NVEO
Natural Volatiles &
Essential Oils

Classification of Indian English Poetry into Pre-Independence and Post-Independence Eras using Combination of Semantics, Topics and Style features

K.Praveenkumar[1], Venkata Naresh Mandhala [2], Debnath Bhattacharyya[3], Debrup Banerjee[4]

[1]*Research Scholar, Department of CSE, VFSTR Deemed to be University, Guntur, India*
[2]*Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, AP, India*
[3]*Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, AP, India*
[4]*Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, AP, India*
*Corresponding author. Email: *mvnaresh.mca@gmail.com*; azlan@fsmt.upsi.edu.my, *debnathb@kluniversity.in, mail2debrupbanerjee@gmail.com*

Abstract

Automatic classification of poetry era is a challenging task. In Indian English Poetry, the poems are categorized into two eras named Pre-Independence and Post-Independence. The poetry style and themes are changed from one era to another era depending on the authors era that he/she belongs to. Hence, this study is testing of different feature selection methods and ensembled features to identify the poems era automatically. The poetry classification can be carried out on semantics, topics and style features. In this experiment, we have used Latent Semantic Analysis (LSA) to find semantic features, Latent Dirichlet Allocation (LDA) topic modeling to find topic features, along with these phonemics, syntactic elements and structure of poetry as style features. The experiment is carried out on 760 poems written by 28 authors, in this, 344 belongs to Pre-Independence era and 416 belongs to Post-Independence era. The classification accuracy 91.20% is achieved using Random Forest classifier with combination of LSA and LDA feature set. Further with the combination of style features to LSA and LDA features the classification result achieved is 92%. The study showed that the poetry can be classified into different eras with decent accuracy based on the combination of topics, words and style of poetry as features.

Keywords: Classification of Poetry, Indian English Poetry, Pre-Independence and Post-Independence

Introduction

Literature of Indian English poetry analyzes the poetry in two dimensions one is before the independence(Pre-Independence) and another is after the independence(Post-Independence). Indian poetry has more than 175 years of history. The Pre-Independence period is measured from 1820 to 1947/50 and 1950 onwards considered as Post-Independence period. In Pre-Independence era considerable amount of poetry is expressed on topics of Indian philosophy, culture, bhakti etc and further Pre-Independence era is a mix of romantism and patriotism the notable poets are Sarojini Naidu, Rabindranath Tagore and Arabindo. In Post-Independence poetry more Indianness is witnessed, authors vividly expressed societal issues and their own experiences, amongst A.K Ramanujan, R. Parthasarathy and Arun Kolhatkar are notable poets.

Due to complex nature of poetry classification of timeline is a challenging task. For instance, the verse *"I have flung to the East and the West Priceless treasures torn from my breast, And yielded the sons of my stricken womb To the drum-beats of the duty…"* taken from The Gift of India(Sarojini Naidu) conveys the greatness of India and Indian Soldiers at the time of 1915(Pre-Independence). The manual strategy for

detecting the timeline of a poem based on poetry text is difficult and time consuming.

Identifying the changes happened over the years in literature is a challenging task. Indian English poetry changed over the years in terms of topics covered, language used, and the style of expressing the ideas. This study aims to test the applicability of different classification algorithms and ensembled features to classify the poetry with respect to the changes in the language over the 175 years in to two classes Pre-Independence and Post-Independence.

To our knowledge, performing such experiment using supervised machine learning techniques is first time with Indian poetry. The poems used for experiment are downloaded from *poemhunter.com* website. Each poem is kept in a single text file and in total 760 poems are used as data set in this experiment, in these, 344 poems are taken from Pre-Independence era and 416 from post-independence era.

Related work

In this section, a brief review of relevant studies on classification of poetry on different contexts is presented. Jasleen Kaur et.al[1][2] classified Punjabi poetry using text features based on weighted scheme Term Frequency and Inverse Document Frequency (TF-IDF). The authors classified poetry in to 4 classes using Naïve Bayes, Support Vector machine and K Nearest Neighbor classifiers, amongst SVM achieved highest accuracy. ZHONG-SHI HE et.al.[3] classified Chinese poetry into two styles named Bold-and-Unconstrained and Graceful-and-restrained, for this authors used Support Vector Machine classifier with Style features and the results achieved are satisfactory. For this experiment Authors used Vector Space Model and feature selection methods Chi-square, Mutual Information and Cross Entropy. Noraini Jamal et.al[4] classified Malaya Poetry called Pantun in to two classes poetry and non-poetry using SVM classifier. Further authors classified the poetry in to 10 themes. Andr´es Lou et.al[5]. classified the poetry based on their subject category. Authors used TF-IDF weightage and Latent Dirichlet Allocation topic modeling for feature generation. For feature extraction chi-square method is used and for classification SVM algorithm is used.

