International Journal of Innovative Research in Engineering and Management (IJIREM)
ISSN: 2350-0557, Volume-10, Issue-2, April 2023
https://doi.org/10.55524/ijirem.2023.10.2.24
Article ID IJIR2364, Pages 121-124
www.ijirem.org

Deep Learning Approaches for Twitter User Classification

J Krishna Kishore[1], Jayasankar Sai Krishan[2], SK Arafath[3], CH Ashok[4], and K Nithin Reddy[5]

[1,2,3,4,5] Department of Computer Science &Engineering, PACE Institute of Technology & Sciences, Ongole, Andhra Pradesh, India

Correspondence should be addressed to J Krishna Kishore; kk.jandrajupalli@gmail.com

ABSTRACT- Twitter, a popular social media platform, has become a rich source of user-generated content. The classification of Twitter users based on their characteristics and behavior has gained significant attention. Deep learning techniques, with their ability to capture complex patterns and representations, have emerged as powerful tools for Twitter user classification. This research article presents a comprehensive review of deep learning approaches for Twitter user classification. We discuss various deep learning architectures, pretraining techniques, and transfer learning strategies used in the classification task. Through a thorough analysis of existing studies, we highlight the strengths and limitations of deep learning approaches and provide recommendations for future research in this field.

I. INTRODUCTION

Twitter is a widely popular social networking platform that has revolutionized the way people communicate and share information online. Launched in 2006, Twitter allows users to post short messages called "tweets" of up to 280 characters, along with various multimedia elements like photos, videos, and GIFs. It has become a hub for real-time news updates, social commentary, and global conversations [1].

The value placed on succinct communication is one of Twitter's defining characteristics. Because of the character limit, users are encouraged to communicate their opinions in as few words as possible, making it a platform that thrives on interactions that are brief and immediate. Microblogging, in which users post brief morsels of information, views, or observations on a wide variety of subjects, has become increasingly popular as a result of this trend. Microblogging has contributed to the emergence of the social networking site Twitter [2].

A wider audience uses Hashtags, which are used on Twitter, to organize users' tweets and make them searchable. Twitter is famous for its usage of hashtags. Users are able to follow certain subjects or events, join groups that debate issues that are of interest to them, and engage in trending discussions thanks to hashtags. Because of this feature, Twitter has become an effective instrument for organizing social movements, engaging in activism, and raising awareness about a variety of causes.

Real-time conversations and direct engagement between individuals, including celebrities, public figures, corporations, and regular users, are encouraged by the open and public aspect of the platform, which also makes the site accessible to the general public. It has evolved into a venue for establishing connections with individuals who have similar perspectives, participating in discussions, requesting assistance, and forming communities based on common passions. Twitter has also evolved into a medium for customer service, allowing consumers to communicate with companies and brands for help or feedback.

The influence of Twitter extends well beyond the realm of personal relationships, as the platform has been essential in the evolution of the media environment. Journalists, news outlets, and influencers utilize twitter for the purpose of disseminating news updates, sharing articles, and engaging with one's audience. The platform has made it possible for anybody with access to the internet to have their thoughts heard, to dispute narratives, and to contribute to public conversation. This has the effect of democratizing the flow of information [3].

Twitter, however, has its own unique set of difficulties, just like any other social network. The brevity of tweets can occasionally lead to misunderstandings or a lack of context, which can result in misunderstood meanings or even online harassment. Both of these issues can be caused by a lack of information. The platform has implemented capabilities that allow users to mute, block, and report content that is abusive in an effort to counteract these concerns and address them head-on.

In general, Twitter has developed into a powerful and prominent social networking platform, making it possible to have conversations all over the world, breaking news, and connecting people from a wide variety of backgrounds. Its influence on society and culture is ever expanding, and it continues to be an important factor in the digital environment, helping to mold public debate and fostering connections in the virtual world.

II. LITERATURE REVIEW

There are multiple steps involved in the process of applying machine learning to create a poll on Twitter. The following is a high-level summary of the various ways in which you can approach it:

Define the questions for the survey as follows: Make a decision on the kinds of questions that will be included in your survey. Because tweets are limited to a certain number of characters, make sure they are clear and succinct.

