FAKESWARM: IMPROVING FAKE NEWS DETECTION WITH SWARMING CHARACTERISTICS

Jun Wu[1]and Xuesong Ye[2]

[1]College of Computing, Georgia Institute of Technology, Atlanta, United States
[2]College of Graduate and Professional Studies, Trine University, Phoenix, United States

ABSTRACT

The proliferation of fake news poses a serious threat to society, as it can misinform and manipulate the public, erode trust in institutions, and undermine democratic processes. To address this issue, we present FakeSwarm, a fake news identification system that leverages the swarming characteristics of fake news. We propose a novel concept of fake news swarming characteristics and design three types of swarm features, including principal component analysis, metric representation, and position encoding, to extract the swarm behavior. We evaluate our system on a public dataset and demonstrate the effectiveness of incorporating swarm features in fake news identification, achieving an f1-score and accuracy over 97% by combining all three types of swarm features. Furthermore, we design an online learning pipeline based on the hypothesis of the temporal distribution pattern of fake news emergence, which is validated on a topic with early emerging fake news and a shortage of text samples, showing that swarm features can significantly improve recall rates in such cases. Our work provides a new perspective and approach to fake news detection and highlights the importance of considering swarming characteristics in detecting fake news.

KEYWORDS

Fake News Detection, Metric Learning, Clustering, Dimensionality Reduction

1. INTRODUCTION

1.1. Background and Motivation

Fake news poses a significant risk to users of social media platforms, as it can cause psychological and health-related issues, increase polarization, and erode public trust. The vast amounts of false information available on these platforms, coupled with their ability to spread rapidly and without restriction, can lead to widespread panic and anxiety among users. Studies have shown that exposure to fake news can cause stress, depression, and other mental health problems. Furthermore, the use of text generation technologies and DeepFake techniques, have made it increasingly challenging to differentiate between authentic and fake news. As such, there is a growing need to develop machine learning algorithms that can detect and combat fake news. Recent research has shown that natural language processing techniques, social network analysis, and deep learning can be used to identify fake news with high accuracy. However, much work remains to be done in this field.

1.2. Fake News Detection

Several surveys and overviews have been conducted to provide a comprehensive understanding of the fake news phenomenon, its characteristics, detection methodologies, and future research opportunities. Zhou *et al.* [1] provide a survey covering fundamental theories, detection methods

David C. Wyld et al. (Eds): NLPML, AIAP, SIGL, CRIS, COSIT, DMA -2023
pp. 175-187, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130815

including machine learning and NLP techniques, and future research opportunities for fake news, emphasizing the impact of fake news on users and the importance of developing innovative detection methods. Zhang *et al.* [2] offer a comprehensive analysis of online fake news, discussing its characteristics and detection methodologies, and fostering academic discussion around potential solutions, providing insights into the evolving landscape of fake news and its consequences. Shu *et al.* [3] emphasize the importance of social context in fake news detection by examining the impact on users and exploring a range of detection methods that consider user behavior and interactions, highlighting the role of social dynamics in identifying deceptive content.

The existing fake news detection research can be roughly divided into two categories: user perspective and content based. The first category focuses on user behavior, network structure, and user profile features to model the correlation between user activities and the spread of fake news. The second category leverages various types of data, such as text, images, and metadata, to provide a more accurate and holistic understanding of potentially deceptive content.

User Perspective: User perspective features, such as user behavior and network structure, play an essential role in detecting fake news on social media platforms. They consider the user's role in spreading fake news, user interactions and relationships within social networks, and user profile features, including demographics, interests, and history. By analyzing these features, researchers can gain a deeper understanding of user behavior and more accurately identify potential fake news content. Shu *et al.* [4] examine user perspective features, such as user behavior and network structure, to detect fake news on social media using data mining techniques and various recognition models, considering the user's role in spreading fake news. Monti *et al.* [5] utilize geometric deep learning to analyze user social graphs for fake news detection, emphasizing user interactions and relationships within social networks, enabling a more accurate assessment of the likelihood of fake news based on users' involvement. Shu *et al.* [6] develop a fake news detection model focusing on user profile features, including demographics, interests, and history, which allows for a deeper understanding of user behavior and more accurate identification of potential fake news content. Ruchansky *et al.* [7] propose a hybrid deep learning model that analyzes temporal patterns in user activities to effectively detect fake news on social media platforms, considering the timing of user interactions and engagements for fake news identification. Sahoo *et al.* [8] create a robust fake news detection model on social networks by combining deep learning with user involvement features, such as connections and interactions with others, emphasizing the importance of understanding user dynamics and relationships.

