Event Detection on Social Media

Hussien Ibrahem Hassan
Information System department
Ain-Shams University
Cairo,Egypt
hussien20191700217@cis.asu.edu.eg

Sama Hesham Sayed
Information System department
Ain-Shams University
Cairo,Egypt
sama20191700292@cis.asu.edu.eg

Riham Mohsen Sayed
Information System department
Ain-Shams University
Cairo,Egypt
reham20191700258@cis.asu.edu.eg

Ahmed Hamdy Abdelsattar
Information System department
Ain-Shams University
Cairo,Egypt
ahmed20191700027@cis.asu.edu.eg

Rana Hamdy Abdelkader
Information System department
Ain-Shams University
Cairo,Egypt
rana20191700249@cis.asu.edu.eg

Marwan Ahmed Mohamed Hafez
Information System department
Ain-Shams University
Cairo,Egypt
marwan20191700615@cis.asu.edu.eg

***Abstract*—Event detection on social media has become a crucial research area due to the vast amount of information generated on these platforms. This paper presents a review of the techniques and challenges that we have used to try and improve event detection on social media. We begin by discussing the characteristics of social media data and the unique challenges associated with event detection that have faced us during our project .We then provide an overview of the different approaches used that we used for event detection on social media , including keyword-based like NLP and machine learning-based technique as Naïve bayes , Random Forest , Logistic Regression, and SVM. By applying those algorithms we achieved an accuracy equals to 90% iin event detection on social media using Twitter data.**

***Keywords* –** *SVM, NLP, Tokenization, Stemming, Irrelevant Content.*

I. Introduction

Social media platforms such as Twitter , Facebook , and Instagram have become an integral part of our daily lives, providing us with a platform to share our thoughts, opinions, and experiences with a global audience. The vast amount of user content generated on these platforms has led to an explosion of data that can be used to gain insights into various events and phenomena occurring in the real world. Event detection on social media refers to the process of identifying and tracking events as they unfold using social media data.[2]

Event detection on social media has become an increasingly important research area due to its potential applications in various domains such as crisis management, public health, and marketing.[4] For example, during natural disasters, social media platforms are often the first source of information about the event, and event detection techniques can be used to identify important information such as the location, severity, and impact of the event. Despite the potential benefits of event detection on social media, there are several challenges associated with this task. The high volume, velocity, and variety of social media data make it difficult to identify relevant events and filter out noise and irrelevant information[9]. In addition, the informal and unstructured nature of social media data presents challenges for traditional natural language processing techniques.[11]

In this paper, we provide a comprehensive review of the techniques and challenges involved in event detection on social media. We begin by discussing the unique characteristics of social media data and the challenges associated with event detection on these platforms.[10] We then provide an overview of the different approaches used for event detection, including keyword-based and machine learning-based techniques. We also discuss the importance of feature selection and dimensionality reduction in improving the accuracy and efficiency of event detection on social media. [12]

II. Background

A. Definition of Event Detection on Social Media

Event detection on social media typically involves the use of natural language processing, machine learning, and social network analysis techniques to filter and analyze social media data. [6] These techniques help in identifying relevant events and track their evolution over time.[5] By leveraging social media data, event detection on social media can provide a more timely and comprehensive view of events than traditional news sources, as event detection on social media has the potential to revolutionize the way we monitor and respond to real-world events.[8] It provides a unique opportunity to understand the mind of the people and how they think , which can be valuable for decision making and crisis management.

B. Historical Background of Event Detection

The historical background of event detection can be traced back to the early 2000s with the rise of social media platforms such as Twitter and Facebook. With the growth of these platforms, there was a rapid increase in the amount of user generated content being produced, including text, images, and videos. [4]

Researchers quickly recognized the potential of social media data for detecting and tracking events occurring in the real world. Early work in the field focused on using keyword-based methods to identify relevant content related to specific events [13] . For example, researchers used keywords such as "earthquake" or "flood" to identify tweets related to natural disasters.

