Original Article

EXTREMISM DETECTION ON SOCIAL MEDIA USING SVM TEXT CLASSIFIER

Vijay[1], Dr. Pushpneel Verma[2]

[1]Research Scholar, Department of Computer Science and Engineering, Bhagwant University, Sikar Road Ajmer, Rajasthan.
Email: vijay.movfrose@gmail.com

[2]Department of Computer Science and Engineering, Bhagwant University, Sikar Road Ajmer, Rajasthan.
Email: pushpneelverma@gmail.com

DOI: 10.47750/pnr.2021.12.01.18

Abstract

Spread of extremism on social media is major issue nowadays. Extremism can be expressed in the form of hate speech. It is necessary to distinguish hate speech from offensive language. In text documents hate speech can be detected by text classification. Text classifiers based on supervised machine learning can be used for hate speech detection. In this paper we discussed a method for hate speech detection by performing text classification with SVM. The dataset we used for experiments contained text tweets having hate speech or offensive language or neither hate speech nor offensive language. We used SVM classifiers with different kernels i.e. linear, sigmoid and RBF kernel. The highest classification accuracy was delivered by SVM with RBF kernel i.e. 89.04%.

Keywords: hate speech, SVM, linear kernel, RBF kernel and sigmoid kernel.

Introduction

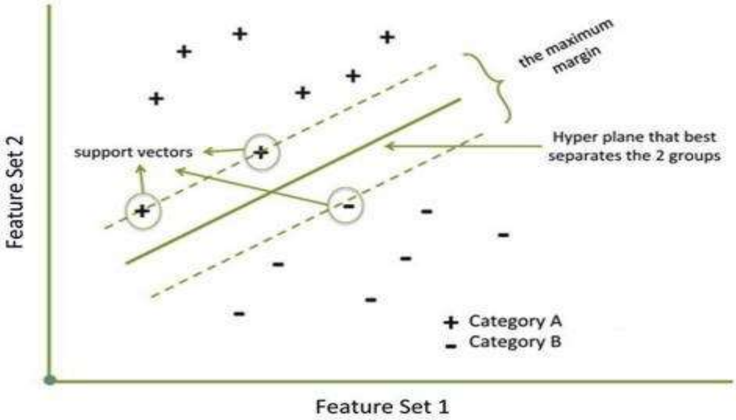
Support vector machine(SVM) is an important supervised machine learning technique. It is used for regression and classification. Vapnik introduced SVM in 1995 [1]. In several studies it was found that SVM was able to deliver better classification accuracy than other algorithms of data classification [2]. SVM has been emerged from SLT (statistical learning theory).SVM has a variety of applications. In image processing SVMs can be used for image clustering, object recognition, recognizing hand-written characters and recognizing face and facial expressions. SVMs can also be used for feature extraction, data classification and text classification. We will use SVM for text classification for detecting extremism in text documents. The main advantages of SVM are: it can effectively handle high dimensional and sparse data and it overcomes the problem of overfitting. The major drawbacks of the SVM algorithm are that it requires large memory and it runs very slow. The SVM algorithms are computationally expensive.

SVM is a supervised machine learning algorithm. It is first trained with an already classified or labeled dataset. The trained SVM classifier can be used to classify a new data object.

SVM can be classified as a linear model or a nonlinear model [3]. The linear model of SVM divides the data domain linearly (e.g., hyperplane or straight line) so that in original domain classes are segregated [4]. When it is not possible to divide data domain linearly then nonlinear SVM transforms data domain into such a space where data domain is linearly divided to segregate the classes. This space is known as feature space.

In linear SVM the data domain is mapped into a response set and then divided. In SVM algorithm an optimal hyper plane is constructed to classify the patterns which are linearly separable. Optimal hyper plane which is used for classification is a maximum margin hyper plane. The hyper plane’s margin signifies the distance between nearest points of each class and hyper plane and these nearest points are called support vectors. The basic goal of SVM algorithm is the maximization of margin [5]. SVMs are also known as maximum margin classifiers. It is assumed the hyper planes with more margin give better classification accuracies.

