Polynomial Neural Networks versus Other Arabic Text Classifiers

Mayy M. Al-Tahrawi*
Computer Science Department, Faculty of Information Technology, Al-Ahliyya Amman University, Amman, Jordan

* Corresponding author. Tel.: 00962-795414927; email: mtahrawi@ammanu.edu.jo
Manuscript submitted December 10, 2015; accepted February 10, 2016.
doi: 10.17706/jsw.11.4.418-430

Abstract: Many Text Classification (TC) algorithms have been proposed for Arabic TC. Polynomial Neural Networks (PNNs) were used recently in English TC, and have proved to be competitive to the state of the art text classifiers in this field. Lately, they were proposed for classifying Arabic documents. In this research paper, an experimental study that directly compares PNNs against five famous classification algorithms in TC is conducted on Aljazeera-News Arabic dataset. All experiments use the same TC settings, like preprocessing, Feature Selection (FS) and reduction criteria, feature weighting and classifier performance evaluation measures. These algorithms are: SVM (Support Vector Machines), NB (Naive Bayes), kNN (k-Nearest-Neighbor), LR (Logistic Regression) and RBF (Radial Basis Function networks). Results reached in this study reveal that PNN are competitive classifiers in the field of Arabic TC.

Key words: Polynomial neural networks, Arabic text classification.

1. Introduction

Text Classification (TC) is to determine, automatically, the topic (class or category) of a new text, such as Sport, Religion or Law. The huge amount of text documents which are available online needs fast, accurate and efficient automatic classification; such need grows rapidly with the dramatically increasing widespread of internet content and usage all around the world, using several human languages. TC applications include, but are not limited to, spam filtering, information retrieval, topical crawling, digital libraries and online news.

Interest of researchers in Arabic TC started to grow in the last decade as Arabic is a major language world-wide, which has hundreds of millions of native speakers. In fact, Arabic language is one of the six official languages in the United Nations. Add to this, a large percentage of Arabic speakers don't have the ability to read or understand English. Furthermore, more than 3% of the internet content is Arabic content, which puts it in the fourth rank [1]. In order to make use of such content worldwide, it has to be categorized and retrieved efficiently; thus, the need for efficient and accurate automatic Arabic TC systems has grown rapidly in the last decade.

Several English TC algorithms were proposed for Arabic TC; examples include (NB) Naïve Bayes algorithm [2], [3] , kNN (k-Nearest Neighbor) [3], [4], SVM (Support Vector Machines) [5]-[9] , Decision Tree [5], [9], [10], Manhattan Distance and Dice Measures [11] and Rocchio [12]. More recently, Polynomial Neural Networks (PNNs) are investigated in Arabic TC [13].

Polynomial Neural Network (PNN) is a supervised machine learning technique which uses mathematical methods and evolutionary concepts to gradually develop a network of polynomial functions that can approximate continuous multivariate functions from input-output collections of data. PNNs were first used in English TC in 2008 [14], and have proved to be competitive to a set of well-known English text classifiers [14]-[17].

In this research, PNNs are directly compared to five well-known Arabic text classifiers: NB, kNN, SVM, RBF (Radial Basis Function Networks) and LR (Logistic Regression). Experiments are conducted on Aljazeera-News Arabic dataset, using the same TC settings for the six classifiers.

The paper is organized in six sections: Related work on Arabic TC is briefed in Section 2, experimented TC algorithms are presented in Section 3, the Dataset and data preprocessing are summarized in Section 4, while experiments results are presented and analyzed in Section 5 and conclusions take place in Section 6.

2. Related Work

In the last decade, Arabic TC research work has grown rapidly. A group of researchers addressed the issue of Arabic TC using different classification techniques, datasets and pre-processing operations, as no standards exist regarding free benchmark Arabic corpora or pre-processing tools. This section presents an overview of a number of studies in the field of Arabic TC.

