WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Automatic Text Classification on Blogs Using Support Vector Machines (SVM)

Asogwa D.C[1], Efozia F.N[2],Chukwuneke C. I[3], Nnaekwe K.U[4]

Department of Computer Science, Faculty of Physical Sciences, NnamdiAzikiwe University, Awka. Anambra state, Nigeria.[1, 3, 4].
Prototype Engineering Development institute (PEDI), Ilesha, Osun State, Nigeria[2].
dc.asogwa@unizik.edu.ng[1], efozia23@yahoo.ca[2],ku.nnaekwe@unizik.edu.ng[3],ci.chukwuneke@unizik.edu.ng[4]

ABSTRACT

Automatic Text Classification is a machine learning task that automatically assigns a given text document to a set of pre-defined categories based on the features extracted from its textual content. Most online communication forums, including social media, enable users to express themselves freely, and most times, anonymously. The ability to freely express oneself is a human right that should be cherished, but people always induce and spread hate or illegal words towards another group as an abuse of this liberty. For instance many online forums such as Facebook, YouTube, and Twitter consider hate speech harmful, and have policies to remove hate speech content. This paper attempts to automatically classify the textual entries made by bloggers on various topics into hate speech and non-hate speech. This was achieved by following steps like pre-processing, feature extraction and support vector machine classification. Empirical evaluation of this binary classification has resulted in anaccuracy of approximately 83% over the test set. In addition to classifying the textual entries of the blogs, it is proposed that the extracted features themselves be further classified under more meaningful heads which results in generation of a semantic resource that lends greater understanding to the classification task. This semantic resource can be used for data mining requirements that arise in the future.

Keywords: machine learning, text classification, feature extraction, pre-processing, algorithm, supervised learning.

INTRODUCTION

Text classification is the process of assigning tags, classes or categories to text according to its content. It is the task of assigning a set of predefined categories to free-text. Text classification can be done in two different ways: manual and automatic classification. In manual classification, a human annotator interprets the content of text and categorizes it accordingly. This method usually can provide quality results but it is time-consuming and very expensive. Automatic text classification applies machine learning, natural language processing, and other techniques to automatically classify text in a faster and more cost-effective way.

Text classification with machine learning learns to make classifications based on past observations. By using pre-labeled examples as training data, a machine learning algorithm can learn the different associations between pieces of text and that a particular output (i.e. tags) is expected for a particular input (i.e. text). Support vector machine (SVM) is a supervised machine learning algorithm which is an effective technique for classifying high dimensional data. Support vector machines (SVM) are also ranked as one of the best off the shelf supervised learning algorithm as they provide superior generalization performance, requires less examples for training and can tackle high dimensional data with the help of kernels (YinglieTian et al,2012). SVM also tends to perform well for classification of text because of its ability to generalize into high dimensions, which is often the case with text categorization.

Unstructured text is a written content that lacks metadata and cannot readily be indexed or mapped onto standard database fields. It is often user generated information such as e mail or instant messages, social media postings, blogs and so on. Text can be an extremely rich source of information, but extracting insights from it can be very difficult and time-consuming due to its unstructured nature. Businesses are turning to text classification for structuring text in a fast and cost-efficient way to enhance decision-making and automate processes

The widespread of hate speech on the internet has given a strong motivation to study automatic detection of hate speech. By automating its detection, the spread of hateful content can be reduced.Online social networks (OSN) and micro-blogging websites are attracting internet users more than any other kind of websites.These websites offer an open space for people to discuss and share thoughts and opinions and as such, the nature and the huge number of posts, comments and messages exchanged make it almost impossible to control their content. Also given the different backgrounds, cultures and believes, many people tend to use hate speech when discussing with people who do not share the same backgrounds. Hate speech is a speech that is intended to insult, offend or intimidate a person because of some trait (as race, religion, sexual orientation, national origin and so on). Collecting and

analyzing these data allows decision makers to study the escalation of such comments/messages.

