International Journal of Advanced and Innovative Research (2278-7844)/
Volume 11 Issue 1

An LDA Based Context Arima Based Model for Cyber Spam Tweet Analysis

Yamini Patel[1]
M.Tech Schalor, Department of Computer Science & Engineering
Mittal Institute of Technology, Bhopal, Madhya Pradesh, India

Yamini30sept@gmail.com

Jayshree Boaddh[2], Jashwant Samar[3]
Assistant Professor, Department of Computer Science & Engineering
Mittal Institute of Technology, UIT RGPV, Bhopal, Madhya Pradesh, India

Jayshree.boaddh@gmail.com, jashwantsamar.samar2@gmail.com

Abstract: Twitter and other social media are huge source of data generation and communication. Various open communications is being made on the platform which is used by different community of understanding purpose. Cyber recruitment is one of part which takes place through social media platform and hence understanding such spam content is continuous recruitment in platforms. Many algorithms for understanding spam content and finding the relation between the cyber terminologies. The algorithm worked with either in specific text words, limited iteration or phases. Thus they have limited accuracy, mean square error etc, and hence in order to avoid the limitation, here further is done with Arima based model along with SVM kernel functions to understand the cyber tweet spam content. This work performed on twitter dataset with NLP library and Java platform with tweet API. The observed result shows the performance enhancement up to 6% than existing technique.

Keywords: Cyber recruitment, twitter spamming, spam content, text analysis, NLP, text detection, LDA, CTM, SVM.

I Introduction

In now days, social media and social blogs and other such components of web provide an easy and powerful platform to share information and communicate with other peoples. But some radical and terrorist organizations uses these techniques to influence people and recruit youth, to conduct their activities. There are several virtual communities are presents on the web which used to conduct such activities. Like an online magazine published by al-Qaeda called Arabian Peninsula, such magazines or other web contents spread wrong perception among a certain community and used to influence people to join in their terrorist movements. There are many activities are conducted to for online cyber recruitment for terrorist activities.

Azan, is presented. In [5] effect of the counter terrorism policies over the youth, how youth join in such activities and role of policies to which generate wrong perception among the youth, is presented. [10] a review over the various techniques which are used to detect radicalization of the youth, is presented. In [11] a technique is presented which uses to identifies the intensions of posts in the forums, in [12] [13] [14] detection scheme to detect cyber terrorism activities are presented.

To detect such violent extremist recruitment activity a [1] LDA (Latent Dirichlet Allocation) based technique is used which uses ARIMA model to forecast such activity. For that a social media contents are used to analyze such activity. But LDA uses dirichlet distribution to map content to predict such activities which is not able to provide accurate result, because it not take relation in count to map topics. To reduce this defect a CTM (Correlated Topic Model) based technique is presented which used logistic normal distribution to map topics. That technique provides better and accurate performance to map topics. A result analysis for the propose technique is presented in result and analysis section, which shows that proposed technique provides better result as compare to the existing technique.

II Related Work

In [1] a forecasting technique which used to forecast the activity of violent extremist recruitment in forum is presented. In that technique a SVM based model used to detect recruitment post in the forum. A LDA (latent dirichlet allocation) is used to analyze content of the post in the forum. Put that in to different time series model to forecast that recruitment process. That technique provides less no of forecasting error as compare to the other existing technique. In this paper an ARIMA (Auto regressive integrated moving average), PCR

(Principal component regression) is used in the existing technique but there is a problem in generation of the prediction of the recruitment. In this paper an ARIMA and ETS (Exponential smoothing)

is used to provide the better forecasting results. Comparison with other technique shows that this technique provides better results as compare to the other technique.

In [2] a SVM (Support Vector Machine) based classification technique is presented which used to identify the post related to the violent extremist recruitment. To analyze that process a dark web portals data is used which contains data of more than 28 social websites. That data poses the content related to violent extremist activity and that data also contains religious (Islamic) discussion. In that an analysis process on the basis of the different factors likes the time frame, data sources and some other factors. SVM classification technique is used to classify that data which helps to identify the violent extremist recruitment activity. In previous technique a naïve based classification technique is used to classify that data which is not efficient to provide better results.