Many researchers have classified poetry using phonemic (sound devices) features such as rhyme, rhythm, meter, assonance and alliteration. This section discuss about the classification task performed using phonemic features. Chris Tanasescu et.al.[6]classified poetry using meter as a feature. The authors for their experiment used an application called scandroid as a feature extractor. Finally, authors proposed a new rhyme detection method. Timo Baumann[7] classified read out poetry using Neural Network method. Authors compared automatic features selection with manual features selection methods and found that automatic feature selection using Neural network technique is producing better classification results. For this, authors used the grouping structure, the metrical structure, the time-span-variation, and the prolongation as features.

A notable number of works are reported on poetry classification using state-of-art deep learning techniques. In this section we briefly review the classification tasks reported using deep learning methods. SHAKEEL AHMAD et.al.[8]proposed C-BiLSTM technique to classify the poetry based on the emotions. Emotions are extracted from the verse of poetry, in this work authors considered 9 emotions and shows that this method better classifies the poetry than the conventional classification method such as Decision Tree, Random Forest and KNN. Pengfei Liu et.al[9]. in their research paper developed a multitask learning framework where a neural network learns for multiple tasks at a time. authors tested the method on 4 benchmark data sets. This is a multilayer model in these base layers are common layers which will learn commonly after that layers will be divided to learn specific task. Rie Johnson et.al.[10] proposed a method named low-complexity word-level deep convolutional neural network (CNN) for text classification. This method captures the global representation of text by analyzing deeply the word-level CNN.

Background

This section discuss briefly about the technologies and classifiers used in this experiment.

Vector Space Model

To classify the documents understanding the contents of the documents is very important. One simple method is to represent entire document with some important words which describes the meaning of whole document[11]. But to compare these words with other documents need to be represented in a vector space model, where each word is quantified and whole document is represented as vector of these quantified values.

Vector space is a group of vectors. To get a vector space , a term by document matrix is made in that rows represent words and columns represent documents, each document is represented as a vector and each value in vector is a quantified value of a word in the document. Commonly used quantifying scheme of a word is Term Frequency and Inverse Document Frequency (TF-IDF) and mathematically it can be represented as shown in equation (1)

$$W_{t,d} = tf_{t,d} \,.\, log \frac{|D|}{|\{d' \epsilon D | t \epsilon d'\}|} \qquad (1)$$

$tf_{t,d}$is term frequency of term *t* in the document *d* and $. log \frac{|D|}{|\{d' \epsilon D | t \epsilon d'\}|}$ is inverse document frequency where $|D|$ is number of documents and $|\{d' \epsilon D | t \epsilon d'\}|$ is the number of documents containing term *t*. Here if a term *t* appears frequently in a document the term *t* is important, but if the term *t* appears in many documents then the term *t* is a common term which does not contribute much in distinguishing the document from other documents. To classify the documents each document is represented with TF-IDF weighted terms as shown in equation 2.

$$d_j = (w_{1,j}, w_{2,j}, \dots \dots w_{n,j}) \qquad (2)$$

Latent Semantic Analysis (LSA)

In VSM, every document is expressed as the weighted terms, but due to polysemy and synonymy VSM cannot express semantics of document in a better manner. To express semantics of document in a better manner LSA algorithms is used[12]. The typical LSA process is: Document Term Matrix M is constructed from n number of documents with m terms. The Matrix M is resolved by Singular Value Decomposition(SVD) into Term-Vector matrix U, Document-Vector matrix D and a diagonal matrix S.Among these U and D matrices are ortho normal. From these k dimensions are selected to reduce the sparseness of matrix and finds the truncated matrices U_k, D_k and S_k, by multiplying these truncated matrices results into a new reduced dimensional matrix Mk, which is a least squares best fit approximation of given matrix M with k singular values.

