Finding Inconsistency of Security Information from Unstructured Text

Punitha G, Sanggami R, Saranya R, K.Sangeetha

[1-3]*Student, Dept of Computer Science Engineering, PEC College, Tamil Nadu, India*
[4]*Associate Professor, Dept of Computer Science Engineering, PEC College, Tamil Nadu, India*

Abstract - Open source intelligence textual data has become a big topic in a variety of fields, including industry, law enforcement, military, and cyber security. Millions of people use social networking sites all over the world. The user's experiences with social media platforms like Twitter and Face book have a huge effect on their everyday lives, with some detrimental effects. The identification of scams and phony Social media users has become a common research area in the field of modern online social networks (OSNs). We show that Supervised Machine learning can overcome all the bottlenecks. We here find the conflict, disagreement or mismatch of the secured information from the unstructured data. Text analyses and classification is made to detect the Cyber bullying activities, Fake News Identification and Spamming.

Key Words*: *Machine Learning, Cyber bullying, Fake News, Spamming, Online Social Networks.

1. INTRODUCTION

One of the most effective and efficient ways to communicate today is through the Internet and social media. It has made it extremely simple for anyone to publish content on their website, blog, or social media profile and potentially reach large audiences. The internet allows us to connect with a wide range of people and read news and information from all over the world, whether through Face book, Twitter, Yahoo, or any other website. Because there is such a high demand for resources, online networking sites have evolved into a source of real-time data about any user. As a result, it provides commentary on a previously unseen method of information discrimination. Fraudsters can easily deceive many people who have little knowledge of online networking. Although the Internet has many positive aspects and can be extremely beneficial, it can also be a source of danger. With the growth of Online networking sites, there needs to be analysis and study of user behavior. As the data available and shared on the internet can be in the form of text, video, audio, and so on. We here focus only on the text. One of the most crucial aspects of text analysis is text classification. Since text can be a very rich source of knowledge, but due to its unstructured nature, extracting insights from it can be difficult and time-consuming. So, we assign tags or categories to text according to its content, helping structured and analyze the text, quickly and costeffectively, to automate processes and enhance data-driven decisions. Here text analysis is done to detect Cyber bullying activities, Fake news (False Information), and spamming. Many people get their news from social media sites and networks, and determining whether or how reports are accurate can be challenging. Information overload and a general lack of understanding about how the internet works by people have also contributed to an increase in fake news, Bullying, and spamming activities. Sending, publishing, or sharing negative, damaging, misleading, or mean material about someone else is considered cyber bullying. It can also include sharing personal or private information about someone causing embarrassment or humiliation. This crosses the line into unlawful or criminal behavior. False information is news, reports, or hoaxes that are intended to mislead or misinform readers. These stories are usually created to either influence or confuse people, and they can be a lucrative business for online publishers. Spamming is the practice of sending unsolicited or misleading messages across the Internet. It is often used for commercial advertising. All these issues can be resolved by text classification and analysis are made and detection is performed. In this, real-time unstructured data later modified to a structured format is collected from Kaggle. Then, preprocessing of the text is made by bringing it into the corpus (collection of data), and then vectorize the corpus by feature extraction. With the help of a supervised machine learning model, the classification and analysis are made. Then the accurate prediction for each model is determined. By doing this we can predict which model will give a higher accuracy rate. The model with high accuracy can help to detect fake news, spamming, and cyber bullying activities effectively. Maybe this will save a life or reduce the number of teens touched by cyber-bullying activities, fake news, Spamming, or any other internet dangers.

2. RELATED WORKS

The following sub-sections discuss some of the research that has been done by researchers in the areas of cyber bullying, fake news, and spam detection.

2.1 Gender and Age Detection in Cyber bullying Texts Using Computational Stylometry and Machine Learning

The purpose of this paper is to demonstrate the value of Computational Stylometry (CS) and Machine Learning (ML) in detecting author gender and age in cyber bullying texts. We created a cyber bullying detection platform and show the results of gender and age detection in cyber bullying texts we collected in terms of Precision, Recall, and FMeasure.

