*FUDMA Journal of Sciences (FJS)*
*ISSN online: 2616-1370*
*ISSN print: 2645 - 2944*
*Vol. 3 No. 2, June, 2019,* pp 195- *200*

# A FILTER MODEL FOR TEXT CATEGORIZATION AGAINST ONLINE HATE SPEECHES

**[1]Obunadike Georgina N., [2]Emeka Ogbuju &[3]Mukhtar Abubakar**

[1,2]Department of Computer Science and Information Technology, Federal University Dutsin-Ma, Katsina State, Nigeria
[3]Department of Computer Science, Federal University Lokoja, Kogi State, Nigeria

Corresponding Author's Email: gobunadike@fudutsinma.edu.ng

**ABSTRACT**
Text classification is a method of grouping a document text into different predefined categories. This method has been applied in different areas such as classification of scientific articles, spam filtering, and classification of document genre. Text classification is a popular task in data mining because of its level of accuracy and easy application. The Internet is a common message transmission medium among many people, billions of messages move around the internet on a daily basis through different platforms on the internet such as e-mail, Facebook, Twitter, etc. Some of these messages are being transmitted with wrong motives, thus it became imperative to design a model for filtering some of these messages using data mining algorithms to sieve away the unwanted messages from circulation. In the light of this, this paper applied three data mining techniques namely: Support Vector Machine (SVM), Naïve Bayes and K-Nearest Neighbour (KNN) to develop models that can be applied to filter messages from Facebook and e-mail to counter circulation of online hate speeches on these platforms. It also compared the performance of these models against collected data to identify the state of the art text classifier. It was observed that the Naïve Bayes algorithm performed better than the other two with an accuracy of 61.5 and ROC of 0.66.

**Keywords**: Text categorization, Hate Speech, Classification Techniques, Data mining

## INTRODUCTION

Hate speech is a word that defames person or group of people based on their nationality, race or religion. It takes different forms such as pictures, songs, drama as well as speech (Bonnell, 1997; David, 1995; Whillock, 1995; Bakewell, 1998). Indulging in hate speech cannot be claimed to be freedom of speech (Bakewell, 1998).Literature has revealed several cases where hate speeches have resulted in disastrous and deadly situations. These include the Rwanda genocide in 1994 where the massacre of over 800,000 people took place. Also during the presidential election in Kenya in 2007, violence broke up between the three major ethnic groups and more than 1,100 people lost their lives (Bakewell, 1998).Text categorization is the task of classifying text documents (Drucker *et al*, 1999 and Dumais *et al* 2000). A number of algorithms have been applied in this area in recent times; these algorithms include support Vector Machines, Decision Trees, Neural Network, Bayesian Classifiers, and K-Nearest Neighbour. With increase in available methods or techniques for text classification, it has become increasingly important to evaluate the performance of these algorithms to identify the state of the art text classifier. The rest of the sections discussed major classification algorithms, related works of literature, the methodology applied, evaluation metrics employed and the results discussion.

### Classification Techniques

A. **Support Vector Machine (SVM)**: SVM is a machine learning algorithm implemented by Joachims (1998, 1999) and has been applied widely in text classification since that time (Drucker *et al.*, 1999 and Dumais *et al.*, 2000). In mathematical term it is a method of finding among variables *v1, v2…* in N-dimensional space, the variable *vi* that separates the positive from negatives using the widest margins; that is to say that the minimum distance between the hyper plane and dataset is maximum. Thus this method tends to minimize the generalizing error that is the error that was generated by the classifier. Though SVM is mostly suited for binary classification problems, it has been recently applied in solving multiclass problems (Crammer and Singer, 2001). One major advantage of SVM in text classification application is the fact that dimensionality reduction is in most cases not needed. SVM is usually robust to over fitting thus can handle multidimensional data very well (Joachims 1998). Literature also reveals that feature selection is detrimental to SVM performance; thus the use of SVM for classifying high dimensional problem such as text classification is no longer an issue in terms of computational cost (Joachims 1998).

