Python Based Naive Bayes Classifier for Spam Comment Detection

Chandana K C
Dept of Computer Applications
JNN College of Engineering,
Shimoga, Karnataka , India
chandanakc05@gmail.com

Adarsh M J
Dept of Computer Applications
JNN College of Engineering,
Shimoga, Karnataka, India
adarshmj@jnnce.ac.in

Abstract: One of the most complicated issues that topic managers encounter is comment spam. Detecting and banning comment spam may reduce server burden, enhance user experience, and clean up the network. The identification of comment spammers is the topic of this research. Spammer behavior, and also spam material, as well as spam material, were investigated. Two categories of awesome features are retrieved from the findings, which may be used to better describe spammer behaviors. In addition, the comment spam detector was built using a vector support tree method based on the retrieved characteristics. The suggested technique is tested Kaggle spam dataset, and the results show that the method outperforms the prior method in terms of detection accuracy. Furthermore, the CPU time is kept track of to show that the time being spent on training and testing is kept to a minimum.

Keywords:

Spam, Ham, Machine Learning, Natural Language Processing, Social media, Text mining, Naive Bayes Classifier.

1. Introduction:

People are ready to express their thoughts on the Internet because it is so convenient, and social media have been one of the most popular forms of communication throughout recent years. Others can respond on social media, to communicate, and express their thoughts. Commenters might contribute their information, express their support or opposition to social media, and express their thoughts. As a result, comments are becoming increasingly popular on social media.

Experts will be able to choose the best spam detection and control strategies based on the results of the suggested properly. Academicism will be able to compare the merits, limitations, techniques, and datasets used in the various existing spam detection research based on the proposed. This study may help researchers identify present study possibilities, troubles, and the specific text message feature extraction, as well as details on different datasets used by many spam text recognition researchers.

Comment spam is obvious to be worthless or has a negative impact. Similarly, the actions of Spammers are not the same as regular users. As a result, proposed a base feature extraction approach, as well as the extracted features, which may properly characterize comment spam As a result, the retrieved data is mixed with a gradient-enhancing classification model characteristics for detecting bogus comments.

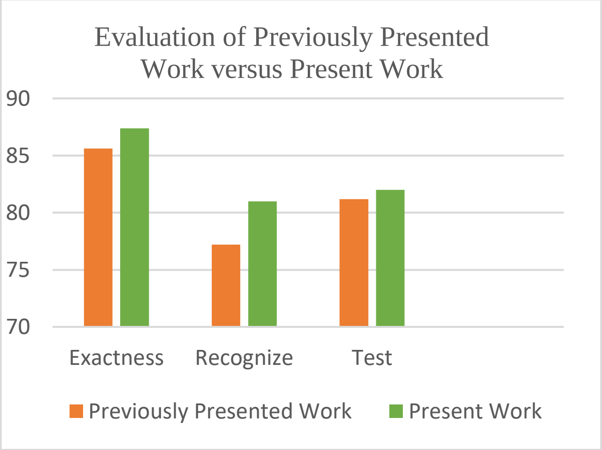
2. Evaluation of Presented Work about Present Work.

To identify spammers on the most popular social networking sites. The process included pre-processing, feature identification, and classification using a combination of Nave Bayes classification.

Previously Presented Work		
Exactness	Recognize	Test
85.6	77.2	81.17

Present Work		
Exactness	Recognize	Test
87.38	80.99	84.01

Table 1: Evaluation of Previously Presented Work versus Present Work.



The above column chart depicts a comparison of the previously presented work and the present work variable. The Orange bar shows the outcome of quality analog sis the previous work, while the Green bar line output of control fact the of current work. It is identified that the exactness value is 2.08 percent, the recognized value is 15.24 percent, and the Test value is 3.5 percent of the academic research has increased by 2.08 percent, 15.24 percent, and 3.5 percent, respectively.

3. TECHNIQUES FOR RECOGNITION OF SPAM COMMENT

Text classifiers can organize and classify the comments virtually including any type of record and internet text is an example of material. Parsing is an essential phase in natural language, with applications ranging from sentiment analysis to subject labeling and spam detection. Text categorization can be done manually or automatically; however, a human auto-complete feature evaluates the content of a text and accurately categorizes it in the procedure. Computer science methods, and other innovations, are used to automatically detect text more quickly and effectively manner for automatic text classification.

Users work by classifying content based on handcrafted linguistic criteria. The mode is classified using semantic variables based on its content. Certain phrases can help you tell if a text is spam or not. The spam text contains a few key terms that help differentiate this from non-spam language. Whenever the percentage of malicious web word vectors is higher than others.

