International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 3 Mar. 2021, pp: 444-451 www.ijaem.net ISSN: 2395-5252

Spam mail Detection using Classification and Machine Learning Techniques

M Kirubha[1], S Gowthami[2], M Mohana[3], B Kathiravan[4]
[1]*Assistant Professor, Department of CSE, Sri Ramakrishna Institute of Technology,*
[2,3,4] *Student, Department of CSE, Sri Ramakrishna Institute of Technology,Coimbatore, Tamil Nadu, India*

Submitted: 01-03-2021 Revised: 09-03-2021 Accepted: 12-03-2021

ABSTRACT: Fake news detection is the huge task of detecting forms of news consisting of deliberate fake information or hoaxes spread through social media or traditional news media. With the widespread effects of the fast dissemination of faux news, efforts are made to automate the method of faux news detection. Due to the exponential growth of information online, it is impossible to identify the true from the false. In this project, we'll specialize in the "fact" and indirectly on the "truth" which is somehow associated with the emotion and private belief within the age of "post-truth". The three interactive functions of social media - tracking, sharing and creating - empower users to keep abreast of popular information, to repost any news (e.g., special reports, news videos, latest research, etc.), to express subjectively determined facts, and to even host the live broadcast. Through the tracking, sharing and creation interaction among a large user base, social media has been able to shape the major civic activities.

Key words: Fake news, SVM, Social media.

I. INTRODUCTION

Opinions are at the core of almost all human behaviours and are a central predictor of our actions. Our confidence in reality, our perceptions, and our actions are all shaped by how others view and value the world. As a consequence, when we need to make a decision, we always seek the guidance of others. Individuals aren't the only ones impacted. It also refers to companies. In the real world, corporations and organisations are still interested in hearing from customers and the general public about their goods and services. Individual buyers want to know what other people think about a product before purchasing it, or what other people think about political parties before voting in elections. individuals and organisations must also take into account the content of social media as a result of increased power in decision-making due to social media (for example, criticism, forum discussions, blogs, microblogs, Facebook comments, and posts on social networking sites) and as a result of increased power in decision-making due to social media. In recent years, sentiment analysis has become increasingly popular in industry. In recent years, sentiment analysis-related industrial activities have expanded rapidly. In this region, a slew of new initiatives have sprung up. Many major companies have built their own sentiment analysis programmes to determine the efficiency of on-site facilities, raising visibility in the market and social arenas.

Social networking is important for marketing and establishing consumer relationships. Small companies are starting to use social media as a marketing tool due to the low barrier to entry. Unfortunately, several small companies struggle with social media and lack a social media plan. As a result, without a basic understanding of the benefits of social media and how to use it to engage customers, innumerable opportunities are missed. The aim of the study is to gain a basic understanding of how a small business that is known for using social media to expand its business uses it to reach customers.

The four main themes found in current research studies are Virtual Brand Communities, Customer Attitudes and Motives, User Created Content, and Viral Ads. This paper starts with an overview of terms that describes social media marketing, followed by a review of the four main themes found in current research studies: Virtual Brand Communities, Consumer Attitudes and Motives, User Generated Content, and Viral Advertising. 2 Although social media marketing is a well-researched subject, it has only been studied through experimental, theoretical and scientific research. The advantages that retailers benefit from this marketing strategy are never completely outlined in studies. After reviewing a wide body of multidisciplinary literature, it's clear that studies are primarily concerned with identifying what social media marketing is and evaluating what factors influence consumer behaviour in relation to social

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 3 Mar. 2021, pp: 444-451 www.ijaem.net ISSN: 2395-5252

networking, Despite the initial progress made by researchers, development in this area of study has been limited.

II. LITERATURE SURVEY

Rozita Talaei Pashiri et. al., Spam emails are fraudulent emails that are distributed around the Internet for a number of purposes, including data theft, advertising, and malware distribution. Artificial neural networks can be used in classification problems as well as for efficient spam detection as a data mining and learning tool. The SCA was used to choose key features in this paper, which were then fed into the MLP neural network for training. The results of the tests revealed that the proposed approach has higher accuracy, precision, and sensitivity in spam detection than other learning methods like the MLP neural network, Bayesian network, decision tree, and random forest. In this respect, the MLP neural network came in second. It is suggested that the proposed algorithm's detection error be minimised by improving the SCA. Combining the proposed approach with other learning methods such as the support vector machine (SVM) and decision tree is also suggested[1].