Reference [2] classified Al-Jazeera Arabic News using NB algorithm. Cross validation experiments were used in this research to evaluate the NB classifier. The corpus documents went through several pre-processing procedures, including root extraction. TF.IDF (Term Frequency. Inverse Document Frequency) was used to score document terms. The average accuracy recorded in this research was 68.78%.

The author in [3] compared three popular TC algorithms: KNN, NB and the Distance-Based classifier. She used an in-house collected corpus of 1,000 text documents belonging to 10 classes. She preprocessed the corpus by removing punctuations and stop words. Root extraction was applied in addition. NB was the top performer in this research with a class-level recall performance varying from 22% up to 98%. Distance-Based came last with micro average recall, precision, error rate and fallout of 62.8, 74.0, 7.4 and 4.1 respectively.

The author in [5] compared six classifiers on Arabic documents from Aljazeera Arabic news channel: ANN (Artificial Neural Networks), Maximum Entropy , KNN , NB , SVM and Decision Tree. His corpus has six classes: Economy, Health, Culture and Arts, Politics, Sports and Science and Technology. The author applied stop words removal, normalization and stemming on the corpus documents. He used Information Gain (IG) for FS, Normalized Frequency for term weighting and ten-folds cross-validation for testing the six classifiers. SVMs and NB were the best classifiers in terms of F-Measure. When no FS was conducted, NB recorded a 91% F-measure and SVMs recorded 88%. Using Information Gain (IG) for FS, NB recorded 83% F1 and SVMs recorded 88%.

The authors in [6] investigated the performance of CBA (a rule based classification method based on association rule mining), NB and SVMs on classifying Arabic texts from Newspapers websites. Stop words were excluded and normalization was applied as text pre-processing. Their results showed that CBA had a better performance compared with NB and SVM with an average precision, recall and F1-measure of 80.5, 80.7 and 80.4 respectively.

Reference [7] used SVM to classify a group of Arabic news datasets (Al-Nahar , Al-Ahram, Al-Dostor, Al-Jazeera and Al-Hayat). Pre-processing (excluding stemming) was applied on the documents and infrequent terms were removed. Chi Square (CHI) was used for FS while TF.IDF was used for term weighting. SVM recorded a macro-Average F1 of 88.11% in this research.

Reference [8] demonstrated a combination of Binary Particle Swarm Optimization and KNN as a feature selection criterion in classifying three Arabic corpora: "Al-jazeera-News", "Akhbar-Alkhaleej" and "Alwatan Arabic". The authors used three classification algorithms: J48 Decision Tree, SVM and NB. TF.IDF was used for term weighting in the three algorithms. The best performance was on Al-jazeera-News: SVM was the best performer with 93.7, 93 and 93.1 for precision, recall and F-measure respectively. NB recorded 85.8, 84.3 and 84.6, whereas J48 recorded 74.7, 72.3 and 72.9 for precision, recall and F-measure.

The authors in [9] studied the performance of SVMs and C5.0 in categorizing Arabic documents. Seven Arabic datasets from different sources were used in their research work: Saudi Press Agency, Internet articles, discussion forums and Saudi newspapers. CHI was used for FS and binary score was used for term weighting. The average accuracies achieved by C5.0 and SVMs were 78.42% and 68.65% respectively.

Reference [11] compared Manhattan Distance and Dice Measure algorithms in classifying a news corpus collected from a group of Jordanian newspapers. Several pre-processing operations were applied on the corpus documents. No feature selection was used and features were scored using Term Frequency (TF). Dice Measure outperformed Manhattan Distance, recording an average precision of 88.75% and recall of 83%, compared with Manhattan Distance which recorded 66.5% and 56% for precision and recall respectively.

The authors in [12] compared the three well-known classification algorithms KNN, NB and Rocchio on an Arabic corpus from several online newspapers (Al-Hayat, An-Nahar, Al-Jazeera, Ad-Dostor and Al-Ahram) which consists of 1,445 Arabic text documents belonging to 9 classes. They applied light stemming for feature reduction, several term weighting schemes for term scoring and 4-fold cross-validation for evaluation. The best results were recorded by NB, then kNN and finally Rocchio.

