IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 4, April 2024

DOI: 10.17148/IARJSET.2024.11440

SPAM DETECTION USING MACHINE LEARNING

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

Student, Computer Science and Engineering, B.M.S College of Engineering, Bengaluru, India[1-4]

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 research, we propose a unique machine learning classification algorithm-based spam message detection and filtering method. Ten factors that can effectively separate SMS spam messages from ham messages were discovered after a thorough analysis of the traits of spam messages. The Random Forest classification technique yielded a 1.02% false positive rate and a 96.5% true positive rate when using our suggested approach.

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

I. INTRODUCTION

Short Message Service (SMS) is one of the popular communication services whereby an electronic message is transmitted. An rise in SMS usage has resulted from telecom firms lowering the cost of their services. This increase drew assailants which have caused an issue with SMS spam. Any unsolicited message received by a user on their mobile device is considered spam. Advertisements, free services, promotions, awards, and so forth are all included in spam messages. Instead of sending emails, people are utilizing SMS texts since sending them is easy and efficient and doesn't require an internet connection.The SMS Spam problem is increasing day by day with the increase in the use of text messaging. There are various security measures available to control SMS Spam problem but they are not so mature. Many android apps are also on play store to block spam messages but people are not aware of these apps due to lack of knowledge.

Other than apps the filtering techniques available mainly focuses on email spam as email spam is one of the oldest problem but with the popularity of mobile devices, SMS spam is the one of the major issue these days. SMS is one of the cheapest ways to communicate and can be considered as the simplest way to perform phishing attacks as mobile devices contain sensitive and personal information like card details, username, password, etc. Attackers are finding different ways to steal this information from mobile devices. SMS spammers can buy any versatile number with any zone code to send spam messages so that it gets to be troublesome to recognize the aggressor. US tatango learning center given the list of beat 25 SMS Spam range codes utilized by spammers National Extortion Insights Bureau (NFIB) distributed a media report almost the most recent tricks which was analyzed by activity extortion in 2016. Spammers are focusing on bank clients these days by sending spam messages for inquiring their bank account points of interest, ATM number, watchword, etc. and the client considers that the message is coming from the bank and he/she may grant all the subtle elements to the spammer. A report was distributed by ACMA that how bank clients are getting to be the casualty of SMS Spam assaults. In our proposed approach fundamental point is to channel the spam and ham SMS utilizing machine learning calculations.

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 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.
Creating a framework for identifying SMS spam utilizing machine learning addresses rising protection concerns and security dangers on versatile gadgets. With the far reaching utilize of SMS as a communication channel, the multiplication

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 4, April 2024

DOI: 10.17148/IARJSET.2024.11440

of spam messages undermines client involvement and postures dangers to delicate data.

Leveraging machine learning methods guarantees to upgrade spam sifting precision and minimize wrong positives, guaranteeing imperative messages are not erroneously classified. This approach points to address restrictions of existing arrangements by fitting highlights and calculations particularly to the interesting characteristics of SMS messages. Eventually, the inspiration lies in progressing in general security, security, and client fulfillment in versatile informing.

The main objective of this study is to develop an efficient system for detecting and filtering sms spam messages on mobile devices using machine learning techniques. Specifically, the aim is to achieve high accuracy in distinguishing spam messages from legitimate ones, while minimizing false positives to ensure that important messages are not incorrectly classified as spam.

II. LITERATURE SURVEY

1. Dixit S, Agrawal AJ (2013) Survey on review spam detection. Int J Comput Commun Technol ISSN (PRINT) 4:0975–7449

Publication Year: 2024

Summary: Reviews that are displayed online is a key position in customers choice to buy an item or facilities. They are vital and worthy source of information that can be used to evaluate public suggestions / points of view regarding items or services. Due to its influence, producers and distributors are deeply worried with its inputs or reviews / opinions from customers. Such opinions are based solely on either consumer or reviewer's feelings or perceptions, thereby giving rise to potential concerns that wrongdoer may generate bogus reviews to unfairly support or bring down the reputation of an item or services. As false reviews become more common on the web and can be an issue to it. These fake reviews are called as spams. A method for distinguishing between truthful and untruthful reviews is therefore necessary.