4. Systematic guide to spam detection in social networking sites text.

The initial step in spam detection is to gather textual data from social networking sites like Twitter and Facebook, as well as online reviews, hotel assessments, and e-mails, to identify spam and non-spam (ham) material. Spam is retrieved using suitable datasets, like the Social Media giant Facebook and Tweets, both of which give a free that allows users to search and gather data from multiple accounts. They also allow data to be captured using a "hashtag" or "keyword," as well as data to be collected over time. We may classify the data as spam or ham based on the text content, and official social media sites could flag some identities or messages as spam.

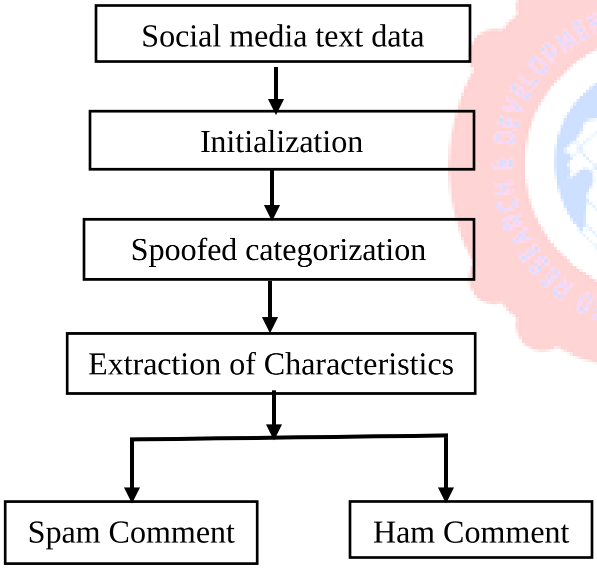


figure1: Step-by-step guide to spam detection

In this figure1, here the comments from the social media extracted those comments are initialized to the datasets then the spoofed categorization is done. Then characters are converted to binary format as 0’s and 1’s. Identification and categorization necessitate various steps.

Following data collection, pre-processing occurs, which uses a variety of natural language processing methodologies to eliminate unwanted files. These feature extraction/encoding methods turn words/text into a numeric vector that may be classified.

Online reviews of a product, resort review, or movie review may be found on sites like TripAdvisor, Amazon, Yelp, and some other sites which hold data of previous consumers who have purchased the product or stayed overnight and contributed to these reviews. Spammers combine spam content with all these evaluations to create an unfavorable image of a product or function, resulting in financial loss for the company.

5. Difficulties in detecting and classifying spam from social media content.

Spam on social media keeps rising as people's use of social networks expands drastically. The technology driving spam spread is incredible, and several social media sites were unable to appropriately identify spam spammers. Some legitimate social media users create duplicate counter-fit accounts in the ability to talk with a set of known friends. Spammers also use a few phony identities to transmit hazardous and false material, making it more difficult to trace them down.

A spammer also may take social bots to post messages automatically based on the user's interests. Many businesses employ "crowdsourcing" to develop quality, where some people are compensated to provide false feedback about a bad product. The machine learning method for spam identification suffers from over-fitting and, in some cases, a shortage of training samples. They may also have challenges if the spammers are knowledgeable and quick to adapt. When the input set is quite large, ML techniques suffer from temporal complexity, as well as memory constraints. When there are undesired features inside the dataset, the classifier's performance falls, necessitating the use of an effective feature selection procedure.

Semi-supervised learning suffers from a lack of storage space as well as an absence of effective spam detection tools. As a result, there is a great need to seek a flexible and efficient method, such as Deep Learning, to address the issues experienced by conventional Machine Learning methodologies. To order to create spam, spammers also use Deep Learning techniques to influence social media content. These fake Deep Learning algorithms-created contents are harder to detect, needing additional effort to oppose them. If there is a lack of correctly annotated data, the concept of transfer learning could be employed instead of Machine Learning.

6. Classification using Naive Bayes

Naive Bayesian spam filtering is a fundamental strategy for dealing with spam that can adjust specific users while providing low false positive spam detection capabilities that are usually acceptable to users. It is among the earliest spam filtering methods, dating back to the 1990s.

In machine learning, naive Bayes classifiers are a type of "probabilistic classifier" that is built on practically Applying' theorem with strong (naive) independence assumptions across the features. However, they could be combined with Kernel density estimation to attain better levels of accuracy.

Naive Bayes has been intensively researched since the 1960s. It was introduced into the information extraction community in the early 1960s and remains a common (benchmark) technique for text summarization, the problem of evaluating texts as being to one class or the other (document categorization) (including such trash or legit, games or politics, etc.) using frequent patterns as the features. It competes in this arena with more advanced technologies such as support vector machines with proper pre-processing. It is also used in automated medical diagnosis.

The machine learning at work is the Naive Bayesian Classifier, which does an excellent job of predicting spam classification.