In [3] perspective over the virtual communities supporting terrorist and violent extremist activities is presented. These virtual communities play a vital role in such extremist recruitment processes. Virtual communities or group over social networks poses components like white power music and white supremacist games which used to promote the racism against the non-whites are used to conduct activities to recruit peoples for violent extremist activities. There are many other means by which radical perception is created, which used to influence the peoples for the terrorist activities.

PROBLEM DEFINITION

In this section the discussion on the problem formulation discussion from the literature survey is done in the previous section.

1. A previous technique such as SVM utilized ARIMA (Autoregressive integrated moving average),PCR (Principal component regression) for the prediction model generation, but still, the obvious problem occurs with the technique is in generating a prediction for the cyber activity. This technique persists better result than existing but still enhancement is required which is provided by the proposed procedure.

2. Previous technique Naïve based classification doesn't perform a better recruitment and violence classification due to a limited number of rules, thus a better probability model can't get generated using the technique.

3. Another problem happens due to data scarcity. For any possible value of a feature, you need to estimate a likelihood value by a frequents approach. This can result in probabilities going towards 0 or 1, which in turn leads to numerical instabilities and worse results.

4. One more problem is that the Naive Bayes classifier makes a very strong assumption on the shape of your data distribution, i.e. any two features are independent given the output class. Due to this, the result can be (potentially) very bad - hence, a naive classifier.

5. In LDA a Dirichlet distribution is used because of that, the data classification occurs in a single direction with single category detection.

6. Thus, in order to propose a better prediction model using classification and further combine approaches requirement is to further acquire a scheme which contributes to getting a better outcome and prediction system, here our proposed methodology LDA is utilized scheme in place of CTM SVM.

PROPOSED METHODOLOGY

As per discussion between the different literature survey and problem associated with them. The further requirement is to find a combination algorithm to observe the outcome.

1. Extracting and loading of all the available data from the created blog which are participating in the communication.

2. Loading the complete violent data dictionary pair from the dataset.

3. Sorting all the values in decreasing order and for which the frequency count enhancement and sorting are performed. Finally, the new subset is getting performed using the NLP sentence processing and matching with the multilevel data according to violence keyword bag and its frequency.

4. Perform the particular algorithm as per selected by the user for further execution, such as LDA or CTM SVM.

5. Perform LDA and matching operation if any single match is obtained and conclude that further using the ARIMA model for the prediction either it is the violent violence post or not.

6. Perform CTM SVM model and match operation if at least 2 or more dictionary match is performed by the system, further perform the ARIMA model date wise by obtaining data.

7. Plotting a date wise graph for the violent keyword history model.

8. Observing the values and thus it affects accuracy and efficiency for the complete scenario.
9. Exit.

Algorithm Pseudo Code:

Input: Dictionary pair Qi, Dataset tables violence post blog.

Output: algorithm process, Metadata, prediction values.

Steps:

Active either LDA or ASVM

While(true) do{

Postextraction{p1,p2…..pN};

dictionaryPairRequest();

If(scorematching()==1)

{

LDARecognition();

Perform ARIMA model;

Compute the prediction values;

{

Plotting a date wise violence post prediction;

}

Set status=finish and exit;

}If(scorematching()>=2)

{

ASVM Recognition();

Perform ARIMA model;

Compute the prediction values;

{

Plotting a date wise violence post prediction;