Preprocessing the data involves cleaning and organizing the information obtained from Twitter users. In order to accom-

plish this, it may be necessary to eliminate responses that are irrelevant, deal with missing data, and standardize the structure of the answers.
Analysis of sentiment (optional): If you wish to examine the sentiment of the responses, you can use machine learning techniques like Natural Language Processing (NLP) to determine whether or not the responses are positive, negative, or neutral. This will allow you to evaluate whether or not the responses should be analyzed.
Model of machine learning Select an appropriate machine learning model by taking into consideration the nature of the survey questions you are asking and the kind of data you will be collecting. You may use a categorization model, for instance, if your survey consists of questions with several possible answers. Text analysis methods such as recurrent neural networks (RNNs) or transformer-based models could be a good choice for situations in which open-ended queries are being asked.
Training the model: When training the machine learning model, use data that has been labeled. If you plan to perform sentiment analysis as part of your process, you will also require labeled data for that stage.
During this step, you will deploy the trained model to a server or cloud platform that is able to process incoming requests from Twitter users.
Integration of the Twitter API: If you want to collect responses to your survey questions, you can use the Twitter API. You are able to create a tweet containing the survey and ask users to react with their responses to that tweet.
Processing responses: In order to process the responses, first collect them from Twitter and then preprocess them using the same methods that were used when preprocessing the data for the training of the model. Utilize the trained machine learning model to make predictions about the replies based on the data that has been preprocessed. In order to properly portray the findings of the study, it is necessary to first perform an analysis of the survey results and then produce visualizations of those results.
Ethical considerations: Ensure that you follow ethical rules while collecting data from Twitter users and that you utilize the data responsibly, respecting users' privacy and consenting to the collection of their data.
Please be aware that developing a comprehensive survey system for Twitter that is based on machine learning can be a challenging endeavor that may call for expertise in machine learning, data processing, and the Twitter Application Programming Interface (API). In order to ensure that your model generates results that are relevant, it is vital to test and validate it in great detail.

III. RELATED WORK

Twitter user classification using deep learning techniques is an area of research that aims to categorize and understand users based on their behavior, interests, or other characteristics. Deep learning, a subset of machine learning, utilizes neural networks with multiple layers to learn complex patterns and representations from large amounts of data [4].
User classification on Twitter can serve various purposes, such as targeted advertising, content personalization, identifying fake accounts or bots, sentiment analysis, and understanding user demographics. Here's a general overview of how deep learning can be applied to Twitter user classification:
Data Collection: The first step is to gather a large dataset of Twitter users and their associated attributes, such as profile information, tweet history, follower/following relationships, and engagement metrics. This dataset forms the foundation for training and testing the deep learning model.
Feature Extraction: Deep learning models require numerical input, so the collected data needs to be preprocessed and transformed into suitable features. This can involve techniques like word embedding or converting categorical data into numerical representations. Text-based features can be extracted using methods such as word2vec, GloVe, or BERT to capture semantic information from tweets.
Model Architecture: Deep learning models like convolutional neural networks (CNNs), recurrent neural networks (RNNs), or transformer models can be utilized for Twitter user classification. CNNs are effective for processing text and extracting local patterns, while RNNs capture sequential dependencies in temporal data. Transformer models like BERT have been successful in understanding context and semantics in text.
Training and Validation: The prepared dataset is divided into training and validation sets. The deep learning model is trained on the training set, optimizing its parameters using backpropagation and gradient descent. Validation data is used to monitor the model's performance and prevent overfitting. Hyperparameter tuning, such as learning rate, number of layers, or dropout rate, can be explored to enhance the model's accuracy.
Evaluation: Once the model is trained, it is evaluated on a separate test dataset to assess its performance. Metrics such as accuracy, precision, recall, and F1 score can be used to measure the model's effectiveness in classifying Twitter users into the desired categories.
Deployment and Application: The trained deep learning model can be deployed to classify new, unseen Twitter users in real-time. This can be integrated into existing systems or applications to provide insights or enable personalized experiences based on user classification results.
It's important to note that the success of deep learning for Twitter user classification relies heavily on the quality and diversity of the dataset, as well as careful feature engineering and model design. The field of deep learning is continually evolving, with new architectures and techniques emerging, offering promising avenues for improving Twitter user classification accuracy and expanding its applications

IV. DEEP LEARNING FOR TWITTER ANALYSIS

Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are both popular deep learning architectures that can be used for Twitter user classification tasks. Let's explore how each of these models can be applied in this context:

CNN for Twitter User Classification: Convolutional Neural Networks are widely used for image recognition tasks, but they can also be adapted for text classification, including Twitter user classification. Here's how CNNs can be utilized:

- Text Preprocessing: The text data, such as tweets or user profiles, needs to be preprocessed before feeding it into a CNN. This involves steps like tokenization, removing stop words, and converting words into numerical representations.
- Embedding Layer: An embedding layer is utilized to represent words or tokens as dense vectors with semantic information. Popular word embedding techniques like word2vec or GloVe can be used to initialize this layer.
- Convolutional Layers: The main component of a CNN, convolutional layers, apply filters over the embedded text, capturing local patterns or n-grams. These filters slide across the text representation, performing convolutions and generating feature maps.
- Pooling Layers: Pooling layers, such as max pooling or average pooling, reduce the dimensionality of the feature maps. They select the most important information from the convolved features, improving computational efficiency and extracting relevant features.
- Fully Connected Layers: The output from the pooling layers is flattened and passed through fully connected layers, followed by activation functions like ReLU or sigmoid. These layers learn the high-level representations and make predictions based on the extracted features.
- RNN for Twitter User Classification: Recurrent Neural Networks are well-suited for sequential data, making them suitable for capturing the temporal nature of Twitter user behavior or tweet sequences. Here's how RNNs can be applied:
- Sequence Preparation: In the case of Twitter user classification, the tweet history of each user can be considered as a sequence. The tweets are ordered based on time, and the sequence is created by concatenating the textual information.
- Embedding Layer: Similar to CNNs, an embedding layer is used to convert words into dense vectors that capture semantic information. The embedding layer can be pre-trained or learned during the training process.
- RNN Layers: The core component of an RNN is its recurrent layer, such as Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU). These layers maintain a hidden state that captures the context of previous words in the sequence and helps make predictions based on that context.
- Fully Connected Layers: The output from the RNN layers can be passed through fully connected layers, similar to the CNN approach. These layers learn higher-level representations and make predictions based on the sequential information extracted by the RNN layers.
- Handling Variable-Length Sequences: Since tweets can have varying lengths, techniques like padding or truncation can be used to ensure uniform input length for efficient processing by the RNN layers.

Both CNNs and RNNs have their strengths in Twitter user classification. CNNs excel in capturing local patterns and textual features, while RNNs are effective in modeling temporal dependencies and sequential information. Depending on the specific task and available data, researchers and practitioners can choose the most suitable architecture or even combine CNNs and RNNs in hybrid models for improved performance in Twitter user classification.

V. PROPOSED APPROACH AND ANALYSIS OF RESULTS

Acquiring a suitable Twitter dataset for classification tasks can be challenging due to privacy concerns and data usage restrictions. However, here are a few publicly available datasets that can serve as starting points for Twitter classification tasks:
Datasets:
The analysis of structured data has seen widespread use. In this scenario, the conventional Relational Database Management System (RDBMS) is able to handle the data. A single computer processor is incapable of processing such a large amount of data due to the growing amounts of unstructured data on numerous sources (such as data from the web, social media, and blogs), which are collectively referred to as Big Data. Therefore, the RDBMS is unable to deal with the unstructured data; in order to process the data, you will need a nontraditional database, which is known as a NoSQL database.
The majority of studies concentrated on instruments, such as R (which is both a programming language and a software environment for data analysis). R is not very effective when dealing with big volumes of data and has several limitations when it comes to processing data from Twitter. An open source Java framework that is used for the processing and querying of enormous volumes of data on huge clusters of commodity hardware is known as Apache Hadoop [5]. A hybrid big data framework, such as this one, is typically used to solve this challenge. In addition to unstructured and semi-structured data, such as XML and JSON files, Hadoop can also process structured data. The use of Hadoop is beneficial because it allows for the storage and processing of a huge volume of data, but the use of R is beneficial because it allows for the analysis of data that has already been processed.
A user's profile and their tweets are two examples of the various forms of data that may be found on Twitter. While the latter is believed to be dynamic, the former is thought to be static. Tweets may be in the form of text, photos, videos, URLs, or even spam tweets.
Because spam tweets and robotic tweeting engines can frequently alter the accuracy of analysis results, as well as add noise and bias to those results, the majority of studies typically do not take into consideration any of these factors. The mechanism of the FireFox add-on and the Clean Tweet filter were utilized to remove users who have been on Twitter for less than a day. Additionally, they removed tweets that contain more than three hashtags.