Nikhil Kumar et.al., author describes a The Multinomial Nave Bayes algorithm provides the best results, but it suffers from class-conditional independence, causing the computer to misclassify some tuples. Ensemble methods, on the other hand, have been shown to be useful because they use several classifiers to predict class. Nowadays, a large number of emails are sent and received, making it impossible to evaluate emails with our project's small corpus. As a result, our spam detection project is capable of filtering emails based on the content of the email rather than two domain names or some other criterion. As a consequence, the email's body is currently small. [2].

Nikhil Govil et.al., author proposed a Spam emails are becoming more common every day, posing a challenge for users. Using a spam detector, we can decide whether a message is spam or not, improving user productivity. We're using the Nave Bayes classifier, which will send you a probabilistic index and decide whether or not the email is spam based on the display results. [3].

Aditya Shrivastava et.al., author describes a The Nave Bayes, Random Tree, REP Tree, Random Forest, and J48 classifiers are spam mail classification algorithms that use different algorithms to classify spam mails. For the testing of the spam base dataset, the Weka tool employs two classification procedures: cross validation and **training** collection. The same data will be used for training and testing in the training package. Training data is often segmented in a variety of folds for cross validation. Following implementation and research, it was determined that using the training set, the Random Tree classifier provides approximately 100 percent accuracy with a time requirement of just 0.01 second. As a result, Random Tree is the approach that provides the best results for spam classification.[4].

Tushaa Gangavarapu et al., author described In order to construct any intelligent device, feature engineering and machine learning are required. In general, the amount of current literature reviewed in this study corroborates the substantial progress made and anticipated in the field of spam and phishing email detection. We used forty insightful and discriminative content-based and body-based features in this analysis, which were chosen based on the underlying email corpus. Second, they clarified how the discriminative function space was extracted from the raw email corpus. [5].

Joseph Stephen et al., author suggested a potential of spam messages to evade filters has been investigated over time. The basic architecture of an email spam filter, as well as the processes involved in spam email filtering, were investigated. The research looked at some of the publicly accessible datasets and efficiency measures that can be used to determine the efficacy of spam filters. The difficulties of machine learning algorithms in effectively coping with the spam threat were highlighted, and comparative reviews of machine learning techniques available in the literature were performed. As a result, academics and industry practitioners studying machine learning techniques for efficient spam filtering will continue to be involved in the production of spam filters. This paper will serve as a springboard for qualitative research in spam filtering using machine learning, deep learning, and deep adversarial learning algorithms by the research students. [6].

G Pavan et.al., author proposed a detect spam mails coming from various sources, an integrated mail system is being developed. The ESPOT is a proposed framework that will be incorporated into an email system to measure a spam score based on the mail and deliver it to the inbox or spam inbox depending on the score. The spam score is assigned a threshold, which means that if the score is greater than 70%, the mail sent by the sender will go to the receiver's spam mailbox, and if the score is less than 20%, the mail will go to

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 3 Mar. 2021, pp: 444-451 www.ijaem.net ISSN: 2395-5252

the inbox [7].

Priti Sharnal et.al., author developed a spam email is one of the most demanding and problematic internet problems. Spammers misuse this contact method by sending spam emails, which has a negative effect on companies and many email users. This paper presents a Spam Mail Detection method that uses a hybrid bagged approach to implementation. Nave Bayes and J48 are the classification algorithms used in this process. The Nave Bayes and J48 algorithms achieve 83.5 percent and 91.5 percent accuracy, respectively. The hybrid bagged approach based SMD method achieved an overall accuracy of 87.5 percent, suggesting that the experimental results are better when performed on only J48 algorithm**.** [8].

V.K.Singh et.al., author suggested a Classification techniques must first be trained to separate spam emails from other emails before they are actually used. To train these techniques, a data set called a training set is used. Thousands of samples are used in this training set to enable the classifier to separate the spam mail. But even after a lot of work, spam mail still persists. They persist because a new type of spam mail is introduced every 5 days. This algorithm is expected to increase the efficiency of other techniques by some margin, depending on the technique. . If it is successful in doing so we will have the spam mail dealt with before it reaches our mailbox. This will also save our time and inbox will be less crowded thus making it easier to find useful emails [9].