As social media platforms became more popular and the amount of user-generated content grew, researchers began to explore more sophisticated techniques for event detection. [7] This led to the development of machine learning-based approaches, which could automatically learn the patterns and characteristics of relevant content and improve the accuracy of event detection.

In recent years, there has been a growing interest in the use of social network analysis for event detection. This approach leverages the structure of social networks to identify the key actors and communities involved in specific events and track their evolution over time.

III. RELATED WORKS

Event detection on social media has gained significant attention in recent years due to the widespread use of social media platforms and the large amounts of data generated by them. The use of social media data for event detection and tracking can be particularly useful, where social media can provide valuable insights into attendee engagement, feedback, and event-related discussions. Several studies have explored the use of social media data for event detection and monitoring. These studies have proposed various techniques and algorithms for event detection, including keyword-based filtering, clustering, and machine learning. The use of these techniques can help to identify key topics and events related to a specific conference, and provide real-time monitoring and analysis of social media data.

For example A tutorial on event detection using social media data analysis (2022).[1]. This paper proposes an event detection system that analyzes social media data from twitter and to detect and track events in real-time. The system uses a combination of keyword-based filtering as Natural Language processing (NLP) and machine learning algorithms as Recurrent neural networks (RNNs) ,Long Short-Term Memory network (LSTM) and the Gated Recurrent units (GRU) .[13].They have been shown to be useful to detect and classify events and it studies the classification of various events including natural disasters , politics , outbreaks, and traffic events .

Road traffic event detection using twitter data (2019).[14]. This proposes a method for detecting road traffic events using Twitter data. It aims to explore the potential of social media data for detecting road traffic events in real-time, which can be useful for traffic management and emergency response .It first collected a large dataset of geotagged tweets related to road traffic events in the United Kingdom. They then used various techniques for data preprocessing, such as removing stop words and stemming, to clean the data and prepare it for analysis.[11]. It used a machine learning algorithm called Support Vector Machines (SVM) to classify tweets related to road traffic events. The SVM algorithm was trained on a labeled dataset of tweets, where each tweet was labeled as either related to a traffic event or not related to a traffic event. The SVM algorithm was then used to classify new, unlabeled tweets as either related to a traffic event or not related to a traffic event.

M.Avventui et al in Real Time Event Detection in Twitter (2014) [2].This paper proposes an event detection system that analyzes social media data from twitter and Facebook to detect and track events in real-time.[9] The system uses a combination of keyword-based filtering and machine learning algorithms to detect and classify events as Document Frequency Support Vector Machine (SVM).[8].

IV. METHODOLOGY

A. Data Collection

In data collection we used Twitter API to collect data through Twitter as a lot of people uses Twitter to publish the important events more than any other social media platform. Our data consists of 35,000 tweets, and 4 categories sports, finance, politics, and business and it includes events and non-events tweets.

B. Data Preprocessing

Once the data is collected, it is preprocessed to remove noise and irrelevant information. This involves stop words removal, lemmatization and stemming, tokenization, irrelevant content and removing special characters as shown in *fig.1*.

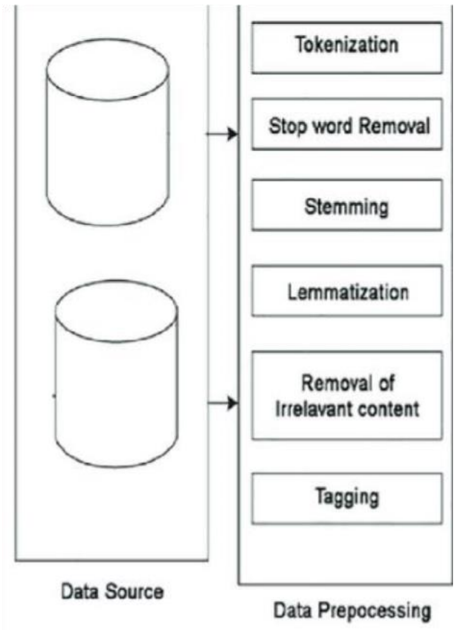

Fig.1.Data preprocessing steps.