Fig. 1 SVM with maximum margin hyperplane



The Fig. 1 shows the SVM having maximum margin hyper plane separating data objects of two different categories i.e. A and B. The data objects of category A are represented by + sign and data objects of category B are represented by – sign.

When a dataset is not linearly separable then it is usually mapped into a higher dimensional space where it can be separated linearly. The kernel function is used to map the dataset into higher dimensions. The kernel functions which are commonly used are: SIGMOID, POLY, LINEAR and RBF.

The setting of SVM's learning problem can be described as: y =f(x) is an unknown nonlinear function where x denotes an input vector of higher dimension and y signifies the output which can be a vector in case of SVM is for multiclass classification otherwise y is scalar. A distribution free learning needs to be performed as the information of underlying joint probability functions is not available. A training dataset D is given. D = {(xi, yi)∈ X x Y}. Often yi is denoted as desired target value [6].

For large datasets the computational cost of SVM is high as the growth of training kernel matrix is quadratic with the dataset's size. Therefore, with large datasets training of SVM gets very slow. A separation hyperplane is found in the training of SVM which implicates an n x n density matrix, where n signifies the count of points present in dataset. SVM's training complexity depends on the dataset's size, so large datasets require large amount of memory and computational time [8]. There are some methods to reduce SVM's training time, for example, decomposition, chunking, data reduction, sequential minimal optimization (SMO) and shrinking. In data reduction method the data with low possibility of being support vectors is removed and data with high possibility of being support vectors is kept for training the SVM. In chunking method an optimization problem is considered as solving a series of sub-problems. In this method for SVM a chunk which is a data's arbitrary subset is selected and then next chunk is acquired and the process continues until complete training data is covered and all the support vectors are got by chunk. In this method a large problem is reduced to a series of small-sized optimization problems, repetitively getting the support vectors. Decomposition techniques and chunking techniques are similar to each other. In decomposition techniques the sub problems are fixed sized. Decomposition techniques can be categorized as dual and primal methods [8].

For large datasets there are some implementations of SVM i.e., SVMLigth, SVMTorch, Pegasos, LIBSVM and Incremental SVM. SVMLigth is a very fast algorithm. It is used for classification and regression. Shrinking and Working selection methods are used SVMLigth. Shrinking and Working selection methods are also used by SVMTorch. According to authors regression problems of large scale can be solved efficiently by SVMTorch.

Decomposition methods are used by Pegasos for training time reduction. LIBSVM is based on SMO, however, it has an advanced work set selection algorithm. Incremental SVM is an incremental learning framework.

Our purpose is to use SVM for extremism detection or hate speech detection in text documents. Our problem is a kind of text classification problem. Many researchers have claimed that SVM gives significant performance in text classification [7].

Mohammad et al. [9] performed text classification of Arabic Texts with Multilayer Perceptron Neural Network, Naïve Bayes and Support Vector Machine. Dataset they used for experiments contained 1400 Arabic text documents. This dataset was collected from Alhayat, Saudi Press Agency and Aljazeera news website. In their experiments best results were given by SVM.

Liwei et al. [10] proposed a new method called SVM-MK (SVM with mixture of kernel) for text classification. In this technique kernel was a convex combination of several finitely basic kernels. They performed experiments with a Chinese corpus which was created by Fudan University. SVM- MK delivered significant performance for text classification in their experiment.

Gharib et al. [11] performed text classification of documents in Arabic using SVM. They compared the performance of SVM with Rocchio, KNN(K- nearest neighbor) and Bayes classifiers. They evaluated the performances of classifiers by performing two experiments. In one experiment training set was used as test set but in another experiment Leave one testing method was used. The dataset they used for performing experiments contained 1,132 documents and it was created by collecting documents from Egyptian newspapers namely ElAkhbar, ElGomhoria and ElAhram in the time period between August 1998 and September 2004. In their experiments for small sized feature set the better results were given by Rocchio classifier and for large sized feature set the performance of SVM was better than other classifiers.

