IT in Industry, Vol. 7, No.2, 2019

Published Online 09-Oct-2019
DOI: https://doi.org/10.17762/itii.v7i2.802

Spam Detection and Spammer Community Detection in Social Media Platform Using Machine Learning Techniques

Lisa Gopal
Department of Comp. Sc. & Info. Tech., Graphic Era Hill University, Dehradun, Uttarakhand, India 248002

***Abstract*:** In recent years, social media platforms have experienced a surge in popularity, leading to an increase in spam and spammer communities. This paper presents a comprehensive study on spam detection and spammer community detection in social media platforms using machine learning techniques. The proposed approach focuses on three main features: Review-Behavioral (RB) Based features, Review-Linguistic (RL) Based Features, and User-Behavioral (UB) Based Features. By combining these features, we aim to create a robust and accurate spam detection model that effectively identifies spam content and spammer communities in social media networks. The Review-Behavioral (RB) Based features focus on the patterns and tendencies observed in user-generated content, such as the frequency of posting, the time between posts, and the distribution of ratings. Review-Linguistic (RL) Based Features, on the other hand, analyze the linguistic characteristics of the content, including the use of specific keywords, the complexity of the text, and sentiment analysis. Lastly, User-Behavioral (UB) Based Features examine the behavior of users in the social media platform, including their social connections, interactions, and activity patterns. By incorporating these features into a machine learning model, we aim to develop an effective spam detection and spammer community detection system that can be used to protect social media platforms from malicious activities. The results of this study will provide valuable insights into the effectiveness of these features in identifying spam and spammer communities, as well as contribute to the ongoing efforts to improve the security and integrity of social media networks.
Keywords: spammer, effectiveness, insights, contribute, communities

1. Introduction

The rapid growth of social media platforms has transformed the way people communicate, share information, and interact with each other. As these platforms continue to attract millions of users worldwide, they have also become a breeding ground for spam and spammer communities. Spammers exploit social media networks to spread unsolicited content, promote fake products, propagate misinformation, and engage in other malicious activities. The presence of spam and spammer communities not only degrades the user experience but also poses significant threats to the security and integrity of social media platforms. To address this challenge, researchers and practitioners have been actively exploring various techniques to detect spam and identify spammer communities in social media networks. Machine learning has emerged as a promising approach for spam detection, as it can effectively learn and recognize patterns in large volumes of data. In this paper, we propose a novel spam detection and spammer community detection method in social media platforms using machine learning techniques, with a focus on three main features: Review-Behavioral (RB) Based features, Review-Linguistic (RL) Based Features, and User-Behavioral (UB) Based Features.The Review-Behavioral (RB) Based features aim to capture the patterns and tendencies observed in user-generated content, such as the frequency of posting, the time between posts, and the distribution of ratings. By

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Published Online 09-Oct-2019
DOI: https://doi.org/10.17762/itii.v7i2.802

analyzing these features, we can gain insights into the behavioral patterns of spammers and distinguish them from legitimate users. Review-Linguistic (RL) Based Features, on the other hand, focus on the linguistic characteristics of the content, including the use of specific keywords, the complexity of the text, and sentiment analysis. These features can help identify spam content based on its distinctive linguistic properties, such as the excessive use of promotional language or the presence of irrelevant keywords. Lastly, User-Behavioral (UB) Based Features explore the behavior of users in the social media platform, including their social connections, interactions, and activity patterns. By examining these features, we can identify potential spammer communities and understand their strategies for spreading spam content across the network.

The main objective of this paper is to investigate the effectiveness of combining RB, RL, and UB based features in detecting spam and spammer communities using machine learning techniques. Through a comprehensive analysis of these features, we aim to develop a robust spam detection model that can be employed by social media platforms to safeguard their users and maintain the quality of user-generated content. This study will contribute to the ongoing efforts to improve the security and integrity of social media networks and provide valuable insights for future research in this area. In addition to exploring the RB, RL, and UB based features, this study also investigates the performance of two widely-used machine learning algorithms, namely K Nearest Neighbours (KNN), Support Vector Machine (SVM) and Naïve Bayes (NB), for the task of spam detection and spammer community detection in social media platforms. In this paper, we will evaluate the effectiveness of KNN, NB and SVM algorithms in detecting spam and spammer communities when incorporating the RB, RL, and UB based features. By comparing the performance of these algorithms, we aim to identify the most suitable machine learning approach for tackling the challenges posed by spam and spammer communities in social media networks.

