Journal of Physics: Conference Series

An Improved Feature Selection Method for Short Text Classification

An Improved Feature Selection Method for Short Text Classification

Olusola Abayomi-Alli[1], Sanjay Misra[1,2], Victor O Matthews[1], Modupe Odusami[1], Adebayo Abayomi-Alli[3], Ravin Ahuja[4] and Rytis Maskeliunas[5]

[1]Department of Electrical and Information Engineering, Covenant University, Nigeria
[2]Department of Computer Engineering, Atilim University, Turkey
[3]Department of Computer Science, Federal University of Agriculture Abeokuta, Nigeria, [4]University of Delhi, Delhi, India, [5]Kaunas University of Technology, Kaunas, Lithuania

{olusola.abayomi-alli;sanjay.misra;moduupe.odusami} @covenantuniversity.edu.ng; abayomiallia@funaab.edu.ng; ravinahujadce@gmail.com; rytis.maskeliunas@ktu.lt

Abstract. Text has become one of the widest means of communication on mobile devices due to cheap rate and convenience for instance short text, web document, emails, instant messages. The exponential growth of text documents shared among users globally has increased the threat of misclassification associated with mobile devices such as Spam, Phishing, License to kill, Malware and privacy issues. Existing studies have shown that the major problem associated with text message classification is the poor representation of feature thus reducing accuracy and increasing f-measure rate. Thus, a modified Genetic Algorithm (GA) for improve feature selection and Artificial Immune System (AIS) algorithm was proposed for effective text classification in mobile short messages. The system will be deployed on an Android OS.

1. Introduction

New generation mobile devices are smart media which add notable value to individual and corporate thereby increasing productivity [1]. The positive impact of mobile devices to users have made information literally available through digitized text such as SMS, emails, web pages, MMS, Instant messages, online advertisement, social media, etc. [2]. On this note, classifying text is a major concern, as lots of activities are migrating from conventional computer based-platform to the mobile device-platform [3]. Text classification (TC) is the act of assigning text documents to a predefined category [4] as depicted in Figure 1while Feature Selection (FS) is the most vital and crucial techniques in data pre-processing for text classification systems.

FS technique has become an essential element in machine learning process [5][6] with its main focus on minimizing datasets spatiality through choosing the most distinguished features thereby improving classifier performances [7]. The significant drawback of TC include: high spatiality of the feature space, increased computational cost, low accuracy, high F-score measure and performance degradation of existing short message spam filters [8][9]. Hence, improving feature selection method helps to choose right input features and removal of less predictive ones, thereby achieving optimal dimensionality reduction, improving learning result, efficiency and performance [10].

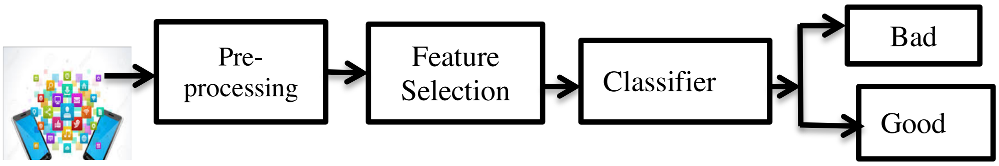

Figure 1: A typical text classification process

Despite research efforts, there is still no definite solution to improve feature selection techniques in short messages hence, degrading the performance of existing text classification methods [11].The motivation of this study is based on the noisy and irrelevant features associated with text, less effective result of existing short text classifier [12], high computational complexity of existing classifier [13, 14] thus making text classification an active research area. Therefore, there is an urgent need for an improved feature selection method to enhance effectiveness and efficiency of short text spam classification system. The rest of the paper is as follows: Section 2, detailed literature review. Section 3 gave a broad description of the proposed methodology and Section 4 explained the proposed system implementation. The paper concludes with Section 5.

2. Literature Review

Authors [7] presented an improved feature section using normalized difference measure for text classification. The performance of the proposed metric based on multinomial Naïve Bayes and SVM was compared with seven established metrics on seven datasets. Authors [15] proposed an evolutionary instance selection for text classification using biological-based genetic algorithm (BGA) while [16] presented a text classification filter for domain-specific search engines. The latter shows a reduction in quantity of annotated training data using cost-efficient filters while the former result shows that BGA outperforms five other classifiers based on largest dataset reduction rate and least computational time. Authors [4] improved text classification using semi-supervised clustering approach. The method uses labeled text for clustering and unlabeled text was used to adapt to centroids.

