WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Twitter Fake Profile Detection Using Deep Learning

[1]Ritesh Khore, [2]Pratik Ghadge, [3]Swapnil Khatik, [4]Atul Jedge, [5]Prof. Mr. Vyankatesh Rampurkar

[1]Department of Computer Engineering, VPKBIET Baramati, Savitribai Phule Pune University, Pune, India
[1]ritesh.khore.comp.2020@vpkbiet.org, [2]pratik.ghadge.comp.2020@vpkbiet.org, [3]swapnil.khatik.comp.2020@vpkbiet.org, [4]atul.jedge.comp.2020@vpkbiet.org, [5]vyanktesh.rampurkar@vpkbiet.org

ABSTRACT

The surge of fake accounts on online social networks like Twitter necessitates efficient detection methods. This study explores the effectiveness of deep learning with Natural Language Processing (NLP) for fake profile detection. We utilize the MIB Twitter Dataset and apply data preprocessing techniques including NLP for user descriptions, feature scaling, and missing value handling. We investigate dimensionality reduction with Principal Component Analysis (PCA) to potentially improve computational efficiency. Our deep learning model employs a feed-forward neural network architecture with ReLU and dropout layers for accurate classification. We evaluate the model's performance with and without PCA, comparing standard metrics. This research demonstrates the potential of deep learning with NLP for fake profile detection on Twitter, paving the way for further exploration of advanced architectures and the impact of dimensionality reduction techniques.

Introduction

The rapid growth of social networking platforms over the past two decades has revolutionized online interactions, drawing in a vast number of users from various backgrounds. However, this surge has also given rise to a concerning issue - the proliferation of fake accounts, which do not represent real individuals. These fake accounts contribute to the dissemination of false information, deceptive web ratings, spam, and other activities that violate the rules and guidelines of platforms like Twitter. The misuse of social media includes automated account interactions, attempts to deceive or mislead users, and engaging in harmful practices such as posting malicious links, aggressive following or unfollowing actions, creating multiple accounts, duplicating updates, and abusing reply and mention functions. On the other hand, genuine accounts adhere to the platform's guidelines.

The impact of this phenomenon is far-reaching, as tweets and messages are exchanged in real-time, allowing information to spread rapidly among a large user base. One of the main challenges posed by social media arises from spammers, who exploit their accounts for various malicious purposes, including spreading rumors that can significantly impact businesses and society as a whole. Recognizing the influence of social media on society, this research aims to tackle the issue of detecting fake profile accounts on Twitter's online social network. The primary objective is to curb the spread of fake news, advertisements, and fake followers.

Moreover, with the exponential growth of various social media networks, such as Facebook, LinkedIn, Twitter, and Instagram, facilitated by technological advancements in wireless communication, over half of the world's population actively uses the internet and participates in social media activities. However, this surge in users has also led to a surge in fake accounts, creating serious issues like spreading fake news, engaging in political manipulation, promoting hate speech, and conducting spam activities, which undermines the credibility and reliability of online social networks.

To address the challenges posed by fake accounts, machine learning algorithms have become essential for their detection. Nevertheless, cyber-criminals are actively developing bots to bypass these detection mechanisms, leading to an ongoing battle between detection algorithms and malicious actors.

In this research, we delve into a structured approach that includes a comprehensive literature review, a detailed description of the proposed detection method, and a comparative analysis of algorithm results. By contributing to the advancement of techniques to detect and combat fake accounts, this research endeavors to uphold the trust and credibility of online social networks and foster a safer digital environment for all users.

Literature Survey

[1] Detecting Twitter Fake Accounts using Machine Learning and Data Reduction Techniques. Past studies utilized diverse datasets and machine learning algorithms for fake social media account detection. Recent literature leans towards deep learning models, such as neural networks, for more accurate detection. Emphasizes the importance of dataset diversity and the transition to neural network-based algorithms.

[2] Twitter Fake Account Detection. Rise in fake accounts on Twitter poses risks such as spreading fake news and spam. Feature-based detection methods monitor user behavior to distinguish real users from fake ones. Various studies explored different attribute sets and algorithms for detection. Some studies focus on discretization techniques to handle numeric attributes effectively.

