Mehran University Research Journal of Engineering & Technology
Vol. 38, No. 4, 935-944 October 2019
p-ISSN: 0254-7821, e-ISSN: 2413-7219
DOI: 10.22581/muet1982.1904.07

Identification of Urdu Ghazal Poets using SVM

NIDA TARIQ*, IQRA EJAZ*, MUHAMMAD KAMRAN MALIK*, ZUBAIR NAWAZ*, AND FAISAL BUKHARI*

RECEIVED ON 08.06.2018 ACCEPTED ON 30.10.2018

ABSTRACT

Urdu literature has a rich tradition of poetry, with many forms, one of which is Ghazal. Urdu poetry structures are mainly of Arabic origin. It has complex and different sentence structure compared to our daily language which makes it hard to classify. Our research is focused on the identification of poets if given with ghazals as input. Previously, no one has done this type of work. Two main factors which help categorize and classify a given text are the contents and writing style. Urdu poets like Mirza Ghalib, Mir Taqi Mir, Iqbal and many others have a different writing style and the topic of interest. Our model caters these two factors, classify ghazals using different classification models such as SVM (Support Vector Machines), Decision Tree, Random forest, Naïve Bayes and KNN (K-Nearest Neighbors). Furthermore, we have also applied feature selection techniques like chi square model and L1 based feature selection. For experimentation, we have prepared a dataset of about 4000 Ghazals. We have also compared the accuracy of different classifiers and concluded the best results for the collected dataset of Ghazals.

Key Words: **Text classification, Support Vector Machines, Urdu poetry, Naïve Bayes, Decision Tree, Feature Selection, Chi Square, k-Nearest Neighbors, Ghazal, L1, Random Forest.**

1. INTRODUCTION

Poetry is a form of literature in which a poet expresses his feelings in a rhythmic manner. Great writers and poets of every era have contributed to the literature on different themes with their unique writing style. Since the 17th century, when Urdu was declared the official language, we have seen many famous poets as the likes of Ghalib, Meer, Anees, Dard, Daag, Zauq, Iqbal, Akbar, Josh, Firaq, Faiz and Faraz to name a few [1]. Urdu has a rich tradition of poetry and has many forms. Ghazal is one of the most famous forms of poetry in Urdu, which describes love in a rhythmic way [2]. Like poetry in other languages, love has been the most popular topic amongst Urdu poets. An interesting work is to identify poets from the given poetry, since each poet has a different style of writing and content. Poet identification may help us to fight plagiarism.

A very little work has been done in poetry classification in any language. In the Malay language, Noraini et. al. [3] have classified the theme of the poetry. The apparent reason for little work done in Urdu and Urdu poetry is its different and complex language structure. As poetry can

Authors E-Mail: (mscsf16m005@pucit.edu.pk, iqraijaz44@gmail.com, kamran.malik@pucit.edu.pk, z.nawaz@pucit.edu.pk, faisal.bukhari@pucit.edu.pk)
* College of Information Technology, University of the Punjab, Lahore, Pakistan.

be very different in terms of lines, stanzas, rhyme and style of rhythm. Classifying Urdu poetry thus is much trickier.

This paper presents the evaluation experiments on the performance of various machine learning algorithms as the likes of SVM [4-5], KNN [6], Random forest [7], Decision Tree [7] and Naïve Bayes [8-9] for Urdu poet identification. The application of machine learning is now very common in classification and identification of text, e.g. classifying emails as spam [10], identifying authors in English text [11] and text classification of news in Urdu [12]. In this study, we have focused on poet identification in Ghazals only. Ghazals, in Urdu Poetry, consists of a complex structure. Each sher/verse of a Ghazal can contain two different thoughts which make it difficult to classify. Two main factors that can be used for categorization of Ghazal text, as presented in [13-14], are content and the style. In Urdu poetry, each poet can have a varying degree of styles and themes. In this study, we have applied five different classification models and compared their results according to their accuracy scores. Also, we have analyzed the classifier performance by applying chi-2 and L1 based feature selection and observed changes in performance scores.