D Karthika et.al.,author concluded a comprehensive analysis of the different classifiers using WEKA has been carried out on a common dataset. The results were compared on the basis of the evaluation criteria mentioned above. The study found that the same classifier performed in a different way when running on the same dataset but using different software tools. Some of these classifiers to different software tools for one would expect the classifiers to be consistent as the test was done on the same dataset. [10].

III. EXISTING SYSTEM

Long before the awareness of the Internet became widespread, some of our friends asked me to recommend a television, ask who they planned to vote for in local elections, ask colleagues to ask for reference letters about business owners for job applications or what dishwashers they wanted to buy. Today, however, the development of internet technologies has given us the opportunity to discover the views and experiences that we both have both personal and well known professional critics. Such studies show that more and more people are starting to present their opinions for foreigners on the internet. Ideas such as opinions, measurements, evaluations, attitudes, interpretations and concepts related to them are the areas of study of sentiment analysis and opinion mining. The rapid growth of workspaces has helped to increase the use of forums, discussion and dating pages, blogs, micro blogs and other social media tools among people The social sharing sites that emerged with the development of information technologies have had an important place in human life.

Among the most popular social networking sites on the planet are web sites and applications like Facebook, Facebook, Instagram, YouTube, Google Play, Vine, blog, micro blog, social networking and social bookmarking services. The "Social Media" system is exposed when all of these services are integrated.

In addition to real-world applications, research papers in the field of sentiment analysis have been published. For example, Leilei and his team conducted a sentiment analysis study using Facebook data to predict the election results.Ozel has conducted a survey using a software tool called "Limesurvey", a web based survey interface. In this study, the effect of using Facebook of employees on company profile was analyzed. The survey was tweeted and retweeted by 10 different Facebook accounts to reach two thousand Facebook users. The obtained data were analyzed by statistical analysis. Sentiment analysis studies using supervised learning approach from machine learning methods in social networks. In doing this study, the data set consisting of the interpretations of various products of some food companies on Facebook manually are obtained. Oguz and his fellow researchers the identification of influenza-like illnesses via social media using Facebook messages and newspaper websites. Facebook data were collected using free "Topsy" real-time search engine application developed for social media. Yazan and Uskudarli have made earthquake detection through social networks.

Disadvantages of the existing system

• Event detection and summary, opinion mining, sentiment analysis, and many others.

• Since tweets are limited in length (i.e., 140 characters) and writing styles are unregulated, posts often contain grammatical errors, misspellings, and informal abbreviations.

• On the other hand, despite the chaotic existence of posts, the core semantic knowledge in the form of named entities or semantic phrases is well preserved.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 3 Mar. 2021, pp: 444-451 www.ijaem.net ISSN: 2395-5252

IV. PROPOSED SYSTEM

In this study, By using Facebook data sentiment analysis study was done . Data collection is the initial step in this study. It is known that collecting the data and the data sets require the most time and power for the researchers who work on social media. There are many different types of data collection on social networks. When we look at literature, there are many tools and methods for collecting data. Among these, the most commonly used are custom designed APIs, web crawling, web scraping operations and scripts. In this study, it is aimed to obtain the data sets in a meaningful and regular manner based on Facebook data and to carry out sentimental analysis work.

In this study, tagged Facebook data set named Sentiment140 was used. This set was created by Stanford University's Computer Science graduate students Alec Go, Richa Bhayani and Lei Huang. This data set contains about 1.6 million positive and negative tagged Facebook data. For example, "smiley" is considered to have a positive tag because it is an emotion expressing happiness. It is likewise classified as negative because it is an emotion containing the phrase "□" sadness. In this study we used totally 10 thousand facebook data. Two different data sets were created.In facebook totally it contains 5 thousand data collect in both the sets. The number of Facebook data with positive and negative tags in each set was calculated.

Advantages

- Reduces noisy and irrelevant words that are not associated to the users and maintains their privacy.
- Data and information are maintained with better security and dilutes the negative words up to an extent.

