Social Networking News Feeds Classification Using Naïve Bayes and Support Vector Machine and Watchdog

Gargi Bhosale[1], Kiran Jain[2], Aboli Kalyankar[3], Susmita Patange[4]

B.E. Student, Department of Computer Engineering, P.E.S. Modern College of Engineering, Pune, India [1]

B.E. Student, Department of Computer Engineering, P.E.S. Modern College of Engineering, Pune, India [2]

B.E. Student, Department of Computer Engineering, P.E.S. Modern College of Engineering, Pune, India [3]

B.E. Student, Department of Computer Engineering, P.E.S. Modern College of Engineering, Pune, India [4]

ABSTRACT: Classification is a data mining technique in which the various kinds of data are classified into the pre-defined categories. In this paper, we attempt to devise techniques to classify the users news feeds or posts taken from Facebook, into various categories using Naive Bayes classification algorithm. The posts are classified into five different categories. After the classification of posts into one of the specified categories sentiments of the posts is to be identified and categorized. Sentiment Analysis is done on the classified post to determine the opinion of that post. Here sentiments of the post are identified and classified into positive, negative or neutral category. For this purpose, Support Vector Machine(SVM) algorithm is used.

In the next part a way to implement Facebook Watchdog application, which aims to detect the explicit images and to avoid them from being uploaded on the server, is given. Explicit images are detected using an Adult Image Detection Algorithm.

KEYWORDS: Facebook news feed, Text classification, Sentiment analysis, Watchdog,Naive Bayes, SVM.

I. INTRODUCTION

At present the use of social networking is increasing at a great speed. Almost everyone uses it , from kids to adults. But it has some issues that we are trying to solve through this paper. The first issue that we felt is that on Facebook, if one has to find a news feed/post of a particular category or interest in it difficult to find it in ones timeline , probably because it is gone way down the timeline or is too old. So categorizing the post is a good option for it. Hence it becomes easier to search for a particular interest wise post. For this we are using Naive Bayes algorithm for classification. Secondly, we are also finding the opinion the posts hold. The opinion or sentiment of the post is also very important at time. For this we are using Support-Vector Machine algorithm to define the sentiment. The third issue is about images that are explicit and not apt for all to view, it indirectly leaves a negative impact on the viewers. For solving this we are performing an Adult Image Detection Algorithm on them. Just so that it is a moresafer and secure platform for all to use.

II. RELATED WORK

The aim of this project is to classify the news feeds into various categories and perform sentiment analysis on the classified posts. In [2] different methods are described for classification of facebook news feeds anddetection of sentiments of the post. The posts of the registered users are takenlive from their account, which is created on thefacebookdeveloper's site[1]. A classification algorithm is used, which classifies the posts taken from the registereduser's accounts. Sentiment analysis on the classified posts is done to detectthe sentiments of the post as positive, negative or neutral[8]. To determine whether the posts are classified correctly, different performance measures are used. The performancemeasure we have used is recall and precision[4]. The Naive Bayes algorithm is a simple

probabilistic classifier that calculates a set of probabilities by counting the frequency and combinationsof values in a given data set[6],[9].This project focuses on classifying sentiment ofclassified news feeds[7]. In Adult Image Detection, before uploading the image on our server, Adult Image Detection algorithm is applied.Ifthe image is detected as explicit it is not uploaded on our server and if it is notadult image it is allowed to be uploaded [3],[10]. To able to determine if the input image is explicit a mathematic expression is proposed, described in[5].

III. PROPOSED ALGORITHM

First one has to register on the Facebook developer's website and create an app, so that you get the permission to get live news feeds of the users of Facebook registered on the app. These posts are then used for the further classification and sentiment analysis.

The posts are classified into five different categories that are politics, sports, entertainment, education and history. The algorithm used for this is Naive Bayes Classification Algorithm.

Naive Bayes algorithm builds a conditional probability model for classification and it is supervise learning. The only assumption that the algorithm uses is that the class labels, categories in our case, are independent features. Before performing classification, pre-processing on the input posts is done. In this stop-word removal and tokenization is done. After this the conditional and posterior probabilities are calculated for the post, for each of the category. The output of applying NB is that all the posts are categorised into one of the five categories.

The next part is to apply sentiment analysis on the category-wise classified posts.

