing a large number of false negatives. On the other hand, the precision value of 0.94 indicates that NB had a low occurrence of false positives, meaning that when it predicted positive instances, it was highly likely to be correct.

3) DT Model

Achieving an accuracy of 79%, which is similar to that of the NB model, the DT model also performed fairly well. The F1-score of 0.80, exhibits a decent balance between precision and recall. The recall value of 0.94 disclosed the effectiveness of catching the positive instances. Similarly, the precision value of 0.699 indicated that the model had a moderate false-positive rate.

As evidenced in Table I, the NB model had the highest accuracy of 92.67%, which suggests that it was able to correctly classify a great number of instances. The SVM model had a strong balanced F1-score of 0.912, indicating that it performed well with a balanced approach between precision and recall.

However, the NB model slightly outperformed the SVM model, with an F1-score of 0.92. The DT model had the lowest accuracy of 78.67%, but it had the highest recall of 0.94, meaning it correctly identified a large proportion of the positive instances. In contrast, the DT model had the lowest precision of 0.70, suggesting that it had a higher rate of false positives. The choice of the most appropriate model depends on the specific requirements and goals of the application. If correctly catching positive instances is the highest priority, then the DT model may be the best choice due to its high recall. If minimizing false positives is more important, then the NB model, with its high precision, could be the best option. The SVM model provides a balance between precision and recall, making it a candidate for many applications. Further analysis and experimentation, such as cross-validation and feature engineering, could potentially improve the performance of these models. Exploring ensemble methods or other algorithms might also help provide even better results.

TABLE I. MODEL RESULTS

Model	Accuracy	F1 Score	Recall	Precision
SVM	92%	91%	90%	93%
NB	93%	92%	90%	94%
DT	79%	80%	94%	70%

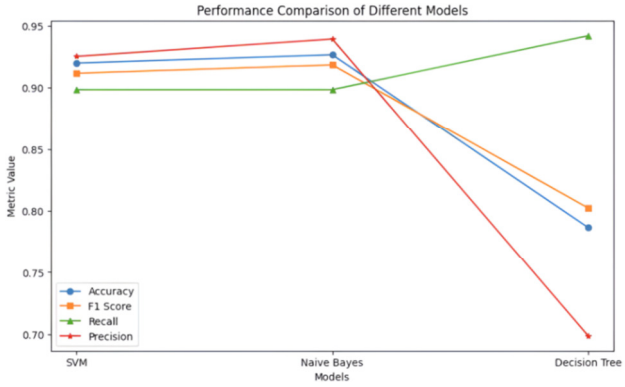

Fig. 4. Machine learning models' results.

Confusion matrices give a brief overview of how the three ML models perform. Even though it does quite well in terms of precision, the SVM model sometimes misclassifies. The Naive Bayes model performs at a relatively higher level with not so many errors. As for the false alarm rate, the DT model outperforms other algorithms. These findings can assist in selecting or enhancing an optimal model for the given classification work at hand.

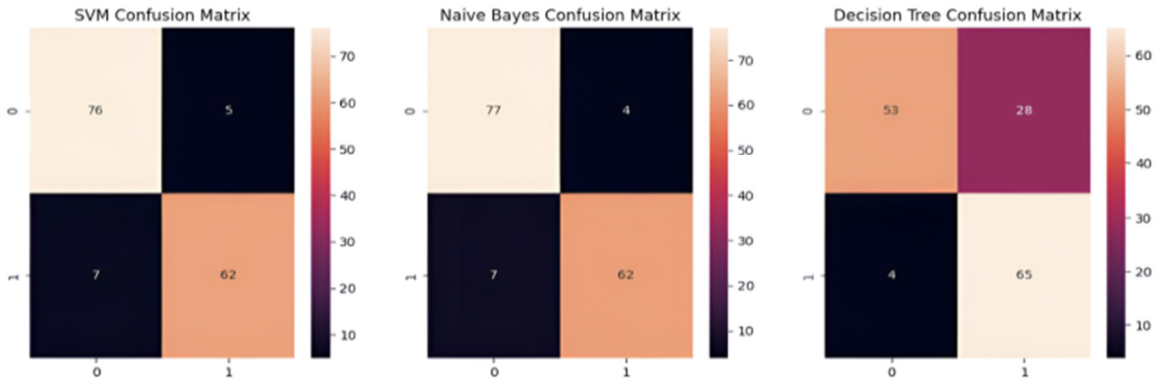

Fig. 5. Confusion matrices of the utilized machine learning models.

F. Threats to Validity

In this research, there are some threats to validity, such as the size of the dataset and the inability to confirm whether some of the accounts relied on tweets created by real humans or not.

V. CONCLUSION AND FUTURE WORK

This paper discusses Generative Artificial Intelligence (GenAI) in the form of content prediction within social networks using the X platform, Tweeter, and Arabic content. It deployed a number of Machine Learning (ML) models, including Decision Trees (DTs), Naive Bayes, and Support Vector Machines (SVMs), which were compared to distinguish between AI-generated and human-made tweets.

This study's results indicate that the SVM model achieved an accuracy of 92% and the Naive Bayes model achieved an accuracy of 93%. The accuracy of the DT model was 79%. Such results indicate the possibility of using ML algorithms for the detection of GenAI content. However, more sophisticated pre-processing techniques and the integration of more diverse datasets are needed to achieve better performance.