2. Li F, Huang M, Yang Y, Zhu X (2011) Learning to identify review spam. In: IJCAI Proceedings-International Joint Conference on Artificial Intelligence, vol 22, No. 3., p 2488

Publication year: 2011

Summary: In the past few years, sentiment analysis and opinion mining becomes a popular and important task. These studies all assume that their opinion resources are real and trustful. However, they may encounter the faked opinion or opinion spam problem. In this paper, we study this issue in the context of our product review mining system. On product review site, people may write faked reviews, called review spam, to promote their products, or defame their competitors' products. It is important to identify and filter out the review spam. Previous work only focuses on some heuristic rules, such as helpfulness voting, or rating deviation, which limits the performance of this task. 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. Jindal N, Lui B. Opinion spam and analysis. In: Proceedings of the 2008 international conference on web search and data mining.

Publication year: 2008

Summary: Evaluative texts on the Web have become a valuable source of opinions on products, services, events, individuals, etc. Recently, many researchers have studied such opinion sources as product reviews, forum posts, and blogs. However, existing research has been focused on classification and summarization of opinions using natural language processing and data mining techniques. An important issue that has been neglected so far is opinion spam or trustworthiness of online opinions. In this paper, we study this issue in the context of product reviews, which are opinion rich and are widely used by consumers and product manufacturers. In the past two years, several startup companies also appeared which aggregate opinions from product reviews. It is thus high time to study spam in reviews. 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.

DOI: 10.17148/IARJSET.2024.11440

4. Mukherjee A, Venkataraman V, Liu B, Glance N. What yelp fake review filter might be doing? In: Seventh international AAAI conference on weblogs and social media.

Publication year: 2008

Summary: Online reviews have become a valuable resource for decision making. However, its usefulness brings forth a curse – deceptive opinion spam. In recent years, fake review detection has attracted significant attention. However, most review sites still do not publicly filter fake reviews. Yelp is an exception which has been filtering reviews over the past few years. However, Yelp's algorithm is trade secret. In this work, we attempt to find out what Yelp might be doing by analyzing its filtered reviews. The results will be useful to other review hosting sites in their filtering effort. There are two main approaches to filtering: supervised and unsupervised learning. In terms of features used, there are also roughly two types: 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. Ott M, Cardie C, Hancock JT .Negative Deceptive Opinion Spam. In: HLTNAACL., 497–501, 2013.

Publication: 2013

Summary: The rising influence of user-generated online reviews (Cone, 2011) has led to growing incentive for businesses to solicit and manufacture DECEPTIVE OPINION SPAM—fictitious reviews that have been deliberately written to sound authentic and deceive the reader. Recently, Ott et al. (2011) have introduced an opinion spam dataset containing gold standard deceptive positive hotel reviews. However, the complementary problem of negative deceptive opinion spam, intended to slander competitive offerings, remains largely unstudied. Following an approach similar to Ott et al. (2011), in this work we create and study the first dataset of deceptive opinion spam with negative sentiment reviews. Based on this dataset, we find that standard n-gram text categorization techniques can detect negative deceptive opinion spam with performance far surpassing that of human judges.

6. Feng S, Xing L, Gogar A, Choi Y.Distributional footprints of deceptive product reviews. ICWSM 12:98–105, 2012.

Publication: 2012

Summary: This paper postulates that there are natural distributions of opinions in product reviews. In particular, we hypothesize that for a given domain, there is a set of representative distributions of review rating scores. A deceptive business entity that hires people to write fake reviews will necessarily distort its distribution of review scores, leaving distributional footprints behind. In order to validate this hypothesis, we introduce strategies to create dataset with pseudo-gold standard that is labeled automatically based on different types of distributional footprints. A range of experiments confirm the hypothesized connection between the distributional anomaly and deceptive reviews. This study also provides novel quantitative insights into the characteristics of natural distributions of opinions in the Trip Advisor hotel review and the Amazon product review domains.

7. Xie S, Wang G, Lin S, Yu PS. Review spam detection via temporal pattern discovery. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, Beijing, China 823–831, 2012.