C. Feature Extraction

After preprocessing, features are extracted from the data using Natural Language Processing (NLP) technique to represent the content of the posts. Features can include text based features such as keywords, hashtags, and named entities, as well as metadata features such as the user who posted the content or the location of the post and for example *Fig.2*. presents a word frequency analysis of tweets that pertain to events, highlighting the most commonly used words in the dataset.

Fig.2.Most frequent words.

After applying NLP to detect events it achieves an accuracy equals to 83% in detecting true events and it the percentage of true events in our data sets equals 10.7% and non-events equals 83.3% as shown in *fig.3*.

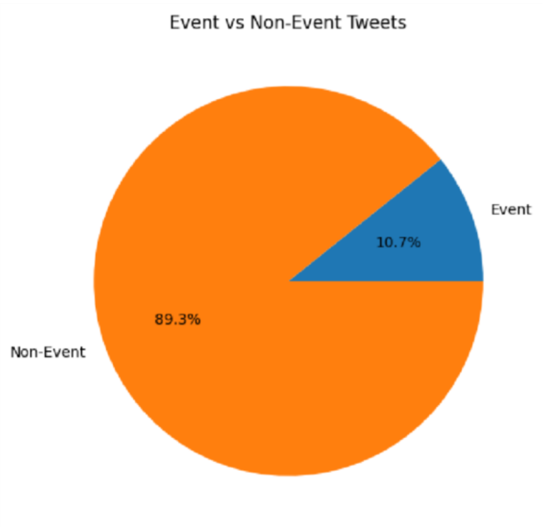

Fig.3 .Event and Non-Events in our data.

D. Machine Learning Algorithms

i- ***Naïve Bayes:*** is a simple but effective machine learning algorithm for classification tasks. It is based on Bayes' theorem and assumes that all features are independent of each other, which is why it is called "naïve". It is used to predict the probability of a particular class given a set of input features.

$$P(C \mid F1, F2, ..., Fn) = (P(C) * P(F1 \mid C) * P(F2 \mid C) * ... * P(Fn \mid C)) / P(F1, F2, ..., Fn) \quad (1)$$

Where:

- *P(C | F1, F2, ..., Fn)* is the probability of the class C given the input features F1, F2, ..., Fn.
- *P(C)* is the prior probability of class C.
- *P(Fi | C)* is the conditional probability of feature Fi given class C.
- *P(F1, F2, ..., Fn)* is the probability of the input features F1, F2, ..., Fn. And after applying it on our data it gives us accuracy equals 83.22%.

ii- ***Logistic Regression:*** is a popular machine learning algorithm for binary classification tasks. It models the probability of a binary response variable (y) as a function of one or more predictor variables (x) using a logistic function.

$$p(y=1|x) = 1 / (1 + exp(-(\beta 0 + \beta 1x1 + \beta 2x2 + ... + \beta n*xn))) \quad (2)$$

Where:

- p(y=1|x) is the probability of the response variable y being 1 given the predictor variables x.
- β0 is the intercept term.
- β1, β2, ..., βn are the coefficients associated with predictor variables x1, x2, ..., xn.
- exp() is the exponential function.

And after applying it on it gives us accuracy equals 85.55%

iii- ***Support Vector Machine (SVM):*** are a popular machine learning algorithm used for classification and regression tasks. SVMs are based on the concept of finding the hyperplane that best separates the data points into different classes.

Equation of SVM :F(x)=sign(w^Tx+b) *(1)*

where:

- x is the input feature vector representing a social media post
- w is the weight vector that defines the position and orientation of the hyperplane
- b is the bias term that controls the position of the hyperplane along the y-axis
- sign() is the sign function that returns -1 or 1 depending on whether the argument is negative or positive, respectively.