The benefits are as follows:

1. We can categorize and analyze cyber bullying text by grouping it into taxonomy (linguistic rules) to detect cyber bullies' gender and age.

The disadvantages include:

1. Focusing on categories based solely on written text
2. Developing a smaller number of stylistic features.

2.2 An integrated approach for malicious tweets detection using NLP

Detection of malicious user accounts has been the focus of many previous studies. Detecting spam or spammers on Twitter is a relatively new area of social network research. However, we present a method based on two new aspects: the detection of spam-tweets without knowing the user's previous background, and the other based on language analysis for detecting spam on Twitter in topics that are currently trending. Topics of discussion that are popular at the time are known as trending topics. Spammers benefit from the growing micro blogging phenomenon. Our research uses language tools to detect spam tweets. We began by gathering tweets related to a variety of popular topics and categorizing them as malicious or non-malicious. We extracted a number of features based on language models using language as a tool after a labeling process. We also assess performance and categorize tweets as spam or not. As a result, our system can be used to detect spam on Twitter, with a focus on tweet analysis rather than user accounts.

The merits are as follows:

1. We detect spam accounts using the SVM algorithm, which yields a standard result.

The Demerits, on the other hand, are

1. It can only be used to detect spam on Twitter, with a focus on tweet detection.

2.3 Unsupervised cyber bullying detection in social networks

Modern young people (also known as "digital natives") have grown up in a world dominated by new technologies, in which communication is pushed to near-real-time levels and there are no boundaries to forming relationships with other people or communities. However, due to the rapid pace of evolution, young people are unable to distinguish between consciously acceptable and potentially harmful behaviors, and a new phenomenon known as cyber bullying is gaining traction, attracting the attention of educators and the media. “Willful and repeated harm inflicted through the use of electronic devices” is what cyber bullying is defined as. Using techniques derived from NLP (Natural Language Processing) and machine learning, we propose a possible solution for automatic detection of bully traces across a social network in this paper. We will create a model based on Growing Hierarchical SOMs that can efficiently cluster documents containing bully traces and is based on semantic and syntactic features of textual sentences. We fine-tuned our model to work with Twitter, but we also put it to the test with other social media platforms like YouTube and Form spring. Finally, we present our findings, demonstrating that the proposed unsupervised approach can be used effectively in some scenarios with good results.

The merits are as follows:

1. Use of an unsupervised approach to analyze cyber bullying with several handcrafted features that were used to catch the cyber bullies' semantic and syntactic communication behavior.

The disadvantages are as follows:

1. Less accuracy and performance results.

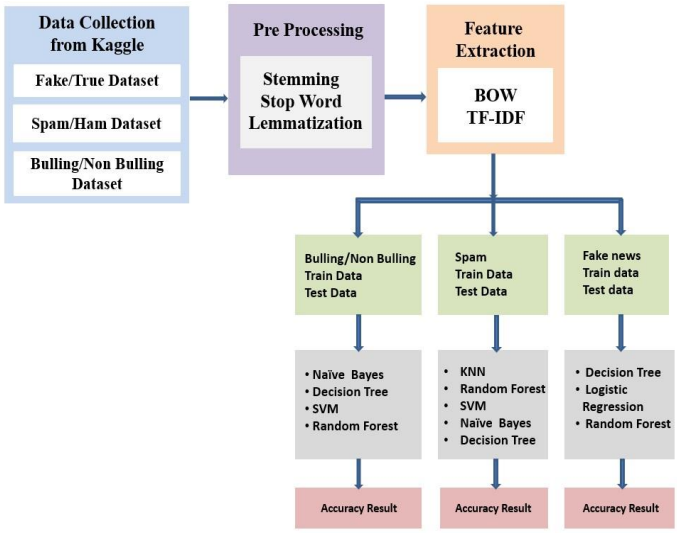
2. PROPOSED SYSTEM

Fake news on the Internet and cyber bullying on social media are major concerns for everyone in society, including the government, policymakers, organizations, businesses, and citizens. Fake news spreading on social media has become a significant problem in this world, with the potential to lead to mob violence. Bullying can have a negative impact on a person's physical, mental, and emotional well-being, leading to depression. As a result, a recipient of spam emails risk having their computers infected with a malicious program. Spam will clog mail servers, slowing down all email and putting a strain on the Server. Because spam email, fake news detection, and cyber bullying activities all fall under the category of text analysis, we've combined all of these domains and developed a system that can detect such texts. Using a supervised machine learning algorithm, we perform text classification on the dataset to identify fake/true news, spam/ham in the mail, bullying and non-bullying activity in our proposed system. We clean and prepare the real-time dataset before applying it directly to the algorithm, using pre-processing techniques such as stemming, lemmatization, and feature extraction to derive values (features in vector format) that are intended to be informative and nonredundant. When features are applied to various supervised

Machine learning models, the accuracy of each model is predicted, and the model with the highest accuracy is most effective in detecting the problem.

4. SYSTEM ARCHITECTURE

When features are applied to various supervised machine learning models, the accuracy of each model is predicted, and the model with the highest accuracy is most effective in detecting the problem.



Real-time Unstructured data is made into structured format are collected from Kaggle. We are collecting datasets specifically for Bullying, Fake News, and Spamming. Then the datasets collected will be subjected to preprocessing like Stemming, Stop Word and tokenization if needed lemmatization. To analyze preprocessed data, it must first be converted into features, or the text must be converted into numerical vectors. A technique for extracting features from text is by BOW and TF - IDF. A bag of words is a method to represent text into numerical features. Bag of Features can be easily created by using the count Vectorizier function. This process is called feature extraction (or Vectorization). It is used to convert a collection of text documents to a vector of token counts. The TF-IDF is a statistical measure that assesses the relevance of a word to a document in a set of documents. Then we will be building predictive models for Bullying, Fake News, and Spamming. For bullying, we will be using Decision Tree, Linear Regression, and Random Forest. For Spamming, we will be using Support Vector Machine, Naïve Bayes, Decision Tree, Random Forest, and K - Nearest Neighbor. Then for fake news, we will be using Random Forest, Linear Regression, and Decision Tree. At last, based on the outcomes we will predict the probability of occurrence of an event by evaluating Metrics – Accuracy. We conclude that the model with good accuracy is chosen as the best one.

DATASETS

Kaggle is also used to collect real-time unstructured data that has been converted to a structured format. For each (i.e., bullying and non-bullying, fake news, and spamming), unique datasets are collected. The most important task of any machine learning project is data collection. Because the data we feed the algorithms is the input. As a result, the algorithms' efficiency and accuracy are determined by the accuracy and quality of the data collected.

PREPROCESSING

Data Preprocessing is the process of converting raw data into a usable and efficient format by removing noise and inconsistencies such as punctuation, date, title, Stop Words, Stemming, Lowercase conversion, and, if necessary, lemmatization. As a result, it generates consistent and reliable data, which improves the efficiency of the training data for analysis and allows for accurate decision-making.

	title	text	subject	date	target
0	Croatia wants to adopt euro within 7-8 years: ...	ZAGREB (Reuters) - Croatia aims to become a eu...	worldnews	October 30, 2017	true
1	House narrowly passes measure paving way for T...	WASHINGTON (Reuters) - The U.S. House of Repre...	politicsNews	October 26, 2017	true
2	Oklahoma rejects Russian request to monitor el...	(Reuters) - Oklahoma voting officials have den...	politicsNews	October 21, 2016	true

1. Removing Date (attribute not used for analysis)

	title	text	subject	target
0	Croatia wants to adopt euro within 7-8 years: ...	ZAGREB (Reuters) - Croatia aims to become a eu...	worldnews	true
1	House narrowly passes measure paving way for T...	WASHINGTON (Reuters) - The U.S. House of Repre...	politicsNews	true
2	Oklahoma rejects Russian request to monitor el...	(Reuters) - Oklahoma voting officials have den...	politicsNews	true