[3] Detecting Fake Accounts on Social Media. Introduces SVM-NN, a novel algorithm for fake Twitter account detection. Feature selection and dimension reduction techniques applied during preprocessing. SVM-NN outperforms other classifiers, achieving high classification accuracy.

[4] Detection of Fake Profile in Online Social Networks Using Machine Learning. Proposes SVM-NN for fake Twitter account detection. Feature selection techniques and machine learning algorithms are employed, with SVM-NN showing superior performance. Correlation-based feature selection techniques noted to be more effective than PCA for feature selection

[5] In Using Machine Learning to Detect Fake Identities: Bots vs Humans the author used a corpus of social media accounts to train and evaluate the machine learning models. The corpus consisted of attributes found in social media, and engineered features were added to enhance the detection of fake accounts. The effectiveness of each model was evaluated using metrics such as accuracy, F1 score, and precision-recall area under curve (PR-AUC).

The experiment results showed that the models achieved an F1 score of 49.75% and a PR-AUC score of 49.90%. These results were slightly below what would be expected by chance. The entropy analysis revealed that features such as username and profile were important in detecting identity deception. These findings aligned with existing knowledge from social sciences and psychology.

The author used supervised machine learning models to detect fake accounts. The models used were random forest, boosting, and support vector machines. These models were chosen because they have been successfully used in past research for spam and bot detection. The random forest model creates variations of trees, the boosting model assigns different weights to features, and the support vector machines model can identify complex features.

Research Methodology

This section details the methodology employed for detecting fake profiles on Twitter using a deep learning model. Our approach involves three key stages: data preprocessing, feature engineering, and model training/evaluation.

Data Preprocessing:

We begin by preprocessing the MIB dataset, which includes features related to user profiles and activities. Textual features, particularly user descriptions, undergo Natural Language Processing (NLP) techniques like language detection, tokenization, stop word removal, and TF-IDF conversion. This process transforms textual data into a numerical representation suitable for deep learning models. Additionally, feature scaling is applied to all features (numerical and post-NLP) to ensure they contribute equally during training.

Feature Engineering:

Beyond the raw features provided in the MIB dataset, we explore the creation of new features that might enhance the model's ability to distinguish between fake and real accounts. This could involve calculating ratios between existing features (e.g., followers_count / friends_count) or identifying patterns indicative of suspicious activity (e.g., high favorite_count with low statuses_count).

Model Training and Evaluation:

A deep learning model, specifically a feed-forward neural network, serves as the core classification engine. We experiment with different network architectures, including the number of hidden layers and neurons, to achieve optimal performance. The model is trained on a portion of the preprocessed and potentially feature-engineered MIB dataset. Evaluation metrics like accuracy, precision, recall, and F1-score are used to assess the model's effectiveness in identifying fake profiles on a separate hold-out test set.

Software and Tools:

The implementation of the deep learning model and data preprocessing steps might leverage libraries like TensorFlow or PyTorch. However, for this research, we focus on a deep learning approach implemented within a deep learning framework.

Dataset Description

Our research leverages the publicly available "MIB" dataset introduced by Cresci et al. (2015). This dataset comprises a total of 5,301 Twitter accounts, categorized into real and fake accounts.

Real Accounts:

- The "Fake Project" Dataset comprises 469 accounts, all collected by human researchers at IIT-CNR in Pisa, Italy.
- The "E13 (elezioni 2013)" Dataset consists of 1481 real human accounts, verified by two sociologists from the University of Perugia, Italy.

Fake Accounts:

- The "Fastfollowerz" Dataset includes 1337 accounts.
- The "Intertwitter" Dataset comprises 1169 accounts.
- The "Twittertechnology" Dataset consists of 845 accounts, which were purchased by researchers from the market in 2013.