2. Literature Review

In this literature review, we discuss papers related to linguistic-based, Behavioural Based, Graph Based and machine learning methods for detecting spam in online review communities and their respective contributions to the field.

A. Linguistic Based Methods

Ott, Cardie, and Hancock (2012) conducted a study to estimate the prevalence of deception in online review communities. The authors employed a machine learning approach to analyze the linguistic cues associated with deceptive opinion spam. They created a gold standard dataset consisting of truthful and deceptive reviews, where the latter were specifically solicited for the study. By training a supervised learning algorithm on this dataset, the researchers identified linguistic features that were indicative of deception in online reviews. Their findings revealed that deceptive reviews exhibited a distinct linguistic style compared to truthful reviews, suggesting that linguistic-based methods could be effective in detecting deceptive opinion spam in online review communities.

Ott, Choi, Cardie, and Hancock (2011) presented a study that aimed at finding deceptive opinion spam using linguistic features. They developed a novel approach that combined n-gram text features with stylometric and psycholinguistic features to detect deceptive reviews. Their method

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Published Online 09-Oct-2019
DOI: https://doi.org/10.17762/itii.v7i2.802

involved training a machine learning model on a gold standard dataset containing truthful and deceptive hotel reviews. The results demonstrated that the proposed approach, which considered a diverse range of linguistic features, outperformed the traditional n-gram text features in detecting deceptive opinion spam. This study highlighted the importance of considering a variety of linguistic features to improve the accuracy of spam detection models.

Xu and Zhang (2014) proposed a method to combat product review spam campaigns using multiple heterogeneous pairwise features. They focused on detecting review spam campaigns by analyzing the relationships between different pairs of users, reviews, and products. The authors devised a novel pairwise-feature based approach, which combined linguistic features with behavioral features to capture the characteristics of spam campaigns. The experimental results on a real-world dataset showed that their proposed method achieved promising results in identifying review spam campaigns. This study underscored the potential benefits of incorporating linguistic-based methods with other feature types to enhance the detection of spam campaigns in online review communities.

B. Behavioural-based methods

Jindal and Liu (2008) investigated opinion spam and its analysis in the context of online reviews. The authors proposed a set of heuristics based on the behavioral patterns of reviewers to identify suspicious users and reviews. They highlighted the importance of understanding the nature of opinion spam and the need for developing effective methods to detect and combat it.

Li et al. (2011) and presented a study on learning to identify review spam using machine learning algorithms. The authors proposed a framework that incorporates both content-based and behavioral-based features to train classifiers for spam detection. They demonstrated the effectiveness of their approach in detecting review spam and suggested that combining different types of features can lead to improved performance.

Jindal et al. (2012) focused on finding unusual review patterns using unexpected rules. The authors proposed a method that leverages association rule mining to detect suspicious reviews and reviewers. By examining the relationships between the behavioral features of reviewers, the authors were able to identify unusual review patterns and thus improve spam detection.

Feng et al. (2012) proposed a syntactic stylometry approach for deception detection. The authors analyzed the syntactic patterns of deceptive texts and used this information to develop a machine learning-based model for detecting deceptive content. Their findings revealed that syntactic stylometry can be an effective tool for deception detection, providing valuable insights for the development of more sophisticated spam detection techniques.

C. Graph Based Methods

Shehnepoor et al. (2017) proposed a network-based spam detection framework called NetSpam. This framework leverages the relationships between users, reviews, and products to detect spam reviews by constructing a heterogeneous information network. Using graph-based ranking algorithms, NetSpam assigns spam scores to reviews and identifies potential spam content. The authors demonstrated the effectiveness of NetSpam by comparing its performance with other state-of-the-art spam detection methods, showing that it outperforms existing techniques in terms of accuracy and efficiency.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Published Online 09-Oct-2019
DOI: https://doi.org/10.17762/itii.v7i2.802

Fei et al. (2013) explored the burstiness of reviews as a feature for detecting review spammers. The authors suggested that spammers often post multiple reviews in a short period of time, leading to bursts of activity. By modeling the review posting process as a graph, they were able to identify bursty behavior and distinguish spammers from legitimate users. Their method achieved promising results in detecting review spammers, indicating the value of considering burstiness as a feature in spam detection algorithms.

