International Journal of Research in Engineering, Science and Management
Volume 4, Issue 6, June 2021
https://www.ijresm.com | ISSN (Online): 2581-5792

A System of Content Analysis of Social Media using AI and NLP

Tuhina Jayanta Banerjee*
Department of Computer Engineering, Mukesh Patel School of Technology Management and Engineering, NMIMS University, Mumbai, India

***Abstract*: This paper talks about the proposed solution towards the monitoring and analysis of social media content. Social media is a platform for interchanging thoughts, views and individual perspectives but without affecting the sentimental, religious, or personal feelings of the crowd. Also, the spread of fake news has been a trend on social media. Social media platforms also provide the functionality of hiding the user credentials of the account holder and thus the spread of ill-intentioned content is on a rise with the rise in usage of such platforms. This paper talks about a step taken towards the control of such ill-intentions by designing a system that analyzes and detects fake news, toxic comments, or posts in a text or any multimedia format. Also, it reads an ill-intended chat held in a group and monitors accordingly. We have currently worked on three social media platforms which are WhatsApp, Twitter and Instagram.**

***Keywords*: Social media, toxic comment, fake news, multimedia, machine learning.**

1. Introduction

Social media has been an important part of our daily routine. A social media platform is used for content sharing and building virtual networks and communities. Among many social media platforms, Twitter, WhatsApp and Instagram are currently the most trending platforms in major parts of the world for content posting. There are several types of content posted by several account holders, a few being informative, publishing one's achievements, etc.

As per the studies, 1.8 billion photos are posted and shared over various social media platforms daily [1]. It is also studied that all the photos or content posted doesn't seem to be verified and may lead to a hoax.

Moreover, there are many snide comments under the posts which might affect the mental health of a few users. As per reports, most of the suicides are committed by mentally disturbed people who accept death to be the only solution to their problems. Problems including molestation, exploitation, bullying, untreated depression, unplanned future, or absence of healthy outlets for a child's physical or emotional needs [2]. 31% of juveniles consider social media to be a massive platform that can be used for exploring and expressing their interests, communicating with the world, gaining access to immense amounts of information and much more. Also, social media gives a broad scope of creating a feigned account which then becomes an easy way to approach such activity over these platforms being anonymous.

As per the problems discussed above, it clearly pictures the requirement of an intelligent system that finds and flags suspicious content on different social media and networking sites. This paper talks about such a system, its design, architecture and output along with the effects in the current scenario.

2. Related Work

Several studies and researches have shown that the upsurge of fake news emphasizes the degradation of long-standing constitutional bulwarks against deceptions in the internet age [3]. However, the vulnerabilities of individuals, institutions and society leading to exploitations by malicious actors remain concealed. Article "The Legal Perspective of Mental Harassment" talks about **"**Section 67: Punishment for publishing or transmitting obscene material in electronic form", which affirms, Transmission of any offensive material over an electronic platform, possessing the potential to create unfavorable conditions, or pursue people to demonstrate iniquitous behavior in the society is forbidden.

Any person trying to violate the law will be subjected to three-year imprisonment along with a fine of five lakh rupees. Any further violation and the person will be subjected to five-year imprisonment along with penalty charges of up to ten lakh rupees [4].

Few papers have been published on systems and projects similar to our system. Figure 2.1 lists the details about several papers and compares the proposed system.

Wesam Alruwaili, Bradley Protanob, Tejasvi Sirigirirajuc and Hamed Alhoorid proposed, ways to reduce illegal drug dealing on Twitter by classifying Arabic tweets related to the selling of illegal drugs using machine learning classifiers [5]. They used Twitter's API to collect 200,000 Arabic tweets, as a dataset, tweeted from October 1st to December 31st, 2016. They manually labeled the tweets, divided the dataset into 70% training and 30% testing set and applied four classifiers: Random Forest (RF), Naive Bayes (NB), Support Vector Machine (SVM) and Decision Tree (DT).

Nicole O'Brien worked on, a Machine-Learning model that

*Corresponding author: banerjeetuhina.tech@gmail.com

determines the attestations of real and fake news. They also built an application that provides a visual demonstration of the classification decision [6]. As a dataset, a paper 'liar, liar pants on fire: A new benchmark dataset for fake news detection' and its dataset was used. This dataset included 12,800 manually labeled short statements on a variety of news topics.

