# Linear Discriminant Analysis for Hate Speech Text Classification

1Vijay, 2Dr. Pushpneel Verma
1Research scholar, 2Associate Professor
Bhagwant University

*Abstract* - **Social media has enabled the people to share their ideas widely online. Social media has many advantages. It provides a platform to people to express their talent. It provides a way to communicate with large number of people. Many people use social media to grow their network and strengthen their business. As the numbers of users are increasing on social media, the problem of hate speech is also increasing on social media. Hate speech on social media can provoke violence. There are many supervised machine learning based algorithms which can be used to detect hate speech on social media. In this paper we have proposed a method to detect hate speech texts by using Linear Discriminant Analysis (LDA). LDA is a dimensionality reduction technique. In this paper we will use LDA to classify a text document as containing hate speech or not containing hate speech.**

*keywords* - **Linear Discriminant Analysis, text classification, hate speech, dimensionality reduction, document term matrix.**

## I. INTRODUCTION

Social media provides a platform where people can interact with others and express their ideas. Hateful texts on social media can create social disturbances. Before publishing a text message on social networking websites there is a need to detect whether it is hateful or not [1]. In this paper we have discussed a method of hate speech detection on social media using Linear Discriminant Analysis (LDA). LDA requires already classified data for training as it is a supervised technique. It is also known as Fisher's LDA. In simple words in this paper we will use Linear Discriminant Analysis for text classification. First an LDA classifier was trained with already classified text documents and then it predicted whether a new text document is containing hate speech or not. For pattern recognition problems (like face recognition) LDA has shown good performance [2]. In this paper text documents are classified into two classes: "hate" and "nothate".

Linear Discriminant Analysis is a dimensionality reduction and feature extraction technique [3]. It was proposed in 1936 by R. Fischer [4]. In Linear Discriminant Analysis a projection hyperplane is found which fulfils two objectives: 1) interclass variance should be minimized, 2) projected means of classes should have maximum distance between them [4]. For dimensionality reduction and classification this hyperplane can be used. In LDA data vectors belonging to c different classes are projected into (c-1) dimensional space in such a manner so that between class variance is maximized and within class variance is minimized. The LDA obtains new low dimensional space by eigenvalues decomposition of within-class covariance matrix and between-class covariance matrix. The dataset which we used in this paper consist of text documents belonging to one of these two classes: hate and nothate.

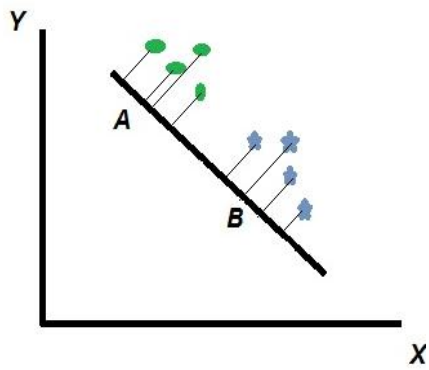


Figure 1 An example showing how data samples in two dimensional space are projected by LDA in one dimensional space

For text classification usually dataset needs to be transformed into document term matrix. Document term matrices usually have high dimensionality. Suppose a dataset contains n documents and transformed into a document term matrix of nxm dimension. Each document is represented by m features. There are two ways to reduce the dimension of document term matrix: 1) select some columns (features) from the document term matrix for the learning purpose and reject others, 2) project feature vectors of high-dimensions into lower dimensions. LDA uses second approach. The purpose of LDA is that

documents are projected into lower dimensional space in a way that classes are separated well [5]. Since in high dimensional space it is difficult to analyze and compare data points and data points are also far apart, learning is a challenging task in high-dimensional spaces [6]. There is a phenomenon called curse-of-dimensionality, according to which to make accurate predictions on high dimensionality problems a large number of samples are required [7]. The curse of dimensionality problem can be handled by dimensionality reduction. Dimensionality reduction facilitates data interpretation and data visualization also. Figure 1 shows how data samples (belonging to two different classes A and B) in two dimensions are projected by LDA in one dimension space.

## II. Linear Discriminant Ananlysis

Linear Discriminant Analysis is used not only for dimensionality reduction but it is also used for classification. The dataset which we used in this paper contains documents belonging to two classes. Therefore, in this section we will discuss how Linear Discriminant Analysis can be used for two class classification problem.

Consider a dataset in which data objects belong to either one of two classes $C_1$ and $C_2$.

Let $m_1$ is the mean vector of class $C_1$ and $m_2$ is the mean vector of class $C_2$

For this two class problem within-class scatter matrix ($S_w$) can be given by:

$$S_w = \sum_{i \in I_1} (x_i - m_1)(x_i - m_1)^T + \sum_{i \in I_2} (x_i - m_2)(x_i - m_2)^T$$

Here $I_1$ is the index set of $C_1$ class and $I_2$ is the index set of class $C_2$.