Content Based: In recent years, researchers leveraged text, image and multimodal data to detect fake news. These methods combine various types of data, such as text, images, and metadata, to provide a more accurate and holistic understanding of potentially deceptive content. Khattar *et al.* [9] introduce a multi-modal variational autoencoder (MVAE) model that fuses textual and visual information for fake news detection, offering a comprehensive approach to understanding and detecting misleading content using both text and image data. Singhal *et al.* [10] propose a multi-modal framework utilizing text, images, and metadata in combination with deep learning models for effective fake news detection, leveraging various types of data to provide a more accurate and holistic understanding of potentially deceptive content. Granik *et al.* [11] employ a Naive Bayes classifier for fake news detection based on textual features and probabilistic reasoning, considering the likelihood of text being fake news and offering a statistical method for identification. Kaliyar *et al.* [12] present a deep convolutional neural network (CNN) model for fake news detection that analyzes text with high accuracy, leveraging the strengths of CNNs in text analysis and feature extraction to detect deceptive content. Karimi *et al.* [13] develop a model using advanced NLP and machine learning techniques to detect fake news by combining multiple sources and classes of information, recognizing the importance of considering various

aspects of content and user behavior. Wang *et al.* [14] propose an event adversarial neural network that leverages multi-modal data (text, images, and metadata) to improve fake news detection performance, utilizing multiple data types to offer a comprehensive understanding of content for more accurate detection. Kaliyar *et al.* [15] utilize a BERT-based deep learning model for effective detection and classification of fake news in social media content, leveraging the power of BERT for a highly accurate and context-aware fake news detection method.

1.3. Attack and Defense in Other Layers

Most research proves that sophisticated attacking methods exist in layers beyond the fake content itself. These methods involve the utilization of social bots to breach social media platforms and the use of adversarial learning and multimedia data generation techniques to confuse detection models.

Social Bots: Online social applications are vital platforms for people to share and read news. However, bot accounts are flooding almost every popular social media application, and many research works have started to reveal their harmful activities, such as spreading fake news, rumors, and misinformation. Emilio *et al.* [16] revealed the rising process of social bots, which are software agents developed by the black industry to mimic humans. Stefano *et al.* [17] investigated account performance discrimination between genuine accounts and bot accounts on Twitter. Giovanni *et al.* [18] explored the negative influence of social bots spreading misinformation on Facebook.

Detecting and deleting fake news from the platform is a direct defense strategy, but discovering and banning bot accounts can solve the problem at its root. Maryam *et al.* [19] built account embeddings to transform metadata, such as age, gender, and personality, into a feature space for bot classification. Shangbin *et al.* [20] investigated heterogeneity structures in the account relation graph, spreading information among nodes, and discriminating between genuine users and social bots. BotShape [21] built a novel and accurate detection system based on the disparity of account behavioral patterns between bots and real humans. BotTriNet [22] is a content model for bot detection that first applies metric learning to increase the distance between bots and normal users in the embedding space, resulting in significant accuracy improvement for content-less bot categories. BotMoE [20] is a Twitter bot detection system that uses multiple modalities, such as textual content and network topology, to improve detection accuracy.

Model Security: We consider two aspects of model security: (i) the robustness of the detection model against evasion attempts and (ii) the speed of detection to prevent the spread of fake news. Model attacks have dual purposes: they provide a way to bypass model identification and also offer opportunities to enhance model robustness. The black industry leverages multimedia data generation and adversarial learning technologies to evade detection.

For textual data generation, Rik *et al.* [23] used a Transformer to generate text from a knowledge graph. Zhuoyi *et al.* [24] designed a novel rephrase detection system based on contextual content in dialogue scenes. Image generation and synthesis techniques are widely utilized. Tao *et al.* [25] provided a systematic introduction to DeepFake generation and detection. LiveBugger [26] systematically studied the security of facial liveness verification systems and improved the attack success rate by up to 70%. Qingzhao *et al.* [27] investigated adversarial examples to exploit prediction errors in vehicle trajectories. Jiachen *et al.* [28] proposed a novel black-box adversarial sensor attack targeting the security of autonomous driving perception models. To defend against adversarial examples, Changjiang *et al.* [29] designed a transferability-based approach for both black and gray box attacks on deep neural networks.

Real-time detection of fake news is crucial because online users can immediately consume them once they are published. Knowledge distillation is the process of transferring knowledge from a complex model to a simpler model, especially for deep models. Jianping *et al.* [30] provide a comprehensive survey of knowledge distillation techniques. Yuke *et al.* [31] proposed a distillation-based inverse-network attack via partitioning the neural network model. Souvik *et al.*[32] designed a model compression method by removing ReLU layers and merging them with preceding layers.

1.4. Contributions

To summarize, this paper makes the following contributions:
– We designed FakeSwarm, a fake news identification system that incorporates an innovative concept of fake news swarming characteristics. The system includes three types of generic features and extraction methods, namely principal component analysis, metric representation, and position encoding.

– We evaluated the three types of swarm features on a public dataset and analyzed their contributions to the classification accuracy of the original text. By combining all three types of swarm features, our system achieved an f1-score and accuracy of over 97%, demonstrating the effectiveness of incorporating swarm features in fake news identification.