All research papers presented in the introduction section presented distinctive approaches and strategies for distinguishing the contents generated by GenAI tools in various fields. However, there is not enough demand for AI-generated Arabic content, particularly on the X platform. That is why this study's focus was placed on finding the AI-generated Arabic tweets in the X platform.

The conducted research has implications in education, finance, information security, legal documents, and any other areas, where identifying AI-generated content is of paramount importance. In light of this, responsible and ethical use of AI technology must include the ability to distinguish between AI-generated and human-created content for the mitigation of the risks of disinformation and potential harm. Future studies in this regard will be carried out with larger data sets and more advanced techniques, including Deep Learning (DL) algorithms, in order to make the predictive models more accurate. Besides, the practical application of this research in developing tools for identifying and countering the false information generated by AI will come in handy.

This work should be able to contribute to the mounting pile of knowledge relating to GenAI and its impact on social networks. Better admissibility of GenAI content will build up a safer and more knowledgeable online environment while reaping the benefits of AI technologies in positive applications.

REFERENCES

[1] H. Yu and Y. Guo, "Generative artificial intelligence empowers educational reform: current status, issues, and prospects," *Frontiers in Education*, vol. 8, Jun. 2023, Art. no. 1183162, https://doi.org/10.3389/feduc.2023.1183162.

[2] I. Augenstein *et al.*, "Factuality Challenges in the Era of Large Language Models." arXiv, Oct. 09, 2023, https://doi.org/10.48550/arXiv.2310.05189.

[3] R. Shijaku and E. Canhasi, "ChatGPT Generated Text Detection," Jan. 2023, https://doi.org/10.13140/RG.2.2.21317.52960.

[4] I. Katib, F. Y. Assiri, H. A. Abdushkour, D. Hamed, and M. Ragab, "Differentiating Chat Generative Pretrained Transformer from Humans: Detecting ChatGPT-Generated Text and Human Text Using Machine Learning," *Mathematics*, vol. 11, no. 15, Jan. 2023, Art. no. 3400, https://doi.org/10.3390/math11153400.

[5] R. Tang, Y.-N. Chuang, and X. Hu, "The Science of Detecting LLM-Generated Text," *Communications of the ACM*, vol. 67, no. 4, pp. 50–59, Nov. 2024, https://doi.org/10.1145/3624725.

[6] N. Islam, D. Sutradhar, H. Noor, J. T. Raya, M. T. Maisha, and D. M. Farid, "Distinguishing Human Generated Text From ChatGPT Generated Text Using Machine Learning." arXiv, May 26, 2023, https://doi.org/10.48550/arXiv.2306.01761.

[7] Y. Chen, H. Kang, V. Zhai, L. Li, R. Singh, and B. Raj, "GPT-Sentinel: Distinguishing Human and ChatGPT Generated Content." arXiv, May 17, 2023, https://doi.org/10.48550/arXiv.2305.07969.

[8] N. Islam, D. Sutradhar, H. Noor, J. T. Raya, M. T. Maisha, and D. M. Farid, "Distinguishing Human Generated Text From ChatGPT Generated Text Using Machine Learning." arXiv, May 26, 2023, https://doi.org/10.48550/arXiv.2306.01761.

[9] C. Vasilatos, M. Alam, T. Rahwan, Y. Zaki, and M. Maniatakos, "HowkGPT: Investigating the Detection of ChatGPT-generated University Student Homework through Context-Aware Perplexity Analysis." arXiv, Jun. 07, 2023, https://doi.org/10.48550/arXiv.2305.18226.

[10] S. Chakraborty, A. S. Bedi, S. Zhu, B. An, D. Manocha, and F. Huang, "On the Possibilities of AI-Generated Text Detection." arXiv, Oct. 02, 2023, https://doi.org/10.48550/arXiv.2304.04736.

[11] K. Hayawi, S. Shahriar, and S. Mathew, "The Imitation Game: Detecting Human and AI-Generated Texts in the Era of Large Language Models," Jul. 2023, [Online]. Available: https://www.researchgate.net/publication/372583505_The_Imitation_Game_Detecting_Human_and_AI-Generated_Texts_in_the_Era_of_Large_Language_Models.

[12] M. Perkins *et al.*, "GenAI Detection Tools, Adversarial Techniques and Implications for Inclusivity in Higher Education." arXiv, Mar. 28, 2024, https://doi.org/10.48550/arXiv.2403.19148.

[13] M. Fattah and M. A. Haq, "Tweet Prediction for Social Media using Machine Learning," *Engineering, Technology & Applied Science Research*, vol. 14, no. 3, pp. 14698–14703, Jun. 2024, https://doi.org/10.48084/etasr.7524.

[14] M. Madhukar and S. Verma, "Hybrid Semantic Analysis of Tweets: A Case Study of Tweets on Girl-Child in India," *Engineering, Technology & Applied Science Research*, vol. 7, no. 5, pp. 2014–2016, Oct. 2017, https://doi.org/10.48084/etasr.1246.

[15] H. Zhang and H. Shao, "Exploring the Latest Applications of OpenAI and ChatGPT: An In-Depth Survey," *Computer Modeling in Engineering & Sciences*, vol. 138, no. 3, pp. 2061–2102, 2024, https://doi.org/10.32604/cmes.2023.030649.

[16] K. Darwish and H. Mubarak, "Farasa: A New Fast and Accurate Arabic Word Segmenter," in *Tenth InternationalConference on Language Resources and Evaluation*, Portoroz, Slovenia, Dec. 2016, pp. 1070–1074.

[17] "Word Segmentation Module," *Farasa*. https://farasa.qcri.org/segmentation/.