The authors in [13] proposed using PNNs for Arabic TC. They have shown that PNN is a fast and highly accurate Arabic text classifier. PNNs recorded 89.3% as an average performance on the widely used Arabic TC corpus: Aljazeera-News (Alj-News), using just 1% of the corpus features in their experiments.

Reference [18] conducted document clustering and classification experiments on the Arabic NEWSWIRE corpus using Maximum Entropy. These methods have shown to be very robust and reliable, in their experiments, although no morphological analysis was applied on the corpus documents.

The author in [19] used Maximum Entropy to classify Alj-News and other Arabic corpora. He pre-processed documents and used stemming. Results recorded in his research were 80.48 for recall, 80.34 for precision and 80.41 for F-measure.

Reference [20] used SVM to classify Alj-News and Alj-Magazine Arabic datasets. Four FS metrics were experimented: Document Frequency (DF), MI (Mutual Information), IG and Correlation Coefficient (CC). Stop words were removed and three different stemmers were compared: RDI MORPHO3, Sebawai Root Extractor (SR) and Light Stemmer (AS). AS with MI or IG recorded the best performance in their research.

RBF was used for Arabic TC by [21] and [22]. The authors in [21] presented a NN-based model for Arabic TC: the Singular Value Decomposition (SVD). They tested their model on MLP (the Multi-Layer Perceptron) and the RBF. A locally collected corpus of Arabic documents was used to test their model. They showed that their proposed model could improve the effectiveness of both classifiers with better results recorded using the MLP classifier. On the other hand, the authors in [22] implemented a FS algorithm using Particle Swarm Optimization, and tested it on RBF as a TC algorithm. They compared their proposed algorithm with other FS methods and showed that the proposed algorithm is superior to DF, tf.idf and CHI FS methods.

Reference [23] investigated NB and SVM on the Saudi Newspapers (SNP) Arabic dataset. Some data pre-processing operations were applied on the dataset like removal of digits, punctuations, stop words and non-Arabic texts, as well as normalization of Al Hamza. SVM outperformed NB with an average precision of 77.9, recall of 77.8 and F-measure of 77.8, whereas NB recorded 74.1 for precision and 74 for recall and

F-measure.

Reference [24] conducted a comparison of six TC algorithms: SVM, kNN, NB, C4.5, C5.0 and MLP (MultiLayer Perceptron neural networks) on King Abdul-Aziz city for Science and Technology Dataset. Simple text pre-processing was applied using the ATC tool (an Arabic Text Classification tool that removes numbers, punctuations, diacritics, kashida and stopwords as well as normalization of Al Hamza). Two term weighting schemes were used (TF and DF) and two FS methods were experimented (CHI and IG). The best accuracy was recorded by SVM, then C4.5 and NB.

The authors in [25] compared three different approaches of Arabic TC: Artificial Neural Networks (ANN), SVMs and BSOCHI-SVM on the Open Source Arabic Corpora (OSAC). Two stemming approaches were used: light and root-based stemming. TF.IDF was used for term weighting and CHI was used for FS. They have shown that light stemming recorded better performance than the root-based one and SVMs outperform ANN. The best recorded accuracy in their research was 95.67, which was achieved using the BSO-CHI-SVM approach.

The authors in [26] proposed FRAM (the Frequency Ratio Accumulation Method) as a new Arabic TC approach. FRAM is tested in this research versus three classifiers: NB, MNB (Multi-variant Bernoulli Naïve Bayes) and MBNB (Multinomial Naïve Bayes). Text-preprocessing and light stemming were conducted on the corpus documents, TF was used for term weighting and CHI was used for FS. FRAM recorded the best performance (95.1 macro-F1 value) in their research.