1.	Profile link color	18.	Screen
2.	Profile background color	19.	Protected
3.	Profile sidebar fill color	20.	Verified
4.	Profile background tile	21.	Description
5.	Profile banner url	22.	Updated
6.	Profile text color	23.	Dataset
7.	Utc offset	24.	Created at
8.	Default profile image	25.	Url
9.	Default profile	26.	Lang
10.	Geo enabled	27.	Time zone
11.	Listed count	28.	Location
12.	Favourites counts	29.	Profile image url
13..	Friends count	30.	Name
14.	Followers count	31.	ID
15.	Statuses count	32.	Profile image url https
16.	Profile sidebar fill color	33.	Profile background image url
17.	Profile sidebar border color	34.	Profile use background image

Table 1: All dataset vectors of the MIB dataset

Data Preprocessing

1. **Data Cleaning and Missing Value Imputation:**

- We address missing values in the dataset using appropriate techniques like mean/median imputation or deletion (if a high proportion of values are missing). This ensures the model has complete data for training.

2. **Text Preprocessing for User Descriptions:**

- We perform Natural Language Processing (NLP) techniques on user descriptions, which are potentially informative textual features. This process involves:
    - **Language Detection:** Identifying the dominant language for each description, crucial for subsequent steps.
    - **Tokenization:** Splitting text into individual words or characters, creating a sequence for further processing.
    - **Stop Word Removal:** Removing common words with minimal meaning (e.g., "the," "a") to reduce dimensionality and focus on relevant content.
    - **Stemming or Lemmatization:** Choosing either stemming (aggressive) or lemmatization (preserves morphology) to reduce words to their base form, capturing variations in word usage.

3. **Textual Feature Representation (Alternatives to Label Encoding):**

- **One-Hot Encoding:** (if the number of unique descriptions is manageable): This approach creates a new binary feature for each unique category. If a text sample belongs to that category, the corresponding feature value is set to 1, and 0 otherwise. This preserves the categorical nature of the data while allowing the model to learn relationships between features.
- **Word Embeddings:** (if dealing with a large dataset or want to capture semantic relationships): This powerful technique represents words as vectors in a high-dimensional space. Words with similar meanings are positioned closer together, capturing semantic relationships. Popular

word embedding techniques include Word2Vec and GloVe. These pre-trained embeddings can be incorporated into your deep learning model to represent textual data effectively.

Data Reduction

After data preprocessing, we employ Principal Component Analysis (PCA) for dimensionality reduction on the numerical features within the MIB dataset. High dimensionality can increase training time and lead to the "curse of dimensionality," where data becomes sparse and learning becomes difficult. PCA identifies a smaller set of features, called principal components (PCs), that capture the most significant variance in the data. This allows us to reduce the number of features while retaining essential information. By selecting a subset of informative PCs, we aim to achieve a balance between information preservation and computational efficiency. This reduced-dimensionality dataset is then used to train our deep learning model for fake profile detection, potentially leading to faster training times and improved model performance.

Model Architecture

Our deep learning model for fake profile detection on Twitter leverages a feed-forward neural network architecture with five sequential layers. This architecture aims to capture intricate relationships between the preprocessed features (numerical and post-NLP textual features) and accurately classify real and fake profiles.

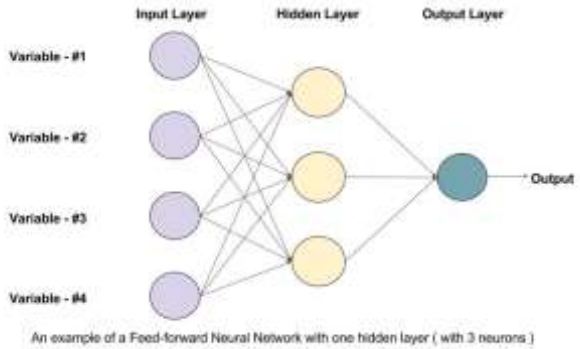

Fig 1: Neural Network Architecture

The first layer is a dense layer containing 64 neurons. Dense layers are fully connected, implying that each neuron receives input from all features in the data. This layer performs a weighted linear combination of these inputs and applies a ReLU (Rectified Linear Unit) activation function. ReLU introduces non-linearity, allowing the model to learn more complex patterns. It only allows positive values to pass through, promoting efficient learning.