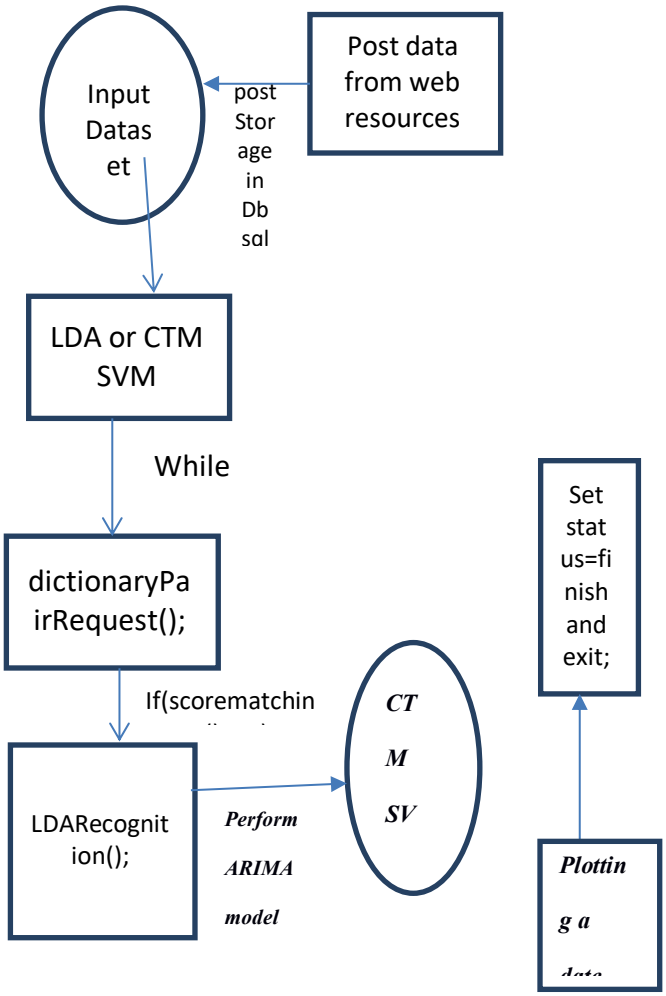
}

Set status=finish and exit;

}

The Figure below represents the complete flow of the proposed scenario which represents our work and computes parameters efficiently.

Figure 1: Flow diagram of complete approach.



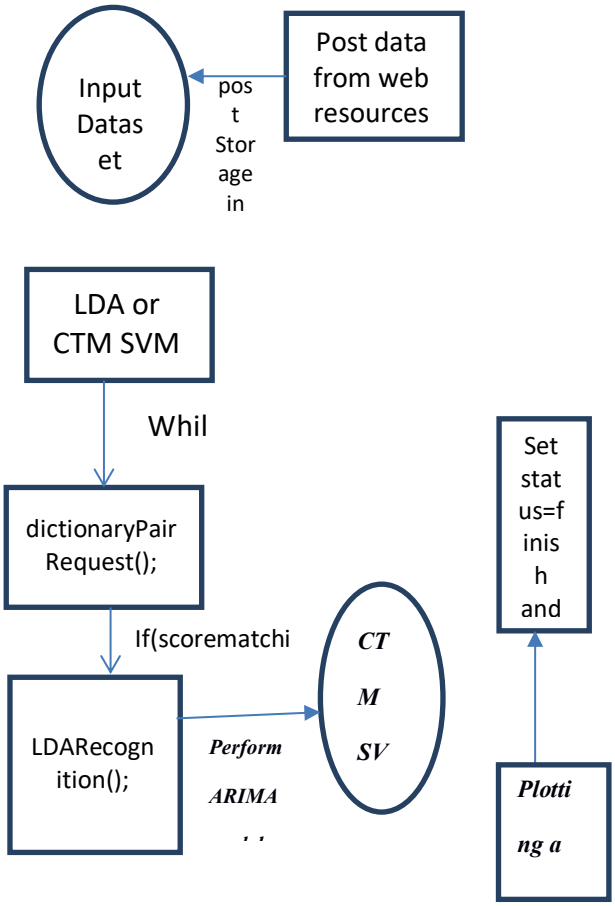


Figure 2: Block Diagram of LDA.

In this section, our work is to define the problem definition of the existing presented algorithm in the scenario and the proposed work derived from our work definition. Thus the work presented by us is highly efficient for prediction model, Further performed work is investigating dataset, query input for the processing with system available and outperform result parameter monitoring.