2. Removing the title (we will only use the text)

	text	subject	target
0	ZAGREB (Reuters) - Croatia aims to become a eu...	worldnews	true
1	WASHINGTON (Reuters) - The U.S. House of Repre...	politicsNews	true
2	(Reuters) - Oklahoma voting officials have den...	politicsNews	true

3. Convert to Lowercase

	text	subject	target
0	zagreb (reuters) - croatia aims to become a eu...	worldnews	true
1	washington (reuters) - the u.s. house of repre...	politicsNews	true
2	(reuters) - oklahoma voting officials have den...	politicsNews	true

4. Remove Punctuation

	text	subject	target
0	zagreb reuters croatia aims to become a euro ...	worldnews	true
1	washington reuters the us house of representa...	politicsNews	true
2	reuters oklahoma voting officials have denied...	politicsNews	true

5. Remove Stop Words

	text	subject	target
0	zagreb reuters croatia aims become euro zone m...	worldnews	true
1	washington reuters us house representatives he...	politicsNews	true
2	reuters oklahoma voting officials denied reque...	politicsNews	true
3	world reeling today britain made shocking deci...	News	fake

6. Stemming

Stemming is the process of removing or stemming the last few characters of a word, which frequently results in incorrect meanings and spelling.

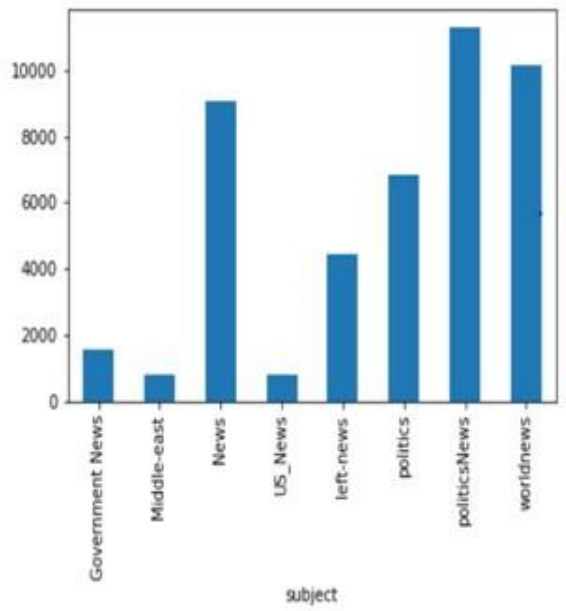
e.g., adjustable - > adjust
formality - > formal

7. Lemmatisation
Lemmatization takes the context into account when converting a word to its meaning base form.

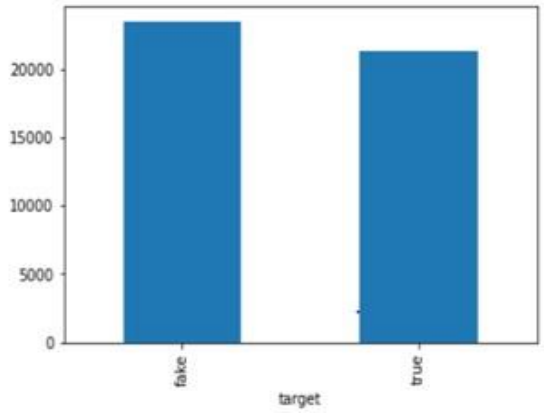
e.g., playing, plays, played - > play
am, are, is -> be