VI. DATA RETRIEVAL

Before retrieving the data, some questions should be addressed: What are the characteristics of the data? Is the data static, such as the profile user information “name, user Id, and bio”; or dynamic such as user’s tweets, and user’s network? Why is the data important? How is the data will be used? And how big the data
is? It is important to note that it is easier to track a certain keyword attached to a hashtag rather than a keyword not attached to it.
Twitter-API is a widely used application to retrieve, read and write twitter data. Other studies, as in [6], have used

GNU/GPL application like YourTwapperKeeper tool, which is a web-based application that stores social media data in MySQL tables. However, YourTwapperKeeper in storing and handling large size of data exhibits some limitations in using, as MySQL and spreadsheets databases can only store a limited size of data. Using a hybrid big data technology might address such limitations as we suggested above.

First, we undertake an evaluation of TWIROLE using a variety of classifiers. We also measure the performance of each individual model as well as the performance of our hybrid model. Concerning a multi-classifier that takes into account both the basic and advanced features (BF, AF), we conduct an experiment in which we compare traditional individual classifiers, such as the decision tree and the support vector machine (SVM), with ensemble classifiers, which include the random forest, AdaBoost, and GradientBoosting. For the CNN model, the default architecture that we utilize is Res-Net-18.

The accuracy of TWIROLE's modules when used with a variety of classifiers is outlined in Table 1. The CNN model on its own is successful, but using both models together is even more effective. Gradient-Boosting performs the best out of the five different classifiers for both sets of features (AccBF = 0.816, AccAF = 0.738), while random forest has the highest accuracy (Acc = 0.899) considering the total model. In addition, the performance of the ensemble classifiers is superior to that of the conventional individual classifiers in both the individual models and the hybrid model. Specifically, the accuracy of SVM with the enhanced features is just 0.352, which is merely a marginal improvement over random results. When we compare the performance of every single model to that of the hybrid model, we find that the hybrid model is always superior to every single model with different classifiers, with the exception of the decision tree (which is tied), which is where we find that there is no clear winner. As a consequence of this, the usage of random forests will be the default option in subsequent evaluation studies and a hybrid overall model will be favored.

Table 1: Various classification scores

Classifier Type	Accuracy			
	BF Multi-classifier	AF Multi-classifier	Profile Image CNN	Overall
Decision Tree	0.721	0.618	0.790	0.721
SVM	0.739	0.352		0.800
AdaBoost	0.790	0.704		0.850
GradientBoosting	**0.816**	**0.738**		0.842
Random Forest	0.796	0.708		**0.899**

For determining each role's recall (R) [7], precision (P) [8], and F1 score [9], we make use of a confusion matrix. The three values are as follows for a specific role r, the result present in

Table 2: Recall, Precision and F1 Score on Twi Role

Tool	Male			Female			Brand			Acc
	R	P	F1	R	P	F1	R	P	F1	
TWIROLE	0.885	0.922	0.903	**0.920**	**0.897**	**0.908**	**0.891**	**0.879**	**0.885**	**0.899**

VII. CONCLUSION

Utilizing Twitter as a source of information for data analysis has become possible thanks to the platform's vast quantity of data, its many distinct forms of data, and the public nature of tweets. First, by assessing the life cycle of a certain issue by counting the number of tweets over a period; second, by measuring the sentiment of users towards a specific topic through the application of NLP and ML algorithms. Our goal is to improve the analysis of Twitter data for specific events in order to assess the impact of those events on user behavior and to categorize those effects into different event types. An additional piece of work will concentrate on analyzing the data and the properties it possesses,

As well as researching the modeling techniques, that can be used to determine the frequency distribution for each event.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] Twitter. From https://en.wikipedia.org/wiki/Twitter

[2] Kumar, Shamanth, Fred Morstatter, and Huan Liu. Twitter data analytics. New York: Springer, 2014.

[3] Gaglio, Salvatore, Giuseppe Lo Re, and Marco Morana. "A framework for real-time Twitter data analysis." Computer Communications 73 (2016): 236-242.

[4] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521.7553 (2015): 436-444.

[5] Prajapati, V. (2013). Big Data Analytics with R and Hadoop. Packet Publishing.

[6] Bastos, M. T., Travitzki, R., & Raimundo, R. (2012). Tweeting political dissent: Retweets as pamphlets in #FreeIran, #FreeVenzuela, #Jan25, #SpanishRevolution and #OccupyWallSt. University of Oxford.

[7] Recall. From https://en.wikipedia.org/wiki/Precision_and_recall#Recall

[8] F-Score. From https://en.wikipedia.org/wiki/Precision_and_recall#F1_score

[9] Precision. From https://en.wikipedia.org/wiki/Precision_and_recall#Precision