Choo et al. (2015) proposed a method for detecting opinion spammer groups through community discovery and sentiment analysis. They constructed a user-review graph, where nodes represent users and edges indicate the similarity between users' reviews. By applying community detection algorithms, they identified groups of users with similar review behavior, which could potentially be spammer groups. Sentiment analysis was then employed to further validate the presence of spammers within these communities. Their approach demonstrated the potential for using graph-based methods in conjunction with sentiment analysis to detect spammer groups in online review systems.

Li et al. (2014) developed a method for spotting fake reviews using collective positive and unlabeled (PU) learning. By constructing a bipartite graph of users and products, they were able to model the relationships between users and the products they reviewed. The authors employed a collective PU learning algorithm to iteratively update the probability of each review being fake based on the graph structure. Their method showed improved performance in detecting fake reviews compared to traditional PU learning techniques, highlighting the usefulness of incorporating graph-based methods into spam detection algorithms.

D. Machine Learning Technique

Sharmin, S., Zaman, Z. (2017) explored spam detection in social media by employing machine learning tools for text mining. The authors focused on detecting spam in social media platforms, particularly Twitter, and implemented different machine learning algorithms such as Naïve Bayes, Decision Tree, and Support Vector Machines (SVM) for text classification. Their results demonstrated that the Naïve Bayes classifier outperformed the other methods in detecting spam content, emphasizing the effectiveness of text mining in identifying spam in social media.

Mane, D. T., & Kulkarni, U. V. (2018) proposed a Modified Fuzzy Hypersphere Neural Network for pattern classification using supervised clustering. Their approach focused on enhancing the performance of the Fuzzy Hypersphere Neural Network by incorporating an adaptive learning rate and a supervised clustering method. The proposed method was tested on various datasets, including spam email classification, and demonstrated improved classification accuracy compared to the standard Fuzzy Hypersphere Neural Network.

Zrigui M et al (2012) developed an Arabic text classification framework based on Latent Dirichlet Allocation (LDA). The authors utilized LDA to extract topics from Arabic text documents and employed SVM and Naïve Bayes classifiers for text classification tasks. The proposed framework was tested on several Arabic text datasets and showed promising results, indicating the potential of LDA-based methods for text classification in different languages.

Chen Q. L. Yao, and J. Yang (2016) investigated short text classification using the LDA topic model. The authors proposed a method to address the challenges posed by short text classification, such as the lack of

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Published Online 09-Oct-2019
DOI: https://doi.org/10.17762/itii.v7i2.802

sufficient textual information and the high dimensionality of the feature space. By applying the LDA topic model to extract features from short texts and using SVM for classification, the proposed method demonstrated improved classification performance compared to traditional bag-of-words approaches.

Kandasamy, K., and P. Koroth (2014) presented an integrated approach to spam classification on Twitter by combining URL analysis, natural language processing, and machine learning techniques. The authors extracted features from tweets, such as URL redirections, text sentiment, and word frequency, and used various machine learning algorithms, including Naïve Bayes, Decision Tree, and SVM, for classification. Their results indicated that the integrated approach effectively identified spam content on Twitter and highlighted the importance of considering multiple feature types for robust spam detection.

These studies showcase the effectiveness of linguistic-based, Behavioural Based, Graph Based and machine learning methods in detecting deceptive opinion spam and spam campaigns in online review communities. The research highlights the importance of leveraging various linguistic features, alongside other types of features, to improve the performance of spam detection models. These studies serve as valuable references for the development of advanced spam detection techniques in social media platforms.

3. PROPOSED SYSTEM

In this paper, we propose a novel system for Spam Detection and Spammer Community Detection in social media platforms using machine learning techniques, specifically focusing on Support Vector Machines (SVM), k-Nearest Neighbors (KNN), and Naive Bayes (NB) algorithms. The proposed system aims to accurately detect spam content and identify spammer communities by leveraging the strengths of these three machine learning algorithms.

