IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁑ Peer-reviewed & Refereed journal ⁑ Vol. 12, Issue 5, May 2024

DOI: 10.17148/IJIREEICE.2024.12529

SPAM DETECTION USING VARIOUS MACHINE LEARNING ALGORITHMS

Sowmya T[1], Suvarna M[2], Sanjeev J R[3], Kiran K[4], Ganjendran[5]

Assistant Professor, Computer Science and Engineering, BMS College of Engineering, Bengaluru, India[1]

Student, Computer Science and Engineering, BMS College of Engineering, Bengaluru, India[2]

Student, Computer Science and Engineering, BMS College of Engineering, Bengaluru, India[3]

Student, Computer Science and Engineering, BMS College of Engineering, Bengaluru, India[4]

Student, Computer Science and Engineering, BMS College of Engineering, Bengaluru, India[5]

Abstract: The popularity of mobile devices is increasing day by day as they provide a large variety of services by reducing the cost of services. Short Message Service (SMS) is considered one of the widely used communication service. But this has also resulted in a rise in attacks on mobile devices, such as SMS spam. In this study, we suggest an innovative machine learning spam message detection and filtering technique based on classification algorithms. After a careful examination of the characteristics of spam messages, ten parameters were found to be useful in distinguishing SMS spam messages from ham messages. When our recommended method was applied, the Random Forest classification algorithm produced a 1.02% false positive rate and a 96.5% true positive rate.

Keywords: SMS spam, Mobile devices, Machine learning, Feature Selection

I. INTRODUCTION

Short Message Service (Sms) stands out as a ubiquitous and necessary way to stay connected in today's digital age, as communication is increasingly dependent on mobile devices. SMS has permeated every aspect of our everyday lives, from sending brief updates to pals to getting critical alerts from companies and organizations. A preferred means of communication for billions of people globally, its accessibility, immediacy, and simplicity have made it such.

But in addition to SMS's ease and widespread use, there's a serious threat that never goes away: SMS spam. SMS usage has increased dramatically as a result of telecom companies cutting their rates, drawing not only genuine users but also unscrupulous parties looking to take advantage of the channel's reach for personal gain. The outcome an overload of unsolicited communications pushing everything from shady adverts to scams, flooding users' mobile devices. Spam on mobile devices, or SMS spam, is any unsolicited message that a user receives. It's a complex problem with wide-ranging effects. Spam communications really endanger users' security, privacy, and entire experience, making them more than just a nuisance. The ramifications of falling for SMS spam can be severe, whether it comes from phishing tactics designed to obtain private information or tricks posing as official correspondence from reliable sources.

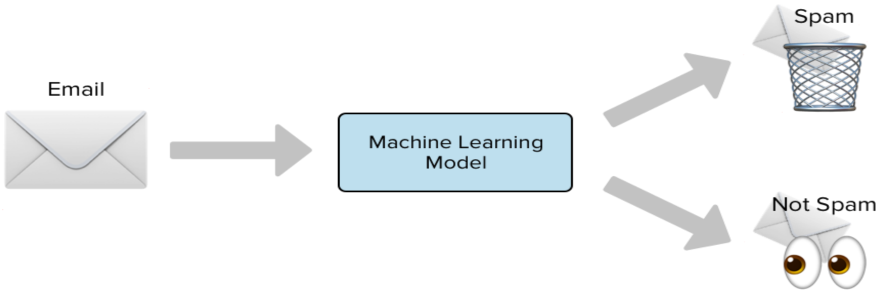

Fig 1. The text is spam or ham.

We have utilized a highlight set of 10 highlights for classification. These highlights can separate a spam SMS from ham SMS. Machine learning strategies were compelling in e-mail spam sifting because it makes a difference in avoiding zero day assaults and gives the tall level of security. The Same approach is being utilized for versatile gadgets in arrange to anticipate from SMS Spam issue but within the case of SMS Spam highlights will be distinctive from mail spam as the estimate of the content

DOI: 10.17148/IJIREEICE.2024.12529

message is little and the client employments less formal dialect for content messages. And content message is straightforward without any realistic substance and connections. The widespread adoption of Short Message Service (SMS) as a primary communication channel, the issue of SMS spam has become increasingly prevalent. Despite various security measures and spam-blocking applications available, the problem persists and poses a significant threat to users' privacy and security. The challenge lies in effectively identifying and filtering out spam messages from legitimate ones in real-time, especially considering the limitations of mobile devices and the unique characteristics of SMS messages.