Dadgar et al. [12] proposed a method for news classification based on SVM and TF-IDF(term frequency- inverse document frequency). For performing their experiments they used two datasets: BBC dataset and 20Newsgroup dataset. BBC dataset consists of 2225 news text documents collected on these five topics: technology, business, sports, politics and entertainment in the time duration from 2004 to 2005. 20Newsgroup is formed of 19,997 news articles of 20 classes taken from the Internet. For their experiments they used only 5070 articles of five classes. First they preprocessed the text, then they extracted features by using TF-IDF and then they performed classification by using SVM. They got classification precision values of 97.84% for BBC dataset and 94.93% on 20Newsgroup datasets.

Rennie and Rifkin [13] performed experiments to compare the performances of SVM and Naïve Bayes for multiclass text classification. It has been found that SVM performs binary text classification effectively. In their experiments they demonstrated that SVM can also perform multiclass text classification effectively by using with Error Correcting Output Coding (ECOC). In ECOC several binary classifiers are learnt and their outputs are used to assign the label to a new example. They used two datasets for performing experiments i.e., Industry Sector and 20Newsgroups. In their experiments they found that the performance of SVM was significantly better than the performance of Naïve Bayes for multiclass classification.

Utomo and Sibaroni [14] developed a system for text classification to classify sentences on the basis of types of English (i.e. American English and British English). They divided data by 10-fold-cross-validation. In their experiments they used SVM classifier with linear kernel and got accuracy of 96.53%.

For the classification of short texts a framework of deep uniform kernel mapping SVM (DUKMSVM) was introduced by Liu et al [15]. They proposed DRKMSVM (deep bidirectional recurrent kernel mapping SVM) which was based on that introduced framework so that the performance of classification of short texts can be improved. In DRKSVM for solving the parameters the kernel trick is not required. In DRKMSVM the kernel mapping is expressed by using bidirectional RNN. In DRKMSVM the context information is obtained effectively by the bidirectional RNN in short texts. They used five datasets for performing experiments i.e., Movie Review (MR), Custom Review (CR), Subject (Subj), Multiple Perspective QA (MPQA) and Text Retrieval Conference (TREC). The results of their experiments demonstrated that best performance was delivered by DRKMSVM in most instances in the comparison of Deep Neural mapping SVM, Naïve Bayes, traditional SVM and CNN in terms of F_1-score, classification accuracy, recall rate and precision.

Hao et al. [16] proposed a new technique for text categorization. They combined Naïve Bayes with SVM. First they preprocessed the text dataset by performing various preprocessing tasks like removing low-frequency words, stop words removal suffix and prefix removal etc. After preprocessing the weight function was used for feature word. They extracted text features and reduced the dimensions. They trained SVM by Naïve Bayes algorithm. In their experiments their proposed method delivered the F score value of 86.89% whereas the traditional SVM delivered the F score of value 83.33%.

Sabbah et al.[17] proposed a method based on machine learning for selection and ranking of features called SVM-FRM (SVM based feature ranking method). In their proposed method SVM is used for the weighting of significant features and their selection so that the classification performance would be better. They also used hybridization methods in some experiments to get better performance of SVM-FRM. They performed text classification on three Arabic datasets i.e., Watan dataset, Abuaiadah dataset and BBC dataset with SVM-FRM and also with its enhancement. The Watan dataset consists of more than 20000 text documents of six catefories : sports, religion, international news, local news, culture and economy. Abuaiadah dataset contains 2700 text documents of 9 categories i.e., technology, economy, sport, literature, law, religion, politics, art and health. BBC dataset contained 4763 text documents classified into seven classes i.e., newspapers highlights, world news, science and technology, business and economy, sports, Misc, and Middle East news. They compared the performance of SVM-FRM with Chi Square and IG which are feature selection (ranking) methods. The results of their experiments demonstrated that SVM-FRM delivered higher F-measure and better accuracy on balanced datasets. On an unbalanced dataset the performance was significantly enhanced by the proposed hybridization methods [19].

Mitra et al.[18] introduced a least square SVM (LS-SVM) for text classification. The system based on LS- SVM delivered the classification the 99.9% of classification accuracy. The LS-SVM array using Gaussian radial basis function kernel was the final classifier. It used the coefficients which were generated by latent semantic indexing algorithm. When they compared their proposed system with Naïve Bayes and K nearest neighbor they found that the performance of their proposed system was better than that of KNN and Naïve Bayes [20].