B. **K-Nearest Neighbour Classifier:** Fix and Hodges (1951) was the first to introduce the Nearest Neighbour algorithm and since then it has been used in classification and regression problems of different forms. Though regarded as a lazy classifier; it has proven to be an effective algorithm that competes favorably with other leading classifiers (Obunadike*et al.*, 2018). It has been applied successfully in different areas such as recommendation system, document classification,

pattern recognition, computer vision and it is a supervised learning algorithm (Obunadike*et al.*, 2018). KNN gets its name from the fact that it uses ideasfrom an item *k*-nearest neighbour to classify unlabeled item (Kataria and Singh, 2013). The letter *k* is variable which means that any number of nearest neighbours can be used. KNN identifies *k*-values in training data that are closest in similarity and assigned to the unlabeled data the class of the closest neighbor (Kubat and Jr, 2000). It uses distance measures to measure similarities between two data items. The common distance measures are Euclidean distance, Manhattan, simple matching, squared Euclidean distance and Minkowski (Parvin*et al.*, 2010).

C. **Naïve Bayes Classifier:** Naïve Bayes algorithm is one of the machine learning algorithms that are statistical in nature. It was developed by Thomas Bayes and it applies Bayes theorem (Obunadike *et al*, 2018). It competes effectively with other popular classification algorithms; it is quite effective for classification of categorical data. It usually gives an impressive performance in classification problems though usually being criticized because of its attribute independent assumption. It is a popular text classification algorithm because of its simplicity and ease of use (Obunadike *et al*, 2018). The classifier performs its classification using equation 2

$$P(doc|c_j) = \prod_{i=1}^{len(doc)} P(a_i = wd_k|c_j) \quad \text{Eqn (1)}$$

where$P(doc|c_j)$ is the probability of a document belonging to a certain class and $P(a_i = wd_k|c_j)$ is the probability that word in position $j$ is $wd_k$ given $c_j$. It also assumes that

$$P(a_i = wd_k|c_j) = P(a_m = wd_k|c_j)\ \forall\ i, m \quad \text{Eqn (2)}$$

**PREVIOUS WORKS**

Due to recent development in technology, cyber space has been the major platform for communication and businesses. It hosts several social networks such as WhatsApp, email, Facebook, Twitter even business platforms. The uncontrolled nature of this platform has opened door for misuse to malicious users. So many works have been done in literature to proffer solutions to this problem. Some studies have been carried out to detect malicious activities, the general characteristics of such activities and proffer some technical and effective techniques to guard against such acts Chew *et al* (2018). Cao *et al* (2008) are of the opinion that software base system are preferred as decision support system for users and that blacklisting is an effective technique. Gupta *et al* (2018) in their work stated that when new solutions are proposed to overcome malicious use, offenders usually come up with a way of overcoming the new solution. They suggested the need to apply effective techniques in combating such crimes. Zang *et al* (2007) developed text based technique called CANTINA which extracts keywords using the term frequency-inverse document frequency algorithm. The keywords were used to search out malicious activities. Buber *et al* (2017) proposed a detection system with 209 word vector features and 17 NLP base features. They suggested the need to increase the number of NLP base features and word vectors. This was handled in their later work and they got better result with higher accuracy value. Du*et al* (2013) proposed a text classification algorithm that determined the pages of a site that must be blocked. Theirs was a web filtering system that uses text classification approach to classify web pages into desirable and undesirable ones. Though not in the domain of malicious activities, the work of Yindalon *et al* (2005) was on text categorization models for high-quality article retrieval in internal medicine. The work showed the possibility of using machine learning to automatically build models for retrieving high-quality, content-specific articles in a given time period in internal medicine which perform better than the 1994 PubMed clinical query filters. One of the popular methods of checking malicious users and activities is the use of machine learning algorithms by using simple classification techniques. With the use of machine learning algorithm it's usually easy to detect malicious activities. This paper is focused on evaluating the performance of some selected machine learning algorithms usually applied to checkmate malicious activities.