Publication year: 2012

Summary: Online reviews play a crucial role in today's electronic commerce. It is desirable for a customer to read reviews of products or stores before making the decision of what or from where to buy. Due to the pervasive spam reviews, customers can be misled to buy low-quality products, while decent stores can be defamed by malicious reviews. We observe that, in reality, a great portion (> 90% in the data we study) of the reviewers write only one review (singleton review). If not, how to detect spam reviews in singleton reviews? We call this problem singleton review spam detection. To address this problem, we observe that the normal reviewers' arrival pattern is stable and uncorrelated to their rating pattern temporally. In contrast, spam attacks are usually bursty and either positively or negatively correlated to the rating.

DOI: 10.17148/IARJSET.2024.11440

8. Jindal N, Liu B. Review spam detection. In: Proceedings of the 16th international conference on World Wide Web, ACM, Lyon, France pp. 1189–1190,2007.

Publication Year: 2007

Summary: It is now a common practice for e-commerce Web sites to enable their customers to write reviews of products that they have purchased. Such reviews provide valuable sources of information on these products. They are used by potential customers to find opinions of existing users before deciding to purchase a product. They are also used by product manufacturers to identify problems of their products and to find competitive intelligence information about their competitors. Unfortunately, this importance of reviews also gives good incentive for spam, which contains false positive or malicious negative opinions. In this paper, we make an attempt to study review spam and spam detection. To the best of our knowledge, there is still no reported study on this problem.

9. Jindal N, Liu B.Opinion spam and analysis. In: Proceedings of the 2008 International Conference on Web Search and Data Mining, ACM, Stanford, CA 219– 230,2008.

Publication Year: 2008

Summary: Evaluative texts on the Web have become a valuable source of opinions on products, services, events, individuals, etc. Recently, many researchers have studied such opinion sources as product reviews, forum posts, and blogs. However, existing research has been focused on classification and summarization of opinions using natural language processing and data mining techniques. An important issue that has been neglected so far is opinion spam or trustworthiness of online opinions. In this paper, we study this issue in the context of product reviews, which are opinion rich and are widely used by consumers and product manufacturers. In the past two years, several startup companies also appeared which aggregate opinions from product reviews.

10. Fei G, Mukherjee A, Liu B, Hsu M, Castellanos M, Ghosh R. Exploiting Burstiness in reviews for review spammer detection. ICWSM 13:175–184,2013.

Publication Year: 2013

Summary: Online product reviews have become an important source of user opinions. Due to profit or fame, imposters have been writing deceptive or fake reviews to promote and/or to demote some target products or services. Such imposters are called review spammers. In the past few years, several approaches have been proposed to deal with the problem.

In this work, we take a different approach, which exploits the burstiness nature of reviews to identify review spammers. Bursts of reviews can be either due to sudden popularity of products or spam attacks. Reviewers and reviews appearing in a burst are often related in the sense that spammers tend to work with other spammers and genuine reviewers tend to appear together with other genuine reviewers. This paves the way for us to build a network of reviewers appearing in different bursts.

11. Li J, Ott M, Cardie C, Hovy E.Towards a general rule for identifying deceptive opinion spam. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland, USA, June 23-25 2014. ACL 1566– 1576,2014.

Publication Year: 2014

Summary: Consumers' purchase decisions are increasingly influenced by user-generated online reviews. Accordingly, there has been growing concern about the potential for posting deceptive opinion spam— fictitious reviews that have been deliberately written to sound authentic, to deceive the reader.

In this paper, we explore generalized approaches for identifying online deceptive opinion spam based on a new gold standard dataset, which is comprised of data from three different domains, each of which contains three types of reviews, i.e. customer generated truthful reviews, Turker generated deceptive reviews and employee (domain-expert) generated deceptive reviews. Our approach tries to capture the general difference of language usage between deceptive and truthful reviews.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 4, April 2024

DOI: 10.17148/IARJSET.2024.11440

III. METHODOLOGY

In our methodology, we point to create an proficient arrangement for identifying and sifting SMS spam messages on versatile gadgets utilizing machine learning strategies. The framework starts with the collection of a comprehensive dataset comprising labeled SMS messages, recognizing between spam and genuine (ham) messages. After information collection, pre-processing procedures are connected to clean and change the information, planning it for highlight extraction. We distinguish key highlights that can successfully separate between spam and ham messages, considering variables such as word recurrence, message length, and the nearness of particular watchwords demonstrative of spam.