Sentiment analysis is used to find the emotional tone of a series of words. For this Support-Vector Machine Algorithm is used. Here unilateral SVM is used, which is a simple linear classification algorithm. Here also first pre-processing is performed on the posts. After this the posts are classified as to whether they are neutral (0), positive (1), negative (-1). Thus predicting the sentiments or emotional tone of the post.

In the last part Adult Image Detection algorithm is applied on the profile image. If the image is detected as explicit, it is not allowed to be uploaded on our server and if it is not adult image, it is allowed to be uploaded.
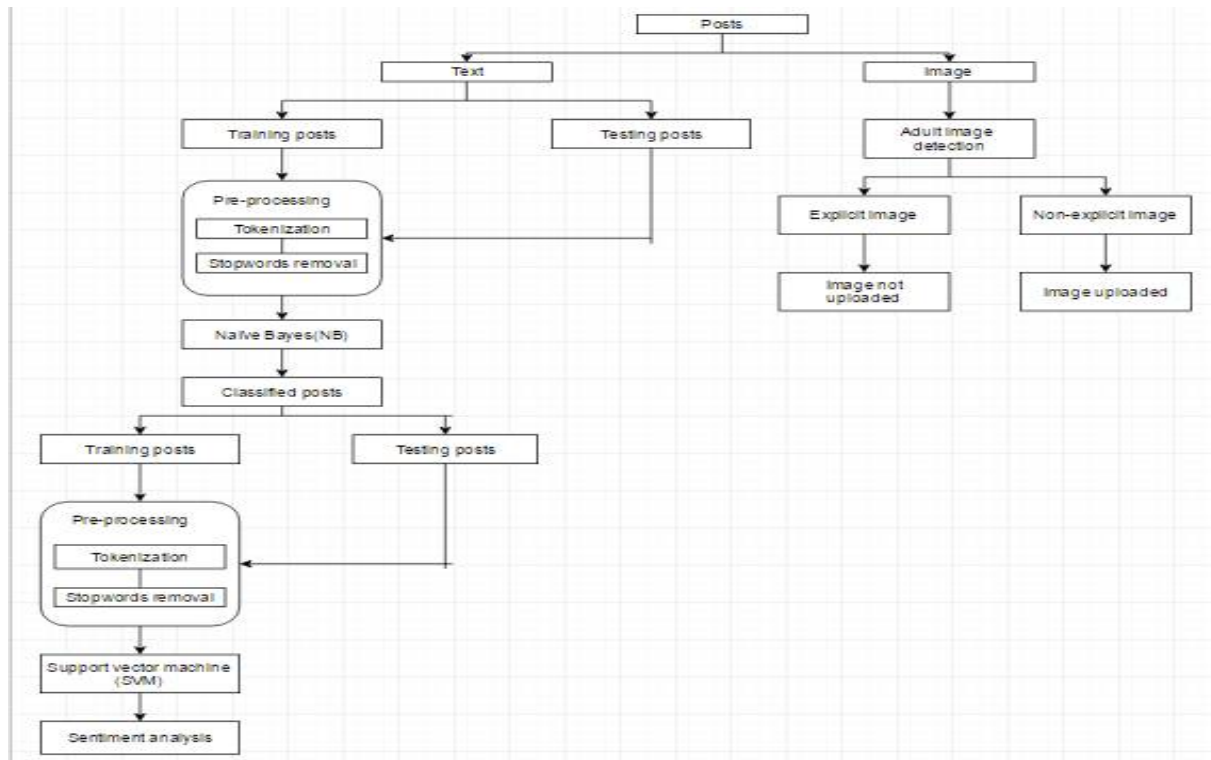

Fig1: architecture diagram

IV. PSEUDO CODE

Naïve Bayes Algorithm
The data is divided into training and testing data.

For the training data, the prior probability of each class is calculated using the formula,
    $P(c) = Nc/N$ where,

        Nc = Number of words belonging to that class and
        N = Total number of words.

In the next step, the likelihood of the word in the class is calculated using the formula,
    $P(w|c) = count (w, c) + 1/count (c) + |V|$ where,

        count (w, c) = count of that word occurring in that class count(c) = count of all words in that class.

- And then to decide in which class the testing data will fall, posterior probability is calculated.

- The post belongs to the class with highest probability.

- After classification, precision and recall is calculated to evaluate whether the posts are correctly classified.