Authors [17] presented an effective integrated learning framework for summarizing and categorizing text using pseudo-relevance feedback method while [2] proposed a term-based discrimination information space for text classification. The former adopts a binary independent model using query weighting method and category-based smoothing method for solving sparsity in data issues. [18] proposed a method to reduce dimensionality in text representation based on clustering document using Hidden Markov Model (HMM). The proposed method was applied to kNN and SVM classifiers and it outperforms method based on information gain. [19] compared different term weighing scheme ranging from Term Frequency/ Inverse Document Frequency (TF-IDF) to Term Frequency/ Inverse Gravity Moment (TF-IGM). Authors concluded that their scheme performed better when comparing with the popular TF-IDF using SVM and kNN classifiers.

Authors [20] proposed a modified frequency-based term weighting schemes on SVM and kNN classifiers while authors [21] exploited the efficiency and effectiveness of NB solutions. The latter proposed four lazy semi-NB approaches to overcome the problems of over-fitting. Authors [22] presented an efficient text classification scheme for clustering using similarity measure for text processing (SMTP). The result gave a better accuracy when compared with other similarity measure such as Euclidean distance, cosine and dice coefficient. While [23] proposed a text classification method based on self-training and LDA (ST LDA) in a semi-supervised manner. Their result shows that combining the proposed method with NBMN gave a better accuracy and it can help in text classification with small set of labeled instance.

3. Proposed Design Methodology

This study proposed an evolutionary system based on modified genetic algorithm and Artificial Immune System for improved feature selection model for short text classification thus developing a lightweight obfuscation resilient system. The modified genetic algorithm is designed for creating word-clusters model which will enhance an effective text classification as against the conventional Term Frequency (TF), Term Frequency/ Inverse Document Frequency (TF-IDF), Information Gain (IG), etc.

3.1. Data Collection

Based on existing recommendation on the need to build a more robust short text database to aid further future research in this area, a new dataset was collected for experimentation purposes consisting of 42,020 text messages. For further validation of our work, we will be using UCI corpus, NUS corpus, Reuters-21578, spam email dataset, 20 newsgroup and YouTube spam comment databases.

Table 1. Dataset for validation of the proposed method

S/N	Corpus	Size (text)
1	UCI (University of California, Irvine)	5,574
2	SMS spam corpus v.0.1 (big) (NUS)	1,324
3	Reuters-21578	21,578
4	Spam email dataset	4,601
5	20 newsgroup	20,000
6	YouTube Spam comment	1,956

3.2. Proposed Methodology

The proposed study is divided into four major modules which include the following and the flow diagram of the proposed system is shown in Figure 2.

1. Innate immunity module
2. Preprocessing module
3. Feature selection module
4. Classification module using Artificial Immune System

3.2.1. Innate Immunity Module. This is the preliminary classification phase of the system and it utilizes the whitelist and blacklist techniques to determine whether a sender's number is legit or not. Thus classifying the text into the appropriate category.

3.2.2. Preprocessing Module. The pre-processing module involves the tokenization, removal of stop word and using porter stemming algorithm.

3.2.3. Feature Selection Module. This study integrates a modified genetic algorithm for improving feature selection (mGA_FS). Considering constant obfuscation of keywords and the unstandardized acronyms associated with text messages, there is a need for developing a model that will understand the relationship between key features. The modified Genetic Algorithm adopts the heuristic nature of traditional GA based on natural selection from the population members, and tries to find high-quality solutions to large and complex optimization problems [15]. It aim is to improve feature selection in generating best population clusters. The mathematical expression for proposed GA is depicted in Figure 3.