The authors in [27] investigated three variations of the vector space models using KNN algorithm to classify the Saudi Newspapers (SNP) dataset: Cosine, Jaccard, and Dice coefficients. Simple pre-processing was applied on the corpus. The best performance in their experiments was achieved by Cosine coefficients: 91.7 for Precision, 97.9 for Recall and 94.7 for F1.

Recently, the author in [28] was the first researcher to test Logistic Regression (LR) in Arabic TC. She tested the algorithm on Alj-News dataset and proved that Logistic Regression is an efficient Arabic text categorization algorithm.

Apparently, no agreement either on the corpus or on the preprocessing operations applied on the corpus documents exists. As a result, direct fair comparisons between such studies are impossible. A fair direct comparison of Arabic TC performance between PNNs and the other algorithms is carried out in this research, by using the same data set, data pre-processing, FS, feature reduction policy, feature weighting and performance evaluation criteria for the six classifiers. A brief overview of each of the six TC algorithms is presented in the next section.

3. Text Classification Algorithms

PNNs are compared directly versus five well-known Arabic TC algorithms in this research. These algorithms are SVMs, NB, kNN, LR and RBF networks. The six algorithms are overviewed in the subsequent subsections.

3.1. Polynomial Neural Networks (PNNs)

PNNs have been used early in several applications [29]-[32]. Recently, PNNs have shown to be one of the top English text classifiers of Reuters and 20News Groups [14]-[17] . More recently, PNNs have been investigated in Arabic TC and have achieved very high classification accuracy [13].

In this comparative study, the Polynomial Neural Networks algorithm proposed by [31] is used to classify Aljazeera News Arabic corpuss. This algorithm uses discriminative training with a mean-squared error objective criterion. Reference [14] explains in detail the PNNs algorithm and how to apply it in TC. It is worth noting that quadratic polynomials are adopted in this research, as they recorded the best results regarding classifier accuracy and computing resources.

3.2. k-Nearest Neighbor (kNN)

kNN is a statistical approach that was applied to TC very early [33]. It is one of the top-performers in English TC [14] , [33], as well as Arabic TC [3],[5], [12] , [24], [27]. The algorithm works simply as follows: Given a new unseen text document, which needs to be classified into one of a collection of pre-determined classes, kNN algorithm searches for the k nearest neighbors of this text document by measuring the cosine similarity between document vectors. The class of this text is chosen based on the majority voting among its neighbors.

3.3. Naive Bayes (NB)

NB is a simple probabilistic classification algorithm which has been commonly used in both English [14], [34]-[36] and Arabic [8], [23]-[26] TC. Given a new unseen document, NB uses the joint probabilities of terms and classes to compute the probabilities of classes.

To use NB in TC, compute the probability of a class c_i, given a text document d_k as follows [37]:

$$P(C_i|d_k)= \frac{P(d_k|C_i)P(C_i)}{P(d_k)} \quad (1)$$

3.4. Radial Basis Function (RBF) Networks

RBF artificial neural network model performed well in English TC [14], [38], [39], as well as in Arabic TC [21], [22], [28]. The RBF network structure consists of one hidden layer of units which are locally-tuned and fully interconnected to the output layer. The output layer consists of linear units. All the hidden units receive the input vector at the same time. The outputs are produced by measuring the distance between the input vector and weight vector of the hidden unit multiplied by a bias value; this bias value allows for the adjustment of the sensitivity of the RBF unit. RBF network is adaptive; hence, a smaller number of local units can be used.

3.5. Support Vector Machines (SVMs)

SVM classifiers were firstly introduced by [40]. SVMs have shown to be the winners in English TC [41], [42]. For the details of SVM application in machine learning and TC, readers can refer to [40], [41], [43].

3.6. Logistic Regression (LR)

LR is a probabilistic statistical model which has been used in machine learning applications, including TC [44]-[46]. Studies have proved that the Logistic Regression English text classifier has a performance which is similar to that of SVMs [46], [47]. LR was used for the first time for Arabic TC by [28] and have proved to be a good performer in this area.