**METHODOLOGY**

Text classification is formalized using the function $F: DXC \rightarrow \{T, F\}$, where $C = \{c_1, c_2 \ldots c_n\}$ is a predefined set of classes. D is a set of documents. If $f(d_j, c_i) = T$, then $d_j$ is called a positive document or a member of $c_i$while $f(d_j, c_i) = F$ is called a negative example of$c_i$. The classes are just symbolic labels, no additional knowledge of their meaning is usually available thus metadata are not usually available. The classification was done based on the knowledge extracted from the documents. Document indexing implies the act of mapping a document $d_j$ into a compact representation of its content that can be easily read by a classifier. The text $d_j$ is represented as a vector $\overrightarrow{d_j} = \{w_{ij} \ldots w|T|\}$ where T is the dictionary, that is a set of features or terms that occur at least once in at least K documents and $0 \leq w_{kj} \leq 1$. The words occurred in the documents were identified with exception of the stop words. The document was stemmed to obtain the morphological roots of the terms or features found in the document. Dimensionality reduction was applied to reduce the size of features or terms that occurred in the word document. This is to avoid the classifier performing better on trained data than on new datasets and to make the problem more manageable for classifiers since many classifiers are known not to scale well on problems with high dimensionality. For the text classification, the document was divided into three namely: a training set, validation set, and test set. The training set was the set of documents that were used to train the classifiers. The validation set was used to fine-tune the classification

models. The test set was used to evaluate the effectiveness of the classifiers. Three classifiers were applied for the text classification they include Support Vector Machine, Naïve Bayes and K- Nearest Neighbour. Classifiers evaluation was based on three metrics precision, recall, and accuracy. The methodology for this work is as represented in Figure 1.

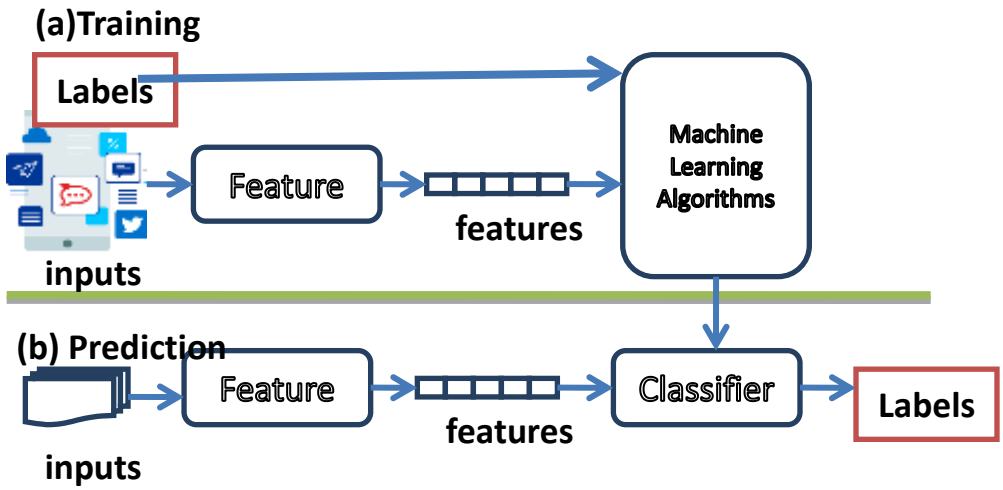


Fig. 1: Work Methodology

**SAMPLE DOCUMENT AND THEIR CLASSES**

The data source for this work is the internet. In text categorization using the arff file, the arff file has two sections: the header and data section. The first attribute of the header section represents the entire document as a single text attribute of type string. The second attribute is the class attribute and will define the class each document belongs. An example of the resulting document arff file use for this work is represented in Listing 1

```
@Relation HateSpch

@Attribute Document String
@Attribute Class {yes, no}

@Data
"Nigerian politicians are liars", yes
"Fulanis are murderers", yes
"Buhari is the sponsor of Boko Haram", yes
"Solution to Nigeria problems is restructuring", no
"Igbo should be allowed to have their own country", no
"Igbo people are thieves", yes
"Nigeria is the giant of Africa", no
"Hausas are lazy", yes
"Killer herbs men are fulanis", yes
"Buhari is fighting corruption", no
"United Nations have accepted Biafra secession", no
"Igbos should leave the northern state or they will be killed", yes
"Igbos will not go with their properties", yes
"Igbo are hardworking people", no
"Corruption is fighting back", no
"Igbos can never be president", yes
"Jonathan government is the most corrupt government in Nigeria", yes
"Hausas are the problem of this nation", yes
"Nigeria is a blessed country", no
```

Listing 1: A Sample Arff file of the Text

Hate statements can be more than this but for this research, these are the text used for the text classification.