Latent Dirichlet Allocation (LDA)

LDA is a popular model for topic modeling. LDA is an unsupervised technique for extracting topics from a set of documents. The Idea behind this method is that the documents are represented as a random mixture of topics where each topic is a probability distribution over words. Classifying documents based on LDA based topic distribution is widely accepted technique[13].LDA process as follows: let D={d1,d2,….dn} document collection and V={w1,w2…..wm} is vocabulary of corpus. A topic Z_j is represented as multinomial probability distribution over words. Further each document is represented as a probability distribution of k number of topics where k is a user given parameter.

Style features

To classify poetry, considering only word occurrences is not enough[14] because poetry has many features such as rhyme, meter, assonance, alliterations etc. This section discuss about the style features considered for poetry classification. Kaplan in his work categories the style features in to Orthographic, Phonemic and syntactic. Orthographic features are computed based on the words appeared in the poetry, which includes word count, number of lines, number of stanzas, number of lines per stanza, frequencies of noun, adjective and verb etc. Under Syntactic features category frequency of Parts of Speech (POS)tags are considered, these POS tags expresses the level of formality in the poem. Under Phonemic category all kinds of sound elements are considered that includes assonance, alliteration, rhyme, semi slant rhyme etc. These metrics are computed using Poetry Analyzer application developed by Kaplan.

Classification techniques

This section provides a brief description of the classification algorithms used in this experiment. In this experiment 4 prominent classification algorithms are used Logistic regression, Random Forest, Support Vector Machine (SVM) and K Nearest Neighbor (KNN).

Logistic Regression

It is a simple and elegant classification algorithm. This algorithm classifies based on the probability of an observation rather than its value. In binary classification the algorithm generates a probability then based on the threshold value given by user it is usually 0.5, if the probability is below 0.5 assigned one class else the other class. This algorithm uses sigmoid function for detecting the class transition.

Random Forest

Random forest is an ensemble method, where n number of decision trees are constructed with randomly selected features and the majority class is considered as the resultant class of the algorithm. This ensemble classifier is more robust than decision tree.

Support Vector Machine

SVM is a supervised non-linear classifier that is applied on nonlinear data classification tasks. In SVM the challenging task is to find the hyperplane that has maximum margin between two classes. With Maximum margin classification of new data becomes easy. Support vectors are the data points that are close to the hyper plane using these support vectors we maximize the margins of classifier.

K Nearest Neighbor

KNN is a simple and basic algorithm that is used for classification in machine learning area. This algorithm classifies new data based on the nearest neighbors for this it used distance function. In this algorithm K is the number of neighbors, it is a user defined parameter. Among the K nearest neighbor's majority class is identified and the new data point is assigned the majority class.

Methodology

Figure 1 shows the experiment procedure, initially collected the poems from poem hunter.com website, later with the help of literature poems are categorized into Pre-Independent era and Post-Independent era. Before 1950 is called as Pre- Independence era in this author themes are mostly covered with romance, bhakti and patriotism. After 1950 poetry is called post-independence era in this author themes are mostly society issues and the style and word usage methods are different in these eras. Because of this we experimented by building the classification models to exhibit the performance of classification algorithms in

classifying these eras.

Total poems are 760 with 28 authors, in this 314 are pre independence poems and 416 are Post independence poems. As a next step we performed preprocessing in this stop words are removed, numbers, special symbols and unwanted spaces removed. Next we have tested the classifier performances with 3 different features semantic, topic and style independently later we combined 2 features in all combinations and finally we tested by combining all 3 features .

Phase 1 Independent features

LSA based features: For this we used SVD function, this is implemented in R programming language. Initially computed term document matrix with weight TF-IDF, this matrix is given to SVD function, it returns 3 matrices USD, among this S is diagonal matrix and U and D matrices are orthonormal, with the help of scree plot we have selected k features in scree plot where the bend occurs there we have considered the k value, we performed this experiment by considering k values 10,15 and 20. To this dimension reduced matrix class labels are allocated and given as input to the classification algorithms.