This work focused on developing a system that can detect hate speech in some selected blogsby using Support vector machines under supervised machine learning with text mining techniques. More efforts were placed in ensuring that the right data was used and also to get the labeling right for more accuracy in the results. According to Grondahl et al., 2018 after reproducing the state of the art in hate speech detection argued that model architecture is less important than the type of data and labeling criteria. These clearly indicate that efforts should not just be made only on the model used but also on the type of data and labeling criteria. Most machine learning techniques used are performance driven and show that roughly equal accuracy were achieved with different datasets provided that the training and testing are based on the same dataset. This clearly indicates lack of effective transferability of machine learning models across datasets.
Hence, automatic detection and filtering of such inappropriate language has become an important problem for improving the quality of conversations with users as well as virtual agents.

2.0 RELATED WORKS:

Yenala, et al (2017) proposed a novel deep learning-based technique for automatically identifying inappropriate languages in a text. They focused on solving the problem in two application scenarios: query completion suggestions in search engines and users conversations in messengers. They proposed a novel deep learning architecture called convolutional bi-directional LSTM (C-BiLSTM) which combines the strengths of both convolution neural network (CNN) and Bi-directional LSTMs (BLSTM).
Georgioset al (2018) addressed the important problem of discerning hateful content in social media. They proposed a detection scheme that is an ensemble of recurrent neural network (RNN) classifiers and it incorporates various features associated with user related information such as the user's tendency towards racism or sexism. The scheme can successfully distinguish racism and sexism messages from normal text and achieve higher classification quality.
Arthur (2018) discussed about the implementation of a simple anti-spam control based on the famous naïve Bayesian classification algorithm that can be actively used to locate and filter out those texts (E-mails, SMS, …) from a local messages database that most likely might contain spam or other unsolicited data.
Davidson T. et al, 2017 automatically classified tweets into hate speech, offensive and neither of the above. They compared among logistic regression, naïve Bayes, decision trees, random forest and SVM. They concluded that logistic regression gave the best prediction. They never combined any of the algorithms.
Hajime Watanabe et al, 2018, proposed a new method to detect hate speech in twitter and classified the tweets into hateful, offensive and clean. They algorithm used was J48graft.
Georgios K. Pitsilis et al, 2018 used deep learning algorithm to address the important problem discerning hateful contents in social media. They were able to distinguish racism and sexism messages from normal text.
Grondahl et al, 2018 reproduced seven state-of-the-art hate speech detection models from prior work, and showed that they perform well only when tested on the same type of data they were trained on. Based on these results, they argued that for successful hate speech detection, model architecture is less important than the type of data and labeling criteria. They further showed that all proposed detection techniques are brittle against adversaries who can (automatically) insert typos, change word boundaries or add in-nocuous words to the original hate speech.

Durairaj M. et al, 2017 combined K nearest neighbor, Support vector machine (SVM) and Naïve Bayes with text mining techniques to classify some text document. This system performed very well with the accuracy of 83.5% but still had some challenges. These challenges include addressing the problems like handling large text corpora, similarity of words in text documents and association of text documents with a subset of class categories.

3.0 METHODOLOGY:

The methodology adopted in this work is the Object-Oriented Analysis and Design (OOADM) with text mining techniques. The model was developed using WEKA tools which is a collection of machine learning algorithms of which support vector machine (SVM) is one of them for data mining tasks. It also contains tools for data pre-processing and classification which makes it very suitable for the development of the new model. The codes were developed in JAVA programming language and a confusion matrix used to evaluate the accuracy

3.1 SYSTEM DESIGN AND IMPLEMENTATION:

Text classification involves preprocessing of the unstructured text documents retrieved from the different web blogs used. After the preprocessing comes the feature extraction of labeled corpus (machine learning classifier for training data) followed by the model selection and classifier. These steps can be seen in figure 1.0