Methodology

We performed experiments on hate speech and offensive language dataset. This dataset contains 24783 tweets. The text tweets in this dataset are classified into three classes 0,1 and 2. The classes 0, and 1 signify the presence of hate speech and offensive language respectively in text documents. Class 2 signifies that the text tweet is neutral and contains neither offensive language nor hate speech. In this dataset 1430 text tweets belong to the class 0, 19190 belong to class 1 and 4163 belong to class 2.

The SVM is a supervised machine learning algorithm and it needs to be trained. First we shuffled the text documents of the dataset. We preprocessed the dataset by converting it into lowercase and then removing punctuations, numbers, stopwords and URLs. Finally we stripped the white space and formed document term matrix of text documents. The dataset contained 24783 text documents, we used 20000 text documents for training the SVM model. The trained SVM model was tested on remaining 4783 text documents of dataset. We performed experiments with SVM with different kernels (i.e. linear, RBF and sigmoid).

Results

SVM with linear kernel delivered the accuracy of 87.27%. Fig. 2 shows the confusion matrix and statistics for the SVM with linear kernel.

Fig. 2. Confusion matrix and statistics for SVM with linear kernel

```
> conf.mat
Confusion Matrix and Statistics

          Reference
Prediction    0    1    2
         0   19   25    3
         1  237 3554  162
         2   37  145  601

Overall Statistics

               Accuracy : 0.8727
                 95% CI : (0.8629, 0.882)
    No Information Rate : 0.7786
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6138

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: 0 Class: 1 Class: 2
Sensitivity          0.064846   0.9544   0.7846
Specificity          0.993764   0.6232   0.9547
Pos Pred Value       0.404255   0.8991   0.7676
Neg Pred Value       0.942145   0.7952   0.9587
Prevalence           0.061259   0.7786   0.1602
Detection Rate       0.003972   0.7430   0.1257
Detection Prevalence 0.009826   0.8265   0.1637
Balanced Accuracy    0.529305   0.7888   0.8696
> |
```

SVM with sigmoid kernel delivered the accuracy of 88.65%. Fig. 3 shows the confusion matrix and statistics for SVM with sigmoid kernel.

Fig. 3 Confusion matrix and statistics for SVM with sigmoid kernel

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1    2
         0   46   62    5
         1  191 3487   54
         2   56  175  707

Overall Statistics

               Accuracy : 0.8865
                 95% CI : (0.8771, 0.8953)
    No Information Rate : 0.7786
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6843

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: 0 Class: 1 Class: 2
Sensitivity          0.156997   0.9364   0.9230
Specificity          0.985078   0.7686   0.9425
Pos Pred Value       0.407080   0.9344   0.7537
Neg Pred Value       0.947109   0.7745   0.9847
Prevalence           0.061259   0.7786   0.1602
Detection Rate       0.009617   0.7290   0.1478
Detection Prevalence 0.023625   0.7803   0.1961
Balanced Accuracy    0.571037   0.8525   0.9327
```

The highest classification accuracy was delivered by SVM with RBF kernel i.e., 89.04%. Fig. 4 shows the confusion matrix and statistics for SVM with RBF kernel.

Fig. 4 Confusion matrix and statistics for SVM with RBF kernel.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1    2
         0   45   53    5
         1  199 3536   83
         2   49  135  678

Overall Statistics

               Accuracy : 0.8904
                 95% CI : (0.8812, 0.8992)
    No Information Rate : 0.7786
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6855

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: 0 Class: 1 Class: 2
Sensitivity          0.153584   0.9495   0.8851
Specificity          0.987082   0.7337   0.9542
Pos Pred Value       0.436893   0.9261   0.7865
Neg Pred Value       0.947009   0.8052   0.9776
Prevalence           0.061259   0.7786   0.1602
Detection Rate       0.009408   0.7393   0.1418
Detection Prevalence 0.021535   0.7982   0.1802
Balanced Accuracy    0.570333   0.8416   0.9197
```

Conclusion

The performance of SVM for text classification is remarkable. The hate speech detection in text documents is also a text classification problem. We performed experiments on a dataset of 24783 text tweets. We performed text classification on this dataset by using SVM with linear, sigmoid and RBF kernel. The SVM with RBF kernel delivered the highest accuracy of 89.04%.