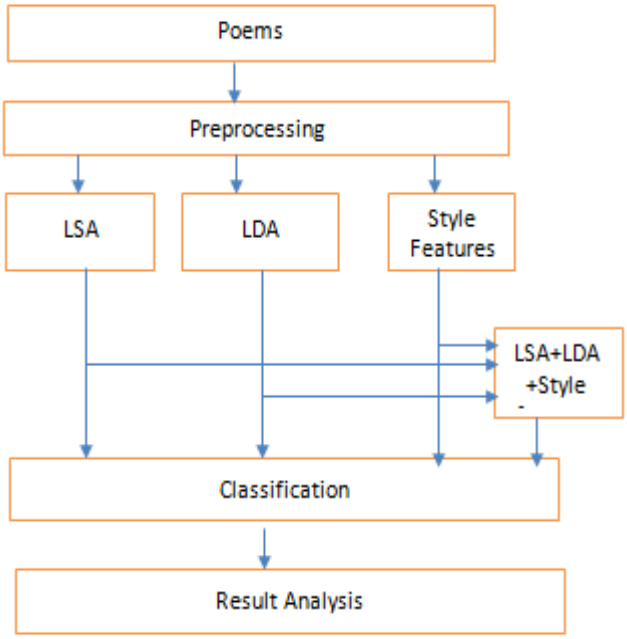

Figure 1. Procedure Followed for the Experiment

LDA based features: To implement LDA topic modeling we have used Gensim package in python. Initially tokenized the poems using word tokenize function then all the words are converted to lower case later stop words are removed for this standard stop words are considered. After that word dictionary created later by suing filter extremes function rare words and very frequent words are removed, then this dictionary of words are given to LDA function, LDA function is implemented by considering topic values 5,10,15 and 20 topics respectively and evaluated the classification methods.

Style features: To compute the style features, POETRY ANALYZER tool is used. The poems are placed in separate text files and given the set of files as input to the tool, for each poem it generates 84 style features which are in 3 categories syntactic, phonemic and structural. While computing, all the values are normalized to be the value range between 0 and 1. To this feature set class labels are assigned and given as input to classification algorithms and analyzed the results.

Phase 2 ensembled features

In this section we describe how we have prepared ensembled feature set for classification of poetry. We have taken all possible combination of all 3 LSA, LDA and Style features.

As a first combination LSA and Style features are combined and given as input to the classification algorithms. Here semantic features and style features combination is tested against classification of poetry. Second combination is taken from LDA topics and style features, in this different topics of poetry and style features tested against classification of poetry and third combination is taken from LSA and LDA here semantic features and topics combined and tested against the classification and finally combined all three features LSA, LDA and style and tested against classification of poetry. the idea is to test how the ensembled feature set improves the classification accuracy.

Results and discussion

Table 1 shows the classification accuracy of 4 classifiers with LDA topic modeling. This experiment is evaluated for different number of topics 5, 10,15 and 20 at topic 20 the classification accuracy is falling so the experiment is stopped. Maximum accuracy achieved with Random Forest classifier with 10 topics is 70% . Here each poem is considered as a document and a poem length can be maximum of 8 stanzas, in this context a poem can contain less number of distinct topics when compared with large size text documents. Because of this the experiment is started with only 5 topics, further a poem very subtly expresses the concept with the help of various literary devices such as metaphor, simily etc. in this regard the vocabulary used by the poet will be in general more distinct. Due to these 10 topics this model could able to classify the poetry with 70% accuracy. Along with this Random forest classifier used randomly chose features and multiple decision trees due to this the classification accuracy is high with this classifier. Other classifiers also performed similar RF classifier.

Table 1. Results of 14760 words using LDA for pre and post-Independence poetry classification

Number of Topics chose for LDA	Classifier	Accuracy	Precision	Recall	F1 score
5	Logistic Regression	59.3	58	59	57
	Random Forest	63	64	63	63
	SVM	60	59	60	58
	KNN(n=3)	62	63	62	63
10	Logistic Regression	69	68	69	67
	Random Forest	69	70	69	69
	SVM	69	69	69	68
	KNN(n=3)	66	66	66	66
15	Logistic Regression	62	61	62	61
	Random Forest	61	62	61	61
	SVM	67	66	67	65
	KNN(n=3)	62	63	62	63
20	Logistic Regression	58	57	58	57
	Random Forest	58	57	58	58
	SVM	60	59	60	59
	KNN(n=3)	59	58	59	58

The same experiment performed after filtering the extreme words using filter extreme function of LDA dictionary. After filtering the dictionary size is reduced from 14,760 words to 4189. The results are shown in table 2. Topic modeling assigns topics based on the frequency of words i.e. if word w1 is assigned a topic with x probability in topic A and same word is assigned with y probability in topic B then based on the majority the word probability is maximized. Because of this rare words and very frequent words are filtered with this the topic distribution among the documents is improved, due to this the classification accuracy with 20 topics increased to 74% with Logistic regression classifier, second highest 71% accuracy is achieved by SVM classifier. In SVM Radial basis Function(RBF) kernel is used.

