Arabic Text Genre Classification

Alaa M. El-Halees
Faculty of Information Technology, Islamic University of Gaza, Gaza, Palestine, email alhalees@iugaza.edu.ps

Abstract— Text genre is a type of written text. Arabic text genre classification predicts genre of specific text document written in Arabic independent of its topic. In this paper, an approach was proposed that takes an Arabic document and classify it into one of four genres which are advertisements, news, subjective and scientific documents. Since the frequency of words approach produces a low performance when used in the genre, an attempted was made to generate attributes based on the style of the text. This approach evaluated using corpus collected for this purpose. Using four machine learning methods, our approach compared with the word frequency approach, and it found that our approach is better than this mainstream approach. It, also, found that predicting subjectivity and scientific genre is more accurate than predicting advertisements and news.

Index Terms— Text genre, text genre classification, Arabic language processing, text mining, machine learning methods.

I Introduction

Text genre classification is concerned with predicting the type of an unknown text correctly, independent of its topic [1]. Genre means kind of text; it is functional role of the text, not its topic. Examples of text genre are scientific articles, news reports, reviews, and advertisements. The importance of text genre comes from that user wants a specific type of text. The typical example is in informational retrieval and search engine where the user may desire to see documents for a specific reason such as a review for some object (i.e. people opinion in a product) or scientific article in some subject [2]. Text classification gene is different from traditional text classification where traditional classification is based on the frequency of certain words in the document using TFIDF representation. In classifying genre, text style is used instead.

Most research in the area of text genre classification deals with English text. Some works deal with other languages, but in Arabic, which is a language for Millions of people, there is no work in text genre classification. Arabic is a challenging language for some reasons. It has a complex morphology as compared to other languages like English. This is due to the unique nature of Arabic language. The Arabic language is an inflectional and derivational language which makes monophonical analysis a very complex task [3].

The first and the most important task of classification genre are to choose genre types. Based in the field of linguistic three abstract and very general classes are used, namely, expressive, appellative, and informative text [4]. Accordingly, the text tagged as subjective (expressive), advertisement (appellative) and scientific papers and news (informative). Then, the text genre needed to identify cues such as the structure of the sentence, the length of sentence, characters used and punctuation are used to generate the features. Then, machine learning methods are used to classify the genre. Four machine learning methods were used which are: Support Vector Machine, Naive Bays, k-nearest neighbors and Decision trees.

To evaluate this approach, corpus was collected from many Arabic websites since no other work was done on this topic. Finally, our method compared with a traditional TFIDF method that used in topic classification.

The remainder of the paper is organized as follows: Section two about related work in this area, section three about genre classification, section four about our methodology, section five about the experiment and results, and finally, this paper closed with a conclusion and an outlook for future work.

II Related Works

In English language, text genre classification was addressed by many works such of Kessler et. al. in [1] who proposed a theory of genres as bundles of facets, which correlate with various surface cues. They argued that genre detection based on surface cues as successful as detection based on deeper structural properties. They developed a taxonomy of genres and facets. Also, they found an effective strategy for variable selection to avoid overfitting during training with neural networks that have higher performance on average. Karlgren and Cutting in [5] used discriminate analysis to categorize texts into pre-determined genre categories. They argued that discriminate analysis make it possible to use a large number of parameters that may be specific for a certain corpus, and combine them into a small number of functions, with the parameters weighted by how useful they are for discriminating text genres. Also, Liu et. al. in [6] discussed the automatic genre classification and its application. They argued that word level features and sentence level features are two important measures which vary in number among different genres. Based on the two aspects of views, they explore an approach where the Co-training method is employed to obtain genre classification. Stamatatos et. al. in [7] took full advantage of existing natural language pro-

cessing tools to propose some style markers including analysis-level measures that characterize the way in which the input text has been analyzed and capture valuable stylistic information. They present a set of small-scale experiments in text genre detection, author identification, and author verification tasks. They showed that the proposed method performs better than the most distributional lexical measures, functions of vocabulary wealth and frequencies of occurrence of the most frequent words. Galitsky et. al. in [8] proposed to use methods based on deep textual parsing, which depends on finding complex features such as syntactic and discourse structures of the text, to improve the quality of genre classification. In their paper they had presented three experiments on style and genre classifications. For the genre classification task they adopted a corpus annotated with 7 different genres and conducted a series of pairwise classification between two genres. Melissourgou and Frantzi in [9] investigated a range of genres involved in writing tasks presented in English language teaching material. They explained how they identified genres based on Systemic Functional Linguistics (SFL) principles. They added another stage which is 'naming' of genre categories mainly based on purpose and mode to guide anyone with a need to understand genre requirements.