– Based on the hypothesis of the temporal distribution pattern of fake news emergence, we designed an online learning pipeline. We validated the pipeline on a topic with early emerging fake news and a shortage of text samples, demonstrating that swarm features can significantly improve recall rates in such cases.

2. DATASET

Ahmed *et al.* [33] and [34] designed text-based fake news detection systems based on a real-world fake news dataset they collected called ISOT FAKE NEWS. The dataset includes news articles from 2015 to 2018, with 23,481 fake news and 21,417 real news articles. The fake news articles were collected from different sources flagged by fact-checking organizations like Politifact and Wikipedia. In contrast, truthful articles were obtained by crawling articles from Reuters.com.

Each news article in the dataset is provided with the title, body text, subject, published date, and a label that distinguishes between fake and real news. The fake news articles cover multiple subjects, including Left News, Politics, Middle-East, News, Government News, and US News, while the real news articles are categorized into Politics News and World News. Categories of fake news and real news are mutually inclusive, meaning that subjects are not useful information for distinguishing between fake and real news.

We observed that the sources of some texts were correlated with the labels. For example, body texts with a beginning of Reuters were always real news. To prevent source information from leaking the labels, we removed the source information from each body text in ISOT FAKE NEWS, making the dataset as objective and unbiased as possible.

3. DESIGN

We present FAKESWARM, a fake news detection system leveraging the swarming characteristics of a few news to improve the detection accuracy. It takes the body texts of news as input data and predicts whether the news is fake or not.

The approach leverages the concept of swarming, which refers to the tendency of fake news to spread quickly and widely across online platforms. A swarm is a group of fake news with similar topics that occur in a concentrated period of time.

To enhance the swarming attributes in the original text embeddings learned from news body texts, FAKESWARM designed three approaches: (i) principal embedding produced by principal components analysis (PCA); (ii) metric embedding produced by contrastive learning; (iii) position embedding produced by a clustering algorithm. Principal embedding is the projection of raw text embedding with the most important and reduced dimensionality, which actively pushes data points in a swarm closer. Metric learning is a representation of raw text embedding learned by unsupervised contrastive learning loss, which actively increases the distance between fake news and realnews and also reduces the distance between fake news and swarms. Position embedding stands for the distances between news to all clustering centers of text embeddings, respectively. If the news is close to an arbitrary clustering center, the probability of belonging to a swarm will increase.

We will first introduce how to transform raw news body texts into text embeddings. And then, we explain how these swarming characteristic embedding mining approaches and their implementation.

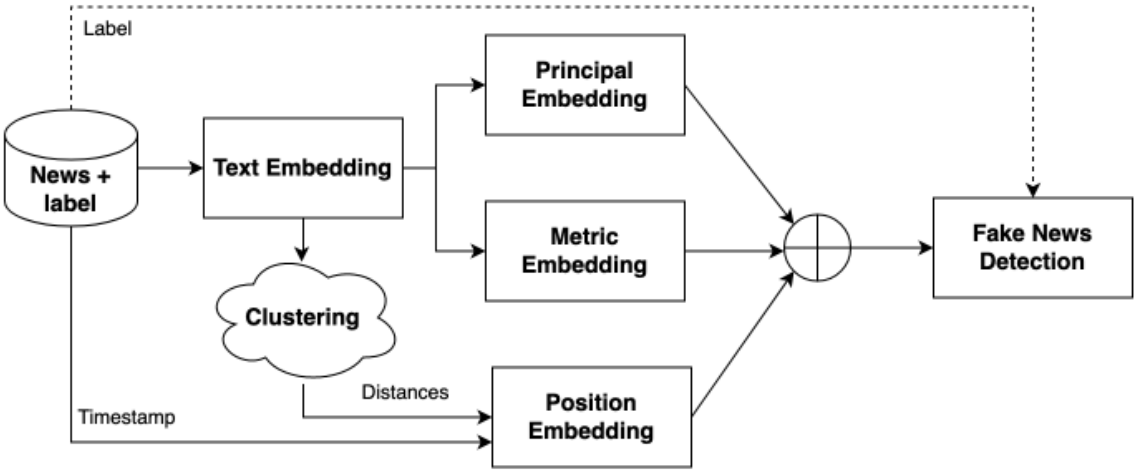

Figure 1. FAKESWARM architecture

3.1. Text Embeddings

Text embedding is a comprehensive task in NLP and includes various targeting embedding content at different levels, such as word embedding, sentence embedding, and document embedding. Word embedding is the basic fuel for all text embedding tasks because it transforms a word in the physical world into a word vector in semantic space. Higher-level embeddings are the aggregation and combination of word embeddings. FAKESWARM considers the text embedding

of news as a document embedding problem and utilizes word embeddings and pooling techniques to aggregate text embedding.