Given a new unseen document y, LR computes the conditional probability of classifying y into a class x by [45]:

$$P(x \mid y) = \frac{1}{1+ \exp(-x\alpha^T y)} \quad (2)$$

Using α as a parameter of the model.

4. The Dataset

Since there are no benchmark Arabic TC datasets, several Arabic Datasets were collected for this purpose. Aljazeera Arabic News Dataset [48], collected from the website of Al-Jazeera Arabic News channel, is used in the experiments of this research. This dataset has five classes, with 300 documents in each class. In each

class, 80% of the texts are used for training and 20% for testing. Many researches in Arabic TC have adopted this dataset using different algorithms [49], [20], [8], [13], [28].

The data pre-processing conducted on the dataset is detailed in Sections 4.1 through 4.3 next.

4.1. Data Pre-Processing

The following pre-processing steps are conducted on Aljazeera Arabic News dataset, in order to minimize data noise and the number of terms, thus memory and processing requirements needed to build the PNN classifier:

1) Tokenization: to partition a text of characters into words by recognizing punctuations, white spaces and other delimiters.
2) Elimination of letters which are not part of the Arabic alphabet.
3) Elimination of digits and diacritics.
4) Elimination of stop words, like pronouns, prepositions and conjunctions. In this research, an extended version of the 168 stop word list used by [50] is widened to include 478 stop words.
5) Stemming: to reduce a word to its stem or root. This step aims mainly to reduce the dataset terms and processing requirements. The stemmer of [50] is used in this research; it can be downloaded at [51].
6) Elimination of any word that is reduced to just one-character after applying the stemming steps above.

Table 1 shows the remaining numbers of words in each topic in the dataset after applying these pre-processing steps.

Table 1. The Number of Words in Aljazeera Arabic News Dataset after Pre-processing

Topic	**Words**
Art	3745
Economic	2178
Politics	2984
Science	2806
Sport	3332
TOTAL	**15045**
TOTAL after filtering duplicates	**8218**

4.2. Feature Selection (FS)

FS is very common in TC, since many classifiers do not have the capability of working with all the corpus terms (features). Furthermore, using the whole corpus terms mostly affects the classifier performance negatively; this burdens the classifier with noisy terms and extra processing requirements. On the other hand, some researchers have proved that FS degrades classification accuracy [41], [52]-[54].

In FS, the strength of each feature in the corpus is measured and the most discriminating ones will be used to build a classifier. Some FS criteria that were used in Arabic TC research include Chi Square (CHI) [13]. [24]-[26], [55], [9], [7], Cross Validation [2], Document Frequency (DF) [24], [20], Information Gain(IG) [24], [20], [5], Correlation Coefficient (CC) [20], Semi-Automatic Categorization Method (SACM), Mutual

Information (MI) [20], BPSO-KNN (Binary Particle Swarm Optimization- K-Nearest-Neighbor) [8] and ACM (Automatic Categorization Method) [56]. Some researchers performed random FS [18]and others haven't used any FS [11].

In this research, Chi Square (CHI) FS is used. CHI has achieved high classification accuracy in English TC [14]-[17], [5], [57], [58], as well as Arabic TC. CHI computes the independence between a term t and a class c_i in the corpus as follows [34]:

$$\chi^2(t, c_i) = \frac{N \times (\text{AD-CB})^2}{(\text{A} + \text{C}) \times (\text{B} + \text{D}) \times (\text{A} + \text{B}) \times (\text{C} + \text{D})} \quad (3)$$

where: N is the number of training text documents in the corpus, A is the number of text documents that belong to class c_i and contain t, B is the number of text documents which belong to class c_i but do not contain t, C is the number of text documents which belong to class c_i and contain t and D is the number of text documents which do not belong to class c_i and not contain t.