In multi-language text genre classification, Petrenz in [10] described a new approach to classifying text genres across languages. It can bring the benefits of genre classification to the target language without the costs of manual annotation to achieve good results. In his experiments, he considered English and Chinese languages, because these languages are very dissimilar linguistically. He expected the approach to work at least equally well for more closely related language pairs.

III Genre Classification

Genre classification is different from the topic classification that most classification research has dealt with. From an information retrieval point of view, a retrieval query about a certain topic would retrieve many documents related to that topic, but they may be of the different genre [11]. For example, if someone searches for a certain product, the retrieved page will be any document that contains the name of that product. However, genre means you can specify if somebody wants for example news, advertisement, or critical review about that product [2].

Genres give a way to describe the nature of a text, which allows for assigning the document to classes. Arabic genre classification is concerned with predicting the genre classes of unknown Arabic documents correctly, independent of its topic. In Arabic genre classification: *Let* $C = \{c_1, c_2, \ldots c_m\}$ be a set of genre classes and $D = \{d_1, d_2, \ldots d_n\}$ a set of Arabic documents. The task of the Arabic genre classification consists in assigning class label c_i to each document d_j , if the document d_j belongs to c_i, which exactly one class must be assigned to each d_j.

Based in the field of linguistic , text genre can be classified into three general classes, namely, expressive, appellative, and informative [4]. Expressive means that text aims to express the attitude, expression of feelings, attitudes, and opinions of a person. According to this definition, opinion mining corpus mapped to the expressive genre. Appellative means appealing to the receiver's experience, feelings, knowledge and sensibility to make him/ her react in a specific way [12]. The best text maps to this genre are an advertisement which used in this research. Finally, the informative text provides information about any topic of knowledge. They identify impersonal, objective, non-emotive style [4]. Two classes were mapped to this genre which are scientific papers and news.

IV Methodology

Our methodology consist of the following steps:

A. Generate Features

Text genre mostly characterized by its text style. To generate features in this work, it concentrated on two levels of text styles: token level and lexicon level. Token level considers the text as a set of tokens grouped in sentences. In this level, features were generated from each document such as average number of words in a sentence, average number of short words in a document where it considered short words are the words with less than six characters, average number of words per phrase and the average number of characters per word. In lexicon level, features were generated from each document such as an average number of nouns, adjectives, and verbs per word. Also, features were added such as an average number of pronouns, coordinating conjunctions, cardinal numbers, and determines per document.

B. Corpus

As stated above, this researh used four text genres: subjective, advertisements, news and scientific. Since no other works in Arabic genre classification, there is no corpus exist in the literature. Therefore, our own corpus was collected. As shown in Table , 78251 documents were used for the four genre types where each genre type contains more than one topic. For example, the subjective genre has positive and negative reviews on topics such as movies, hotels, books…etc. Advertisements have topics from many products such as electronics, furniture, medical and sports equipment. News has topics from culture, economy, international and sports. Finally, scientific papers have topics from medicine, science, economy and literature.

TABLE 1
Corpus used in the experiments

Genre	Type	No. Documents	Total No. of Documents
Subjective	1.Positive 2. Negative	1430 1430	2860
Advertisement	1. Computers and Electronics 2. Furniture 3. Medical Equipments 4. Sports Equipments	340 522 254 342	1456
News	1. Culture 2. Economy 3. International 4. Sports	500 500 500 500	2000
Scientific	1. Medicine 2. Science 3. Economy 4. Literature	512 327 378 292	1509

C. Methods

In our experiments to classify documents to genres using two approaches TFIDF representation and text style extraction, four classifiers were applied, which are Naïve Bayes, k-Nearest Neighbors and Support Vector Machine and Decision Trees.

Naïve Bayes classifier is widely used because of its simplicity and computational effectiveness. The model assigns a class label to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. In the text, It uses training methods consisting of relative-frequency estimation of words in a document as words probabilities and uses these probabilities to assign a class to the document. To estimate the term $P(d \mid c)$ where d is the document and c is the class, Naïve Bayes decomposes it by assuming the features are conditionally independent [13].

k-Nearest Neighbors is a method to in classification. The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples. In the text, the training phase documents have to be indexed and converted to vector representation. To classify new document d; the similarly of its document vector to each document vector in the training set has to be computed. Then its k nearest neighbor is determined by measuring similarity which may be measured by, for example, the Euclidean distance [14].

Support Vector Machine is a classification algorithm proposed by [15]. In its simplest linear form, it is a hyperplane that separates a set of positive examples from a set of negative examples with maximum margin. In the text, test documents are classified according to their positions on the hyperplanes.