Word Embedding: Word embedding is a technique used in natural language processing to create numerical vectors that capture the semantic meaning of words. Word embeddings help to overcome the limitations of traditional bag-of-words models, which treat words as independent and unstructured units. With word embeddings, NLP models can more accurately understand the meaning of the text, leading to improved performance on a wide range of tasks.

FAKESWARM selected Word2Vec [35] as the word embedding algorithm because it performed better on the domain-specific corpus. The system takes the new body texts in ISOT FAKE NEWS as the corpus. Each news is a long text with multiple sentences. The system first splits news into sentences through punctuation and then sets a hyperparameter to limit the maximal count of surrounding words for Word2Vec learning. Under the setting of embedding dimensionality, Word2Vec outputs a word embedding (a numerical vector format) for each word, comprehensive semantic information of all appearances of the words.

Text Embedding: Nal *et al.* [36] achieved state-of-the-art results on several sentence classification tasks by using word embeddings as input and applied *average pooling* to obtain a fixed-size sentence embedding. Therefore, we also applied the average pooling technique to generate the text embedding of news. Each dimension is the average of feature values in that dimension of all words in that news.

3.2. Principal Embedding

Principal Component Analysis (PCA) is a technique used to reduce the dimensionality of high-dimensional data while retaining the most important information. The method works by finding a lower-dimensional representation of the data that captures as much of the variability in the data as possible.

The implementation of PCA involves calculating the eigenvectors and eigenvalues of the covariance matrix of the data. The eigenvectors represent the principal components of the data, while the eigenvalues represent the amount of variance captured by each principal component. The eigenvectors are sorted in descending order based on their corresponding eigenvalues, and the first few eigenvectors are selected to create the lower-dimensional representation of the data.

FAKESWARM leverages PCA to generate new embeddings with a small count of dimensionality, which is called Principal Embedding. Our system sets the default value of reduced dimensionality as a small number of 3. We infer fake news belonging to the same swarm would get closer in the space of *Principal Embedding* because reducing dimensionality will reduce the global distances between every pair of data points and naturally reduces the distance of the data point inside each swarm. Evaluation results in sub section 4.2 prove that *Principal Embedding* effectively improves the detection accuracy.

3.3. Metric Embedding

FAKESWARM adopts contrastive learning to represent raw text embeddings, drawing inspiration mainly from our previous works: BotTriNet [22] and FineEHR [37]. These works employed a metric-based approach to create sentence embeddings, enhancing the performance of downstream text classification tasks and surpassing the performance of the original sentence embeddings.

We consider contrastive learning could adjust the inner distances between fake news in the same swarm and define the represented embedding as the Metric Embedding. Contrastive learning learns a new embedding from raw text embedding based on the optimizing object of reducing the distances between instances with the same label and also increasing the distances between instances having different labels. Naturally, fake news would be actively pushed together, the local density of each swarm would increase, and swarms would get far away from real news.
FAKESWARM uses a multilayer perceptron structure as the embedding network for representing the raw text embedding to the new Metric Embedding and also applies the contrastive loss to optimize the parameters of the perceptron. The contrastive loss is:

$$Loss = \mathrm{Y} * D^2 + (1 - Y) * max(margin - D, 0)^2$$

With Y , the label of two sampled instances (both are fake or real, label = 1; otherwise, label = 0), D is the Euclidean distance between two metric embeddings. margin is a hyper-parameter to adjust the distances between instances with different classes (fake or real).

3.4. Position Embedding

We made an assumption that news in a shared swarm will distribute close in the semantic space, and the distances between their text embedding are close too. However, mining features to measurethe relationship between a data point, and its corresponding swarm is challenging due to several factors. First, the set of swarms (clusters) is varying and hard to decide. In other words, it is a traditional challenge in clustering problems. For example, the classic algorithm K-Means requires setting a hyperparameter of the number of clusters by hand. Second, correspondingly, the belong- ing swarm of a new is hard to decide. We consider simply bonding a news text embedding to the closest swarm (cluster) not enough due to the natural limitation of the clustering algorithm.

To solve the first problem, FAKESWARM chooses DBSCAN as the clustering algorithm. DB-SCAN is particularly useful for datasets with irregular shapes and clusters of different densities, as it can identify clusters of any shape and size. The algorithm is also able to distinguish noise points that do not belong to any cluster. Also, it has the ability to automatically determine the number of clusters. To solve the second challenge, FAKESWARM not only involves the identification of belonging swarm as a positional feature but also extracts the distances from one news text embed- ding to all cluster centers respectively as extra positional features. The swarm identification and the distances make up the *Positional Embedding*.

3.5. Fake News Detection

FAKESWARM uses the concat of text embedding, principal embedding, metric embedding and positional embedding as a complete feature vector for fake news detection. We consider that concat embedding will improve the accuracy because it combines the semantic information and potential swarming characteristic for each piece of news.