Fig 1.1. Flow Diagram

V. MODULE DESCRIPTION

- Data Acquisition
- Pre processing
- Hybrid segmentation
- Named Entity Recognition
- Performance Evaluation

A. Data Acquisition

Facebook is an online social networking service that enables users to send and read messages,images as well as videos posts, Registered users can read and post, but unregistered users can only read them it is also only if the concerned data owner provides permission, then only it is possible It is important to review the user's posts in order to form an opinion about him. Therefore, users post are popped first using the facebook API However, crawling all friends' posts is a huge overload, and misleading since Facebook following mechanism does not show an actual interest every time.

People sometimes tend to follow some users for a temporary occasion and then forget to un-follow. Sometimes they follow some users just to be informed of, although they are not actually interested in. In this module, we can upload the datasets in the form of CSV file. It contains following id, followers id, time stamp, user following, user followers and posts. The data of all Facebook users has been reviewed, and their entities have been evaluated for a better method. The data that has been acquired are the posts that has been done by the users, messages that has been sent and received and so on it continuous.

B. Preprocessing

In this research, we tried to examine the user not only through his own posts, but also through the posts of his fellow students.. Before real data has entered our lives, studies on the area were being conducted on formal texts such as news articles. Basically named entities are considered as words written in uppercase or mixed case phrases where uppercased letters are at the beginning and ending, and almost all of the studies bases on this assumption. In posts like informal messages, however, capitalization isn't always a reliable predictor, and it can even be deceptive.The methods must be updated, as the example of capitalization illustrates.

Following tasks are applied on the data for obtaining minimalism.

• Links and mentions are removed since they cannot be a part of a named entity.

• Conjunctives, stop words, vocatives, and slang words, among other things, are eliminated.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 3 Mar. 2021, pp: 444-451 www.ijaem.net ISSN: 2395-5252

• Although punctuation is not taken as an indicator since posts are informal, still elimination of punctuation is needed. So, smiley's are also removed.
• Repeating characters to express feelings are removed.
• Informal writing style related issues such as mistyping are corrected.
• Asciification related problems are solved since users connecting from mobile devices tend to ignore Turkish characters.

Preprocessing tasks can be divided into two logical classes, as can be seen. Pre-segmenting and Fixing. Removal of links, references, conjunctives, stop words, vocatives, slang words and elimination of punctuation shall be considered as pre-segmentation. It is accepted that parts in the texts before and after a redundant word, or a punctuation mark cannot form a named entity together, therefore every removal of words is behaved as it segments the post as well as punctuation does it naturally. Since posts are pre-segmented before they are handled in post data segmentation process, pre-segmentation tasks reduces the complexity of the text. Whereas the messages that has been sent and received, the requests for the users are also preprocessed. At the same time once the concerned user can block the unwanted users those who send the messages and posts with negative words.

C .Hybrid segmentation

Hybrid Segmentation is capable of learning from both global and local contexts, as well as pseudo feedback. As pseudo feedback, HybridSeg is also designed to iteratively learn from confident segments. Posts are created for the purpose of exchanging knowledge and communication. In posts, called entities and semantic phrases have been well preserved. As a consequence, the global context extracted from Web pages aids in the detection of meaningful segments in posts. The well-preserved linguistic features in these posts allow for high-accuracy named entity recognition. Each called entity is a section that can be used. HybridSegNER is the term for a method that uses local linguistic features. Based on the voting results of several off-the-shelf NER instruments, it generates confident segments. Another approach based on local collocation information, HybridSegNGram, is proposed in response to the observation that several posts published over a short period of time are about the same topic.

By estimating the term-dependency within a batch of posts, HybridSegNGram segments the posts. The segments that are recognised with high confidence based on the local context serve as strong input for extracting more relevant segments. Iterative learning is used to learn from pseudo feedback, and the approach used to execute it is called HybridSegIter.

D Named Entity Recognition

Named Entity Recognition is the process of defining and categorising particular types of data (such as persons, locations, and organisations, as well as date-time and numeric expressions) in a document. Posts, on the other hand, are notorious for being short and noisy. Given the length of a post and the independence with which it is written, named entity recognition on this form of data is difficult. A significant number of named entities in the document, such as personal names, place names, and organisation names, are not properly segmented and recognised after simple segmentation. Speech tagging can be used for a variety of NLP activities, such as named entity segmentation and knowledge extraction. There are three major factors that affect called entity recognition strategies: linguistics, textual genre and domain, and entity type. Language is critical because it has an effect on how people solve issues. Assign the most common POS tag to each word and the most common OOV tag to each Out of Vocabulary (OOV) word. Another term whose ramifications should not be ignored is textual genre.