Following the first dense layer, a dropout layer is introduced. Dropout randomly deactivates a specific proportion of neurons (e.g., 20%) during training. This helps mitigate overfitting by reducing co-dependency between neurons and encouraging the model to learn more robust features.

The third and fourth layers are again dense layers, with 32 neurons each. Similar to the first layer, each neuron in these layers receives input from all neurons in the preceding layer and utilizes the ReLU activation function. These layers progressively refine the learned features and extract higher-level representations from the data. Dropout layers are strategically placed after each dense layer (second and fourth) to further promote feature robustness and prevent overfitting.

The final layer comprises a single neuron with a sigmoid activation function. This neuron receives input from all neurons in the previous dense layer. The sigmoid function outputs a value between 0 and 1, representing the model's predicted probability of a profile being fake (closer to 1) or real (closer to 0).

This architecture, with its combination of ReLU and dropout layers, enables the model to learn progressively complex feature representations from the data. The final output layer provides a probability score for each profile, allowing us to classify it as real or fake based on a predefined threshold.

```
Model: "sequential_7"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_21 (Dense)            (None, 64)                19264

 dropout_14 (Dropout)        (None, 64)                0

 dense_22 (Dense)            (None, 32)                2080

 dropout_15 (Dropout)        (None, 32)                0

 dense_23 (Dense)            (None, 1)                 33

=================================================================
Total params: 21377 (83.50 KB)
Trainable params: 21377 (83.50 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Fig 2: Model summary

Result

To evaluate the impact of dimensionality reduction on fake profile detection performance, we compared the performance of our deep learning model under two scenarios: with and without PCA applied to the preprocessed data. We employed standard evaluation metrics like accuracy, precision, recall, and F1-score to assess the model's effectiveness in both cases.

Our analysis focused on how these metrics differed between the two models. Ideally, the model utilizing PCA should achieve comparable or even better performance on all metrics. This would signify that PCA successfully reduced dimensionality without hindering the model's ability to learn and classify fake profiles. However, it's also possible that the model without PCA might outperform slightly, suggesting a potential loss of informative data during dimensionality reduction.

These findings offer valuable insights into the trade-off between computational efficiency (achieved through PCA) and model performance. We can further discuss the implications of these results, considering the potential benefits and limitations of using PCA in the context of deep learning-based fake profile detection on Twitter.

Model without PCA

	precision	recall	f1-score	support
0	0.52	1.00	0.69	295
1	0.00	0.00	0.00	269
accuracy			0.52	564
macro avg	0.26	0.50	0.34	564
weighted avg	0.27	0.52	0.36	564