The rest of the paper is organized as follows. We discuss the background and related work in Section II, and present the methodology in Section III. Section IV presents the experimental setup, results, and discussions. Finally, the paper is concluded in Section V.

2. BACKGROUND AND RELATED WORK

In particular Urdu Poetry has many Principals such as Ghazal, Nazm, Hamd, Manaqbat etc but in this study we only focus on Ghazal but the same idea can be extended to other principals as well. Ghazal can consist of different, two liner couplets called Sher with same ending rhyme. Ghazals can have different themes which are mainly relevant to authors who have a different writing style for the same theme. There is a minimum of five couplets in each sher and couplets does not always have the same thoughts. Ghazals are one of the most difficult forms of poetry due to its writing parameters.

Research on automatic identification of Urdu poet is little to none. However, there are studies which use classification techniques to classify documents or text according to genres and authors for languages such English [13-15], Malay [3], Chinese [16], and Punjabi [17].

The Urdu language originated from Arabic language and various computational linguistic tasks like name entity recognition, part of speech tagging and machine translation have been performed on Urdu [18-22]. Fewer studies have been done to classify Arabic text of different categories such as sports, politics etc. using Decision Trees [23]. Work related to poetry classification has been done in different languages such as discussed in [23], which use SVM techniques to classify poetry according to themes and differentiate between poetry and non-poetry text. Naïve Bayes [8-9], SVM [4-5], KNN [6], Decision Trees [7], and Random Forest [7] are more widely used supervised learning methods for classification. However, in text classification problem, SVM usually beats them all [4,8,24].

3. THE PROPOSED METHODOLOGY

In order to perform the poet classification, we have adopted the methodology shown in Fig. 1. The Fig. 1 has three main sections namely Data Collection, Data

Preprocessing and Poet Identification. Since no one has done a similar work in Urdu language, we need to prepare such data for the first time. For this, we have scraped Urdu ghazal data from various sites, tagged ghazals with their respective poets and stored in a database in the first section. In second section, we have applied few preprocessing techniques to convert it into an acceptable format. Afterward in Poet Identification section, we have applied some feature selection algorithms to choose only the important features. Finally, we have applied various machine learning classifiers to identify the poets.

Data Collection: In supervised machine learning algorithms, extensive training is required to develop a reasonable model. For this purpose, we wrote scrappers as well as collected data manually from different poetry websites, e.g. rakhta.pk, urdupoint.com. In total, we collected 3967 couplets of four different poets, having more than one million tokens with vocabulary size of 6427. Data were collected in CSF (Comma Separated File) and tagged manually. We collected Ghazals from the following poets: Mir Taqi Mir, Ahmed Faraz, Mirza Ghalib and Zafar Iqbal. These poets are the classes used to identify poets from our couplets data.