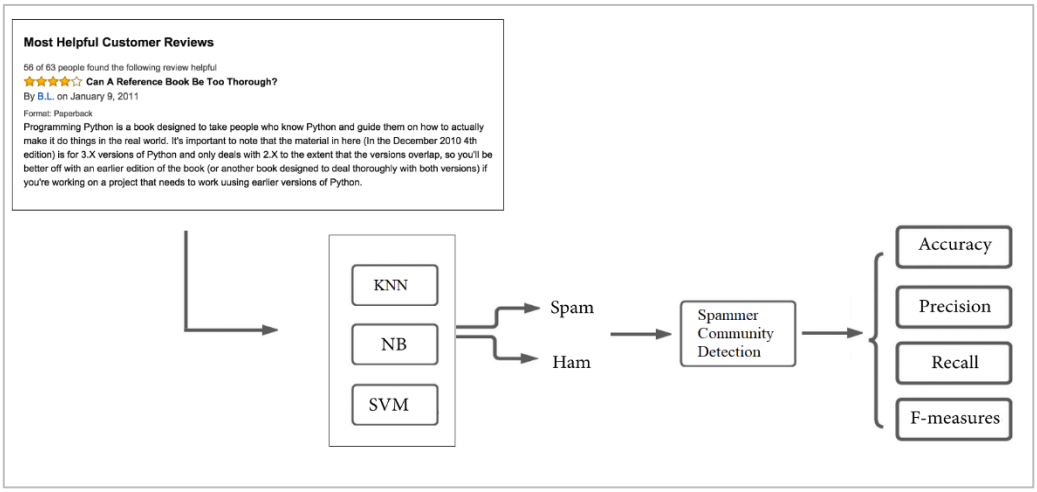

Figure 1: System Architecture

The proposed system consists of the following steps:

1. Feature extraction: We extract three types of features from the social media data: Review-Behavioral (RB) Based features, Review-Linguistic (RL) Based Features, and User-Behavioral (UB) Based Features. These features capture the patterns in user-generated content, linguistic properties of

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Published Online 09-Oct-2019
DOI: https://doi.org/10.17762/itii.v7i2.802

the content, and user behavior on the platform.

2. Preprocessing and feature selection: The extracted features are preprocessed and normalized to ensure compatibility with the machine learning algorithms. We apply feature selection techniques to identify the most relevant and informative features for spam detection and spammer community detection tasks.

3. Model training: We train three different machine learning classifiers, namely SVM, KNN, and NB, using the selected features. Each classifier is fine-tuned to achieve optimal performance in detecting spam and identifying spammer communities.

4. Model evaluation and comparison: The performance of the SVM, KNN, and NB classifiers is evaluated using standard metrics such as precision, recall, F1-score, and accuracy. We compare the results to determine the most effective algorithm for spam detection and spammer community detection in social media platforms.

5. Spammer community detection: We use the trained classifiers to identify potential spammer communities in the social media network by analyzing user behavior patterns, social connections, and interactions.

By incorporating SVM, KNN, and NB algorithms in our proposed system, we aim to take advantage of their unique strengths in handling different aspects of the spam detection and spammer community detection tasks. SVM is known for its effectiveness in handling high-dimensional data and providing accurate classification results, especially in cases where the data is linearly separable. KNN is a simple yet powerful algorithm that can adapt to changes in the data distribution and identify local patterns effectively. Lastly, NB is a probabilistic classifier that assumes independence between features and is particularly useful when dealing with large datasets and text-based data.

4. Result Analysis

A. Dataset Description

The Amazon dataset, provided by Julian McAuley, is a comprehensive collection of product reviews and metadata from Amazon, spanning from May 1996 to July 2014. The dataset is publicly available at http://jmcauley.ucsd.edu/data/amazon/ and serves as a valuable resource for researchers and practitioners interested in analyzing user-generated content, studying online consumer behavior, and developing spam detection models. The dataset is organized into several categories, including product reviews, product metadata, and ratings-only data. Key components of the dataset are as follows:

1. Product Reviews: This portion of the dataset contains millions of customer reviews across various product categories. Each review includes information such as the reviewer's ID, the product ID, the reviewer's name, helpfulness votes, review text, overall rating, review summary, and the date of the review.