The system integrates various classifiers for fake news detection, e.g., Logistic Regression (LR), Multilayer Perceptron (MLP), Random Forest (RF), and Gradient Boosted Decision Tree (GBDT). The classifier uses the concat embeddings as the feature vectors and uses the labels in the training data to fit a correlation function between the features and the label. The prediction process is similar, where the model uses the concat features of news to judge whether it is fake.

4. EVALUATION

4.1. Ground-Truth and Metrics

The concept of ground-truth is vital in assessing the performance of predictive models using various metrics. In the ISOT FAKE NEWS dataset, we divided all fake and real news articles into twoportions using random splitting: a training set (comprising 70 percent) and a testing set (comprising 30 percent). Based on the predicted and actual labels of each sample, there are four possibleoutcomes: (i) True Positive (TP); (ii) False Positive (FP); (iii) True Negative (TN); (iv) FalseNegative (FN). We utilized two widely-accepted metrics, accuracy and f1-score, to assess the performance of FAKESWARM in detecting fake news. Accuracy is calculated as $\frac{TP+TN}{TP+FP+TN+FN}$ and represents the proportion of instances correctly classified, including both true and false instancesin their respective categories. F1-score is a combined measure of precision and recall, computed as $\frac{2 \cdot precision \cdot recall}{precision+recall}$. Precision, defined as $\frac{TP}{TP+FP}$, measures the accuracy of identified positive instances. A higher f1-score indicates improved precision and recall.

4.2. Effects of Swarming Characteristics

To address the unique characteristics of fake news networks, we devised three novel feature embeddings in addition to the baseline text embeddings: principal embeddings, metric embeddings, and positional embeddings. These new features aimed to enhance the recognition capabilities of our detection model. I conducted experiments to validate the effectiveness of each of these four embeddings individually in terms of accuracy and f1-score for fake news detection. For a comprehensive comparison, we applied four classic classification algorithms to test their performance

Table 1. Performance of Swarming Embeddings

	GBDT		**RF**		**LR**		**MLP**	
	Accuracy	**F1score**	**Accuracy**	**F1score**	**Accuracy**	**F1score**	**Accuracy**	**F1score**
Text	96.29%	96.45%	95.63%	95.84%	95.63%	95.84%	95.63%	95.84%
Prinicipal	93.27%	93.55%	92.97%	93.28%	93.00%	93.29%	92.91%	93.21%
Metric	**96.85%**	**96.99%**	**97.33%**	**97.45%**	**96.60%**	**96.75%**	**96.55%**	**96.69%**
Position	87.02%	87.55%	85.42%	86.40%	85.42%	86.40%	85.42%	86.40%

Table 2. Performance of Concat Embeddings and FakeSwarm

	GBDT		**RF**		**LR**		**MLP**	
	Accuracy	**F1score**	**Accuracy**	**F1score**	**Accuracy**	**F1score**	**Accuracy**	**F1score**
Text+Prinicipal	96.31%	96.47%	97.25%	97.37%	95.96%	96.13%	96.47%	96.62%
Text +Metric	96.96%	97.09%	97.69%	97.80%	**96.56%**	**96.71%**	96.55%	96.68%
Text +Position	96.20%	96.36%	97.28%	97.40%	96.05%	96.22%	96.51%	96.65%
FakeSwarm	**96.98%**	**97.11%**	**97.84%**	**97.94%**	96.54%	96.69%	**96.70%**	**96.84%**
Improvement	**0.69%**	**0.67%**	**2.20%**	**2.11%**	**0.90%**	**0.85%**	**1.07%**	**1.01%**

on the testing dataset. Table 1 shows the result, where the performance of metric embedding is the best, even better than text embedding. Principal embedding also performs well. The result is

reasonable because metric embedding and principal embedding are all representation of text embedding and retains semantic information. The performance of position embedding is not so good. But, think about it another way, it still achieved a very high accuracy of 87% using pure swarming characteristic without direct text semantic information, and better than most classifications in [34].

We concatenated the text embeddings with various swarming embeddings and utilized a classifier for fake news detection to evaluate the performance of multimodal features. Incorporating multiple types of information has the potential to improve the detection capabilities of the model. The rationale behind this approach is that diverse embeddings can capture different aspects of fake news patterns, leading to a more comprehensive understanding of the underlying data. By combining text embeddings with swarming embeddings, we aim to extract richer features, thereby enhancing the classifier's ability to distinguish between real and fake news. Based on the experimental results shown in Table 2, we obtain several important conclusions. First, FAKESWARM acquires the best performance on both f1score and accuracy by combining three kinds of swarming embeddings and text embeddings. Second, the concat of text embedding and arbitrary swarming embedding largely improved the performance compared with single text embedding. Third, the performance of position embedding was improved to the average level after combining with text embedding, proving that it is still effective for improving prediction accuracy even having no single-applying performance.