Another, we select suitable machine learning calculations for classification, with a center on Calculated Relapse and XGBoost due to their demonstrated viability in double classification scenarios. The chosen models are prepared utilizing the pre-processed information, and their execution is assessed utilizing measurements such as precision, exactness, review and F1-score. We point to realize tall exactness in recognizing spam messages from genuine ones whereas minimizing wrong positives to guarantee that critical messages are not erroneously classified as spam. Upon accomplishing palatable execution amid demonstrate assessment, the prepared show is conveyed to classify approaching SMS messages in real-time.

This sending may include integration into versatile applications or back-end servers, permitting for consistent sifting of spam messages some time recently they reach the users' inbox. Through ceaseless checking and refinement, we point to adjust the framework to advancing spamming strategies and guarantee its viability in relieving the dangers postured by SMS spam.

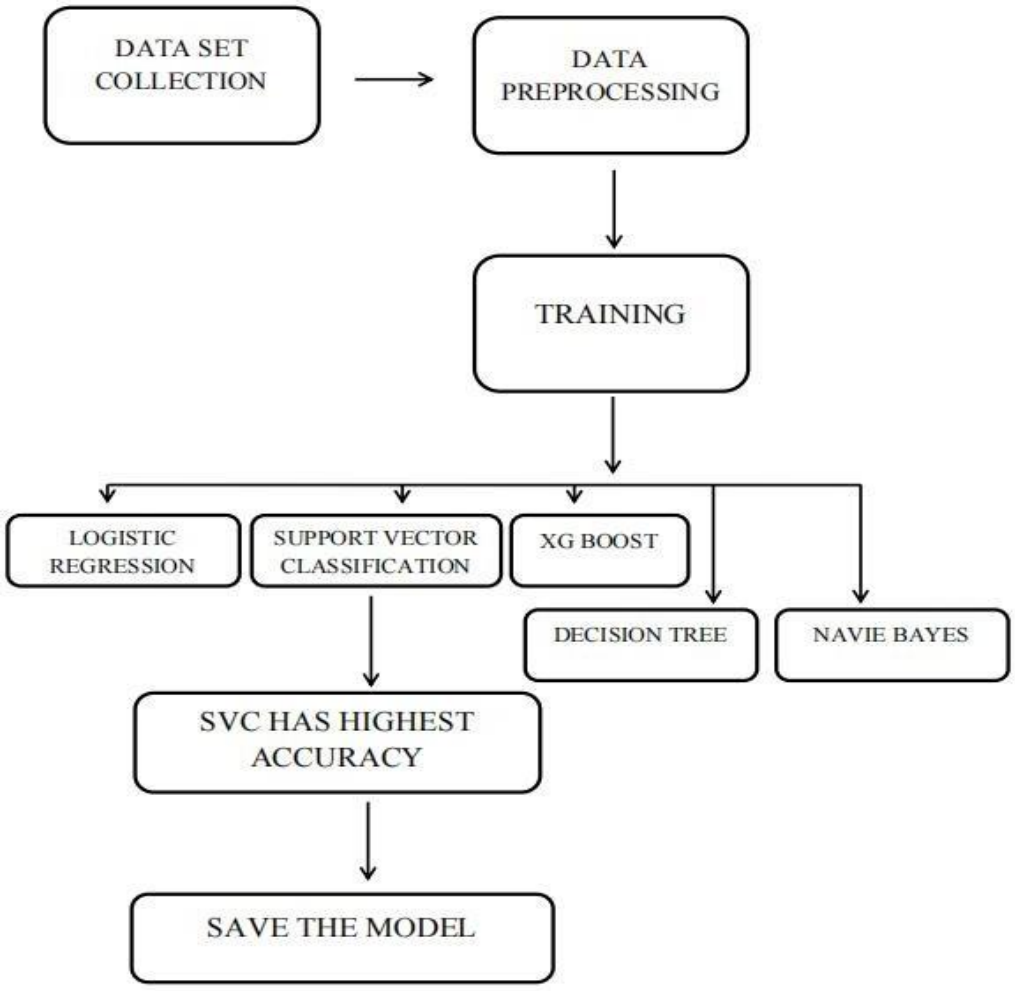

Fig 1.1 Block diagram.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 4, April 2024

DOI: 10.17148/IARJSET.2024.11440

The models such as:

Logistic regression : Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

Logistic Function – Sigmoid Function
The sigmoid function is a mathematical function used to map the predicted values to probabilities. It maps any real value The S-form curve is called the Sigmoid function or the logistic function.