E Performance Evaluation

We can use accuracy rate and normalised usefulness to test the system's process in this module. The accuracy rate and normalised utility of our proposed system are both increased. When the recipient receives the messages from the sender, it analyses the data for negative and positive words and alerts the user if there are any negative words present. If the process continues, more than 3 or 4 times it will make a suggestion and blocks the users who are associated with the negative contents. Additionally, the data owner's posts can be posted as both public and private. The posts in public mode are viewable by all users in the data owner's profile, while in private mode, only the owner's approved users have access to the data owner's posts.

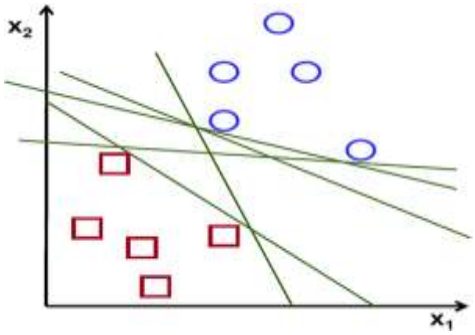
VI. ALGORITHMS, TRAINING AND TESTING

SUPPORT VECTOR MACHINE

I assume you've become familiar with linear regression and logistic regression algorithms by now. If you haven't already, I recommend that you do so before moving on to the support vector

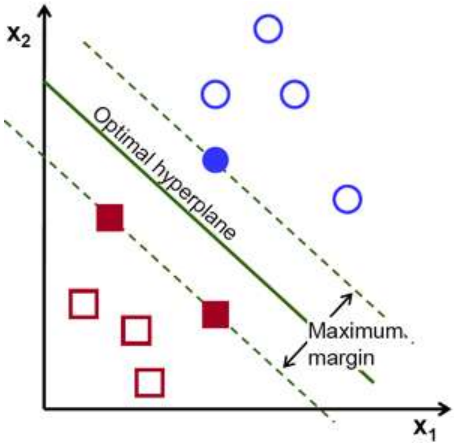
International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 3 Mar. 2021, pp: 444-451 www.ijaem.net ISSN: 2395-5252

machine. Another basic algorithm that any machine learning expert should know about is the support vector machine. Many people prefer the support vector machine because it achieves significant accuracy.

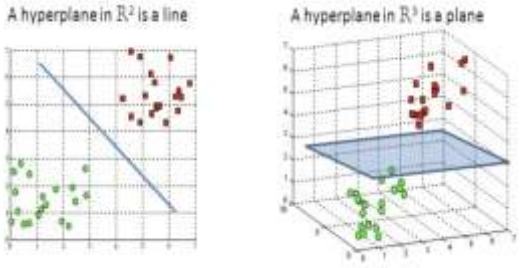


Possible Hyperplanes

There are several hyperplanes from which to choose to distinguish the two types of data points. Our aim is to find a plane with the greatest margin, or the greatest distance between data points from both groups. Maximizing the margin gap provides some reinforcement, making it easier to distinguish potential data points.



Hyperplanes and Support Vectors

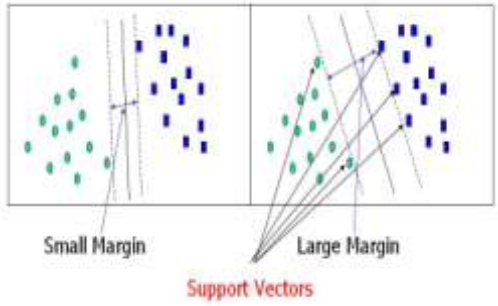


Hyperplanes in 2D and 3D feature space

Hyperplanes are decision boundaries that assist in data classification. Different groups may be allocated to data points on either side of the hyperplane. The hyperplane's dimension is also defined by the number of functions. If there are only two input features, the hyperplane is just a line.

SUPPORT VECTORS

Support vectors are data points that are closer to the hyperplane and have an impact on the hyperplane's direction and orientation. We optimise the classifier's margin by using these support vectors. The hyperplane's location would be altered if the support vectors are deleted. These are the points that will assist us in constructing our SVM.