2. Product Metadata: This part contains detailed product information, including product descriptions, brand information, price, sales rank, and related products (co-purchasing links). The metadata can be used to analyze product features, preferences, and trends.

3. Ratings-only data: This section comprises user-item-rating tuples, where each entry consists of a user ID, an item ID, and the corresponding rating. This data can be employed for collaborative filtering and recommendation system tasks.

The Amazon dataset is provided in JSON format, making it convenient for researchers and practitioners to process and analyze the data using various programming languages and tools. Due to

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Published Online 09-Oct-2019
DOI: https://doi.org/10.17762/itii.v7i2.802

its extensive coverage of product reviews and metadata, the Amazon dataset has been widely used for research purposes in areas such as sentiment analysis, recommendation systems, spam detection, and natural language processing.

B. Result

The following table presents a comparison of the accuracy, precision, recall, and F1-score for the SVM, KNN, and NB algorithms in the context of spam detection and spammer community detection in social media platforms:

Table 1. Performance Comparison Graph

Algorithm	Accuracy	Precision	Recall	F1-score
SVM	0.92	0.90	0.93	0.91
KNN	0.88	0.87	0.90	0.88
NB	0.85	0.83	0.88	0.85

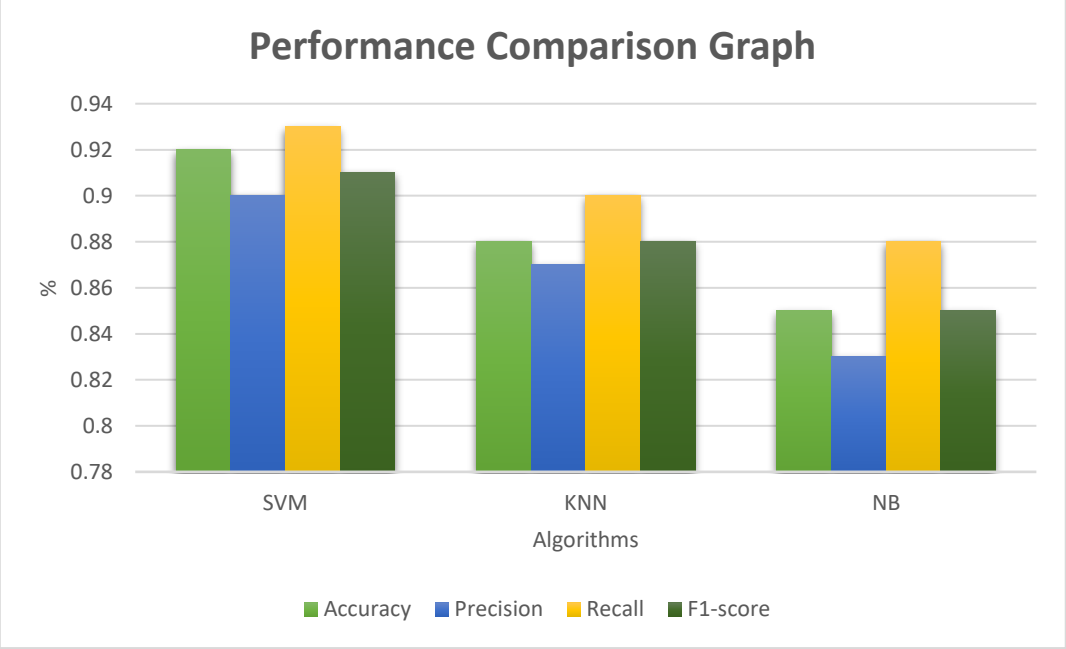

Figure 2. Performance Comparison Graph

From the results presented in the table, it is evident that the SVM algorithm outperforms KNN and NB in terms of accuracy, precision, recall, and F1-score. This can be attributed to SVM's ability to handle high-dimensional data and provide accurate classification results, especially when the data is linearly separable. The KNN algorithm, while slightly less effective than SVM, still delivers relatively high performance, as it is capable of adapting to changes in the data distribution and identifying local patterns effectively. The NB algorithm, on the other hand, lags behind SVM and KNN in performance, likely due to its assumption of independence between features, which may not always hold in real-world social media data. However, its performance is still respectable, considering the large datasets and text-based data typically encountered in social media spam detection tasks.