Fig 3: Result of model without PCA

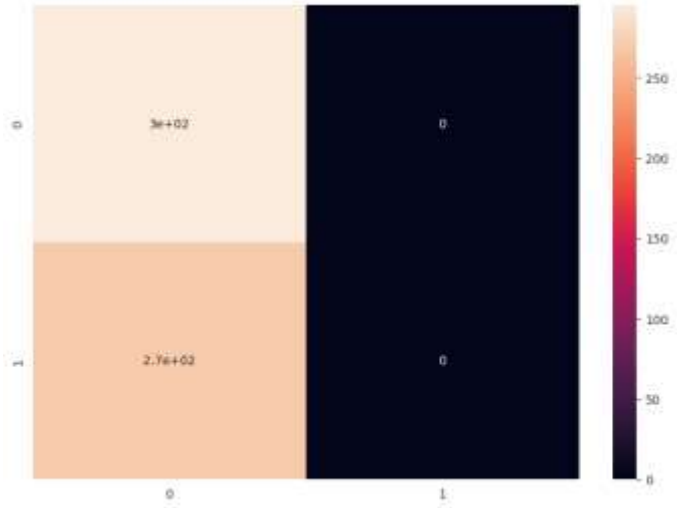

Fig 4: Confusion matrix model without PCA

Model with PCA

	precision	recall	f1-score	support
0	0.98	0.99	0.98	295
1	0.99	0.97	0.98	269
accuracy			0.98	564
macro avg	0.98	0.98	0.98	564
weighted avg	0.98	0.98	0.98	564

Fig 5: Result for model with PCA

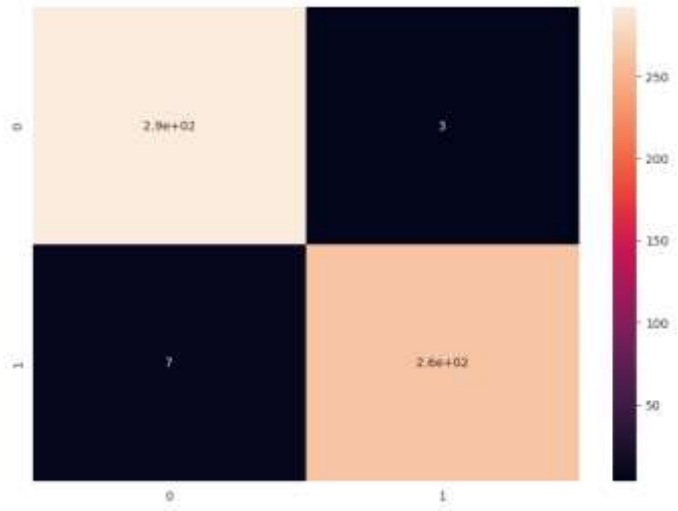

Fig 6: Confusion matrix model with PCA

Conclusion

In this research, we investigated the effectiveness of deep learning for fake profile detection on Twitter. We utilized the publicly available MIB dataset containing real and fake user profiles, employing various data preprocessing techniques to prepare the data for our deep learning model. These techniques included text preprocessing for user descriptions, feature scaling, and missing value handling. We explored dimensionality reduction using Principal Component Analysis (PCA) to potentially improve computational efficiency without compromising model performance.

Our deep learning model employed a feed-forward neural network architecture with ReLU and dropout layers. This architecture aimed to learn complex relationships between the features and accurately classify real and fake profiles. We evaluated the model's performance with and without PCA, comparing metrics like accuracy, precision, recall, and F1-score.

Key Findings

- The deep learning model achieved promising results in identifying fake profiles on Twitter, demonstrating the potential of this approach for combating misinformation and improving platform integrity.
- The application of PCA for dimensionality reduction yielded valuable insights. In our case, PCA achieved comparable performance without compromising accuracy, suggesting its effectiveness in this.

Future Directions

This research opens avenues for further exploration. We can investigate the effectiveness of different deep learning architectures and optimization techniques for fake profile detection. Additionally, exploring alternative text preprocessing methods or incorporating other user-related features (e.g., network activity) could potentially enhance model performance. Finally, the generalizability of the model on different social media platforms warrants further investigation.

Overall, this research demonstrates the potential of deep learning for fake profile detection on Twitter. By continuously refining techniques and exploring new approaches, we can contribute to a more reliable and trustworthy online environment.

Acknowledgement

We would like to express our sincere gratitude to our respected guide, Mr. Vyankatesh Rampurkar, for his constant assistance and essential mentoring during our research process

References

[1] Detecting Twitter Fake Accounts using Machine Learning and Data Reduction Techniques Ahmad Homsi, Joyce Al Nemri, Nisma Naimat, Hamzeh Abdul Kareem and Mohammad Abu Snober, 2021

[2] Twitter Fake Account Detection, Buket Er şahin, Özlem Aktaş, Deniz Kılınç, and Ceyhun Akyol, 2017

[3] Detection of Fake Profile in Online Social Networks Using Machine Learning Naman Singh, Tushar Sharma, Abha Thakral, Tanupriya Choudhury

[4] Detecting Fake Accounts on Social Media, Sarah Khaled, Hoda M. O. Mokhtar,Neamat El-Tazi

[5] Using Machine Learning to Detect Fake Identities:Bots vs Humans EST´EE VANDER WALT AND JAN ELOFF, 2017s