LINE MARGIN INTUTION

The sigmoid function is used in logistic regression to squash the output of the linear function within the range of [0,1]. If the squashed value is greater than a threshold value (0.5), we label it 1, otherwise we label it 0. In SVM, we take the linear function's output and, if it's greater than 1, we assign it to one of two classes; if it's -1, we assign it to the other.

Cost Funqxction And Gradient Updates

The goal of the SVM algorithm is to maximise the distance between the data points and the hyperplane. Hinge loss is a loss feature that aids in margin maximisation.

$$c(x, y, f(x)) = (1 - y * f(x))_+$$

$$c(x, y, f(x)) = \begin{cases} 0, & \text{if } y * f(x) \geq 1 \\ 1 - y * f(x), & \text{else} \end{cases}$$

Loss of Hinge Feature (function on left can be represented as a function on the right) .

If the expected and actual values have the same value, the cost is zero. If they aren't, the loss value is determined. A regularisation parameter is also applied to the cost function. The regularisation parameter's target is to strike a balance between margin maximisation and loss. The cost functions look like this after adding the regularisation parameter.

$$min_w \lambda \parallel w \parallel^2 + \sum_{i=1}^{n} (1 - y_i \langle x_i, w \rangle)_+$$

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 3 Mar. 2021, pp: 444-451 www.ijaem.net ISSN: 2395-5252

Loss function for SVM

To find the gradients, we take partial derivatives with respect to the weights now that we have the loss function..

$$\frac{\delta}{\delta w_k}\lambda \parallel w \parallel^2 = 2\lambda w_k$$

$$\frac{\delta}{\delta w_k}\left(1 - y_i\langle x_i, w\rangle\right)_+ = \begin{cases} 0, & \text{if } y_i\langle x_i, w\rangle \geq 1 \\ -y_i x_{ik}, & \text{else} \end{cases}$$

Gradients

We just need to change the gradient from the regularisation parameter when there is no misclassification, i.e. when our model correctly predicts the class of our data point.

$$w = w - \alpha \cdot (2\lambda w)$$

Gradient Update — No misclassification

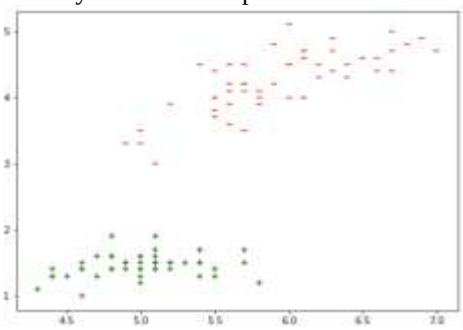
We add the loss along with the regularisation parameter to perform gradient update when there is a misclassification, i.e. our model predicts the wrong class for our data point incorrectly.

$$w = w + \alpha \cdot (y_i \cdot x_i - 2\lambda w)$$

SVM Implementation in Python

The Iris dataset will be used in the implementation of our SVM algorithm. You can download it from this link.

The Iris dataset will be used to evaluate and apply our SVM algorithm. This connection will take you to a download page where you can get it.We'll delete one of the classes from the Iris dataset because it has three. As a consequence, we're left with a binary classification problem.



Visualizing data points

There are also four features that we can use. Just two features will be used: Sepal length and Petal length. We visualise these two features by plotting them. You can see from the graph above that a linear line can be used to distinguish the data points.

We extract the required features and divide the data into training and testing classes. We have used 90% of the data for training and the rest 10% is used for testing. Now we'll use the numpy library to build our SVM model.

The learning rate is (0.0001), and the regularisation parameter is set to 1/epochs. As a consequence, the regularising value decreases the number of epochs.

Since the test data only has 10 data points, we must now clip the weights. We predict the values by extracting features from the test data. We get the forecasts, compare them to the real values, and print the model's accuracy.

ACCURACY 1.0

Another straightforward method for implementing the SVM algorithm exists. To implement the SVM model, we can use the Scikit learn library and simply call the relevant functions. The number of lines of code is greatly decreased, resulting in far too few lines.

The support vector machine is a simple but efficient algorithm.

VII. CONCLUSION

The widespread use of computers and the internet has resulted in a major increase in the methods for collecting information from social media. The literature's knowledge access and interpretation steps were explored in depth in this report. Facebook data was used to perform the sentiment analysis. Obtaining Facebook data, clearing Facebook data, and translating Facebook data to numerical form. The following steps are followed: collecting Facebook data, clearing data, converting data into numerical form, extracting relevant results, and analysing them. KNIME software is used to implement machine learning algorithms.