Regarding the experiments conducted in this paper, normalized frequency weighting is adopted for all classifiers, except NB, which uses binary weighting.

4.3. Feature Reduction

The topmost 1% of the terms of each topic is used to build the six classifiers, as shown in Table 2.

Table 2. The Numbers of Features Used to Build the Classifiers

Topic	**Number of selected features**
Art	37
Economic	22
Politics	30
Science	28
Sport	33
TOTAL	**150**
TOTAL after filtering duplicates	**135**

5. Experiments and Results

The measures used to evaluate the classifiers performance are presented in Section 5.1 and experiments results are presented in Section 5.2 next.

5.1. Evaluation Measures of Classifiers Performance

The six classifiers tested in our experiments are evaluated using recall, precision and F1- measures. The formulae of computing precision and recall are shown below [59], where P_i denotes precision of a class c_i and R_i denotes recall of a class c_i:

$$P_i = \frac{TP_i}{TP_i + FP_i} \quad (4)$$

$$R_i = \frac{TP_i}{TP_i + FN_i} \tag{5}$$

where TP_i denotes true positives with respect to class c_i, FP_i denotes false positives with respect to c_i and FN_i denotes false negatives with respect to c_i.

The *F1* measure is the average of precision and recall introduced by [60]:

$$F1 = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} \times \text{precision}} \tag{6}$$

Finally, the overall performance on all classes are averaged using both microaverage and macroaverage measures.

5.2. Results

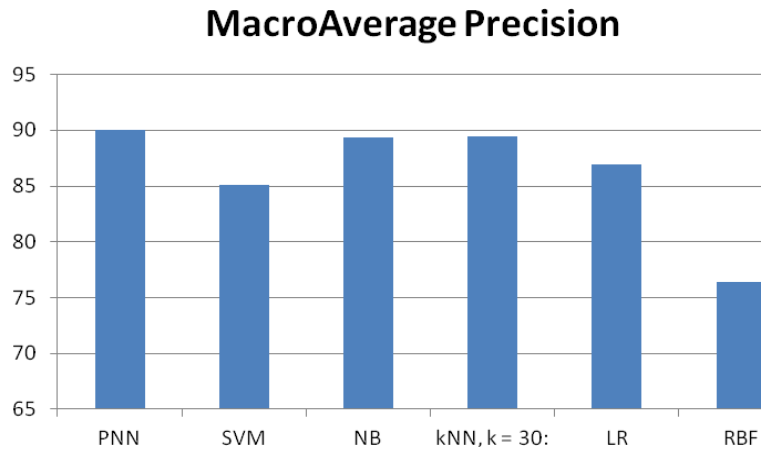

Fig. 1. Macroaverage precision of the six classifiers on Aljazeera Arabic news dataset.

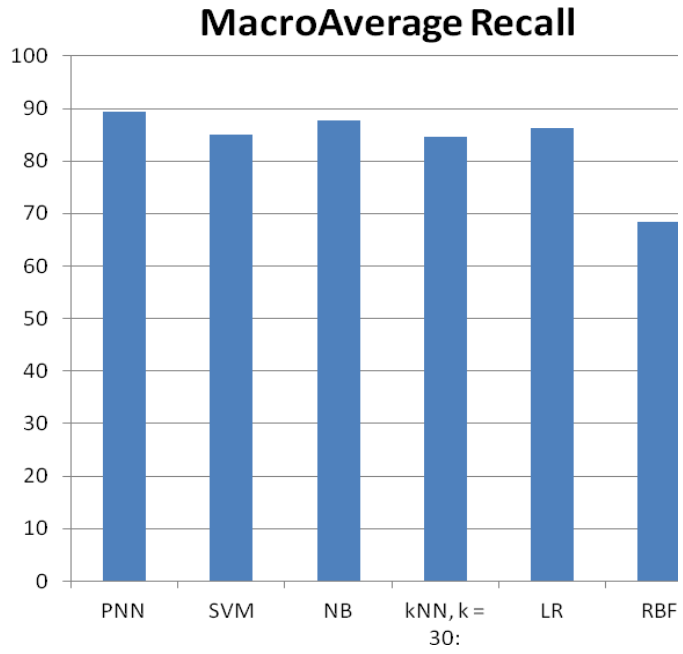

Fig. 2. MacroAverage recall of the six classifiers on Aljazeera Arabic news dataset.