SMS spam has frequently been disregarded or handled as an afterthought, in contrast to traditional email spam, which has been the focus of in-depth investigation and mitigation efforts. Effective spam identification and filtering is made more difficult by the particular qualities of SMS messages, such as their briefness, informality, and lack of complicated formatting. Furthermore, as SMS is widely used as a key communication medium, it is even more critical to discover effective ways to stop the growing problem of SMS spam.

This work suggests a novel method for addressing the SMS spam issue on mobile devices by utilizing machine learning techniques in response to this urgent requirement. Our objective is to create a system that is effective and flexible, able to recognize and eliminate spam communications in real-time, by utilizing customized feature sets and machine learning techniques. In contrast to conventional rule-based methods, which could find it difficult to stay up with the always changing landscape of spamming strategies, machine learning presents the prospect of proactive and dynamic spam detection that can pick up on and adjust to new spam patterns and strategies.

With the ultimate goal of protecting users' privacy, security, and overall mobile messaging experience, we hope to further SMS spam detection and mitigation techniques through this research project. We want to offer users efficient defense against the persistent danger of SMS spam by utilizing machine learning approaches that are specifically designed to cater to the distinct features of SMS messages. This guarantees that their mobile communication stays safe, effective, and hassle-free.

II. LITERATURE SURVEY

[1] When making a decision to purchase a product or service, people heavily consider online reviews. A valuable and respectable source of information, they can be used to assess opinions and suggestions made by the general public about goods or services. Producers and distributors worry a great deal about its contributions, reviews, and customer opinions because of its influence. Since these judgments are based only on the sentiments or perceptions of the reviewer or the customer, there is a chance that someone will fabricate evaluations in order to harm or unfairly support a product or service's reputation. Since fraudulent reviews are becoming more prevalent online, it may become a problem. These phony reviews are referred to as spam.

[2] Opinion mining and sentiment analysis have gained popularity in the last several years and are now significant tasks. It is assumed in all these research that the sources of their opinions are reliable and authentic. They might, nevertheless, run into the issue of opinion spam or falsified opinions. We examine this problem in this research within the framework of our product review mining method. People may post fraudulent evaluations, often known as review spam, on product review websites in an effort to highlight their own products or disparage those of their rivals. Recognizing and eliminating review spam is crucial. Previous research has only looked at identifying review spam. Finally, we use our crawled reviews to manually compile a spam collection. Certain heuristic guidelines, such rating deviation or helpfulness voting, restrict how well this activity can be completed. In this paper, we exploit machine learning methods to identify review spam. Toward the end, we manually build a spam collection from our crawled reviews.

[3] Online reviews of various goods, services, people, events, and other topics have grown in importance. Opinion sources including product reviews, blog entries, and forum discussions have been the subject of a lot of research lately. The classification and summarizing of opinions utilizing data mining and natural language processing techniques, however, has been the main focus of current research. Opinion spam and the veracity of internet opinions are significant issues that have received insufficient attention up to this point. The present study examines this problem within the framework of product reviews, which are highly subjective and often utilized by both consumers and product makers. A number of startup businesses that compile user feedback from product reviews have also emerged in the last two years. So, it's time to research review spam. To the best of our knowledge, there is still no published study on this topic, although Web spam and email spam have been investigated extensively.

[4] Online reviews are now a useful tool for making decisions. But with its usefulness comes a curse: opinion spam that is misleading. Fake review detection has garnered a lot of attention in the last several years. Still, the majority of review

websites do not publicly screen out fraudulent evaluations. Yelp is an exception, having for the previous few years been censoring reviews. Yelp's algorithm is proprietary, though. In this work, we analyze Yelp's filtered reviews in an effort to determine possible actions. Other review hosting sites will find the results helpful in their filtering process. The two primary methods of filtering are unsupervised and supervised learning: linguistic features and behavioral features. In this work, we will take a supervised approach as we can make use of Yelp’s filtered reviews for training.