In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0. into another value within a range of 0 and 1. The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form.

XGBoost : brief for Extraordinary Slope Boosting, may be a capable machine learning calculation that has picked up critical notoriety in later a long time for its uncommon execution over a wide extend of assignments. It has a place to the family of outfit learning strategies and is especially eminent for its viability in managing with structured/tabular information.

XGBoost builds upon the concept of angle boosting, which consecutively trains a arrangement of frail learners (more often than not choice trees) and combines their forecasts to deliver a strong last show. What sets XGBoost separated is its optimization methods, counting gradient-based optimization, regularization, and parallel handling, which make it greatly effective and versatile. Its flexibility permits it to exceed expectations in relapse, classification, and positioning issues, and its capacity to handle lost information and join custom assessment measurements assist upgrades its adaptability.

Support Vector Classification (SVC): it is a powerful algorithm for both classification and regression tasks. It works by finding the hyperplane that best separates the classes in the feature space while maximizing the margin between them. SVC is effective in high-dimensional spaces and is versatile in handling various types of data. It can handle non- linear decision boundaries using techniques like kernel trick. SVC tends to perform well when there is a clear margin of separation between classes in the data.

Navie Baise: Naive Bayes is a probabilistic classifier based on Bayes' theorem with the assumption of independence between features. Despite its simplicity, Naive Bayes can perform well in many complex real-world situations, particularly in text classification tasks like spam detection or sentiment analysis. It is efficient, especially with high dimensional datasets, and requires a small amount of training data to estimate the necessary parameters.

All these algorithms have their strengths and weaknesses, Support Vector Classification (SVC) may exhibit the highest accuracy compared to others in certain scenarios. SVC's ability to find complex decision boundaries and handle highdimensional data effectively makes it a strong contender for classification tasks, particularly when there is a clear margin of separation between classes in the data.

However, the choice of algorithm ultimately depends on various factors such as dataset size, complexity, computational resources, and the desired level of accuracy. Therefore, it is essential to evaluate and compare the performance of different algorithms on the specific task at hand before drawing conclusions about their effectiveness

Decision Tree: A decision tree is a popular machine learning algorithm used for both classification and regression tasks. It works by recursively partitioning the feature space into smaller regions, where each partition is associated with a simple decision rule based on the features` values.

At each step of the tree-building process, the algorithm selects the feature that best splits the data into pure or nearly-pure subsets, typically using metrics like Gini impurity or information gain. This process continues until a stopping criterion is met, such as reaching a maximum tree depth, having a minimum number of samples in each leaf node, or when further splitting does not improve the model's performance.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 4, April 2024

DOI: 10.17148/IARJSET.2024.11440

IV. RESULT & CONCLUSION

The methodical approach to handling the growing problem of spam messages in text messaging is provided by the suggested method for SMS spam filtering. The research attempts to improve the precision and effectiveness of spam identification by utilizing machine learning techniques and a carefully selected set of ten features. The study offers important insights into the efficacy of various classification techniques for this task by comparing five well-known algorithms-Naïve Bayes, Logistic Regression, J48, Decision Table, and Random Forest-on a dataset of 2608 messages, which includes both manually collected samples and publicly available data from the SMS Spam Corpus v.0.1.

The experimental findings' analysis provides insight into the advantages and disadvantages of each method, as well as the importance of the selected feature set for message classification into spam and non-spam. The study provides useful recommendations for practitioners and researchers looking to design strong SMS spam filtering systems by evaluating performance parameters like accuracy, precision, recall, and F1-score.

Essentially, by providing a structured framework for algorithm selection, feature engineering, and evaluation, the suggested technique aids in the ongoing attempts to prevent SMS spam. The project intends to promote the development of more accurate, dependable, and scalable solutions to solve this ubiquitous problem in communication technology by expanding our understanding of effective tactics for spam identification in text messaging.

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
[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/