**Measure of Effectiveness**

A number of different measures of effectiveness can be used in the evaluation of classifiers applied in text classification. A contingency table is simple and widely used effectiveness measures. The contingency table is shown in Table 1.

**Table 1: Contingency Table**

| | **Positive** | **Negative** | |
|---|---|---|---|
| **Positive** | TP | FP | TP + FP Total with positive |
| **Negative** | FN | TN | FN + TN Total with negative |
| | TP + FN Total with actual positive | FP + TN Total with actual negative | |

From the contingency table three important measure of classifier effectiveness are considered:

$$Recall = {}^{TP}/_{TP + FN}$$

$$Precision = {}^{FP}/_{TP + FP}$$

$$Recall = {}^{TN + TP}/_{TN + TP + FN + FP}$$

- Precision is the measure of exactness of the relevant data retrieved.
- The recall is the measure of completeness. That is the percentage of all relevant data that is returned by the model.
- High Recall means that the model returns most of the relevant data.
- High Precision means the model returns more relevant results than irrelevant.
- Receiver Operating Characteristics (ROC) curve is a useful tool for comparing two classification Models

**RESULTS**

Tables 2 and Table 3 display the classification result obtained using the training set and test set respectively for the algorithms. Table 4 to Table 6 shows the individual result of the text classification on the test data set.

**Table 2: Trained Classifier Result using Training Set**

| Metrics | Naïve Bayes | SVM | KNN |
|---|---|---|---|
| Precision | 1 | 1 | 1 |
| Recall | 1 | 1 | 1 |
| Accuracy | 100 | 100 | 100 |
| Time | 0.0 secs | 0.1 secs | 0.0 secs |
| ROC | 1 | 1 | 1 |
| Kappa | 1 | 1 | 1 |

From Table 2, it is observed that the three algorithms performed well on the trained data set.

**Table 3: Revaluation Result on Trained Classifiers using Test Set**

| Metrics | Naïve Bayes | SVM | KNN |
|---|---|---|---|
| Precision | 0.8 | 0.6 | 0.8 |
| Recall | 0.5 | 0.5 | 0.5 |
| Accuracy | 61.5 | 53.8 | 61.5 |
| ROC | 0.66 | 0.50 | 0.56 |
| Kappa | 0.3 | 0.4 | 0.3 |

Table 3 shows that Naïve Bayes algorithms perform better than the other two algorithms on the test data set with higher accuracy and ROC curve of 61.5 and 0.66 respectively.

**Table 4: Result of Re-evaluation of Train SVM on Test Set**

| *inst#* | ***actual*** | ***predicted*** | ***error*** | ***Probdist*** |
|---|---|---|---|---|
| *1* | *1:yes* | *2:no* | *+ 0* | ** 1* |
| *2* | *1:yes* | *1:yes* | ** 1* | *0* |
| *3* | *1:yes* | *1:yes* | ** 1* | *0* |
| *4* | *1:yes* | *2:no* | *+ 0* | ** 1* |
| *5* | *1:yes* | *2:no* | *+ 0* | ** 1* |
| *6* | *2:no* | *2:no* | *0* | ** 1* |

| | | | | |
|---|---|---|---|---|
| 7 | 2:no | 2:no | 0 | *1 |
| 8 | 2:no | 1:yes | +*1 | 0 |
| 9 | 1:yes | 1:yes | *1 | 0 |
| 10 | 2:no | 2:no | 0 | *1 |
| 11 | 1:yes | 1:yes | *1 | 0 |
| 12 | 1:yes | 2:no | +0 | *1 |
| 13 | 2:no | 1:yes | +*1 | 0 |