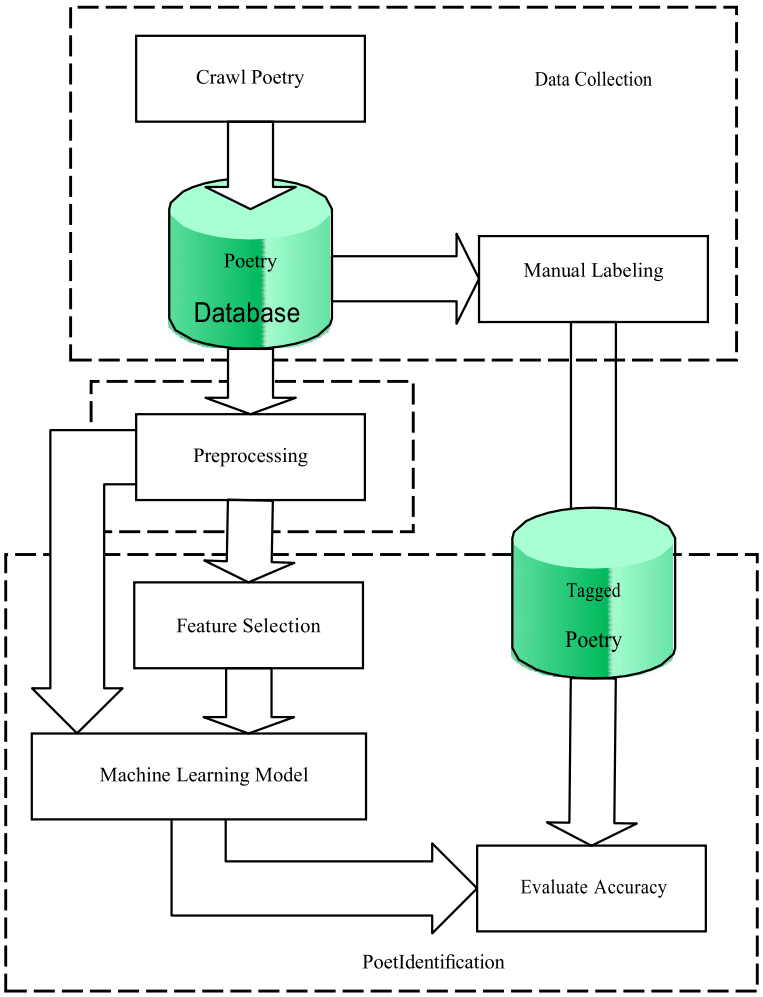

FIG. 1. THE PROPOSED ARCHITECTURE

Preprocessing: We cannot directly apply any machine learning models on the collected data. We need to do some preprocessing to transform into an acceptable format. First, we apply tokenization to separate word from the couplets by using white space (space, tab, newline) as the delimiter. Then, we create the TDF (Term- Document-Frequency) matrix, where each row in the matrix corresponds to a sher and each column corresponds to a term (word) in the sher. The preprocessing step can be explained by the following simple example. Let's say we have three sentences.

(1) He eats apple

(2) I am drinking water

(3) He eats food

As a first step, we build a unique vocabulary He, eats, apple, I, am, drinking, water, food from three sentences, thus the count is 8 unique words. In second step we build a TDF matrix of size 3x 8, where 3 represents total number of documents and 8 represents unique words. In the last step, we will fill the matrix as shown in Table 1. Each entry of the Table 1 shows the number of occurrences of a word in the particular document.

Poet Identification: This section is further subdivided into sub-sections.

Feature Selection: After preprocessing step, we perform feature selection step to select the most relevant features for the poet identification. The feature selection methods applied are chi-square and L1-based. The reason to choose these methods is that the first is based on statistical and second is based on select from model approach. These both are the two widely used high level techniques for feature selection. The Chi-Square (χ^2) [25-27] is a feature selection method that computes the chi square statistics to evaluate the feature importance with respect to the classes. If the term and class are independent then its score is equal to 0, otherwise 1. A term with a higher chi-squared score is more informative. The L1-based feature selection method is based on linear regression where regression coefficients are penalized with L1 penalty forcing many of them to zero. The leftover non-zero coefficients are selected as features [28-29].

Applying Machine Learning Algorithms: Now, the data is in good shape to apply any machine learning algorithm. We have used five different machine learning classifiers for poet identification. The chosen classifiers cover a range of popular modes of classification: Naïve Bayes [8-9], decision trees [7], support vector machines [4-5,30], KNN [7] and Random Forest [7].

2. EXPERIMENTS

We have performed extensive experiments to evaluate our methodology and finalize the best classifier for such problem. The data were gathered both manually and

TABLE 1. TERM DOCUMENT FREQUENCYMATRIX

Document	He	Eats	Apple	I	Am	Drinking	Water	Food
1	1	1	1	0	0	0	0	0
2	0	0	0	1	1	1	1	0
3	1	1	0	0	0	0	0	1

through scrappers. For evaluation, the data is divided into two parts: training data and testing data. Out of 3967 couplets, 80% of the documents are training dataset and 20% as a testing dataset as shown in Table 2. We have performed three types of experiments i.e. poet identification without any feature selection; poet identification with Chi-2 feature selection; and poet identification with L1 based feature selection. The precision, recall and F1-measure scores in percentages are calculated to evaluate our five classifiers. Since we have a multi-class problem, F1-measure is considered as a more representative metric. Implementation of the system is done in Python language using scikit-learn [31] library.