R. A. S. C. Jayasanka, M. D. T. Madhushani, E. R. Marcus, I. A. A. U. Aberathne and S. C. Premaratne in their paper, revealed an approach which is implemented as a tool that can analyze sentiments on Twitter social media and has then developed an application to generate knowledge that can be useful for business environments using people's attitude about their products and services [7]. A crawler was used for making the dataset, which used Twitter API to get access to twitter data. For the crawler, Public Stream (Twitter's streaming API) which is suitable for users or topics and data mining was used. This read the streams and directed that data to a database.

The authors Md. Abdul Awal, Md. Shamimur Rahman and Jakaria Rabbi in their paper, have designed a Naive Bayes classifier for detection of toxic comments written in Bangla language. This was done using a training corpus fetched from "Youtube.com" [8]. The classifier determines whether the input comment is offensive or non-offensive. The performance of the classifier has been evaluated using a 10-fold-cross-validation on the raw and unprocessed data. For training and testing the model, a dataset consisting of comments in the English language was obtained by the authors from YouTube. This dataset was then translated to Bangla Language using 'Google Translator'. The Naive Bayes classifier was then trained on this dataset.

The author Rigan Ap-apid designed an algorithm responsible for detecting nudity present in RGB images [9]. A skin color distribution model based on the RGB, HSV color spaces and Normalized RGB is designed using linear regression and correlation. The skin regions in the image are identified using the skin color model.

Our Inference:

We infer that the papers discussed above, use the following techniques to solve a relevant problem, which is described in their papers respectively, refer table 1.

Table 1
Inference

Paper	Text Classification	Multi-Lingual Text Classification	Image Classification	Video/GIF Classification	News Authentication
[1]	Yes	No	No	No	No
[2]	No	No	No	No	Yes
[3]	Yes	No	No	No	No
[4]	Yes	No	No	No	No
[5]	No	No	Yes	No	No

3. Proposed Solution and Work Done

The previous paper also talked about the type of output which will be produced by the proposed system. The system was expected to accurately and precisely give an output to the concerned social media account holders when the system encountered a fake news, a toxic comment or a post that might have a negative impact. The paper separated the entire proposed system in three modules. The first module dealt with data collection i.e., collection of various categories of news headlines, that will be used as a database for fake news detection. The second module talked about implementing if-else statements and various natural language processing and machine learning models to analyze and classify the input data. The third module dealt with collection of data from social media i.e., scraping. Then this scraped data is passed to the second module as an input. This third module is collectively referred as 'Bot' in the previous paper.

To analyze the input data, the system proposed in the previous paper talked about using News Similarity Check algorithm for data which was classified as News, Image Processing Algorithm for checking of toxic images and toxic text detection algorithm was used for input text data.

The current paper mainly talks about implementing the design and architecture proposed in our previous paper. This paper elaborates more on the actual implementation of the algorithms used in the modules which were stated in the previous paper and also talk about the accuracies obtained after implementing the system proposed in the previous paper. In further sections of this paper, a detailed comparison will be given to show which machine learning algorithm suits best for a task like toxic text detection, toxic image detection or fake news detection.

A. Data and Method

When we solve such using machine learning algorithms the data set becomes a crucial aspect of all components. The heavier the data set, the more is the accuracy.

1) Data Collection and Cleaning

In this proposed solution, we have collected 268 NSFW (Not Safe for Work) and SFW (Safe for Work) images manually to create a dataset for the image classification algorithm. The NSFW images are further classified into 4 categories based on their content namely Nude, Porn, Gore and Vulgar. The SFW images are classified into 2 subclasses namely Non-Vulgar and Safe.

First, we rescaled the RGB coefficients of the images in the range of 0–255 to target values between 0 and 1. Then we applied a range of image augmentation techniques like horizontal flip, vertical flip, zoom range, shear range, width shift range, height shift range, etc. to obtain variations of the images.