A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. Decision tree text classifier is consists of a tree in which internal nodes are labeled by words, branches departing from them are labeled by tests on the weight that the words have in the representation of the test document, and leaf nodes are labeled by categories c_i . Such a classifier categorizes a test document d_j by recursively testing for the weights. That the words labeling the internal nodes have in the representation of d_j , until a leaf node c_i is reached; the label of this leaf node is then assigned to d_j [16].

V EXPERIMENT AND RESULTS

A Experiments

Two sets of experiments have been applied, first experiments for topic classification as a baseline and the second experiments to evaluate generated features.

For the first set of experiments, our baseline used corpus described above with topic base classifications. Before classification, some pre-processing was done such as tokenization, stop words removal and Arabic light stemming. Then, vector representations were obtained for the terms from their textual representations by performing TFIDF weight which is a well-known weight presentation of terms often used in text mining. Some terms with a low frequency of occurrence were removed. For classification, Four methods described above were used which are Naïve Bayes, k-Nearest Neighbors, Support Vector Machine and Decision Trees.

In the second set of experiments, using the same corpus as described above, features generated based on nature and lexicon of the documents in the corpus. Part-of-speech (POS) was used to generate word classes, such as nouns, adjectives, and verbs. Then the four machine learning methods were applied.

Our experiments was evaluated using 10-cross-validation, and then F-measure was computed, which is a combined metric that takes both precisions and recalls into consideration.

B Results

Table 2 and Figure 1 show F-measure for baseline classification which based on TF-IDF and classification based on generated features from text using four machine learning methods for both classifications. It is clear that generated features have better results than baseline in all machine learning methods. However, there is little difference in performance when using naïve Bayes. Moreover, the biggest difference is when using Decision Trees where it is 48.87% using baseline and 100% using generated features. That is mainly because baseline depends on the frequency of the words and the words may be frequent in more than one genre if it is on the same topic (e.g. word in sports topic can be in the news, Advertisements, subjective or Scientific). That is not the case for generated features which depends on style not frequency.

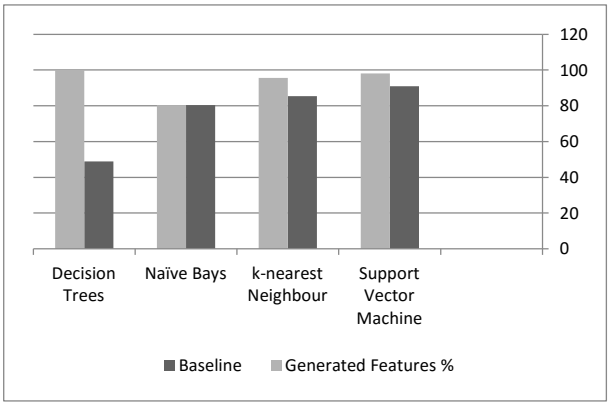

Figure 1: F-measure for Arabic genre classification

TABLE 2
F-measure for baseline and generated features Arabic genre classification

Method	Baseline %	Generated Features %
Support Vector Machine	90.89	98.07
k-nearest Neighbour	85.42	95.58
Naïve Bays	80.37	80.47
Decision Trees	48.87	100

Table 3 shows F-measure for the four selected corpus genre where B.L is for Base-line and F.G. for generated features. It was noted that all methods accurately recognized subjective and scientific. Also, there is a little confusion between Advertisements and news and that is natural because there are many common characteristics between them.

TABLE 3
F-measure for four Arabic text genres

	Support Vector Machine %		k-nearest Neighbour %		Naïve Bays %		Decision Trees %	
	B.L	G.F	B.L	G.F	B.L	G.F	B.L	G.F
Advertisements	84.9	96.3	79.6	90.32	90.7	68.57	28.7	100
News	94.4	96.0	80	88.0	81.2	53.3	67.4	100
Subjective	86.2	100	85.1	100	76.25	100	0	100
Scientific	98	100	97	100	97.23	100	99.3	100

From generated decision tree, as seen in figure 2, , it can seen that *Average words per phrase* ,*Average characters per word* and *Averege number of words per sentence* are the most important attributes that distinguish genre.