Results of applying the six classifiers on Aljazeera Arabic News dataset are presented in Figs. 1 through 6. Direct comparisons of applying PNNs and the other five well-known classifiers on the same dataset using the same settings apparently show that PNNs are superior Arabic TC classifiers on Alj-News dataset using only 1% of each class terms. PNNs outperformed the other state-of-the-art classifiers using all evaluation measures, with a macroaverage precision of 90% and an average F1 of 89.3%. The next top performer was NB with a microaverage F1 of 87.67 and a macroaverage F1 of 87.97. LR got the third rank with a microaverage F1 of 86.33 and a macroaverage F1 of 86.51, then comes SVM, kNN and finally the RBF networks.

In fact, this superior performance of PNNs in Arabic TC is consistent with its performance in English TC as reported by [14]. This superior performance can be attributed to the nature of PNNs which have proved to achieve high classification performance in a non-iterative manner, provided that proper FS and reduction methods are used in selecting a tiny subset of the dataset features to build the PNN classifier. Authors in [14] have also proved that PNNs could achieve distinctive performance even on rare classes (which have a very small number of features), as well as on classes which have very close topics.

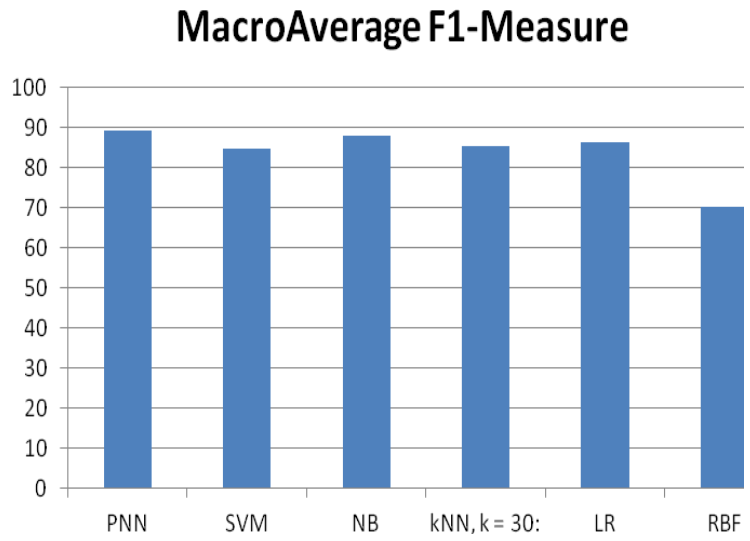

Fig. 3. MacroAverage F1-measure of the six classifiers on Aljazeera Arabic news dataset.

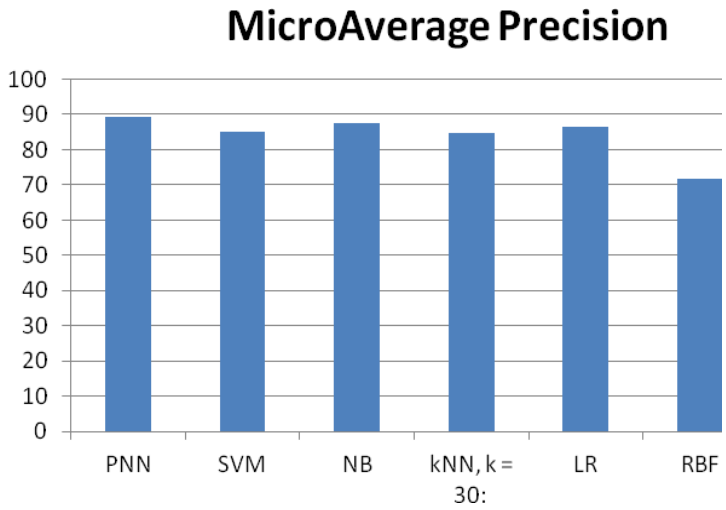

Fig. 4. MicroAverage precision of the six classifiers on Aljazeera Arabic news dataset.