Support Vector Machine Algorithm

- After classification, the classified posts are given as input to the SVM for sentiment analysis.

- In sentiment analysis, the sentiment score is calculated for each word. Each word has a sentimentorientation value, such as -1(negative), 0(neutral) and +1(positive).

- The sentiment score of each word is calculated and their sentiment scores are summed-up.

- The resultant sentiment score is used to detect sentiment of the post.

- Sentiment scores range from [- 1, + 1].

- If the sentiment score of the post is -1, the post is classified into negative category.

- If the sentiment score of the post is 0, the post is classified into neutral category.

- If the sentiment score of the post is +1, the post is classified into positive category.

AdultImage Detection:

- The user will select an image to upload as profile picture.

- A predefined threshold value is specified.

- Total number of pixels in the image are calculated.

- Each pixel is taken one by one and compared with the RGB color model.

- If any pixel matches the skin tone color, then the skinTonePixels count is incremented.

- The skinTonePixels percentage is calculated and if that percentage is greater than the predefined threshold value then the image contains more nuditythan threshold.

- Thus, the image will be an explicit image and will not be allowed to beuploaded on the server.

- If the percentage of skinTonePixels is less then the threshold, then the image does not contain more nudity than threshold and thus is not an explicit imageand will be allowed to be uploaded on the server.

V. SIMULATION RESULTS

The result consist of the following
1. The posts classified into different categories using naïve bayes (NB) that is education, entertainment, sport, history, politics.
2. The sentiment of the classified post as to whether positive, negative and neutral using unilateral support vector machine (SVM).
3. A table showing the accuracy of the algorithms using the measure recall and precision, with respect to number of posts.
4. Detected whether profile image is adult or not, if explicit not uploaded and if not explicit uploaded on our server.

Number of posts	Naïve Bayes Recall	Naïve Bayes Precision
150	89.46%	87.96%
250	90.87%	88.16%
450	85.23%	90.25%
700	85.98%	85.33%
850	87.76%	89.16%
1000	89.91%	88.83%

Fig1: recall and precision of naïvebayes

No of posts	Support vector machine recall	Support vector machine precision
150	93.46%	93.31%
250	91.26%	92.28%
450	92.09%	93.33%
700	92.86%	91.48%
850	93.91%	93.94%
1000	93.75%	93.67%

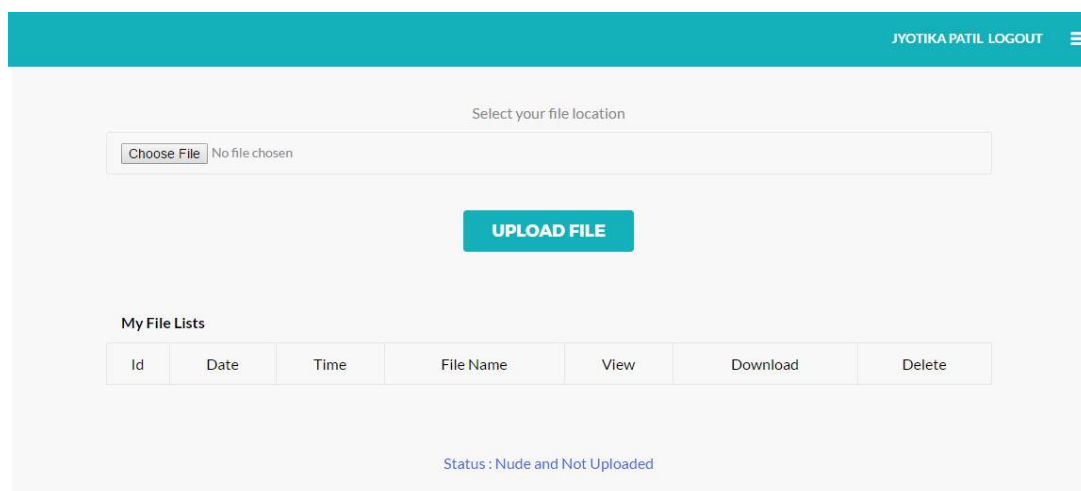Fig2: recall and precision of support vector machine
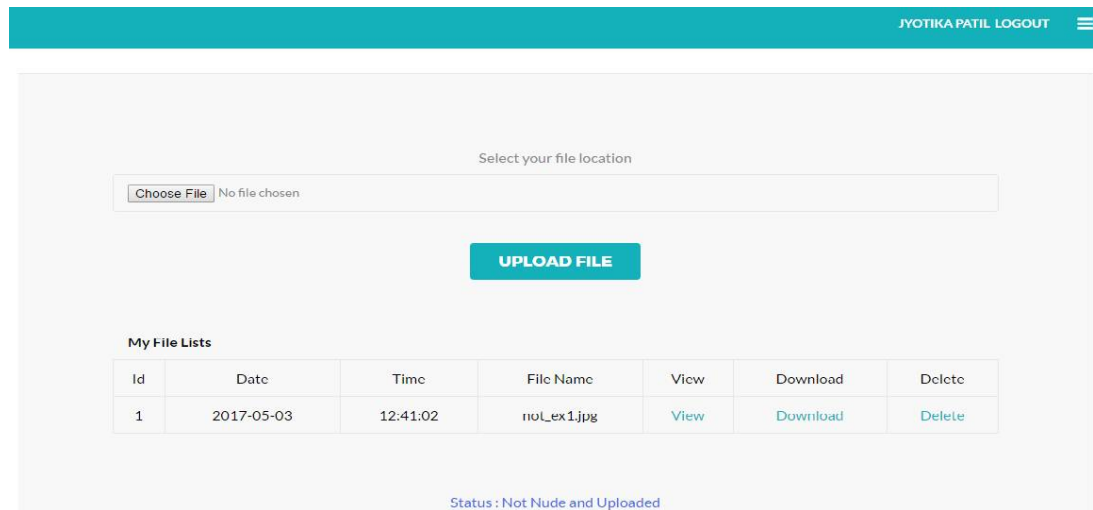
Fig2: nude image and not uploaded