4.3. Compare With Previous Works

We compare FAKESWARM with the previous work [34] that published ISOT FAKE NEWS dataset. This research work applied the N-gram to generate tokens and used the TF-IDF metric to measure the frequency and importance of each gram. It also applied various classifiers to predict fake news and proved the Linear SVM achieved the best accuracy. The authors only show the accuracy score in the paper so we mainly compare it with our system using the accuracy score.
Table 3 shows the results. We conclude that text embedding has already been better than the best classifier in [34]. Especially when combined with three swarming embeddings, FAKESWARM achieved a very high accuracy with a very large improvement compared to the pure text embedding.

Table 3. Performance of Swarming Embeddings

Approaches	**Accuracy**	**F1score**
Ahmed et al. [34] KNN	83.00%	/
Ahmed et al. [34] DT	89.00%	/
Ahmed et al. [34] LR	89.00%	/
Ahmed et al. [34] SGD	89.00%	/
Ahmed et al. [34] LSVM	92.00%	/
FakeSwarmText	95.63%	95.84%
FakeSwarmText+Swarming	97.84%	97.94%

4.4. Fake News Detection in Early Stage

Under the assumption that fake news is swarming appearing through the date time, we hope to simulate the fake news detection process in a real production environment. We construct streaming-format train and test data sets. In detail, we set a time parameter month, referring to the

index of the month from the first detection date to the current detection date. We pick up 14 complete month from ISOT FAKE NEWS with stable news collection. The training set is all the news from the first month to the current month, and the testing set is the news published in the next month. FAKESWARM uses the training set for text and swarming embedding generation and detects fake news appearing in the time window of the next month.

Figure 2 and Figure 3 show the f1score and accuracy score varying through the increasing of *month*. *Baseline* refers to using text embedding for fake detection, and FAKESWARM refers to using text embedding and swarming embeddings for detection. The result shows that our system largely improves the detection accuracy at the early stage without enough training data (Table 4 shows the count of practical instances). And after the data starve stage, FAKESWARM still has a stable improvement of accuracy compared to the baseline approach, which proves that the effectiveness of swarming embeddings is robust through time.

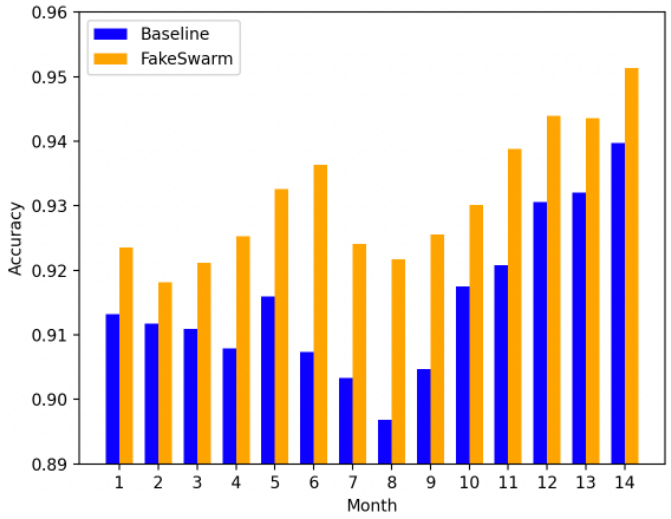

Figure 2. Monthly Accuracy of FAKESWARM and Baseline

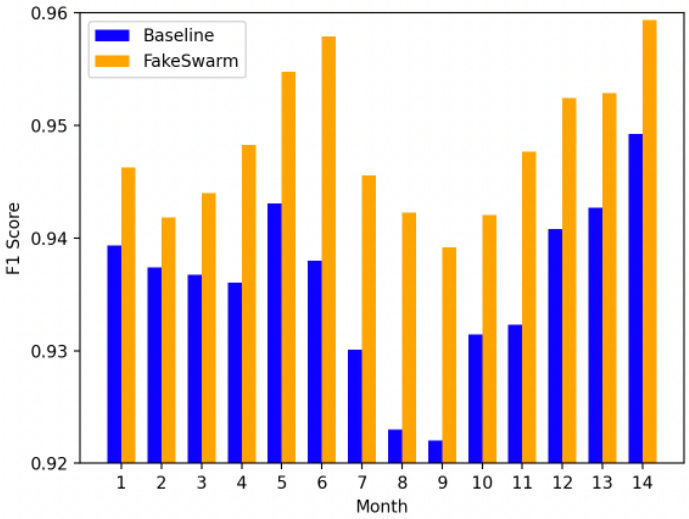

Figure 3. Monthly F1score of FAKESWARM and Baseline

Table 4. The count of instances in monthly training and testing data set

Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Train Instances	1162	2424	3512	4637	5669	6646	7571	8601	9735	11048	12114	13538	14714	16114
Test Instances	4344	4056	3917	3668	3665	3862	4346	4392	4753	4582	4999	4568	4430	4050

5. CONCLUSIONS

We introduced FAKESWARM, a novel fake news identification system that utilizes swarming characteristics to enhance detection accuracy. By incorporating three types of swarm features, namely principal component analysis, metric representation, and position encoding, we demonstrated the effectiveness of considering swarming characteristics in fake news detection.