After applying it on our data it gives us accuracy equals 90%.

Fig. 4: shows the distribution of categories in our dataset, as each section represents the percentage of events that represents each category in our data. The figure provides a clear visualization of how the events are distributed and their relative proportions.

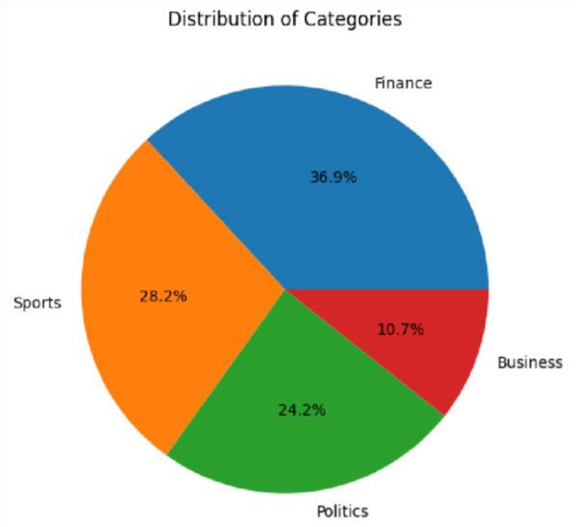

Fig.4. Distribution of events in SVM.

iv- ***Random Forest:***

Random forest is a machine learning algorithm that is used for both classification and regression tasks. It is an ensemble learning method that combines multiple decision trees to improve the accuracy and robustness of the model. In a Random Forest model, multiple decision trees are constructed using random subsets of the training data and random subsets of the features. The output of the model is then determined by aggregating the predictions of the individual trees. First construct the tree, then randomly select a subset of the features from the full set , then make predications

$$p\_k(x) = (1/T) * \Sigma\_t=1\hat{}T\ I(c\_t(x) = k) \quad (1)$$

, where T is the total number of decision trees in the forest and I is the indicator function.

$$y\_pred = argmax\_k\ p\_k(x). \quad (2)$$

Compute the average of the predicted values across all the decision trees:

$$y\_pred(x) = (1/T) * \Sigma\_t=1\hat{}T\ y\_t(x). \quad (3)$$

After applying it on our data it gives us accuracy equals 84%

TABEL I

ALGORITHMS AND THEIR ACCURACIES.

Algorithm	Accuracy
Naïve Bayes	83%
Logistic Regression	85%
Random Forest	84%
SVM	90%

V. SYSTEM ARCHITECTURE

As shown in *fig.5* that System Architecture we have 3 layers. First layer is Presentation Layer, Presentation Layer is the user interface and communication layer of the application, where the end user interacts with the application. Its main purpose is to display information to and collect information from the user so we put in it that we use flutter application as an interface for the user. Second layer is Application Layer, in this layer we apply data preprocessing to dataset we collected to remove noise, unwanted data, stop words, lemmatization and stemming, tokenization, and data cleaning. After the preprocessing the data is ready for feature extraction according to these features the model will consider if this tweet is an event or not. Then the model is build to detect the events in the data according to the features by training and fitting the model on the data. The model is evaluated to predict the data to ensure that it is predicting the right events. The last layer is database layer, Database layer is where we get the data using Twitter platform through API.

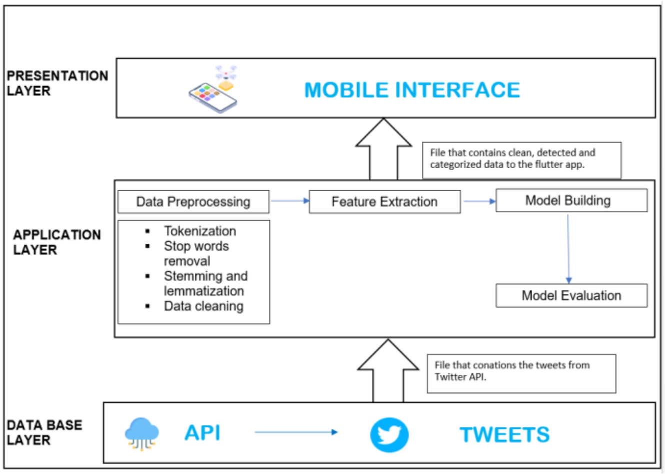

Fig.5.System Architecture.