Results: This section describes the results achieved after applying five different machine learning algorithms on the dataset with and without the two feature selection methods. The purpose of designing the experiments in this way is to evaluate the various set of features and machine learning models. Finally, we evaluate the best set of features along with the best model for poet identification.

The first set of experiments are performed along with the original set of features, that are extracted after applying the preprocessing and on the TDF matrix. The results for this set of experiments are summarized in Table 3.

The second set of experiments are performed after applying the Chi-2 feature selection algorithm to the TDF matrix. The Chi-2 feature selection algorithm selects the best features. The results for this set of experiments are summarized in Table 4.

The third set of experiments is performed after applying the L1 based feature selection algorithm to the TDF matrix. The results for this set of experiments are summarized in Table 5.

The overall F1-measure score for all the sets of experiments is further summarized in the plot shown in Fig. 2. The plot shows that SVM easily beats all the other classifiers, no matter feature selection algorithm is used or not. The second-best performing classifier is Naïve Bayes. The rest of the classifiers performed poorly. Even after applying the feature selection algorithms, the F1-measure score is not improved significantly in the case of Decision Tree, Random Forest and KNN classifier. The feature selection algorithm improved the F1-measure score reasonably for SVM and Naïve Bayes especially Naïve Bayes. Both the feature selection algorithms produce the overall best F1-measure, when SVM is used as a classifier.

TABLE 2. CORPUS DETAIL

Poet	Total Sher	Training	Testing
Mir Taqi Mir	962	770	192
Ahmed Faraz	959	766	193
Mirza Ghalib	1033	826	207
Zafar Iqbal	1013	810	203
Total	3967	3172	795

Performance wise, KNN is the slowest and Naïve Bayes is the fastest classifier. The fast performance of the Naïve Bayes can be attributed to its simplistic computation [7]. SVM performs better and shows higher F1-measure score as it is flexible in term of errors and penalties and scale better for large data. SVM supports dense and sparse input in our context Urdu Poetry. Direct comparison with previous work is difficult due to limited work in Urdu linguistic and Urdu Poetry. However, since Urdu has its roots in Arabic language, therefore Urdu

TABLE 3. POET IDENTIFICATION (%) WITHOUT FEATURE SELECTION

Classification	Poets	Recall	Precision	F1-Measure
SVM	Faraz Ahmed	64	66	64
	Mir Taqi Mir	66	73	69
	Mirza Ghalib	76	66	70
	Zafar Iqbal	68	69	68
	Total	69	69	69
Decision Tree	Faraz Ahmed	40	43	41
	Mir Taqi Mir	48	48	48
	Mirza Ghalib	50	39	43
	Zafar Iqbal	39	46	42
	Total	44	44	44
Random Forest	Faraz Ahmed	45	51	47
	Mir Taqi Mir	47	64	54
	Mirza Ghalib	58	39	46
	Zafar Iqbal	54	46	49
	Total	51	50	50
Naïve Bayes	Faraz Ahmed	45	67	53
	Mir Taqi Mir	56	45	49
	Mirza Ghalib	66	49	56
	Zafar Iqbal	64	63	63
	Total	58	56	57
K-Nearest Neighbor	Faraz Ahmed	36	48	41
	Mir Taqi Mir	43	66	52
	Mirza Ghalib	69	17	27
	Zafar Iqbal	45	42	43
	Total	48	43	45

poetry classification results can be compared with previous research related to Arabic text classification. Mesleh has applied chi-2 feature selection on Arabic text classification and then used SVM to get an average F-measure of 88.11% [26]. Similarly, Thabtah et. al. [32] have achieved an F-measure of 74% on Arabic text classification by using Naïve Bayes as machine learning algorithm and Chi-2 for feature selection. They both used Chi-2 for feature engineering and SVM and Naïve Bayes like us, and their F-measure is also similar to ours. These results show that our study align with the results obtained for automatic Arabic text classification.