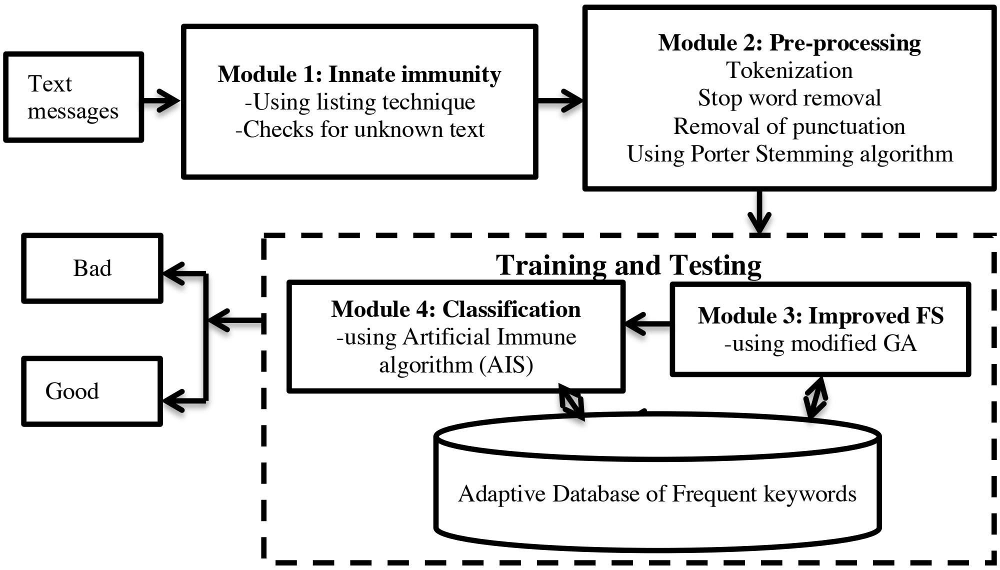

Figure 2. Flow Diagram of the Proposed System

F ϵ {f$_1$, f$_2$, …, f$_n$} //the set of features/ keywords (Chromosomes)
D ϵ {d$_1$, d$_2$, …, d$_n$} // the set of Documents
N = the number of documents

$$idf_k = \log((N - df_k)/df_k) \quad (1)$$

$$\text{Fitness} = \frac{1}{(1+f_k)} \quad (2)$$

$$maxF = \sum_{i=1}^{|F|} \sum_{k=1}^{|E|} (tf_{x_i k} \times idf_k) \quad (3)$$

Where:

$x_i \in F_i$
tf_{ik}: //term frequency of feature $F_i \in F$ in document $D_k \in D$
df_k: //document frequency is the number included $F_k \in F$
idf_k: //inverse document frequency of feature $F_k \in F$ in document $D_k \in D$

Generate Parents:

$$B(S) = (\mu + \lambda) \quad (4)$$

Where:

B(S) = Parent population;
μ = Size of the population
λ = Best individuals

Figure 3. A modified Genetic Algorithm (mGA_FS)

3.2.4. The classification module is based on Artificial Immune System Algorithm (AIS). This module adopts the AIS, a biological theory for observing immune principles, models and functions. It is based on three frameworks which include the vector representation, clonal selection, and affinity measure. The mathematical expression of the three layers within the Artificial Immune System is represented in Figure 4.

1. Vectors Representation

$$P = \{p_1, p_2, ..., p_L\}$$
$$Q = \{q_1, q_2, ..., q_L\}$$

Where:

P = Antibodies
Q = Antigens (Tokens)

2. Clonal Selection Algorithm:

$$N_c = \sum_{i=1}^{n} round(\frac{\beta.N}{i}) \qquad (5)$$

Where:

N_c: is the total number of clones generated for each of Q's
β: Multiplying factor
N: Total number of P
round (): operator that rounds its argument toward
the closest integer.

3. Affinity Measures: This is the related distance between an Antigen (Q) and the corresponding Antibodies (P).

Using Euclidean $$ED = \sqrt{\sum_{i=1}^{L} (P_i - Q_i)^2} \qquad (6)$$

Figure 4. Mathematical expression Artificial Immune System framework

4. System Implementation

This study will be conducted on a personal computer AMD (A6-5200 APU) Quad core, 4.00GB memory (RAM) with 750GB local hard drive running on Microsoft Windows 8 Professional 64-bit. The evaluation will be based on the following metrics: Confusion matrix accuracy based on Specificity, Recall, macroF1, Precision, Error rate, computational time and Accuracy (ACC).

4.1. Software Implementation Requirement

For this study, the software implementation will be done using the following tools, Microsoft office excel 2010, notepad win32, MATLAB Text Analytics toolbox, C++ library, Android version 4.4(KitKat) and validation on Waikato Environment for Knowledge Analysis (WEKA).

4.2. Expected Contribution to Knowledge

The proposed system is expected to contribute to improving feature selection methods for dimensionality reduction in short text classification such as SMS, email, WhatsApp, Microblog, tweet, YouTube comments, etc. In addition, a lightweight Obfuscation-Resilient text message classification system for Android Operating system will be developed.

5. Conclusion

Improving feature selection in text message classification helps to reduce spatiality of datasets thus playing a major role in the effective anomaly detection in text. Several approaches have been applied to classify test messages and reduce spam but the overall accuracy of existing solutions have proven to be computationally expensive for mobile phones as compared to the solution on email platforms. The high spatiality of the Short Message feature space is still a challenging factor for achieving an accurate result thus this research study aims to improve the feature selection method in order to maximize the limited bag of words for optimal result and in turn develop a robust computationally efficient, obfuscation free text classification system.