Our evaluation on a public dataset revealed that combining all three types of swarm features achieved an impressive f1-score and accuracy of over 97%, becoming the start-of-art detection system. We also developed an online learning pipeline to simulate the real production environment, we validated that our system still perform robust and accurate particularly in the early stages with limited training data.

In summary, FAKESWARM introduces a fresh perspective and approach to fake news detection, emphasizing the significance of swarming characteristics in identifying and addressing the challenges posed by the proliferation of false information on social media platforms. We hope that future research will focus more on exploring swarming characteristics and investigating methods to further enhance the effectiveness of fake news detection.

REFERENCES

[1] X. Zhou and R. Zafarani, “A survey of fake news: Fundamental theories, detection methods, and opportunities,” ACM Computing Surveys (CSUR), vol. 53, no. 5, pp. 1–40, 2020.

[2] X. Zhang and A. A. Ghorbani, “An overview of online fake news: Characterization, detection, and discussion,” Information Processing & Management, vol. 57, no. 2, p. 102025, 2020.

[3] K. Shu, S. Wang, and H. Liu, “Beyond news contents: The role of social context for fake news detection,” in Proceedings of the twelfth ACM international conference on web search and data mining, 2019, pp. 312–320.

[4] K.Shu,A.Sliva,S.Wang,J.Tang,andH.Liu,“Fakenewsdetectiononsocialmedia:Adataminingperspective,”ACM SIGKDD explorations newsletter, vol. 19, no. 1, pp. 22–36, 2017.

[5] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, “Fake news detection on social media usinggeometric deep learning,” arXiv preprint arXiv:1902.06673, 2019..

[6] K.Shu,S.Wang,andH.Liu,“Understandinguserprofilesonsocialmediaforfakenewsdetection,”in2018IEEEconference on multimedia information processing and retrieval (MIPR). IEEE, 2018, pp. 430–435.

[7] N.Ruchansky,S.Seo,andY.Liu,“Csi:Ahybriddeepmodelforfakenewsdetection,”inProceedingsofthe2017ACM on Conference on Information and Knowledge Management, 2017, pp. 797–806.

[8] S. R. Sahoo and B. B. Gupta, “Multiple features based approach for automatic fake news detection on social networks using deep learning,” Applied Soft Computing, vol. 100, p. 106983, 2021.

[9] D.Khattar,J.S.Goud,M.Gupta,andV.Varma,“Mvae:Multimodalvariationalautoencoderforfakenewsdetection,” in The world wide web conference, 2019, pp. 2915–2921.

[10] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, “Spotfake: A multi-modal framework for fake news detection,” in 2019 IEEE fifth international conference on multimedia big data (BigMM). IEEE, 2019, pp. 39–47.

[11] M.GranikandV.Mesyura,“Fakenewsdetectionusingnaivebayesclassifier,”in2017IEEEfirstUkraineconference on electrical and computer engineering (UKRCON). IEEE, 2017, pp. 900–903.

[12] R.K.Kaliyar,A.Goswami,P.Narang,andS.Sinha,"Fndnet–adeepconvolutionalneuralnetworkforfakenews detection," Cognitive Systems Research, vol. 61, pp. 32–44, 2020.

[13] H.Karimi,P.Roy,S.Saba-Sadiya,andJ.Tang,"Multi-sourcemulti-classfakenewsdetection,"inProceedingsof the 27th international conference on computational linguistics, 2018, pp. 1546–1557.

[14] Y.Wang, F.Ma,Z.Jin, Y.Yuan, G.Xun, K.Jha, L.Su, and J.Gao, "Eann:Eventadversarialneuralnetworksfor multi-modal fake news detection," in Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining, 2018, pp. 849–857.

[15] R.K.Kaliyar, A.Goswami, and P.Narang ,"Fakebert :Fakenews detection in social media with a bert-based deep learning approach," Multimedia tools and applications, vol. 80, no. 8, pp. 11 765–11 788, 2021.

[16] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," Communications of the ACM, vol. 59, no. 7, pp. 96–104, 2016.

[17] Cresci, Stefano and Di Pietro, Roberto and Petrocchi, Marinella and Spognardi, Angelo and Tesconi, Maurizio, "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race," in Proceedings of the 26th international conference on world wide web companion, 2017, pp. 963–972.