**Table 5: Result of the Re-evaluation of Trained Naïve Bayes on Test Set**

| inst# | actual | Predicted | error | probdist |
|---|---|---|---|---|
| 1 | 1:yes | 2:no | +0.32 | *0.68 |
| 2 | 1:yes | 1:yes | *0.724 | 0.276 |
| 3 | 1:yes | 1:yes | *0.996 | 0.004 |
| 4 | 1:yes | 2:no | +0.035 | *0.965 |
| 5 | 1:yes | 2:no | +0 | *1 |
| 6 | 2:no | 2:no | 0.361 | *0.639 |
| 7 | 2:no | 2:no | 0.037 | *0.963 |
| 8 | 2:no | 1:yes | +*0.828 | 0.172 |
| 9 | 1:yes | 1:yes | *0.989 | 0.011 |
| 10 | 2:no | 2:no | 0.002 | *0.998 |
| 11 | 1:yes | 1:yes | *0.974 | 0.026 |
| 12 | 1:yes | 2:no | +0.361 | *0.639 |
| 13 | 2:no | 2:no | 0.042 | *0.958 |

**Table 6: Result of Re-evaluation of Trained KNN on Test Set**

| inst# | actual | predicted | Error | Probdist |
|---|---|---|---|---|
| 1 | 1:yes | 2:no | +0.009 | *0.991 |
| 2 | 1:yes | 1:yes | *0.979 | 0.021 |
| 3 | 1:yes | 1:yes | *0.96 | 0.04 |
| 4 | 1:yes | 2:no | +0.011 | *0.989 |
| 5 | 1:yes | 2:no | +0.04 | *0.96 |
| 6 | 2:no | 2:no | 0.011 | *0.989 |
| 7 | 2:no | 2:no | 0.04 | *0.96 |
| 8 | 2:no | 2:no | 0.205 | *0.795 |
| 9 | 1:yes | 1:yes | *0.96 | 0.04 |
| 10 | 2:no | 2:no | 0.021 | *0.979 |
| 11 | 1:yes | 1:yes | *0.979 | 0.021 |
| 12 | 1:yes | 2:no | +0.011 | *0.989 |
| 13 | 2:no | 1:yes | +*0.96 | 0.04 |

## DISCUSSION

The individual reevaluation of the trained models reveals that Naïve Bayes classified most of its data correctly compared to KNN and SVM as shown in Tables 4 to Table 6. It also performed comparably better than KNN and SVM with accuracy value of 61.5 as shown in Table 3. The ROC curve which is used to measure effectiveness of classification result also reveals that Naïve Bayes is more effective in its classification given a higher ROC curve value of 0.66 compared to KNN and SVM that gave 0.56 and 0.50 respectively as show also in Table 3.

## CONCLUSION

Text classifications that can make a meaningful distinction between classes of documents have been widely applied in data mining and machine learning. Machine learning algorithms are the most popular techniques usually for checking malicious activities in the internet. The recent growth in online messages and the transmission of hate speeches on social media platforms has called for the application of text classification and the study of various algorithms applied to determine the effectiveness of these algorithms. This paper evaluates three major algorithms for text categorization; it provides both theoretical and empirical evidence of the performance of the algorithms for text categorization. There are so many machine learning algorithms but this work is limited to evaluation of three popularly used machine learning algorithms for text classification. In deploying the model; the algorithm that have given the best performance will be applied and this will be in our next work. Future work includes adding other algorithms in the evaluation, development of hybridized model that would improve the filter model obtained by Naïve Bayes and deployment of the said model on Facebook and email.

**REFERENCES**


Bakewell, L. (1998). Image Acts.*American Anthropologist, **100***(1): 22-32.

Bonnell, V. E. (1997). Iconography of power: Soviet political posters under Lenin and Stalin. Berkeley and Los Angeles: University of California Press.