DATA EXPLORATION

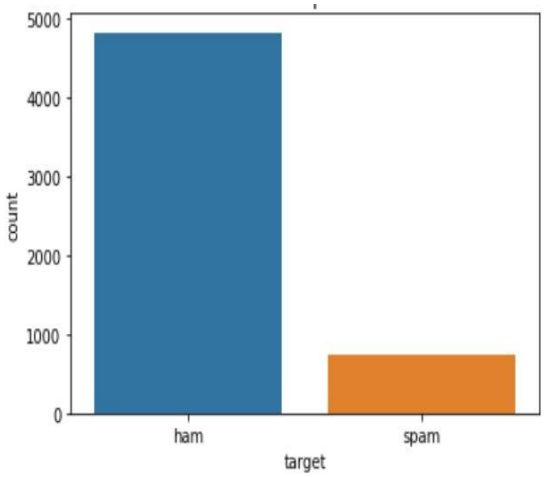
We use visual exploration to figure out what's in a dataset and what the data's attributes are Size, amount of data, completeness of data, correctness of data, and possible relationships are the characteristics here.



Articles per subject



No of Fake and Real Articles



Distribution of Spam and Ham

FEATURE EXTRACTION

Before fitting the train data and test data into a model. We perform feature extraction to vectorize the data.

1. BAG OF WORDS

It's a way of representing text data when using machine learning algorithms to model text. It describes the order in which words appear in a document. The model only cares about whether or not known words appear in the document, not where they appear. It entails two steps:

i) A lexicon of well-known terms.
ii) A metric for determining the presence of well-known words.

Step 1: Collect Data

It was the best of times.
It was the worst of times.
It was the age of wisdom.

Step 2: Design the Vocabulary

We can now create a list of every word in our model vocabulary. The following are the unique words (ignoring case and punctuation):

"it", "was", "the", "best", "of", "times", "worst", "age", "wisdom"

From a corpus of 18 words, that's a vocabulary of 9 words.

Step 3: Create Document Vectors

Create a binary vector from the first document ("It was the best of times").The simplest scoring method is to assign a Boolean value to the presence of words, with 0 indicating absence and 1 indicating presence. The document's scoring would be as follows:

"it" = 1 "was" = 1 "the" = 1 "best" = 1 "of" = 1 "times" = 1 "worst" = 0 "age" = 0"wisdom" = 0

"it was the best of times" = [1, 1, 1, 1, 1, 1, 0, 0, 0, 0]
"it was the worst of times" = [1, 1, 1, 0, 1, 1, 1, 0, 0, 0]
"it was the age of wisdom" = [1, 1, 1, 0, 1, 0, 0, 1, 1, 0]
"it was the age of foolishness" = [1, 1, 1, 0, 1, 0, 0, 1, 0, 1]

2. TF – IDF

The TF-IDF is a statistical measure that assesses the relevance of a word to a document in a set of documents. It incorporates two ideas: i)Term Frequency
ii)Inverse Document Frequency

Text 1	i love natural language processing but i hate python
Text 2	i like image processing
Text 3	i like signal processing and image processing

Step 1: Create a term frequency matrix with documents as rows and distinct terms as columns across all documents. Count the number of times each word appears in each text.

Term frequency = No of Repeatation of words in sentence / No of Words in a sentence

	and	but	hate	i	image	language	like	love	natural	processing	python	signal
Text 1	0	1	1	2	0	1	0	1	1	1	1	0
Text 2	0	0	0	1	1	0	1	0	0	1	0	0
Text 3	1	0	0	1	1	0	1	0	0	2	0	1

Step 2: Compute inverse document frequency (IDF) using the explained formula

IDF = log (No of sentences / No of sentences containing words)

Term	and	but	hate	i	image	language	like	love	natural	processing	python	signal
IDF	0.47712	0.47712	0.4771	0	0.1760913	0.477121	0.1760913	0.477121	0.47712125	0	0.477121	0.477121

Step 3: Multiply TF matrix with IDF respectively

	and	but	hate	i	image	language	like	love	natural	processing	python	signal
Text 1	0	0.47712	0.4771	0	0	0.477121	0	0.477121	0.47712125	0	0.477121	0
Text 2	0	0	0	0	0.1760913	0	0.1760913	0	0	0	0	0
Text 3	0.47712	0	0	0	0.1760913	0	0.1760913	0	0	0	0	0.477121

MODELLING

We divide the data into train and test groups when modeling. This is used to estimate the performance of machine learning algorithms, which are algorithms that are used to make predictions using data that was not used to train the model. The model is fitted with the training data, and the test data is used to predict. It's a quick and painless method for comparing the performance of machine learning algorithms for predictive-based modeling problems.