For toxic comment classification we have used 3 datasets. The first dataset is the "Jigsaw Multilingual Toxic Comment Classification" provided by *Google* which consists of toxic comments classified as Toxic, Severe Toxic, Obscene, Threat, Insult and Identity Hate. We cleaned the dataset by discarding all the numerical and special characters, and then transformed the text to vector format using TFIDF Vectorizer (term frequency–inverse document frequency) for further processing. Secondly, we have a manually created dataset consisting of text messages demonstrating threatening messages, body shamming and drug peddling. Lastly, we have a dataset consisting of 98 abusive Hindi words.

B. Method

1) Module 1: News Scraper

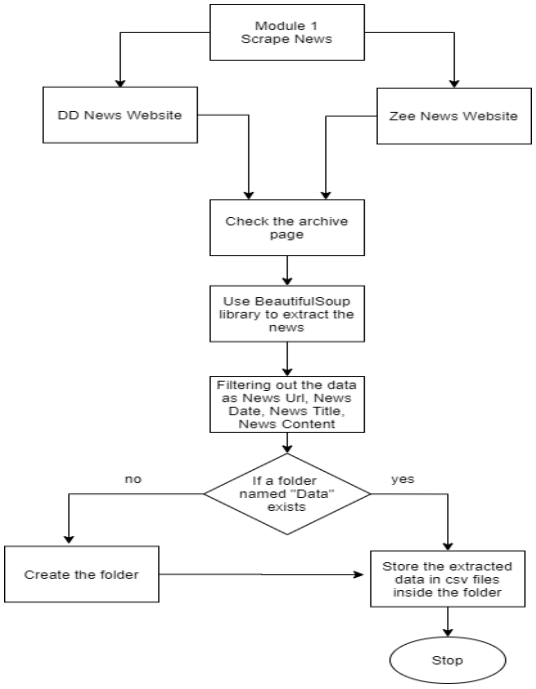

Fig. 1. Flowchart of module 1

Module 1, uses the Python library BeautifulSoup to scrape the news from the News Websites - DD News and Zee News. The data extracted from these websites are then categorized as News URL, News Date, News Title, and News Content and stored in the form of a .csv file inside a folder named 'Data'.

2) Module 2: Social Media Scraper

Module 2, is responsible for scraping the social media – Twitter, Instagram, and WhatsApp for its tweets, posts, and messages. The data scraped is classified as news, text, or image.

If the data represents news, the news similarity check algorithm is applied to get the URL of the most similar news present in the CSV files. This URL is then commented on the post or tweet on Instagram and Twitter respectively or reverted as a message on the concerned WhatsApp group.

If the data is in the form of a text, the toxic text detection algorithm is applied and the toxicity results are obtained. If an image, we check if the image contains any text, if so, then the toxic text detection algorithm is applied, else the image classification algorithm is applied to obtain the results. If the returned result is 1, denoting that the posted text or message is toxic, the concerned authority is notified to delete the message, post, or tweet.

3) Module 3: Text Classification

Module 3, involves the implementation of the "Toxic Text Detection Algorithm". It starts with the training of the Machine Learning model, using two datasets containing toxic and non-toxic comments. These comments are processed and transformed into a TFIDF (term frequency-inverse document frequency) vector to train the SVM (Support Vector Machine) model. The trained model is saved using Python's module - pickle.