Let $n_1$ is the size of $C_1$ class and $n_2$ is the size of $C_2$ class.

For original data the overall mean vector is given by [8]:

$$m = n_1 m_1 + n_2 m_2$$

The between-class scatter matrix $S_b$ for this two-class problem can be calculated by the formula:

$$S_b = n_1(m_1 - m)(m_1 - m)^T + n_2(m_2 - m)(m_2 - m)^T$$

The idea is to find linear transformation W to maximize the fisher criterion J(W).

$$J(W) = \frac{|W^T S_b W|}{|W^T S_w W|}$$

W is the transformation matrix and used for dimensionality reduction [9].

Since in the dataset which we used in this paper contains text documents belonging to two different classes the LDA will project the text documents into one dimensional space.

### *Prediction by LDA classifier*

In this section we will discuss the concept how the class of an observation is predicted by the LDA. Discriminant scores of an observation are computed by LDA classifier for each class to predict the class of the given observation. Suppose in the training dataset observations belong to K classes: $C_1, C_2, \ldots\ldots\ldots\ldots, C_k$.

Consider an observation whose class is to be predicted by the LDA classifier. Let X represents the predictor variables. Suppose X is the single predictor variable i.e. X = x. Let $\delta_k(x)$ represents the estimated discriminant score that the observation belongs to the $C_k$ class. Then, $\delta_k(x)$ can be evaluated by the formula:

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\Pi_k)$$

$\Pi_k$ is the prior probability that the class of observation is $C_k$.

$\mu_k$ represents the average of training observations belonging to class $C_k$.

For each of the K classes the weighted average of sample variances is represented by $\sigma^2$.

The LDA classifier will predict that class *k* for the given observation whose discriminant score is largest.

## III. Methodology

The dataset which we used for experimentation was downloaded from kaggle.com. The downloaded "Dynamically Generated Hate Speech Dataset" had many columns. We considered only two columns for experiment i.e. text and label and deleted all others. In this case column label contained only one of the two values i.e. "hate" and "nothate", specifying whether the corresponding text document contains hate speech or not. This dataset contains 40463 text documents which are classified as hate and nothate. We first shuffled the dataset properly and then took only 8000 text documents from the dataset for experimentation purpose. We used 6000 text documents for training the LDA and 2000 text documents for testing the LDA classifier.

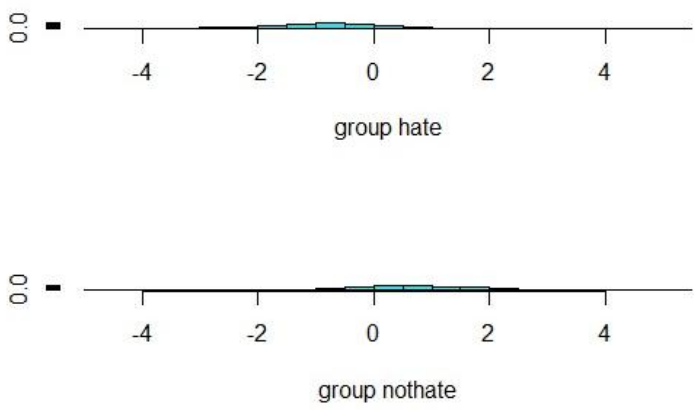


Figure 2 Stacked histogram of the discriminant function values for training dataset

First we cleaned the dataset by converting the texts into lowercase, removing punctuations, stopwords, URLs, and numbers, then we stripped the whitespace. Finally stemming was done. Stemming is a process in which a certain word is reduced to its root [10]. Finally we converted the text documents into 'document term matrix'.