We designed novel features for use in the classification of posts in order to develop a system through which informational data may be filtered from the conversations, which aren't very useful when looking for immediate information for relief efforts or bystanders to use in order to minimise harm.

When it comes to computational resources, the results of our tests show that classifying posts as "rumour" vs. "non-rumor" can be achieved exclusively with the proposed features, since the computing power needed to translate data into features is significantly reduced as compared to a BOW feature set with a much larger number of features. If the processing power and time needed to process incoming Facebook data are not a problem, a feature set of the proposed features combined with the BOW-presence method would improve overall accuracy.

VIII. FUTURE WORKS

In future work, we can extend our approach implement various classification algorithm to predict the attackers and also eliminate the attackers from facebook datasets. And try this approach to implement in various languages on facebook. At the same it can be extended to analyze not only texts but also images, videos and so on. So that the exact scenario of entire users and their entities are managed with proper efficiency and avoids inappropriate medias.

REFERENCES

[1]. Rozita Talaei Pashiri1 ,Mohsen Mahrami1," Spams mail and detection through feature selection using artificial neural network and sine–cosine algorithm" Department of Engineering, Malard Branch, Islamic Azad University, Tehran, Iran Published on April: 2020 in springer https://do.org/10.1007/s40096-020-00327-8.

[2]. Nikhil Kumar, Sanket Sonowal, Nishant ," Email Spam Detection Using Machine Learning Algorithms" Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA-2020) IEEE Xplore Part Number: CFP20N67-ART; ISBN: 978-1-7281-5374-2 .

[3]. Nikhil Govil,Kunal Agarwal, Ashi Bansal , Astha Varshney," A Machine Learning based Spam Detection Mechanism", Proceedings of the Fourth International Conference on Computing Methodologies and Communication (ICCMC 2020) IEEE Xplore Part Number:CFP20K25-ART; ISBN:978-1-7281-4889-2.

[4]. Aditya Shrivastava, Dr. Rachana Dubey," Classification of Spam Mail using different machine learning algorithms", Authorized licensed use limited to: Auckland University of Technology. Downloaded on June 02, 2020 at 07:44:53 UTC from IEEE Xplore. Restrictions apply.

[5]. Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA-2020) IEEE Xplore Part Number: CFP20N67-ART; ISBN: 978-1-7281-5374-2.

[6]. Tushaar Gangavarapu, C. D. Jaidhar1, Bhabesh Chanduka," Applicability of machine learning in spam and phishing email fltering: review and approaches",Artifcial Intelligence Review , Springer Nature B.V. 2020Https://doi.org/10.1007/s10462-020-09814.

[7]. G Pavan, K Lakshmaji, and Dr S Krishna Rao," An ensemble integrated mailing system for detecting spam mails "International conference on computer vision and machine learning, IOP Conf. Series: Journal of Physics: Conf. Series 1228 (2019) 012042 IOP Publishing doi:10.1088/1742-6596/1228/1/012042.

[8]. Emmanuel Gbenga Dada, Joseph Stephen Bassi , Haruna Chiroma , Shafi'i Muhammad Abdulhamid , "Machine learning for email spam filtering: review, approaches and open research problems".https://doi.org/10.1016/j.heliyon.2019.e01802 Received 3 September 2018; Received in revised form 25 February 2019; Accepted 20 May 2019.

[9]. Kabir, Abida Sanjana Shemonti, Atif Hasan Rahman. "Notice of Violation of IEEE Publication Principles: Species Identification Using Partial DNA Sequence: A Machine Learning Approach", 2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE), 2018.A.

[10]. Priti Sharma1,Uma Bhardwaj1, "Machine Learning based Spam E-Mail Detection" International Journal of Intelligent Engineering and Systems, Vol.11, No.3, 2018 DOI: 10.22266/ijies2018.0630.01.

ACKNOWLEDGEMENTS: The authors are deeply grateful to SRIT Coimbatore for providing the necessary facilities for the preparation of the paper.

International Journal of Advances in Engineering and Management
ISSN: 2395-5252

IJAEM

Volume: 03 Issue: 03 DOI: 10.35629/5252

www.ijaem.net
Email id: ijaem.paper@gmail.com