VI. RESULTS

Model:

After collecting data from Twitter using API we applied on it Data preprocessing steps and after getting the clean data we applied NLP on it to detect feature that distinguish the events from non-events and achieved accuracy 83% in the true events we then applied some machine learning algorithms from which was SVM that gives us the highest accuracy from them all in text classification which is 90%.

System:

The development of the Event detection on social application provides the users efficient and satisfied experience in knowing the events on social media. The application includes:

A. *Choosing category:*

The user could choose the category of the events he wants to know.

B. *Search an event:*

The user could search a specific event through search box.

C. *Save an event:*

The user could save an event. *D.*

Share an event:

The user would be able to share the content of the event.

E. Manage his profile:

The user would be able to manage his profile data.

VII. CONCLUSION

We have presented research on event detection using social media data analysis in this tutorial. We have looked into the primary applications of social media data analysis for event detection. As event detection on social media is an important area that has the potential to provide information about events as they happen. The use of machine learning algorithms has made it possible to automate this process. These challenges include dealing with noisy and unstructured data, identifying relevant features and patterns, and handling the dynamic and evolving nature of social media content. It shows the effectiveness of machine learning algorithms for event detection on social media such that Support Vector Machine (SVM) is a very effective machine learning algorithm for event detection on social media. The fact that it achieved an accuracy of 90% suggests that it is capable of accurately identifying events based on patterns and trends in social media data.

VIII. REFERENCES

[1] Beccari, B. A Comparative Analysis of Disaster Risk, Vulnerability and Resilience Composite Indicators. PLoS Currents (2016).

[2] Boettcher, A. and Lee, D. Eventradar: A real-time local event detection scheme using twitter stream. In Green Computing and Communications (GreenCom), (2012) IEEE International Conference.

[3] Ateş, M.A., Purchasing and Supply Management at the Purchase Category Level: strategy, structure and performance, Promoter(s): Prof.dr. J.Y.F. Wynstra & Dr. E.M. van Raaij .

[4] Atefeh, F., Khreich, W.: A survey of techniques for event detection in twitter. Comput. Intell., pp 132–164 (2013).

[5] C. Aggarwal, and P. Yu ,A Framework for Clustering Massive Text and Categorical Data Streams, SDM (2006).

[6] T. Yang, R. Jin, Y. Chi, and S. Zhu ,Combining link and content for community detection: a discriminative approach, KDD Conf. (2009).

[7] H. Sayyadi, M. Hurst, and A. Maykov , Event Detection in Social Streams, AAAI Conf (2009).

[8] H. Schutze, and C. Silverstein ,Projections for Efficient Document Clustering, SIGIR Conf (1997).

[9] N. Panagiotou, I. Katakis, and D. Gunopulos ,"Detecting events in online social networks: Definitions, trends and challenges,", vol. 9580, Springer(2016).

[10] S. B. Kaleel, M. AlMeshary, and A. Abhari , "Event detection and trending in multiple social networking sites," (2013).

[11] W. Dou, X. Wang, W. Ribarsky, and M. Zhou , "Event detection in social media data," in IEEE (2012).

[12] A. Bifet and E. Frank ,"Sentiment knowledge discovery in twitter streaming data," (2010).

[13] Aldhaheri, Abdulrahman, and Jeongkyu Lee. "Event detection on large social media using temporal analysis." 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC). IEEE, (2017).

[14] Batrinca, Bogdan, and P. C.Treleaven. "Social media analytics: a survey of techniques, tools, and platforms." AI & Society pp 89-116, (30.1.2015).