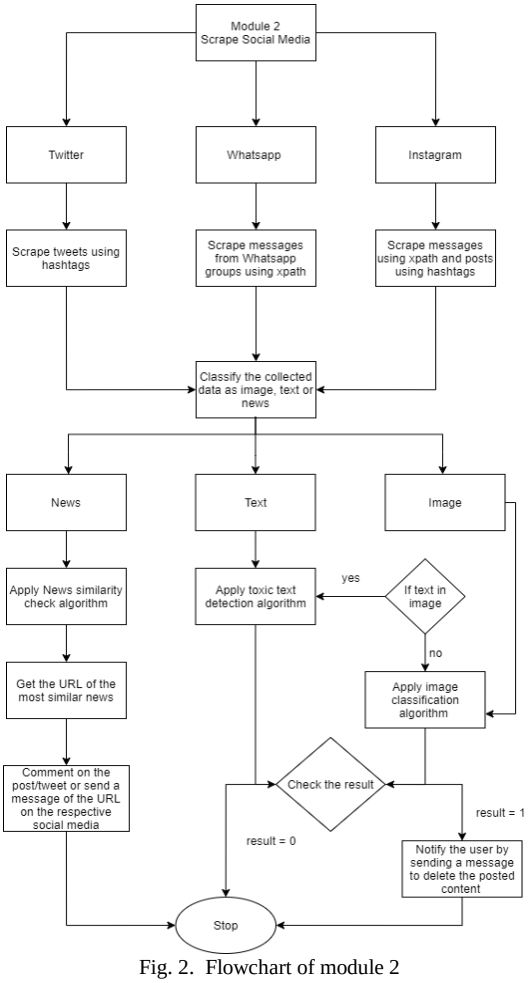

Fig. 2. Flowchart of module 2

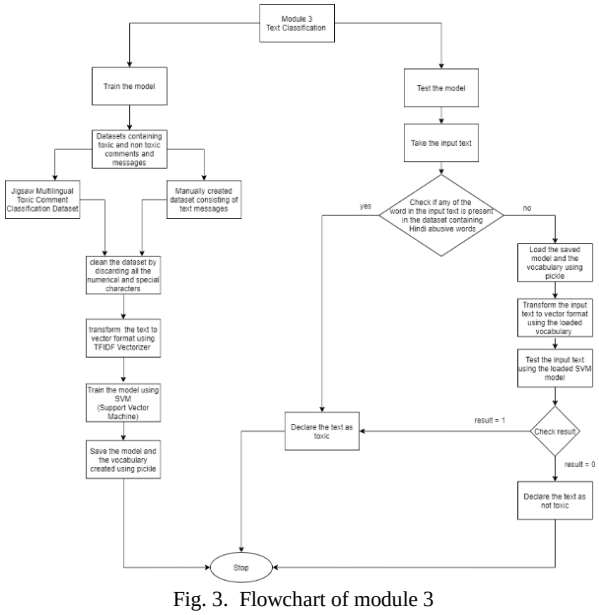

Fig. 3. Flowchart of module 3

Formula Derivation:

*tfidf(t, d) = tf(t, d) * log(N/(df + 1))*

t – term

d – document

df - the count of occurrences of term t in the document set N

When the algorithm is fed with an input text, it first checks whether any word present in the input is a Hindi abusive word, if yes then the text is claimed to be toxic. If no, then the input text is tested using this SVM model. If the result is 1, the text is declared to be toxic.

4) Module 4: Image Classification

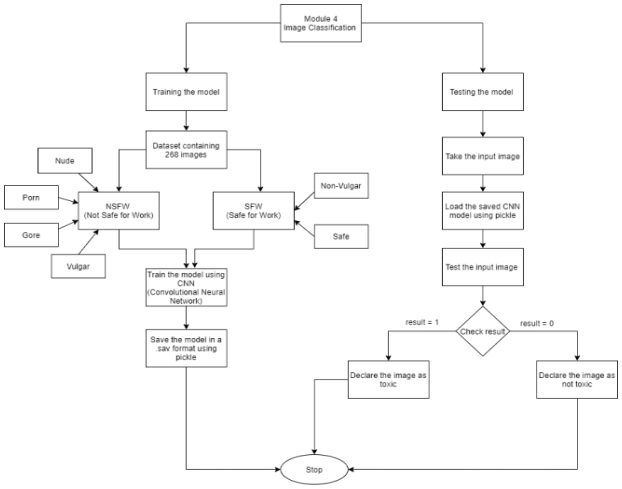

Fig. 4. Flowchart of module 4

Module 4, involves the implementation of the "Image Classification Algorithm". The CNN (Convolutional Neural Network) is trained using a dataset containing SFW (Safe for work) and NSFW (Not safe for work) images. The process of pickling is used to save this model.

The algorithm involves loading the previously saved CNN model and passing the input image through it. The model produces a result of 0 or 1. The value 1 indicating the presence of a toxic image and vice-versa.