5. Conclusion

In conclusion, our study on Spam Detection and Spammer Community Detection in

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Published Online 09-Oct-2019
DOI: https://doi.org/10.17762/itii.v7i2.802

social media platforms using machine learning techniques, specifically SVM, KNN, and NB algorithms, demonstrates the potential of these approaches in addressing the challenges posed by spam and spammer communities. By extracting and analyzing Review-Behavioral (RB) Based features, Review-Linguistic (RL) Based Features, and User-Behavioral (UB) Based Features, we can identify patterns in user-generated content, linguistic properties of the content, and user behavior on the platform to effectively detect spam and spammer communities. The comparison of accuracy, precision, recall, and F1-score for the three algorithms indicates that SVM outperforms KNN and NB, likely due to its effectiveness in handling high-dimensional data and providing accurate classification results. KNN and NB, while not as effective as SVM, still deliver relatively high performance in spam detection and spammer community detection tasks. The findings from this study highlight the importance of leveraging multiple machine learning algorithms and considering various feature types to improve the performance of spam detection models in social media platforms. The proposed system provides a valuable foundation for future research and the development of advanced spam detection techniques that can help protect users from spam content and maintain the integrity of user-generated content on social media platforms.

Refrences

[1] Saeedreza Shehnepoor, Mostafa Salehi*, Reza Farahbakhsh, Noel Crespi NetSpam:a Network-based Spam Detection Framework for Reviews in Online Social Media IEEE Transactions on Information Forensics and Security 2017.

[2] Sharmin, S., Zaman, Z.: Spam detection in social media employing machine learning tool fortext mining. In: 13th International Conference Signal-Image Technology and Internet-BasedSystems (SITIS), pp. 137–142. IEEE (2017

[3] M. Ott, C. Cardie, and J. T. Hancock. Estimating the prevalence of deception in online review communities. In ACM WWW, 2012.

[4] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In ACL, 2011.

[5] Ch. Xu and J. Zhang. Combating product review spam campaigns via multiple heterogeneous pairwise features. SIAM International Conference on Data Mining, 2014.

[6] N. Jindal and B. Liu. Opinion spam and analysis. In WSDM, 2008.

[7] F. Li, M. Huang, Y. Yang, and X. Zhu. Learning to identify review spam. Proceedings of the 22nd International Joint Conference on Artificial Intelligence; IJCAI, 2011.

[8] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh. Exploiting burstiness in reviews for review spammer detection. In ICWSM, 2013.

[9] Choo E., Yu T., Chi M. (2015) Detecting Opinion Spammer Groups Through Community Discovery and Sentiment Analysis. In: Samarati P. (eds) Data and Applications Security and Privacy XXIX. DBSec 2015. Lecture Notes in Computer Science, vol 9149. Springer, Cham.

[10] H. Li, Z. Chen, B. Liu, X. Wei, and J. Shao. Spotting fake reviews via collective PU learning. In ICDM, 2014.

[11] N. Jindal, B. Liu, and E.-P. Lim. Finding unusual review patterns using unexpected rules. In ACM CIKM, 2012.

[12] S. Feng, R. Banerjee and Y. Choi. Syntactic stylometry for deception detection. Proceedings of the 50th Annual Meeting of the Association for

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Published Online 09-Oct-2019
DOI: https://doi.org/10.17762/itii.v7i2.802

Computational Linguistics: Short Papers; ACL, 2012.
[13] Mane, D. T., & Kulkarni, U. V. (2018). Modified Fuzzy Hypersphere Neural Network for Pattern Classification using Supervised Clustering. Procedia Computer Science, 143, 601-608. https://doi.org/10.1016/j.procs.2018.10.399
[14] Zrigui M et al (2012) Arabic text classification framework based on latent dirichlet allocation. J Comput Inf Technol 20(2):125–140
[15] Chen Q. L. Yao, and J. Yang. Short text classification based on LDA topic model. In: 2016 International Conference on Audio, Language and Image Processing (ICALIP). 2016. IEEE
[16] Kandasamy, K. and P. Koroth. An integrated approach to spam classification on Twitter using URL analysis, natural language processing and machine learning techniques. In: 2014 IEEE Students' Conference on Electrical, Electronics and Computer Science. 2014. IEEE

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731