Buber, E. ,Diri, B. , &Sahingoz, O. K. (2017). Detecting phishing attacks from URL by using NLP techniques. In 2017 International conference on computer science and Engineering (UBMK) pp. 337–342.

Cao, Y., Han, W.,& Le, Y. (2008). Anti-phishing based on automated individual white-list. In Proceedings of the 4th ACM workshop on digital identity

Chiew, K. L. , Yong, K. S. C. , & Tan, C. L. (2018). A survey of phishing attacks: Their types, vectors and technical approaches. Expert Systems with Applications, 106 , 1–20 .

Crammer, K., &Singer, Y.(2001).On the algorithmic implementation of multiclass kernel-based Vector Machines.*Journal of Machine Learning Research, **2***: 265–292.

David, S.,&Whillock, R.K. (eds.).(1995). Hate Speech. Thousand Oaks, CA: Sage Publications, Inc. Introduction. pp. ix-xvi; "Symbolism and the Representation of Hate in Visual Discourse." pp. 122-141; "The Use of Hate as a Stratagem for Achieving Political and Social Goals." pp. 28-54; "Afterword: Hate, or Power?" pp. 267-275.

Drucker, H., Vapnik, V.,& Wu, D.(1999). Support Vector Machines for spam categorization. *IEEE Transactions on Neural Networks, **10***(5): 1048–1054.

Du, R., Safavi-Naini, R., & Susilon, W. (2013).Web filtering using text classification. The 11th IEEE International Conference on Networks, 28 September - 1 October 2003, 325-330.

Dumais, S.T., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. Proceedings of CIKM-98, 7th ACM International Conference on Information and Knowledge Management, ACM Press, New York, US: Bethesda, US, pp. 148–155

Dumais, S.T. & Chen, H., (2000). Hierarchical Classification of web content. Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval, ACM Press, New York, US: Athens, GR, pp. 256–263

Fix, E., & Hodges, J. (1951). Discriminatory analysis: Nonparametric discrimination. Consistency Properties, **4**

Gupta, B. B. ,Arachchilage, N. A. G. , &Psannis, K. E. (2018). Defending against phishing attacks: Taxonomy of methods, current issues and future directions. Telecommunication Systems, 67 (2), 247–267

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. Proceedings of ECML-98, 10th European Conference on Machine Learning.

Joachims, T.(1999).Transductive inference for text classification using Support Vector Machines. Proceedings of ICML-99, 16th International Conference onMachine Learning, Morgan Kaufmann Publishers, San Francisco, US: Bled, SL, pp. 200–209

Kataria, A., & Singh, M. D. (2013).A review of data classification using K-Nearest Neighbour Algorithm.*International Journal of Emerging Technology and Advanced Engineering, **3***(6): 354-360

Kubat, M., &Jr, M. (2000). Voting Nearest-Neighbour sub classifiers. Proceedings of the 17th International Conference on Machine Learning, ICML-2000, Stanford, CA, pp. 503-510.

Obunadike, G. N., Dima R., &Abah J. (2018).Empirical evaluation of KNN classifier using various K-Values.Proceedings of the International Conference on Information Technology in Education and Development (ITED, 2018), pp 13-18.

Parvin, H., Alizadeh, H., &Minaei, B. (2010).A modification on K-Nearest Neighbor classifier.*Global Journal of Computer Science and Technology,**10***(14): 37-41.

Obunadike, G., N., Isah, A., &Alhassan, J., K. (2018). Optimized Naïve Bayesian algorithm for efficient performance, *Journal of Computer Engineering and Intelligent System, **9***(3): 8-13.

Yindalon, A., Ioannis, T., Alexander, S., Douglas, H., Constantin, F. A. (2005). Text categorization models for high-quality article retrieval in internal medicine. Journal of the American Medical Informatics Association, 12(2), pp. 207–216. https://doi.org/10.1197/jamia.M1641

Zhang, Y., Hong, J. I., & Cranor, L. F. (2007). Cantina: A content-based approach to detecting phishing web sites. In Proceedings of the 16th International Conference on World Wide Web, WWW '07, ACM, New York, NY, USA (pp. 639–648)