[5] Businesses are more motivated to seek out and create DECEPTIVE OPINION SPAM—fictitious reviews that are purposefully created to look legitimate and trick the reader—as a result of the growing importance of user-generated web reviews (Cone, 2011). Gold standard fraudulent positive hotel reviews comprise the opinion spam dataset that Ott et al. (2011) just introduced. But there is also a lot of research missing on the companion issue of negative, misleading opinion spam meant to disparage rival products. We generate and analyze the first dataset of false opinion spam with negative sentiment reviews in this work, employing a methodology akin to Ott et al. (2011). We discover that typical ngram text categorization algorithms, using this dataset, perform significantly better than human judges in identifying negative deceptive opinion spam.

[6] The hypothesis put out in this research is that opinions in product reviews naturally fall into different groups. We specifically anticipate that a collection of representative distributions of review rating scores exists for a given topic. A dishonest company that employs people to create phony evaluations would inevitably skew the distribution of review scores, leaving distributional fingerprints in its wake. To support this theory, we present methods for building datasets with pseudo-gold standards that are automatically tagged using various kinds of distributional footprints. The proposed relationship between the distributional anomaly and fraudulent reviews is supported by a number of tests. Furthermore, this work offers new quantitative insights into the features of natural distributions of opinions in the areas of Amazon product reviews and Trip Advisor hotel reviews.

[7] In the modern era of electronic business, online reviews are vital. It is preferable for a consumer to peruse product or store reviews prior to deciding what to buy or where to acquire it. Customers may be tricked into purchasing inferior goods by the widespread spam reviews, while negative ratings may damage respectable retailers. We find that the majority of reviewers (> 90% in the data we examine) actually only write one review, or a "singleton review." If not, how can reviews with a singleton be screened for spam? We refer to the issue as spam detection in singleton reviews. We note that the arrival schedule of regular reviewers is consistent and uncorrelated with their rating pattern over time in order to remedy this issue. In contrast, spam attacks are usually bursty and either positively or negatively correlated to the rating.

[8] Electronic commerce websites now routinely allow their users to leave reviews of the goods they have bought. These product reviews are excellent resources for learning more about them. Before making a purchasing decision, prospective customers use them to discover what other users have to say. The manufacturers of products also utilize them to obtain competitive intelligence about their rivals and to pinpoint issues with their products. Regretfully, spam, which contains maliciously bad or falsely positive evaluations, is encouraged to spread due to the value of reviews. We try to examine review spam and spam detection in this research. There hasn't been any documented research on this issue, as far as we are aware.

[9] Online reviews of various goods, services, people, events, and other topics have become a significant resource. Opinion sources including product reviews, blog entries, and forum discussions have been the subject of a lot of research lately. The classification and summarizing of opinions utilizing data mining and natural language processing techniques, however, has been the main focus of current research. Opinion spam, or how reliable internet opinions are, is a significant problem that has received little attention up to this point. We examine this topic in this paper within the framework of product reviews, which are popular among consumers and producers and contain a wealth of opinions. A number of startup businesses that compile product reviews' opinions have also emerged in the last two years.

[10] User opinions can now be found in large quantities through online product reviews. For financial gain or notoriety, fraudsters have been posting false or misleading evaluations in an effort to uplift and/or diminish certain targeted goods or services. Review spammers are individuals who pose as reviewers. To address the issue, a number of strategies have been put out in recent years. We adopt a different strategy in our work, which uses the burstiness of reviews as a feature to detect review spammers. Reviews popping up all of a sudden may be the result of spam assaults or a product's unexpected popularity. There is a common relationship between reviewers and reviews that appear in a burst: spammers typically collaborate with other spammers, while sincere reviewers typically show up alongside other sincere reviewers.