TABLE 4. POET IDENTIFICATION (%) WITH CHI -2 FEATURE SELECTION

Classification	Poets	Recall	Precision	F1-Measure
SVM	Faraz Ahmed	69	70	69
	Mir Taqi Mir	72	75	73
	Mirza Ghalib	79	71	74
	Zafar Iqbal	70	73	71
	Total	73	72	72
Decision Tree	Faraz Ahmed	42	49	45
	Mir Taqi Mir	49	49	49
	Mirza Ghalib	50	41	45
	Zafar Iqbal	40	42	40
	Total	45	45	45
Random Forest	Faraz Ahmed	45	52	48
	Mir Taqi Mir	51	64	56
	Mirza Ghalib	57	43	49
	Zafar Iqbal	50	43	46
	Total	51	50	50
Naïve Bayes	Faraz Ahmed	61	69	64
	Mir Taqi Mir	57	50	53
	Mirza Ghalib	71	48	57
	Zafar Iqbal	62	82	70
	Total	63	62	62
K-Nearest Neighbor	Faraz Ahmed	35	40	37
	Mir Taqi Mir	43	71	53
	Mirza Ghalib	62	32	42
	Zafar Iqbal	53	38	44
	Total	48	45	46

TABLE 5. POET IDENTIFICATION (%) WITH L1 BASED FEATURE SELECTION

Classification	Poets	Recall	Precision	F1-Measure
SVM	Faraz Ahmed	69	69	69
	Mir Taqi Mir	71	78	74
	Mirza Ghalib	79	71	74
	Zafar Iqbal	70	71	70
	Total	72	72	72
Decision Tree	Faraz Ahmed	42	46	43
	Mir Taqi Mir	42	46	43
	Mirza Ghalib	55	46	50
	Zafar Iqbal	43	44	43
	Total	47	46	46
Random Forest	Faraz Ahmed	42	46	43
	Mir Taqi Mir	48	58	52
	Mirza Ghalib	60	43	50
	Zafar Iqbal	51	51	51
	Total	50	50	49
Naïve Bayes	Faraz Ahmed	77	61	68
	Mir Taqi Mir	60	71	65
	Mirza Ghalib	80	55	65
	Zafar Iqbal	66	89	75
	Total	71	69	70
K-Nearest Neighbor	Faraz Ahmed	41	35	37
	Mir Taqi Mir	38	79	51
	Mirza Ghalib	61	25	35
	Zafar Iqbal	52	38	43
	Total	48	44	46

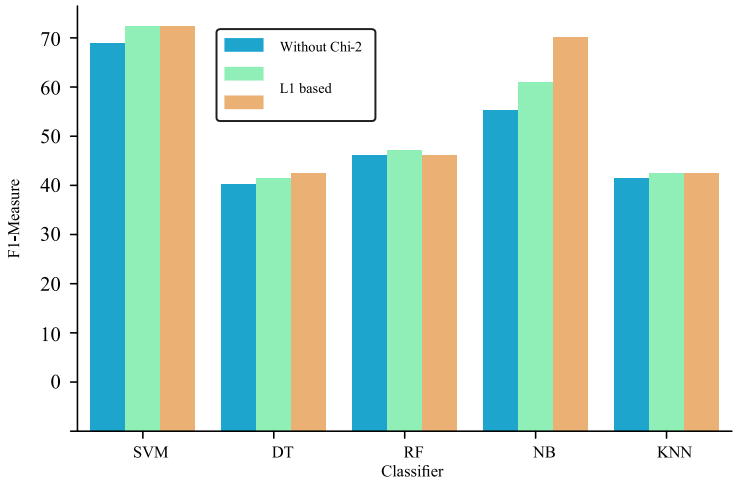

FIG. 2. GRAPHICAL REPRESENTATION OF CLASSIFIERS' PERFORMANCE

3. CONCLUSION

Urdu poet identification is a challenging and interesting research. Unfortunately, it has been largely overlooked in the past studies. This study presents an experimental work on automatic identification of poets from Urdu poetry using five classification techniques i.e. SVM, Random Forest, KNN, Decision