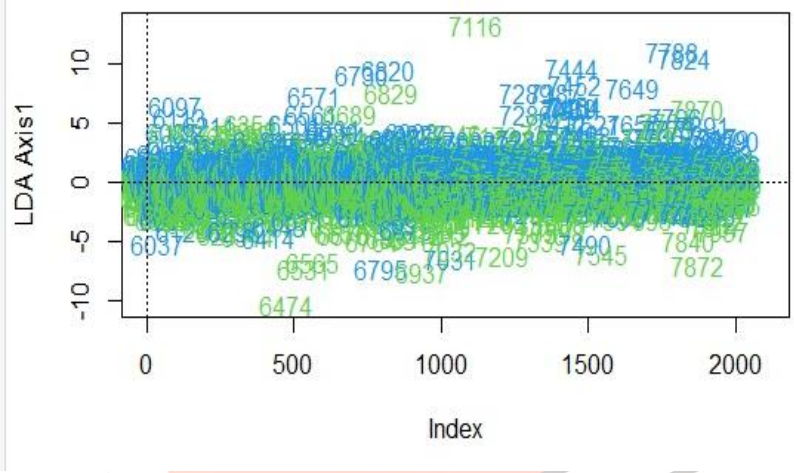


Figure 3 Scatter plot of the discriminant function values for test data

Then we removed the sparse terms from document term matrix having sparsity greater than or equal to 99.9%. In training dataset among 6000 text documents 3291 were labeled as 'hate' and remaining 2709 were labeled as 'nothate'. In testing dataset among 2000 text documents 1085 were labeled as 'hate' and remaining 915 were labeled as 'nothate'. First we trained the LDA with training dataset. After training when LDA predicted the label for test data we got 65.65% accuracy. Figure 2 shows stacked histogram of the discriminant function's values for training observations of two classes. These plots specify the separation between two classes and overlapping areas also where there are chances for mix-ups during prediction of classes. Figure 3 shows the scatter plot of the discriminant function values for test data. Figure 4 shows the confusion matrix and statistics showing the performance of LDA classifier for test data.

```
> conf.mat
Confusion Matrix and Statistics

          Reference
Prediction hate nothate
   hate     758     360
   nothate  327     555

               Accuracy : 0.6565
                 95% CI : (0.6352, 0.6773)
    No Information Rate : 0.5425
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.306

 Mcnemar's Test P-Value : 0.2221

            Sensitivity : 0.6986
            Specificity : 0.6066
         Pos Pred Value : 0.6780
         Neg Pred Value : 0.6293
             Prevalence : 0.5425
         Detection Rate : 0.3790
   Detection Prevalence : 0.5590
      Balanced Accuracy : 0.6526

       'Positive' Class : hate
```

Figure 4 Confusion matrix and statistics showing the performance of LDA classifier

## IV. CONCLUSION

Linear Discriminant Analysis (LDA) is a dimensionality reduction technique but can be used for text classification. In case of text classification higher dimensionality is a problem as it can be observed by seeing the large number of columns in a document term matrix. LDA helps us to manage the curse of dimensionality problem. We did not get very high accuracy in our experiment but found that LDA is quite helpful to cope with curse of dimensionality problem.

## REFERENCES


[1] Georgios K. Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Effective hate-speech detection in Twitter data using recurrent neural networks. Applied Intelligence 48, 12 (December 2018), 4730–4742. DOI:https://doi.org/10.1007/s10489-018-1242-y

[2] Dia AbuZeina and Fawaz S. Al-Anzi. 2018. Employing fisher discriminant analysis for Arabic text classification. Comput. Electr. Eng. 66, C (February 2018), 474–486. DOI:https://doi.org/10.1016/j.compeleceng.2017.11.002

[3] Keinosuke Fukunaga. 1990. Introduction to statistical pattern recognition (2nd ed.). Academic Press Professional, Inc.,USA

[4] Xanthopoulos, P., Pardalos, P. M., & Trafalis, T. B. (2012). *Linear Discriminant Analysis. Robust Data Mining, 27–33.* doi:10.1007/978-1-4419-9878-1_4

[5] Torkkola, Kari. (2001). Linear Discriminant Analysis in Document Classification. IEEE TextDM 2001.

[6] D. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality," Aug. 2000 [Online]. Available: http://www-stat.stanford.edu/~donoho/Lectures/AMS2000/AMS2000.html, American Mathematical Society Lecture-Math Challenges of the 21st Century

[7] S. Ji and J. Ye, "Generalized Linear Discriminant Analysis: A Unified Framework and Efficient Model Selection," in IEEE Transactions on Neural Networks, vol. 19, no. 10, pp. 1768-1782, Oct. 2008, doi: 10.1109/TNN.2008.2002078

[8] Gu Q., Li Z., Han J. (2011) Linear Discriminant Dimensionality Reduction. In: Gunopulos D., Hofmann T., Malerba D., Vazirgiannis M. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2011. Lecture Notes in Computer Science, vol 6911. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-23780-5_45

[9] Nasar Aldian Ambark Shashoa, Nuredin Ahmed, Ibrahim N. Jelta, Omar Abusaeeda. "Classification Depend on Linear Discriminant Analysis Using Desired Outputs". 17th international conference on Sciences and Techniques of Automatic control and Computer Engineering-STA'2016, Sousse, Tunisia, December 19-21, 2016.

[10] Biba M., Gjati E. (2014) Boosting Text Classification through Stemming of Composite Words. In: Thampi S., Abraham A., Pal S., Rodriguez J. (eds) Recent Advances in Intelligent Informatics. Advances in Intelligent Systems and Computing, vol 235. Springer, Cham. https://doi.org/10.1007/978-3-319-01778-5_19