5) Module 5: News Similarity Check

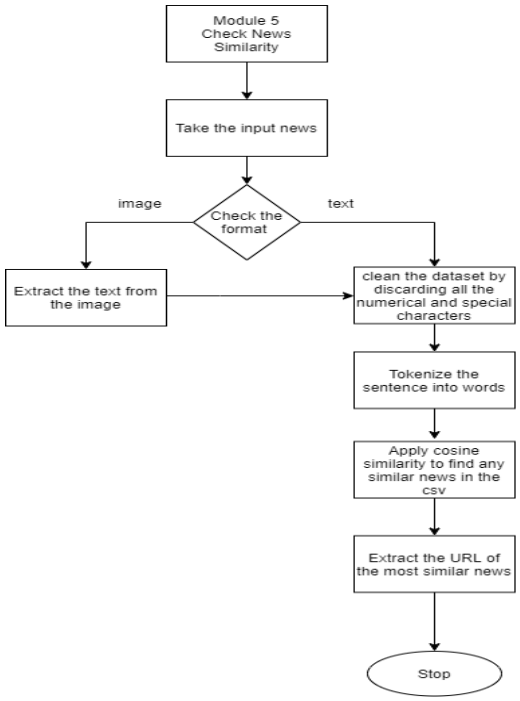

Fig. 5. Flowchart of module 5

Module 5, involves the implementation of the "News Similarity Algorithm" using the cosine similarity formula to find the similarity between two texts. The validity of news is checked by feeding it to the algorithm. When the algorithm is fed with an input image, the text is extracted from the image and fed to the algorithm.

Cosine Similarity algorithm is applied to check the similarity between the cleaned and tokenized input text and the news present in CSV files, scraped in Module 1.

The URL of the most similar news found in the CSV file is presented as the output.

Formula derivation:

$Cos(x, y) = x.y / ||x|| * ||y||$

x.y = product (dot) of the vectors 'x' and 'y'.

||x|| and ||y|| = length of the two vectors 'x' and 'y'.

*||x|| * ||y|| = cross product of the two vectors 'x' and 'y'.*

4. Results and Conclusion

Here are few results which we have received upon research and development of the proposed solution.

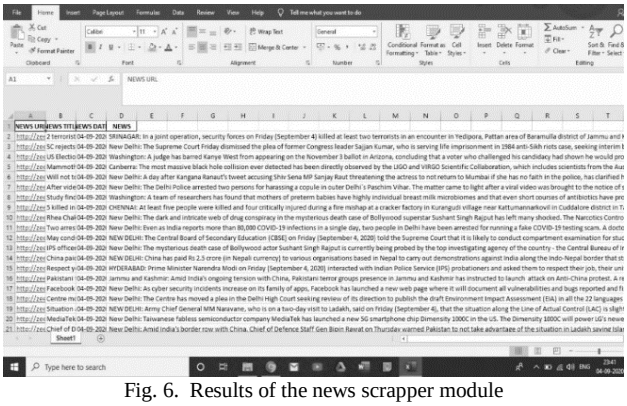

Fig. 6. Results of the news scrapper module

The table shown in the figure above consists of the news scraped from DD News and Zee News websites. The scraped data is tabulated as news URL, news date, news title, and the news.

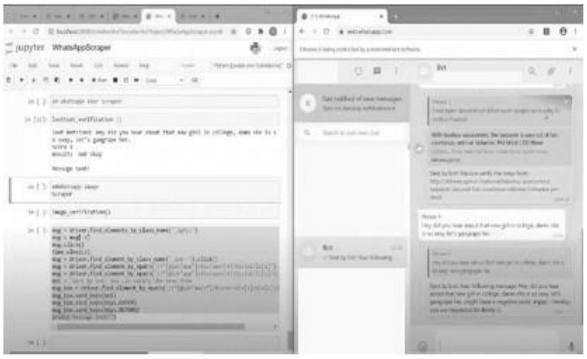

Fig. 7. Results of the social media scraper (WhatsApp)