Trees and Naïve Bayes. A collection of 4000 sher was gathered from various resources for experimenting with the aforementioned classification. Overall, the SVM performed the best and achieved 72% F1-measure score. We have used two widely used feature selection techniques namely L1 based and chi-2 based feature selection algorithms. In the future, research will be carried out to identify suitable poetic features that may improve the reliability of the identification.

ACKNOWLEDGMENT

Authors are grateful to Prof. Dr. Syed Mansoor Sarwar, Principal, Punjab University College of Information Technology, Lahore, Pakistan, for providing infrastructure to perform extensive experiments.

REFERENCES

[1] Wikipedia Contributors, "Ghazal", Wikipedia, The Free Encyclopedia, 2018. [Online; accessed 23-September-2018].

[2] Qureshi, R.B., "Musical Gesture and Extra-Musical Meaning: Words and Music in the Urdu Ghazal", Journal of the American Musicological Society, Volume 43, No. 3, pp. 457-497, 1990.

[3] Jamal, N., Mohd. M., and Noah, S.A., "Poetry Classification Using Support Vector Machines", Journal of Computer Science, Volume 8, No. 9, pp. 1441-1446, 2012.

[4] Joachims, T., "Text Categorization with Support Vector Machines: Learning with Many Relevant Features", 10th European Conference on Machine Learning, Germany, April 21 - 24, *1998.*

[5] Meyer, D., Leisch, F., and Hornik, K., "The Support Vector Machine under Test", Neuro-Computing, Volume 55, No. 1-2, pp. 169-186, 2003.

[6] Ghani, Y.Y.S., "A Study of Approaches to Hypertext Categorization", Journal of Intelligent Information Systems, Volume 18, pp. 219-241, 2002.

[7] Han, J., "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[8] Mccallum, A., and Nigam, K., "A Comparison of Event Models for Naive Bayes Text Classification", Learning for Text Categorization: AAAI Workshop, pp. 41-48, 1998.

[9] Rish, I., "An Empirical Study of the Naive Bayes Classifier", International Joint Conferences on Artificial Intelligence, Volume 3, pp. 41-46, New York, 2001.

[10] Shams, R., and Mercer, R.E., "Classifying SPAM Emails Using Text and Readability Features", IEEE 13th International Conference on Data Mining, pp. 657-666, December, 2013.

[11] Koppel, M., Schler, J., and Bonchek-Dokow, E., "Measuring Differentiability: Unmasking Pseudonymous Authors", Journal of Machine Learning Research, Volume 8, pp. 1261-1276, December, 2007.

[12] Usman, M., Shafique, Z., Ayub, S., and Malik, K., "Urdu Text Classification Using Majority Voting", International Journal of Advanced Computer Science & Applications, Volume 7, No. 8, pp. 265-273, 2016.

[13] Stamatatos, E., Kokkinakis, G., and Fakotakis, N., "Automatic Text Categorization in Terms of Genre and Author", Computational linguistics, Volume 26, No. 4, pp. 471-495, December, 2000.

[14] Finn, A., and Kushmerick, N., "Learning to Classify Documents According to Genre", Journal of the Association for Information Science and Technology, Volume 57, No. 11, pp. .1506-1518, 2006.

[15] Kaplan, D.M., and Blei, D.M., "A Computational Approach to Style in American Poetry", 7th IEEE International Conference on Data Mining, pp. 553-558, October, 2007.