Fig3: not nude and uploaded on our server

VI. CONCLUSION AND FUTURE WORK

Hence in this paper we had devised a method to classify the news feeds/posts taken from Facebook in to different categories and then also detecting the sentiment of the post, to see whether the post has a positive, negative or neutral opinion. Lastly by performing adult image detection on the profile image is performed to see whether it is explicit or not, thus providing better safety for all the users.

The next step as future work could be for performing or using more advanced adult image detecting algorithm that can also perform it on images containing text or bodies with tattoos and finding a method to prevent the image from getting uploaded on the Facebook itself, whereas in our method the image is not uploaded from our server not Facebook itself due to permission issues. Secondly we can use also use bilateral and trilateral SVM for classification

REFERENCES

1. Facebook developers. [Online]. Available: https://developers.facebook.com
2. Shankar Setty , 'Classification of Facebook News Feeds and Sentiment Analysis.' , 2014 International Conference on Advances in Computing,Communications and Informatics (ICACCI).
3. MarliesRybnicek, Rainer Poisel and Simon Tjoa , "FacebookWatchdog: A Research Agenda For Detecting Online Grooming and Bullying Activities." , 2013 IEEE International Conference on Systems, Man, and Cybernetics.
4. Kian Ming Adam Chai, HweeTou Ng, Hai Leong Chieu, "Bayesian Online Classifiers for Text Classification and Filtering", Copyright 2002 ACM 1-58113-561-0/02/0008.
5. Jorge albertomarcialbasilio, Gualbertoaguilartorres, Gabriel sánchezpérez, L. Karina toscano medina, Héctor m. pérezmeana, 'Explicit Image Detection using YCbCr Space Color Model as Skin Detection', ISBN: 978-960-474-270-7.
6. Tina R. Patil, Mrs. S. S. Sherekar, "Performance Analysis of Na¨ıve Bayesand J48 Classification Algorithm for Data Classification".
7. Julie Kane Ahkter, Steven Soria, "Sentiment Analysis: Facebook Status Messages".
8. Bing Liu, "Sentiment Analysis and Opinion Mining", Morgan and Claypool Publishers, May 2012.
9. Na¨ıve Bayes : https://www.youtube.com/watch?v=km2LoOpdB3A
10. Mahmoud A. Mofaddel ,SamySadek, 'Adult Image Content Filtering : A Stastical Method Based on Multi-Color Skin Modeling',978-1-4244-999-1-5/11/2011 IEEE.

BIOGRAPHY

Gargi Bhosale , Kiran Jain , Aboli Kalyankar , Susmita Patange studying at P.E.S Modern College of Engineering in Computer Engineering.