[18] G. C. Santia, M. I. Mujib, and J. R. Williams, "Detecting social bots on facebook in an information veracity context," in Proceedings of the international AAAI conference on web and social media, vol. 13, 2019, pp. 463– 472.

[19] M. Heidari, J. H. Jones Jr, and O. Uzuner, "Online user profiling to detect social bots on twitter," arXiv preprint arXiv:2203.05966, 2022.

[20] S.Feng,Z.Tan,R.Li,andM.Luo,"Heterogeneity-awaretwitterbotdetectionwithrelationalgraphtransformers," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, 2022, pp. 3977–3985.

[21] J. Wu, X. Ye, and C. Mou, "Botshape: A novel social bots detection approach via behavioral patterns," in 12th International Conference on Data Mining & Knowledge Management Process, 2023.

[22] J. Wu, X. Ye, and M. Y. Yuet, "Bottrinet: A unified and efficient embedding for social bots detection via metric learning," arXiv preprint arXiv:2303.03144, 2023.

[23] R.Koncel Kedziorski, D.Bekal, Y.Luan, M.Lapata, and H.Hajishirzi,"Text generation from knowledge graphs with graph transformers," arXiv preprint arXiv:1904.02342, 2019.

[24] Z. Wang, S. Gupta, J. Hao, X. Fan, D. Li, A. H. Li, and C. Guo, "Contextual rephrase detection for reducing friction in dialogue systems," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 1899–1905.

[25] T. Zhang, "Deepfake generation and detection, a survey," Multimedia Tools and Applications, vol. 81, no. 5, pp. 6259–6276, 2022.

[26] C. Li, L. Wang, S. Ji, X. Zhang, Z. Xi, S. Guo, and T. Wang, "Seeing is living? rethinking the security of facial liveness verification in the deepfake era," in 31st USENIX Security Symposium (USENIX Security 22), 2022, pp. 2673–2690.

[27] Q. Zhang, S. Hu, J. Sun, Q. A. Chen, and Z. M. Mao, "On adversarial robustness of trajectory prediction for autonomous vehicles," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 15 159–15 168.

[28] J. S. Sun, Y. C. Cao, Q. A. Chen, and Z. M. Mao, "Towards robust lidar-based perception in autonomous driv- ing: General black-box adversarial sensor attack and countermeasures," in USENIX Security Symposium (Usenix Security'20), 2020.

[29] C.Li,H.Weng,S.Ji,J.Dong,andQ.He,"Det:Defendingagainstadversarialexamplesviadecreasingtransferability," in Cyberspace Safety and Security: 11th International Symposium, CSS 2019, Guangzhou, China, December 1–3, 2019, Proceedings, Part I 11. Springer, 2019, pp. 307–322.

[30] J.Gou,B.Yu,S.J.Maybank,andD.Tao,"Knowledgedistillation:Asurvey,"InternationalJournalofComputer Vision, vol. 129, pp. 1789–1819, 2021.

[31] Y.Zhang,D.Chen,S.Kundu,H.Liu,R.Peng,andP.A.Beerel,"C2pi:Anefficientcrypto-cleartwo-partyneural network private inference," arXiv preprint arXiv:2304.13266, 2023.

[32] S. Kundu, Y. Zhang, D. Chen, and P. A. Beerel, “Making models shallow again: Jointly learning to reduce nonlinearity and depth for latency-efficient private inference,” arXiv preprint arXiv:2304.13274, 2023.
[33] H. Ahmed, I. Traore, and S. Saad, “Detection of online fake news using n-gram analysis and machine learning techniques,” in Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First Interna- tional Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1. Springer, 2017, pp. 127–138.
[34] Ahmed, Hadeer and Traore, Issa and Saad, Sherif, “Detecting opinion spams and fake news using text classifica- tion,” Security and Privacy, vol. 1, no. 1, p. e9, 2018.
[35] T.Mikolov,K.Chen,G.Corrado,andJ.Dean,“Efficientestimationofwordrepresentationsinvectorspace,” arXiv preprint arXiv:1301.3781, 2013.
[36] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, “A convolutional neural network for modelling sentences,” arXiv preprint arXiv:1404.2188, 2014.
[37] J.Wu,X.Ye,C.Mou,andW.Dai,“Fineehr:Refineclinicalnoterepresentationstoimprovemortalitypredicti on,” arXiv preprint arXiv:2304.11794, 2023.

AUTHORS

Jun Wu holds a Master’s degree in computer technology from Tsinghua University and is currently pursuing a Master’s degree in Computer Science (AI-Track) at the Georgia Institute of Technology. Her research interests include machine learning, graph learning, anomaly detection, and their application in social networks and biomedical computing.

Xuesong Ye received Bachelor’s Degree from the Chengdu University of Information Technology in Electronic and Information Engineering. He is pursuing his Master's Degree in Information Science from Trine University. His research interests include Machine Learning, Computer Vision, and Natural Language Processing.