The Naive Bayes Classifiers are supervised learning algorithms based on the Bayes Theorem. And these algorithms performed admirably on data in which each data point or feature is independent of the others. The following is the Naive Bayes Classifier is :

a. **A Multinomial Naive Bayes** is a statistical learning method that is commonly used during Natural Language Processing. The method guesses the tag of a text, such as an email or newspaper article, using the Bayes theorem. It computes the possibility of each tag for each sample and outputs the tag with the highest probability.

$$\boxed{\mathbf{P(c|d) = P(c) * P(d|c)/P(d)}}$$

Here, are calculating the probability of class c when predictor d as the dataset is already provided.
P(d) = prior probability of d
P(c) = prior probability of class c

P(d|c) = occurrence of predictor d given class c probability
This formula helps in calculating the probability of the tags in the text.

Advantages of using Multinomial Naive Bayes Classification

Spam is frequently tied to a user's online behavior. For instance, a user may well have subscribed to a spam online newsletter. This digital newsletter is likely to involve words that are common to everyone stone newsletters, like the name of a newsletter and its originating email address. Depending on the user's distinctive habits, a Bayes spam filter will gradually assign a greater probability.

The legitimate comments that a person gets will be unique. In a corporate setting, for example, the firm name, as well as the name of clients or customers, as well as the name of clients or customers would be mentioned frequently. Emails containing such names will be marked as less likely to be spam by the filter.

When the filter erroneously identifies an email, the keyword chances were specific to every person and can evolve and change with remedial training. As a result, the accuracy of Bayesian spam filtering after training is frequently superior to pre-defined criteria.

7. EXPERIMENTS

All of the evaluations are predicated on the dataset supplied for each social media to identify spam comments, then compute average detector accuracy and, then use using to describe the effectiveness of the features and detectors are as follows:

a. Setup

The dataset was obtained from the website and built detectors using the Python framework. All studies were carried out on such a desktop PC with an 8-core CPU and 12Gb of ram.

b. Training and Testing

The dataset's comments are tagged as 1 or 0 and are classified as spam or non-spam. 1 indicates spam and 0 indicates non-spam. Some sites only have one class commenting, so they are unsuitable for strategy. For training and testing, need both spam and non-spam samples. For that purpose, various social media from the dataset are chosen for testing.

c. Evaluation

To examine the efficacy of the suggested two types of features, first construct detectors using individual content features, next combine attributes features and generate detectors.

8. Conclusion

In comprehensive literature analysis on spam text detection and categorization described various methodologies for fake text identification in detail. The investigation also included strategies for well before, extraction of features, and spam detection. Text categorization This study will aid scholars in their studies on the subject of it highlights some of the top work done in the field of social media spam detection. Also included information on several datasets which can be used to detect spam. Prior projects on the spamming text which was before, feature extraction, and so on This classification will help researchers determine the best techniques for their work in this field like to introduce some more spam in the future, detecting methods, as well as other benefits of spam detection.

9. References

[1]. Naive Bayes Classifier https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[2]. Naive Bayes Spam Filtering https://en.wikipedia.org/wiki/ Naive_Bayes_spam_filtering

[3]. Spam Dataset https://www.kaggle.com/veleon/ham-and-spam dataset#0007.859c901719011d56f8b652e

[4]. Naman Bishnoi's Spamaway project https://github.com/diabloxenon/ Spamaway.git

[5]. Ann Nosseir Khaled Nagati and Islam Taj-Eddin et al;" Intelligent Word-Based Spam Filter Detection Using Multi.

[6]. Rouse M. 2015. Splog (spam social media). Available at http://whatis.techtarget.com/definition/splog-spambl

[7]. Salminen J, Kandpal C, Kamel AM, Jung S, Jansen BJ. 2022. Creating and detecting fake reviews of online products. Journal of Retailing and Consumer Services 64(3):102771 DOI 10.1016/j.jretconser.2021.102771.

[8]. Lara C. (2019) Naïve Bayes Spam Classification. [Online] Available: https://github.com/LeanManager/NLP_Technical_Founders/blob/master/NaiveBayes/NLP_Naive_Bayes .ipynb

[9]. Soliman, A., Girdzijauskas, S.: AdaGraph: adaptive graph-based algorithms for spam detection in social networks. In: El Abbadi, A., Garbinato, B. (eds.) NETYS 2017. LNCS, vol. 10299, pp. 338–354. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59647-1_25

[10]. Ferrara, E.: Measuring social spam and the effect of bots on information diffusion in social media. In: Lehmann, S., Ahn, Y.-Y. (eds.) Complex Spreading Phenomena in Social Systems. CSS, pp. 229–255. Springer, Cham (2018). https://doi.org/10.1007/978-3-319- 77332-2_13

[11]. Vishwarupe, V., Bedekar, M., Pande, M., Hiwale, A.: Intelligent twitter spam detection: a hybrid approach. In: Yang, X.-S., Nagar, A.K., Joshi, A. (eds.) Smart Trends in Systems, Security and Sustainability. LNNS, vol. 18, pp. 189–197. Springer, Singapore (2018). https://doi.org/10.1007/978-981-10-6916-1_17