Table 2. Results of LDA topic modeling with words 4189

Number of Topics chose for LDA	Classifier	Accuracy	Precision	Recall	F1 score
5	Logistic Regression	57	59	57	58
	Random Forest	61	61	61	61
	SVM	58	61	58	58
	KNN(n=3)	54	56	54	55
10	Logistic Regression	57	56	57	56
	Random Forest	56	57	56	56
	SVM	57	58	57	57
	KNN(n=3)	57	58	57	57
15	Logistic Regression	52	55	57	55
	Random Forest	60	60	60	60
	SVM	55	53	55	54
	KNN(n=3)	57	56	57	56
20	Logistic Regression	74	73	74	73
	Random Forest	64	64	64	64
	SVM	71	71	71	70
	KNN(n=3)	60	61	60	60

Table 3 shows LSA feature based classification results. In this experiment number of feature vector selection is performed with the help of scree plot of diagonal matrix which is produced by SVDfunction. A scree plot plots the eigen values of factors in the form of a line, in general this plot is used to determine number of factors . it displays the eigenvalues in a downward curve ordering largest value to smallest value and the “elbow” that is a significant fall of eigenvalues is considered as the point, the factors which are left to this point are considered significant factors. In our context the vector values which are left to the elbow point are considered are distinguishing semantic topics useful to classify the poetry. this elbow point is found at nearly 20, so the experiment is carried out with 10 , 20 and 30 to test the accuracy of classification. The result shows that with 20 and 30 factors the results are similar. With 20 factors RF classifier produced 90% accuracy the same with 30 factors. The SVM classifier produced second highest classification accuracy 88% with 20 factors it is 1% higher when compared with 10 and 30 factors.

Table 3: LSA/SVD based classification result with 10,20 and 30 feature vectors

Number of Vector selected from LSA(SVD)	Classifier	Accuracy	Precision	Recall	F1 score
10	Logistic Regression	68	77	68	61
	Random Forest	89	90	89	89
	SVM	87	88	87	88
	KNN(n=3)	87	87	87	87
20	Logistic Regression	68	78	68	62
	Random Forest	90	90	90	90
	SVM	88	89	88	87
	KNN(n=3)	82	82	82	82
30	Logistic Regression	69	75	69	64
	Random Forest	90	90	90	90
	SVM	86	87	86	86
	KNN(n=3)	85	85	85	85

Table 4 shows the results of Style features-based classification results. In this experiment 84 style features are considered. In this experiment tested the classification of era in poetry style. The maximum accuracy achieved is by Random Forest with 83% classification accuracy, second highest accuracy 74% is achieved

with SVM. These results shows that in comparison with LSA there is less distinction in the style of both pre-Independence and Post-Independence poetry, but when compared with topic distinction style distinction is more.

Table 4: Style features-based classification result

Classifier	Accuracy	Precision	Recall	F1 score
Logistic Regression	72	72	72	72
Random Forest	83	83	83	83
SVM	74	74	74	74
KNN(n=3)	67	69	67	67

Table 5 shows the ensemble features classification result, It shows that combination of LSA and style features can able to classify the poetry with highest accuracy 89% with precision 90%, when compared with LDA and Style combination features LSA and Style better classifies the poetry. LSA and LDA combination produced 91% accuracy, we can understand that the topic distinction and semantic distinction together classifies better than the previous two methods. Finally, when we add the style features to the LSA and LDA features the classification accuracy increased 1% and the classifier Random Forest classifies the poetry with 92% accuracy. Individually recall for this classifier is 96% for post-independence poetry.