[11] User-generated web reviews are having a growing impact on consumers' purchasing decisions. As a result, there is rising concern over the possibility of publishing fake evaluations online that are meant to appear real in order to trick readers. This type of opinion spam is known as deceptive opinion spam. In this work, we investigate generalized methods

for detecting online deceptive opinion spam using a new gold standard dataset. The dataset consists of information from three distinct domains, each of which has three different kinds of reviews: employee (domain-expert) generated deceptive reviews, Turker generated deceptive reviews, and customer generated truthful reviews. With this method, we aim to capture the overall linguistic differences between reviews that are misleading and those that are honest.

III. IMPLEMENTATION

SMS Data Collection involves gathering a dataset of SMS messages, where each message is labeled as either spam or ham (legitimate). The dataset serves as the foundation for training and evaluating the machine learning models.The SMS dataset that we have used for our experiment contains 2608 messages out of which 2408 collected from SMS Spam Corpus publically available and 200 collected manually which consists of 25 spam messages and 175 ham messages. The SMS Spam Corpus v.0.1 consists of following two sets of messages:

- □ SMS Spam Corpus v.0.1 Small - It consists of 1002 ham messages and 82 spam messages. This corpus is useful and has been used in the research.
- □ SMS Spam Corpus v.0.1 Big - It consists of 1002 ham messages and 322 spam messages. This corpus is useful and has been used by the researchers in their research work.

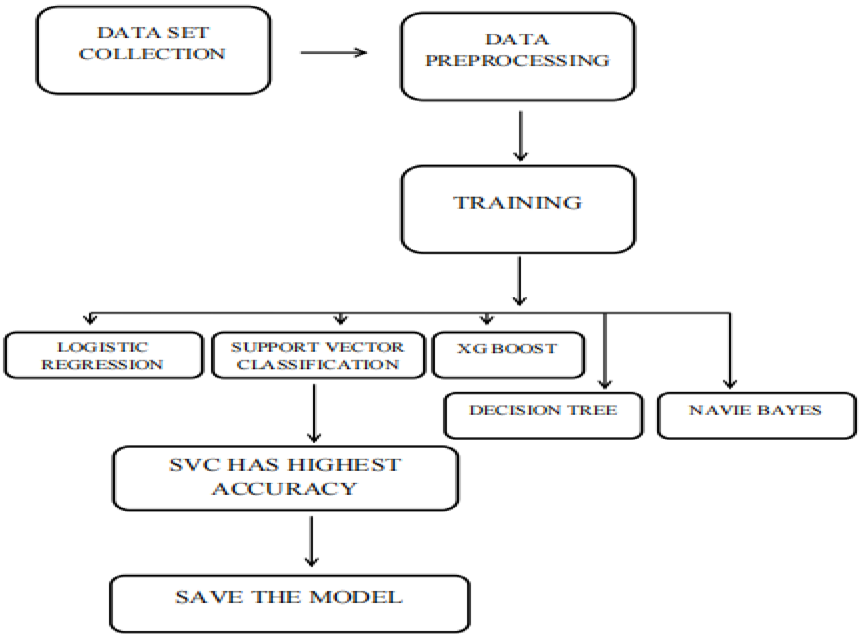

Fig 2. The steps to implement and save model.

Preprocessing is done on the SMS data to clean it up and put it in the right format. This includes:

- Removing noise: Eliminating punctuation, special characters, and other non-textual components will help reduce noise.
- Normalizing : text entails changing every word to lowercase to maintain consistency.
- Tokenization : it is the process of dividing a text into discrete words or units.
- Elimination of stop words: removing common terms (such "the" and "is") that don't really help differentiate spam from ham.
- Lemmatization/Stemming: Reducing words to their base or root form.

Features that aid in differentiating between spam and ham communications are derived from the preprocessed data. Key characteristics include of:

- Word frequency: The quantity of times a word occurs in a communication.
- The total number of letters or words in a message is its length.
- Keyword presence: Particular terms or expressions that are frequently linked to spam, such as "free," "win," or "prize."