Experiment setup & Result Analysis

In order to perform execution and result analysis. The experiment is performed on Java platform using Net beans tool IDE along with Apache server. To work with text processing NLP library with cyber bag word as detection entity is used. Further the extraction of tweets, storing them with database and processing for spam detection is performed. The result computation is made using RMSE, MAE values. The observed result shows the efficiency of proposed algorithm over traditional solution.

Table 1: Violent extremist statics for forecasting values MASE.

Technique Approach Date of data extracted	LDA Technique Model	ASVM Technique Model
#per Day	2.5101	10.2314
%Per Day	0.0470	3.4104

The above table represents the number of forecasting values of the post and comparison is performed using MASE.

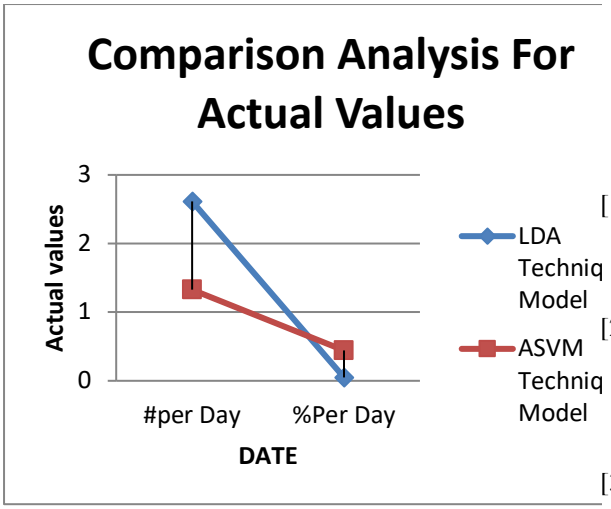
In the above graph drawn x-axis as data from which post was extracted for the query processing for the specified dataset and line, the graph is printed using the chart library provided by the Microsoft and further analysis can easily perform thus the ASVM approach outperforms the best. The graph representation shows the efficiency of our proposed algorithm work and it outperforms the low forecasting value.

In the table present below is a statistical comparison of the actual values which forecasts violent extremist online recruitment in social media or website's posts which are posted by radical or terrorist organizations. There is a post which taken for 5 days and extracted data is processed and further, the following result is computed.

Table 2: Violent extremist statics for forecasting values RMSE parameter.

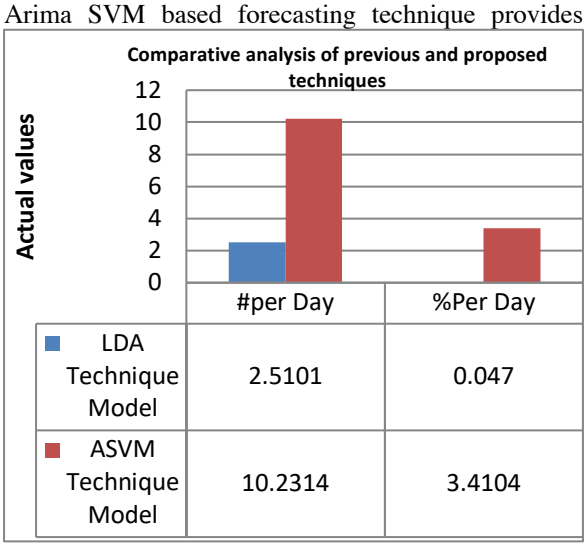
Technique Approach Date of data extracted	LDA Technique Model	ASVM Technique Model
#per Day	2.6138	1.3269
%Per Day	0.0493	0.4423

The above table represents the number of forecasting values of the post and comparison is performed using RMSE.



Graph b: Comparison Line graph for technique analysis RMSE values.

Both graphical and statistical analysis shows that Arima SVM based forecasting technique provides



	#per Day	%Per Day
LDA Technique Model	2.5101	0.047
ASVM Technique Model	10.2314	3.4104

Graph B: Bar graph comparison between the two approaches.