Table 5 Ensemble methods results

Method	Classifier	Accuracy	Precision	Recall	F1 score
LSA+STYLE	Logistic Regression	75	74	75	74
	Random Forest	89	90	89	89
	SVM	78	77	78	77
	KNN(n=3)	68	69	68	68
LDA+STYLE	Logistic Regression	76	76	76	76
	Random Forest	87	87	87	87
	SVM	77	77	77	77
	KNN(n=3)	57	58	57	57
LSA+LDA	Logistic Regression	77	77	77	77
	Random Forest	91.2	91	91	91
	SVM	72	72	72	71
	KNN(n=3)	76	76	76	76
LSA+LDA+Style	Logistic Regression	77	77	77	77
	Random Forest	92	92	92	92
	SVM	78	78	78	78
	KNN(n=3)	60	61	60	60

Conclusion

To classify the Indian English Poetry in to two classes Post-Independence era and Pre-Independence era, we have used semantic, topic and Style based features of poetry. Further we experimented with ensemble of the 3 methods with all combinations. For this experiment the poems are initially preprocessed and for semantic features LSA(SVD) method used by representing the words with TF-IDF weights, for topic distribution LDA topic modeling method is used and style features are computed using Poetry Analyzer tool. The semantic features achieved the classification accuracy 90% and ensembled features of LSA, LDA and Style improved the accuracy to 92% using Random Forest classifier, with this result we can conclude that though semantic features can express the document meaning better, further if we add style features the

classification accuracy can be increased with poetry.

The extension to the experiment can be using different probabilistic method in LDA topic modeling for maximize the approximation and more style features such as simile, metaphor can be computed for better classification. This experiment is performed exclusively on Indian Poetry, the same method can be evaluated on Poetry of other country.

REFERENCES

J. Kaur and J. Saini, "Designing punjabi poetry classifiers using machine learning and different textual features," *Int. Arab J. Inf. Technol.*, vol. 17, no. 1, pp. 38–44, 2020.

J. Kaur and J. R. Saini, "PuPoCl : Development of Punjabi Poetry Classifier Using Linguistic Features and Weighting," *Infocomp*, vol. 16, no. 1–2, pp. 1–7, 2017.

Z. S. He, W. T. Liang, L. Y. Li, and Y. F. Tian, "SVM-based classification method for poetry style," *Proc. Sixth Int. Conf. Mach. Learn. Cybern. ICMLC 2007*, vol. 5, no. August, pp. 2936–2940, 2007.

S. A. N. Noraini Jamal, Masnizah Mohd, "Poetry Classification Using Support Vector Machines Noraini Jamal , Masnizah Mohd and Shahrul Azman Noah Knowledge Technology Research Group , Faculty of Information Science and Technology ," vol. 8, no. 9, pp. 1441–1446, 2012.

A. Lou, D. Inkpen, and C. T. Margento, "Multilabel subject-based classification of poetry," *Proc. 28th Int. Florida Artif. Intell. Res. Soc. Conf. FLAIRS 2015*, pp. 187–192, 2015.

C. Tanasescu, B. Paget, and D. Inkpen, "Automatic classification of poetry by meter and rhyme," *Proc. 29th Int. Florida Artif. Intell. Res. Soc. Conf. FLAIRS 2016*, vol. 5, pp. 244–249, 2016.

T. Baumann, H. Hussein, and B. Meyer-Sickendiek, "Analysis of Rhythmic Phrasing: Feature Engineering vs. Representation Learning for Classifying Readout Poetry," *Proc. Second Jt. {SIGHUM} Work. Comput. Linguist. Cult. Heritage, Soc. Sci. Humanit. Lit.*, no. 4, pp. 44–49, 2018.

S. Ahmad, M. Z. Asghar, F. M. Alotaibi, and S. Khan, "Classification of Poetry Text into the Emotional States Using Deep Learning Technique," *IEEE Access*, vol. 8, pp. 73865–73878, 2020.

P. Liu, X. Qiu, and H. Xuanjing, "Recurrent neural network for text classification with multi-task learning," *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2016-Janua, pp. 2873–2879, 2016.

R. Johnson and T. Zhang, "Deep pyramid convolutional neural networks for text categorization," *ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.*, vol. 1, pp. 562–570, 2017.

R. Ju, P. Zhou, C. H. Li, and L. Liu, "An efficient method for Document categorization based on Word2vec and latent semantic analysis," *Proc. - 15th IEEE Int. Conf. Comput. Inf. Technol. CIT 2015, 14th IEEE Int. Conf. Ubiquitous Comput. Commun. IUCC 2015, 13th IEEE Int. Conf. Dependable, Auton. Se*, pp. 2276–2283, 2015.

F. Wild and C. Stahl, "Investigating Unstructured Texts with Latent Semantic Analysis," pp. 383–390, 2007.

M. Allahyari *et al.*, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques," 2017.

D. M. Kaplan and D. M. Blei, "A computational approach to style in American poetry," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, pp. 553–558, 2007.