DOI: 10.17148/IJIREEICE.2024.12529

Various machine learning algorithms are considered for the classification task. Each algorithm has distinct characteristics, making it suitable for different types of data and scenarios. The algorithms evaluated include:

- Logistic Regression: A linear model that estimates the probability of a message being spam based on the input features.
- XGBoost: A powerful ensemble technique based on gradient boosting that combines multiple weak learners to create a strong predictive model.
- Support Vector Classification (SVC): A model that finds the optimal hyperplane to separate spam and ham messages in a high-dimensional space.
- Decision Tree: A model that splits the data into subsets based on feature values, forming a tree-like structure.
- Naive Bayes: A probabilistic model based on Bayes' theorem, assuming independence between features.

The selected models are trained using the preprocessed dataset and evaluated using cross-validation techniques to ensure robustness. Key evaluation metrics include:

- Accuracy: The proportion of correctly classified messages (both spam and ham).
- Precision: The proportion of true positives (correctly identified spam) out of all predicted positives (messages identified as spam).
- Recall: The proportion of true positives out of all actual positives (all spam messages in the dataset).
- F1-score: The harmonic mean of precision and recall, providing a single metric that balances both.

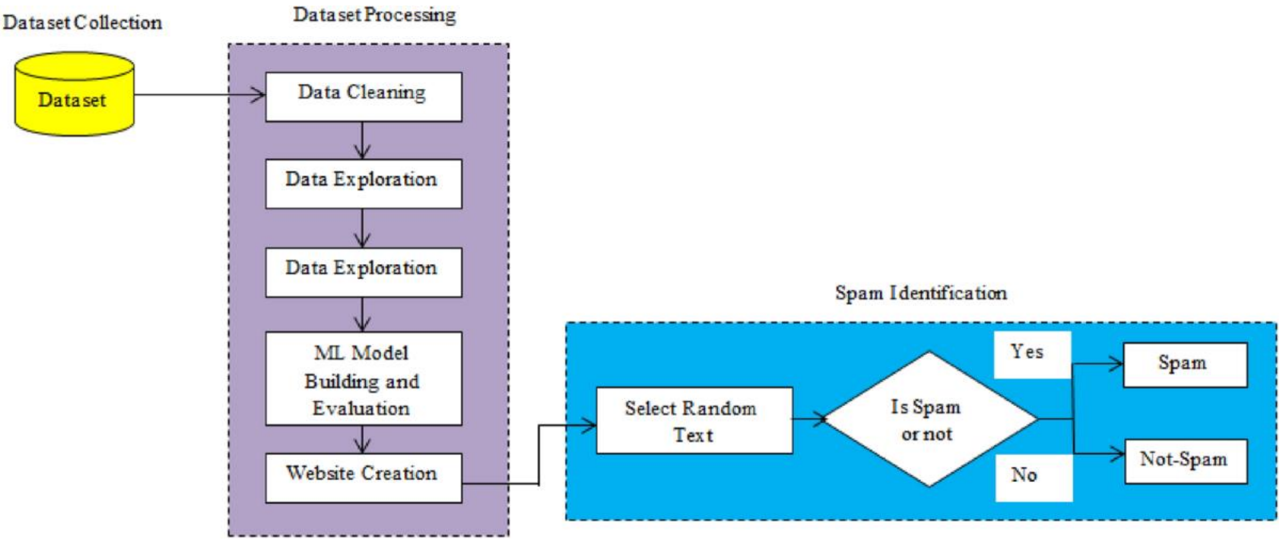

Fig 3. The block diagram of the text detecting spam or ham.

The accuracy results for the models are as follows:

- Logistic Regression: 92%
- XGBoost: 94%
- Support Vector Classification (SVC): 96%
- Decision Tree: 89%
- Naive Bayes: 90%

DOI: 10.17148/IJIREEICE.2024.12529

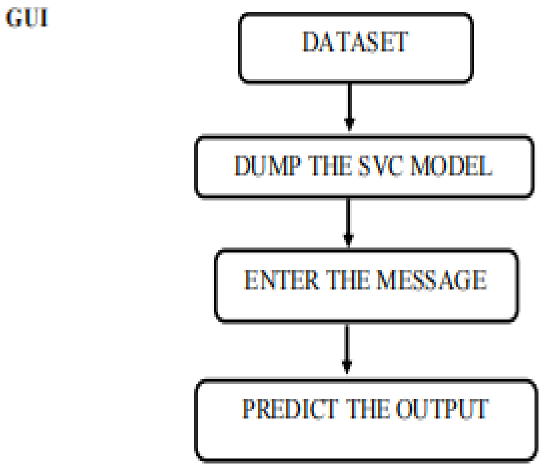

Fig 4. GUI of the implementation.