References

[1] Tully S. and Mohanraj Y. (2017). Mobile Security: A Practitioner's Perspective. Mobile Security and Privacy Advances, Challenges and Future research. Elsevier. Syngress.

[2] Junejo, K. N., Asim K., Malik T. H., and Moongu J. "Terms-based discriminative information space for robust text classification." Information Sciences 372 (2016): 518-538.

[3] Lau L. (2017). Mobile Security: End Users are the Weakest Link in the System. Mobile

Security and Privacy: Advances, Challenges and Future Research Directions. Syngress publications. Elsevier: 56- 66.

[4] Zhang, W., Xijin T. & Taketoshi Y. "Tesc: An approach to text classification using semi-supervised clustering." Knowledge-Based Systems 75 (2015): 152-160.

[5] Bidi N. and Elberrichi Z. (2016). Feature Selection For Text Classification Using Genetic Algorithms. 8th International Conference on Modeling, Identification and Control (ICMIC-2016), Algiers, Algeria. IEEE.

[6] Lu, Y., & Chen, Y. (2017). A Text Feature Selection Method Based on the Small World Algorithm. Procedia Computer Science, 107: 276-284.

[7] Rehman, A., Kashif J., and Haroon A. B. "Feature selection based on a normalized difference measure for text classification." Information Processing & Management 53, no. 2 (2017): 473-489.

[8] Zorarpacı, E., & Özel, S. A. (2016). A hybrid approach of differential evolution and artificial bee colony for feature selection. Expert Systems with Applications, 62: 91-103.

[9] Mahajan A. and Shourya Roy S. (2015). Feature Selection for Short Text Classification usingWavelet Packet Transform. Proceedings of the 19th Conference on Computational Language Learning: 321–326, Beijing, China.

[10] ElAlami M. E. (2009). A Filter Model for Feature Subset Selection based on Genetic Algorithm. Knowledge-based Systems, 22 (2009): 356–362. Elsevier.

[11] Fermandes (2015). SMS Spam Filtering Through Optimum-path Forest-based Classifiers. 2015 IEEE 14th International Conference on Machine Learning and Applications:133-137.

[12] Al-Hasan, A.A. and El-Alfy, E.S.M. (2015). Dendritic cell algorithm for mobile phone spam filtering. Procedia Computer Science, 52: 244-251.

[13] Su, Y., Huang, Y. and Kuo, C.C.J., 2018. Efficient Text Classification Using Tree-structured Multi-linear Principle Component Analysis. eprint arXiv:1801.06607.

[14] Onashoga, A. S., Abayomi-Alli, O. O., Sodiya, A. S. and Ojo, D. A. "An Adaptive and Collaborative Server-Side SMS Spam Filtering Scheme Using AIS." Information Security Journal: A Global Perspective 24(4-6), (2015): 133-145.

[15] Tsai, C.-F., Zong-Yao C., and Shih-Wen K. "Evolutionary instance selection for text classification." Journal of Systems & Software 90 (2014): 104-113.

[16] Schmidt, S., Steffen S., and Christoph R. "Text classification based filters for a domain-specific search engine." Computers in Industry 78 (2016): 70-79.

[17] Jeong, H., Youngjoong K., and Jungyun S. "How to Improve Text Summarization and Classification by Mutual Cooperation on an Integrated Framework." Expert Systems with Applications 60 (2016): 222-233.

[18] Vieira, A. Seara, L. Borrajo, and Eva Lorenzo Iglesias. "Improving the text classification using clustering and a novel HMM to reduce the dimensionality." Computer methods and programs in biomedicine 136 (2016): 119-130.

[19] Chen, K., Zuping Z., Jun L., and Hao Z. "Turning from TF-IDF to TF-IGM for term weighting in text classification." Expert Systems with Applications 66 (2016): 245-260.

[20] Sabbah, T., Ali S., Md H. S., Fawaz S. A.-A., Enrique H. V., Ondrej K., and Hamido F. "Modified frequency-based term weighting schemes for text classification." Applied Soft Computing 58 (2017): 193-206.

[21] Viegas, F., Leonardo R., Elaine R., Thiago S., Wellington M., Mateus F. F., and Marcos A. G. "Exploiting efficient and effective lazy Semi-Bayesian strategies for text classification." Neurocomputing (2018).

[22] Thomas, Anisha Mariam, and M. G. Resmipriya. "An efficient text classification scheme using clustering." Procedia Technology 24 (2016): 1220-1225.

[23] Pavlinek, Miha, and Vili Podgorelec. "Text classification method based on self-training and LDA topic models." Expert Systems with Applications 80 (2017): 83-93.