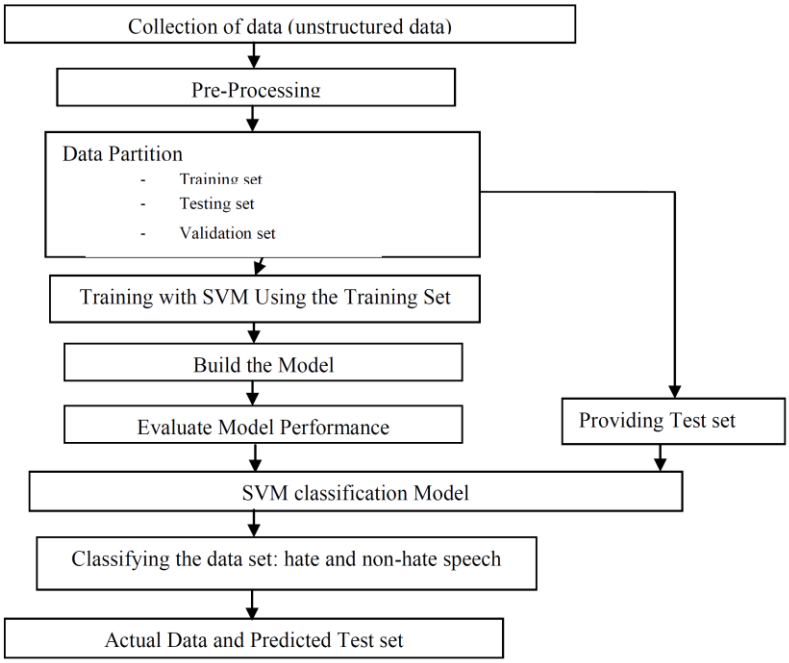

Figure 1.0 Data Flow diagram of the proposed system

i) Data set collection

The dataset was collected from the online social web blogs. These blogs are where people with different backgrounds, culture, beliefs and mind sets read trending news and make comments at any point in time. These comments/messages were retrieved as the unstructured data set. The unstructured data set were collected from the following links

a) https://foreignpolicy.com/comments_view/?view_post_comments=https://foreignpolicy.com/2016/06/14/if-islam-is-a-religion-of-violence-so-is-christianity/

b) http://theconversation.com/challenging-the-notion-that-religion-fosters-violence-85677

c) https://www.nieuwwij.nl/english/karen-armstrong-nothing-islam-violent-christianity/

These web-blogs allow anonymous user comments on articles. Their policy on which comments are allowed is not very restrictive, meaning a lot of hate speech comments are still available online.

i) Preprocessing

The preprocessing stage is a major task of cleaning up the data gathered from unstructured web blog, by labeling those data into their respective classes. This involves 85% of human effort to clean up the dataset in order to be used for building the model and reduce the error during the binary classification

ii) **Feature set Extraction**

Feature extraction is a dimensionality reduction process where the dataset is reduced to more manageable features for processing while still accurately and completely describing the original dataset. It is intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps in some cases leading to better human interpretations.

4.0 RESULTS AND DISCUSSIONS:

Support Vector Machine Classifier

Model Information

==========

Correctly Classified Instances	2004	83.3611 %
Incorrectly Classified Instances	400	16.6389 %
Kappa statistic	0.6628	
K&B Relative Info Score	157853.634 %	
K&B Information Score	1554.2139 bits	0.6465 bits/instance
Class complexity | order 0	2366.6915 bits	0.9845 bits/instance

Class complexity | scheme 429600 bits 178.7022 bits/instance
Complexity improvement (Sf) -427233.3085 bits -177.7177 bits/instance
Mean absolute error 0.1664
Root mean squared error 0.4079
Relative absolute error 34.0062 %
Root relative squared error 82.4705 %
Total Number of Instances 2404

Table 1.0: SVM Detailed Accuracy by Class

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Area**	**PRC Area**	**Class**
	0.830	**0.162**	**0.873**	**0.830**	**0.851**	**0.664**	**0.834**	**0.822**	**Hate speech**
	0.838	**0.170**	**0.786**	**0.838**	**0.811**	**0.664**	**0.834**	**0.728**	**Non-hate speech**
Weighted Avg	**0.834**	**0.165**	**0.836**	**0.834**	**0.834**	**0.664**	**0.834**	**0.782**	

Table 1.2 Confusion Matrix/Contingency Table

A	B	<-- classified as	
1144	234	a|	Hate speech
166	**860**	b|	Non-hate speech

Table 1.0 and 1.2 above shows the general interpretation of results with the confusion matrix for SVM evaluation of results on blogs.