This demonstrates the working of the system. Any text or image found inappropriate by the Bot is flagged, and the concerned person is requested to delete it. Any text or image, if labeled as fake news by the Bot, the concerned person is notified with the URL of the relevant news.

```
In [22]: text = ['i will rape you']
         score = 0
         keyword = re.sub('[^a-zA-Z]',' ',text[0])
         for i in keyword.split(' '):
             for j in hindi_toxic['Words']:
                 if j == i:
                     score = 1
                     break
         if score == 1:
             print('Text: ', text)
             print('Score: ', score)
             print('Result: Text is toxic.')
         else:
                 X = loaded_vec.transform(text)
                 score = clf.predict(X)
                 print('Text: ', text[0])
                 print('Score: ', score[0])
                 if score[0] == 1:
                     print('Result: Text is toxic.')
                 else:
                     print('Text is not toxic.')

         Text:  i will rape you
         Score:  1
         Result: Text is toxic.
```

Fig. 8. Results of text classification

This shows the working of the Toxic Text Detection algorithm that is responsible for detecting offensive content in a text.

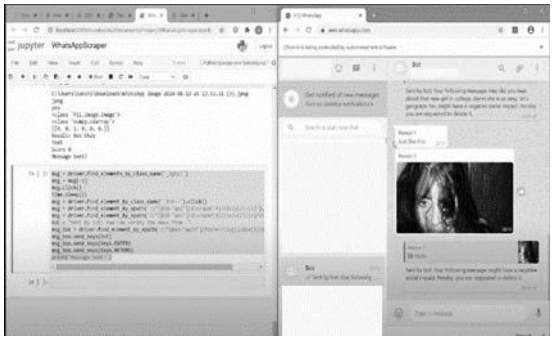

Fig. 9. Results of image classification

This shows the working of the Image Classification algorithm for finding out the presence of obscene content in images.

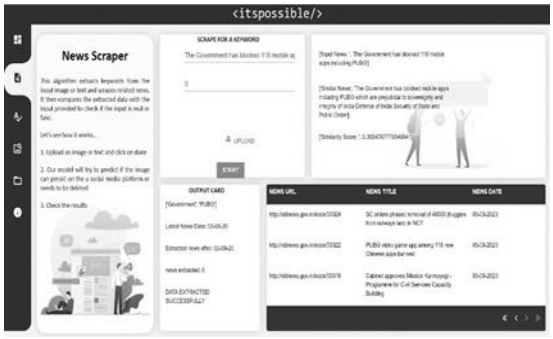

Fig. 10. Results of news similarity check

This figure shows the working of the algorithm used to find out the most relevant news from the database according to the given input news.

5. Conclusion and Future Scope

WhatsApp, Instagram and Twitter are some mainstream social media platforms that have influenced the spread of information massively. Twitter also has information on it that is considered to be authentic since it is uploaded by various verified accounts. The spread of inappropriate information from these highly influential social media platforms would have an enormous consequence on society. The idea proposed in this paper will be implemented using Machine Learning algorithms and other technologies like web-scraping and Natural Language Processing. This will result in the betterment of society. After the successful implementation of our idea in the future, we will be able to create a safety barrier for the users of these social media platforms and hence, will be able to reduce the spread of inappropriate information even before it gets uploaded on social media.

References

[1] Jim Edwards, “We’re now posting a staggering 1.8 billion photos every day,” Business insider, May 2014.
[2] Allie Squires, “Social media self-esteem and teen suicide,” PCC-Pediatric mental health, 2020.
[3] “Fake news and the spread of misinformation: A research roundup” by Denise-Marie Ordway, Journalist Resource, Sept. 2017.
[4] “The Legal Perspective of Mental Harassment” by Neha Gupta, India Legal, Jun. 2020.
[5] Wesam Alruwaili, Bradley Protanob, Tejasvi Sirigirirajuc, Hamed Alhoorid, “Using Arabic Tweets to Understand Drug Selling Behaviors”, Oct. 2019.
[6] Nicole O’Brien, “Machine Learning for Detection of Fake News”, Master’s thesis, June 2018.
[7] R. A. S. C. Jayasanka, M. D. T. Madhushani, E. R. Marcus, I. A. A. U. Aberathne and S. C. Premaratne, “Sentiment Analysis for Social Media,” Nov. 2013.
[8] Md. Abdul Awal, Md. Shamimur Rahman and Jakaria Rabbi, “Detecting Abusive Comments in Discussion Threads Using Naïve Bayes”, Oct. 2018.
[9] Rigan Ap-apid, “An Algorithm for Nudity Detection”, Jan. 2005.