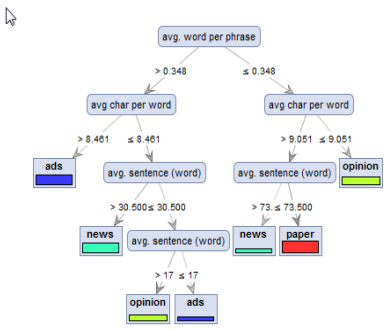

FIGURE 2: DECISION TREE FOR GENERATED FEATURES

VI CONCLUSION AND FUTURE WORKS

There are some significant differences between text topic classification and text genre classification. Text topic classification depends mainly on the frequency of some words in a document to recognize that document. This does not work for text genre classification because words may be frequent in multiple genres. In this paper, Arabic document was classified to a certain genre. Four types were chosen to classify Arabic genre which are an advertisement, news, subjective and scientific. We generated attributes based on Arabic language style. The work evaluated using some corpus collected manually. Using four machine learning methods we found that our generated feature has better performance than the results obtained from using TFIDF method using same machine learning methods and same corpus. Also, we concluded that subjective and scientific genres have better performance than news and advertisements.

In future work, it may use another Arabic genre such as a poem, Islamic Scripts, events, Biography..., etc. Also, it may need to look for other attributes which can recognize the genre such as syntactical level of the Arabic language. Also, the generated feature is done manually, using techniques such as deep learning, it can be generated automatically.

REFERENCES

[1] B. Kessler, G. Numbers, and H. Schütze. "Automatic detection of text genre." In the proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics: 7–12 July, Madrid, 1997.

[2] Y. B. Lee and S. H. Myaeng, "Text genre classification with genre-revealing and subject-revealing features," In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 145-150. 2002.

[3] B. Hammo and S. Lytinen, "QARAB: A Question Answering System to Support the Arabic Language," In the proceedings of Computational Approaches to Semitic Languages., p. 11, 2002.

[4] H. Wachsmuth and K. Bujna, "Back to the Roots of

Genres : Text Classification by Language Function Motivation : Filter search results,". In the proceedings of the 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand, November 8-13, 2011.
[5] J. Karlgren and D. Cutting, "Recognizing Text Genres With Simple Metrics Using Discriminant Analysis," In the proceedings of the 15th Conference of Computer Linguists. - Vol. 2, pp. 1071–1075, 1994.
[6] R. Liu, M. Jiang, and Z. Tie, "Automatic genre classification by using co-training," In the proceedings of the 6th International Conference of Fuzzy Systems and. Knowledge Discovery, vol. 1, pp. 129–132, 2009.
[7] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Automatic Text Categorization in Terms of Genre and Author," Computer Linguists, vol. 26, pp. 471–495, 2000.
[8] B. A. Galitsky , D. A. Ilvovsky, E. L. Chernyak S. and O. Kuznetsov " Style and Genre Classification by Means of Deep Textual Parsing", In the Proceedings of the International Conference Computational Linguistics and Intellectual Technologies: "Dialogue 2016" Moscow, June 1–4, 2016
[9] M. Melissourgou and K. Frantzi "Representation of Text Types and Genres in English Language Teaching Material", Corpus Pragmatics, April, 2017, Springer International Publishing.
[10] P. Petrenz, "Cross-Lingual Genre Classification," *Proceedings of the13th Conference of. European. Chapter Association of Computer* Linguists, no. April, pp. 11–21, 2012.
[11] K. Crowston and B. H. Kwasnik, "Can Document-Genre Metadata Improve Information Access to Large Digital Collections," Library Trends, vol. 1, no. 315, pp. 1–29, 2003.
[12] J. Vaičenonienė, "The Language of Advertising: Analysis of English and Lithuanian Advertising Texts", Studies About Languages. No. 9, pp. 43–55, 2006.
[13] S. L. Ting, W. H. I, and A. H. C. Tsang, "Is Naïve Bayes a good classifier for document classification?" International Journal of. Software. Engineering and its Applications, vol. 5, no. 3, pp. 37–46, 2011.
[14] B. Dasarathy. Nearest neighbor (NN) norms: NN pattern classification techniques. IEEE Computer Society Press, 1991.
[15] C. Cortes and V. Vapnik. "Support-Vector Networks". Machine Learning, 20, 1995
[16] I. Ilovich, and S. Markovitch. "Feature Generation for Text Categorization Using World Knowledge". In the Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, July 30-August 5, 2005.pp 1048-1053.

Alaa M. El-Halees is a professor in computing in the faculty of Information Technology at Islamic University of Gaza, Palestine. He holds a PhD degree in data mining from Leeds Metropolitan University, UK in 2004, MSc degree in Software Engineering from Leeds Metropolitan University, UK in 1998 and BSc in Computer Engineering from University of Arizona, USA. Alaa has more than 24 years of experience including leading a range of IT-related projects. Prof. Alaa supervises M.Sc. students in Information Technology. He also leads and teaches modules at both BSc and MSc levels in Information Technology. His research activities are in the area of data mining, in particular text mining, machine learning and e-learning, Software Engineering and computer ethics.