5.0CONCLUSION:

The propagation of hate speech on social media has been increasing significantly in recent years. This may be as a result of the anonymity and mobility of such platforms, as well as the changing political climate from many places in the world. Despite substantial effort from law enforcement departments, legislative bodies as well as millions of investment from social media companies, it is widely recognized that effective counter measures rely on automated semantic analysis of such content.

A crucial task in this direction is the detection and classification of such messages/comments based on its targeting characteristics.This work makes several contributions to state of the art in this research area. A thorough data analysis was carried out to understand the extremely unbalanced nature and the lack of discriminative features of hate speech content in the unstructured dataset one has to deal with in such tasks. However, it is always difficult to clearly decide on a sentence whether it contains hate speech or not if the message is hiding behind sarcasm or if no clear words showing hate, racism or stereotyping exist. Furthermore, online social networks are full of ironic and joking content that might seem like a hate speech which in reality is not.

A support vector machine (SVM) is used for more accuracy and efficiency. The method was thoroughly evaluated on a large collection of web blog datasets for hate speech to show that they can be particularly effective on detecting and classifying the contents. The results obtained from the algorithm really show that it performed the analysis very well.

5.1 Suggestions for Further Research

The system can be improved and advanced based on the following:

1. To explore other methods or algorithms that aim at compensating the lack of training data in supervised learning tasks. Methods such as transfer learning could be potentially promising as they study the problem of adapting supervised models trained in a resource-rich context to a resource-scarce context.
2. To investigate whether features discovered from one hate speech can be transferred to another thus enhancing the training of each other.
3. To train more models and combine their predictions for more robust and accurate models.

REFERENCES:

Arthur V. Ratz, 11 Mar 2018, “Naïve Bayesian Anti-Spam Filter Using Node.JS, JavaScript And Ajax Requests” under The Code Project Open License (CPOL)

Davidson T, Danawarmsley, Micheal Macy & Ingmar Webar, 2017, 'Automated hate speech detection and the problem of offensive language', proceedi.ngs of the eleventh international association for the advancement of artificial intelligence (AAAI) conference on web and social media (ICWSM), www.aaai.org.

M. Durairaj and A. AlaguKarthikeyan, (2017), "Efficient Hybrid Machine Learning Algorithm for text Classification," International Journal on Recent and Innovation Trends in Computing and Communication, Vol.5 (5), 680-688, 2017. ISSN: 2321-8169. [UGC approved journal list No. 49222, IF: 5.75

Georgios K. Pitsilis, HeriRamampiaro and HelgeLangseth, 2018 "Detecting Offensive Language in Tweets Using Deep Learning" arXiv:1801.04433v1 [cs.CL] 13 Jan 2018

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., &Asokan, N. (2018). All You Need Is "Love": Evading Hate Speech Detection. In Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security (pp. 2-12).NewYork: ACM. https://doi.org/10.1145/3270101.3270103

Hajime watanabe, Mondherbouazizi, &Tomoakiohtsuki, 2018, 'Hate Speech on Twitter: A Pragmatic Approachto Collect Hateful and Offensive Expressionsand Perform Hate Speech Detection' volume 6, 2169-3536, 2018, IEEE ACCESS, http://www.ieee.org/publications_standards

Moorthy U & Gandhi U.D, (2018),"A Survey of Big Data Analytics Using Machine Learning Algorithms" IGI Global desiminator of information and knowledge https://www.igi-global.com/chapter/a-survey-of-big-data- analytics-using- machine-learning-algorithms/187661 p95-123

Yenala H., AshishJhanwari, Manoj K. Chinnakotla& Jay Goyal, 2017, 'Deep learning for detecting inappropriate content in text', International journal of data science and analytics,https://doi.org/10.1007/s41060-017-0088-4

YinglieTian, Yong Shi &Xiaohui Liu, 2012, "Recent advances on support vector machines research", Technological and economic development of economy, volume 18(1), DOI: 10.3846/20294913.2012.661205