The graph shows the accurate result as compare to the LDA based technique. Propose technique is efficient to deal with such violent extremist recruitment activities

CONCLUSION & FUTURE WORK

Cyber tweet analysis and working with their textual data is an important part where the cyber recruitment, text anomaly are common with increasing in trends. In this paper the discussion is mentioned with previous approaches, limitations in them and further finding an approach working closely with ARIMA based model and SVM kernel process over the text data. Algorithm flow is discussed with the problem formulation and proposed solution. Further an implementation is performed using Java and experimental setup is executed. The result comparison is made through and finding the enhance improved parameter value over traditional solutions. An understanding of different language, working with parallel processing for fast execution, working with more parameter computation is left for the future work.

REFERENCES

[1]. Viktor Golem Mladen Karan Jan Snajder, Combining Shallow and Deep Learning for Aggressive Text Detection, 2018.

[2]. Peipei Zhou1,2,3,4, Qinghai Ding1,5, Haibo Luo1,3,4, and Xinglin Hou1,2,3,4, Violent Interaction Detection in Video Based on Deep Learning, 844(1):012044 · June 2017.

[3]. Rishabh Varmaa and Sartaj Ahmadb, Mass Violence Detection Using Data Mining Techniques, EISSN 2392-2192, 31 October 2018.

[4]. Ahmad, Sartaj & Varma, Rishabh. (2018). Information extraction from text messages using data mining techniques. Malaya Journal of Matematik. S. 26-29. 10.26637/MJM0S01/05.

[5]. Ramya, R. S., et al. Feature Extraction and Duplicate Detection for Text Mining: A Survey. Global Journal of Computer Science and Technology 16.5 (2017).

[6]. B. Doosje, F. M. Moghaddam, A. W. Kruglanski, A. de Wolf, L. Mann, and A. R. Feddes, "Terrorism, radicalization and deradicalization," Current Opinion in Psychology, vol. 11, pp. 79–84, 2016.

[7]. Europol, "European union terrorism situation and trend report (te-sat) 2016," Tech. Rep., 2016.

[8]. I. for Economics and Peace, Global Terrorism Index 2014: Measuring and Understanding the Impact of Terrorism, 2016.

[9]. Thompson, Dominic, and Ruth Filik. Sarcasm in written communication: Emoticons are efficient markers of intention. Journal of Computer-Mediated Communication 21.2 (2016): 105-120.

[10]. J. R. Scanlon and M. S. Gerber, Automatic detection of cyber recruitment by violent extremists, Secure. Inform., vol. 3, no. 1, pp. 1–10, Aug. 2014.

[11]. Lisa Kaati, Enghin Omer, Nico Prucha, Amendra Shrestha Detecting Multipliers of Jihadism on Twitter IEEE, 2015.

[12]. W. V. Fitzgerald. (Jun. 2010). Interview With Westboro Baptist Church: Hate Name God. [Online]. Available: http://www.digitaljournal.com/article/2933642470 ieee transactions on information forensics and security, vol. 10, no. 11, november 2015.

[13]. J. Li et al., Social media: New perspectives to improve remote sensing for emergency response, Proc. IEEE, vol. 105, no. 10, pp. 1900–1912, Oct. 2017.

[14]. Sukhjit Singh Sehra 1,2,*, Jaiteg Singh 3 and Hardeep Singh Rai 4, Using Latent Semantic Analysis to Identify Research Trends in OpenStreetMap, : 1 July 2017.

[15]. Ruchika Aggarwal, Latika Gupta, AUTOMATIC TEXT SUMMARIZATION, Ruchika Aggarwal et al, International Journal of Computer Science and Mobile Computing, Vol.6 Issue.6, June- 2017, pg. 158-167.

[16]. Nimai Chand Das Adhikari, Nishanth Domakonda∗, Chinmaya Chandan, An Intelligent Approach to Demand Forecasting, ICICT 2017, ISBN:978-1-5090-6697-1.