The following are some of the supervised machine learning algorithms that were used in this study:

Spamming	Fake News	Bullying/Non-Bullying
Naïve Bayes	Decision Tree	Naïve Bayes
SVM	Logistic Regression	Logistic Regression
KNN	Random Forest	SVM
Decision Tree		Decision Tree
Random Forest		

5. PERFORMANCE ANALYSIS

The following are the predicted evaluation metrics for Fake News, Cyber Bullying and spamming:

FAKE NEWS	
Model	Accuracy
Decision Tree	99.54%
Logistic Regression	98.74%
Random Forest	99.12%

BULLYING & NON BULLING	
Model	Accuracy
Decision Tree	85.82%
Naïve Bayes	76.58%
SVM	90.96%
Random Forest	95.24%

SPAMMING	
Model	Accuracy
Naïve Bayes	98.26%
SVM	83.43%
KNN	86.06%
Decision Tree	97.30%
Random Forest	97.96%

6. CONCLUSION

False information and bullying on social media have become serious issues in recent years, and a system that can detect such texts would undoubtedly be beneficial. As a result, text analysis is performed to identify inconsistencies in the text, which is then subjected to various supervised machine learning algorithms to determine the most precise estimation to resolve the issues. This way, we can assist people in making more informed decisions, as well as ensuring that they are not duped into believing what others want them to believe. Bullying, fake news, and spamming can all be mitigated as a result of this.

FUTURE ENCHANTMENTS

Testing and training corpora can be tailored to a specific domain in the future, as corpus vocabulary varies by domain.

REFERENCES

[1] M. Dadvar, R.B. Trieschnigg and F.M.G. de Jong. "Experts and Machines against Bullies: A Hybrid Approach to Detect Cyber bullies". In 27th Canadian Conference on Artificial Intelligence, University of Waterloo, Montral, Canada, 2014.

[2] A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards. 2013. "Detecting cyber bullying: query terms and techniques". In Proceedings of the 5th WebSci 2013. ACM, New York.

[3] M. Dadvar and F. de Jong. 2012. "Cyber bullying detection: a step toward a safer internet yard". In Proceedings of the 21st International Conference on World Wide Web (WWW '12 Companion). ACM, New York, NY, USA, 121-126.

[4] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the Detection of Textual Cyber bullying," MIT. International Conference on Weblog and Social Media. Barcelona, Spain, 2011.

[5] B. Erçahin, Ö. Aktaş, D. Kilinç, and C. Akyol, "Twitter fake account detection," in Proc. Int. Conf. Comput. Sci. Eng. (UBMK), Oct. 2017, pp. 388–392.

[6] T. Wu, S. Wen, Y. Xiang, and W. Zhou, "Twitter spam detection: Survey of new approaches and comparative study," Comput. Secur., vol. 76, pp. 265–284, Jul. 2018.

[7] N. Eshraqi, M. Jalali, and M. H. Moattar, "Detecting spam tweets in Twitter using a data stream clustering algorithm," in Proc. Int. Congr. Technol., Commun. Knowl. (ICTCK), Nov. 2015, pp. 347–351.

[8] C. Buntain and J. Golbeck, "Automatically identifying fake news in popular Twitter threads," in Proc. IEEE Int. Conf. Smart Cloud (SmartCloud), Nov. 2017, pp. 208– 215.

[9] C. Chen, J. Zhang, Y. Xie, Y. Xiang, W. Zhou, M. M. Hassan, A. AlElaiwi, and M. Alrubaian, "A performance evaluation of machine learning-based streaming spam tweets detection," IEEE Trans. Comput. Social Syst., vol. 2, no. 3, pp. 65–76, Sep. 2015.