[16] Voigt, R., and Jurafsky, D., "Tradition and Modernity in 20th Century Chinese Poetry", 23rd International Conference on Computational Linguistics, 2013.

[17] Kaur, J., and Saini, J.R., "Automatic Punjabi Poetry Classification Using Machine Learning Algorithms with Reduced Feature Set", International Journal of Artificial Intelligent Soft Computer, Volume 5, No. 4, pp. 311-319, January, 2017.

[18] Ali, A., Hussain, A., and Malik, M.K., "Model for English-Urdu Statistical Machine Translation", World Applied Sciences, Volume 24, pp. 1362-1367, 2013.

[19] Ali, W., Malik, M.K., Hussain, S., Siddiq, S., and Ali, A., "Urdu Noun Phrase Chunking: Hmm Based Approach", IEEE International Conference on Educational and Information Technology, Volume 2, pp. V2–494, 2010.

[20] Karamat, N., Malik, K., and Hussain, S., "Improving Generation in Machine Translation by Separating Syntactic and Morphological Processes", IEEE Conference on Frontiers of Information Technology, pp. 195-200, 2011.

[21] Malik, M.K., and Sarwar, S.M., "Urdu Named Entity Recognition System Using Hidden Markov Model", Pakistan Journal of Engineering and Applied Sciences, 2017.

[22] Mateen, A., Malik, M.K., Nawaz, Z., Danish, H., Siddiqui, M.H., and Abbas, Q., "A Hybrid Stemmer of Punjabi Shahmukhi Script", International Journal of Computer Science & Network Security, Volume 17, No. 8, pp. 90-97, 2017.

[23] Al-Diabat, F., "Arabic Text Categorization Using Classification Rule Mining", Applied Mathematical Sciences, Volume 6, No. 81, pp. 14033-4046, 2012.

[24] Wiss, S.M., Aptc, C., Damerau, F.J., Johnson, D.E., Oles, F.J., Goetz, T., and Hampp, T., "Maximizing Text-Mining Performance", IEEE Intelligent Systems, Volume 14, No. 4, pp. 63-69, July, 1999.

[25] Hall, M.A., and Smith, L.A., "Practical Feature Subset Selection for Machine Learning", Australasian Computer Science Conference, 998.

[26] Mesleh, A.M., "Chi Square Feature Extraction Based SVMS Arabic Language Text Categorization System", Journal of Computer Science, Volume *3*, No. 6, pp. 430-435, 2007.

[27] Thabtah, F., Eljinini, M.A.H., Zamzeer, M., and W.M., "Naïve Bayesian Based on Chi Square to Categorize Arabic Data", Communications of the IBIMA, Volume 10, 2009.

[28] Bach, F.R., "Bolasso: Model Consistent Lasso Estimation through the Bootstrap", Proceedings of 25th ACM International Conference on Machine Learning, pp. 33-40, 2008.

[29] Zare H. Haffari, G., Gupta, A., and Brinkman, PO.R., "Scoring Relevancy of Features Based on Combinatorial Analysis of Lasso with Application to Lymphoma Diagnosis", BMC Genomics, Volume14, pp. S1-S14, 2013.

[30] Odeh, A., Abu-Errub, A.M., Shambour, Q., and Turab, N., "Arabic Text Categorization Algorithm Using Vector Evaluation Method", Computing Research Repository, 2015.

[31] Butinick, I., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculaf, F., Prettenhofer, P., Gramfort, A., Grobler, J. Layton, R., Vanderplas, J., Joly, A., Holt, B., and Varoquaux, G., "API Design for Machine Learning Software: Experiences from the Scikit-Learn Project", European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, pp. 108-122, 2013.

[32] Thabtah, F., Eljinini, M., Zamzeer, M., and Hadi, W., "Naïve Bayesian Based on Chi Square to Categorize Arabic Data", Proceedings of 11th International Conference on Innovation and Knowledge Management in Twin Track Economies, pp. 4-6, Cairo, Egypt, 2009.