IV. RESULT

Support Vector Classification (SVC) has the greatest accuracy of 96% according to the evaluation measures. This great degree of precision shows that SVC can reliably identify between spam and ham communications. Thus, in the suggested system, SVC is selected as the recommended algorithm for real-time SMS spam identification. It is appropriate for implementation in a production setting due to its precise message classification capabilities.

Integrating the learned SVC model into the SMS processing pipeline is the deployment phase. Real-time message classification improves user experience by minimizing spam clutter by enabling spam communications to be identified and filtered before they reach users' inboxes.

V. CONCLUSION

Support Vector Classification (SVC) has the greatest accuracy of 96% according to the evaluation measures. This great degree of precision shows that SVC can reliably identify between spam and ham communications. Thus, in the suggested system, SVC is selected as the recommended algorithm for real-time SMS spam identification. It is appropriate for implementation in a production setting due to its precise message classification capabilities.

Integrating the learned SVC model into the SMS processing pipeline is the deployment phase. Real-time message classification improves user experience by minimizing spam clutter by enabling spam communications to be identified and filtered before they reach users' inboxes.

With a high true positive rate, the algorithm is the most effective model for our SMS spam filtering system since it can identify spam messages more precisely. In summary, our suggested method for SMS spam detection—which is based on the Random Forest algorithm—offers a very precise answer. This technique is well-suited for implementation in real-world applications because to its combination of carefully chosen features and Random Forest's strong classification skills, which substantially reduces the negative effects of SMS spam and improves user experience.

The version of this template is V2. Most of the formatting instructions in this document have been compiled by Causal Productions from the IEEE LaTeX style files. Causal Productions offers both A4 templates and US Letter templates for LaTeX and Microsoft Word. The LaTeX templates depend on the official IEEEtran.cls and IEEEtran.bst files, whereas the Microsoft Word templates are self-contained.

REFERENCES

[1]. Mobile Commons Blog. https://www.mobilecommons.com/blog/2016/01/howtextmessaging-will-change-for-the better-in-2016/
[2]. SMS Blocker Award. https://play.google.com/store/apps/details?id=com.smsBlocker&hl=en
[3]. TextBlocker. https://play.google.com/store/apps/details?id=com.thesimpleandroidguy.app.messageclient&hl=en
[4]. Androidapp. https://play.google.com/store/apps/details?id=com.mrnumber.blocker&hl=en

DOI: 10.17148/IJIREEICE.2024.12529

[5]. Puniškis, D., Laurutis, R., Dirmeikis, R.: An artificial neural nets for spam email recognition. Elektronika ir Elektrotechnika 69, 73–76 (2006)
[6]. Jain, A.K., Gupta, B.B.: Phishing detection: analysis of visual similarity based approaches. Secur. Commun. Netw. 2017 (2017). Article ID 5421046. doi:10.1155/2017/5421046
[7]. Gupta, B.B., Tewari, A., Jain, A.K., Agrawal, D.P.: Fighting against phishing attacks: state of the art and future challenges. Neural Comput. Appl. 1–26 (2016). doi:10.1007/s00521- 016-2275-y
[8]. Jain, A.K., Gupta, B.B.: A novel approach to protect against phishing attacks at client side using auto-updated white list. EURASIP J. Inf. Secur. 1–11 (2016). doi:10.1186/s13635- 016-0034-3
[9]. Choudhary, N., Jain, A.K.: Comparative Analysis of Mobile Phishing Detection and Prevention Approaches (Accepted)
[10]. Tatango Learning Center. https://www.tatango.com/blog/top-25-sms-spam-area-codes/