References

[1] Ukil, Abhisek. "Support vector machine." Intelligent Systems and Signal Processing in Power Engineering. Springer, Berlin, Heidelberg, 2007. 161-226.

[2] Durgesh, K. SRIVASTAVA, and B. Lekha. "Data classification using support vector machine." Journal of theoretical and applied information technology 12.1 (2010): 1-7.

[3] T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning. New York: Springer, 2009.

[4] Narayan, Vipul, and A. K. Daniel. "CHOP: Maximum coverage optimization and resolve hole healing problem using sleep and wake-up technique for WSN." ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal 11.2 (2022): 159-178.

[5] Pradhan A. Support vector machine-a survey. International Journal of Emerging Technology and Advanced Engineering. 2012 Aug;2(8):82-5.

[6] Narayan, Vipul, and A. K. Daniel. "Design Consideration and Issues in Wireless Sensor Network Deployment."." Invertis Journal of Science & Technology (2020): 101.

[7] Liu Y, Loh HT, Tor SB. Comparison of extreme learning machine with support vector machine for text classification. InInternational Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems 2005 Jun 22 (pp. 390-399). Springer, Berlin, Heidelberg.

[8] Pramanik, Sabyasachi, et al. "A Novel Approach Using Steganography and Cryptography in Business Intelligence." Integration Challenges for Analytics, Business Intelligence, and Data Mining. IGI Global, 2021. 192-217.

[9] Mohammad, A., Alwada'n, T. & Al-Momani, O. Arabic Text Categorization Using Support vector machine, Naïve Bayes and Neural Network. GSTF J Comput 5, 16 (2016). https://doi.org/10.7603/s40601-016-0016-9

[10] Wei L, Wei B, Wang B. Text classification using support vector machine with mixture of kernel. Journal of Software Engineering and Applications. 2012 Dec 1;5:55.

[11] Gharib TF, Habib MB, Fayed ZT. Arabic Text Classification Using Support Vector Machines. Int. J. Comput. Their Appl.. 2009 Dec 1;16(4):192-9.

[12] Dadgar, S. M. H., Araghi, M. S., & Farahani, M. M. (2016). A novel text mining approach based on TF-IDF and Support Vector Machine for news classification. 2016 IEEE International Conference on Engineering and Technology (ICETECH). doi:10.1109/icetech.2016.7569223

[13] Rennie, Jason DM, and Ryan Rifkin. "Improving multiclass text classification with the support vector machine." (2001).

[14] M. R. Ario Utomo and Y. Sibaroni, "Text Classification of British English and American English Using Support Vector Machine," *2019 7th International Conference on Information and Communication Technology (ICoICT)*, 2019, pp. 1-6, doi: 10.1109/ICoICT.2019.8835256.

[15] Liu Z, Kan H, Zhang T, Li Y. DU KMSVM: A Framework of Deep Uniform Kernel Mapping Support Vector Machine for Short Text Classification. Applied Sciences. 2020 Mar 30;10(7):2348.

[16] P. Hao, D. Ying and T. Longyuan, "Application for Web Text Categorization Based on Support Vector Machine," *2009 International Forum on Computer Science-Technology and Applications*, 2009, pp. 42-45, doi: 10.1109/IFCSTA.2009.132.

[17] Sabbah T, Ayyash M, Ashraf M. Hybrid support vector machine based feature selection method for text classification. Int. Arab J. Inf. Technol.. 2018 Jan 1;15(3A):599-609.

[18] Mitra V, Wang CJ, Banerjee S. Text classification: A least square support vector machine approach. Applied soft computing. 2007 Jun 1;7(3):908-14.

[19] Narayan, Vipul, and A. K. Daniel. "Energy Efficient Protocol for Lifetime Prediction of Wireless Sensor Network using Multivariate Polynomial Regression Model." Journal of Scientific & Industrial Research 81.12 (2022): 1297-1309.

[20] Awasthi, Shashank, et al. "A Comparative Study of Various CAPTCHA Methods For Securing Web."