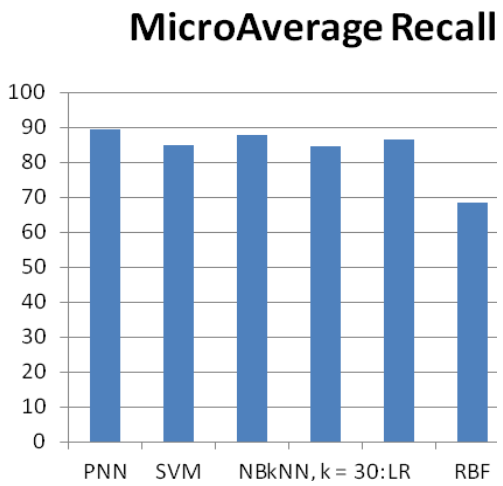

Fig. 5. MicroAverage recall of the six classifiers on Aljazeera Arabic news dataset.

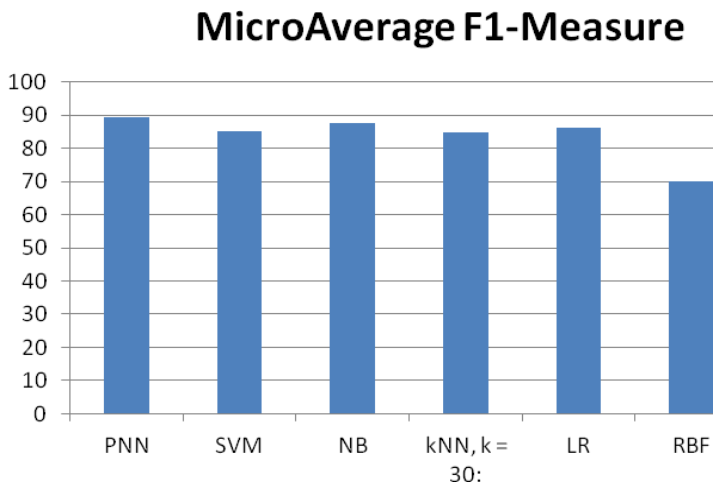

Fig. 6. MicroAverage F1-measure of the six classifiers on Aljazeera Arabic news dataset.

6. Conclusion

In this paper, Polynomial Neural Networks (PNNs) are directly compared to 5 well-known TC algorithms on Aljazeera Arabic News dataset. These algorithms are: SVMs, kNN, NB, LR and RBF. Text pre-processing is applied to the corpus documents, Chi Square FS and a local class-based reduction feature policy are used to select only 1% of each class most discriminating terms to build the six classifiers. Results of the direct comparisons of the six classifiers conducted in this research have proved that PNNs is the best Arabic TC algorithm on Alj-News dataset using these settings. More importantly, PNNs are able to achieve this performance in one shot (non-iteratively) compared to other iterative TC algorithms. PNNs have also proved to be able to record superior results despite all the known weakness points of stemming. My intended near future work is to test PNNs in other related areas, like topical crawling and spam filtering in both Arabic and English.

References

[1] Internet world users by language-top 10 languages. (2015). Retrieved June 2015, fromAvailable from: http://www.internetworldstats.com/stats7.htm

[2] Kourdi, E. M., Bensaid, A., & Rachidi, T. (2004). Automatic arabic document categorization based on the

naïve bayes algorithm. *Proceedings of the Association for Computational Linguistics Workshop on Computational Approaches to Arabic Script-based Languages* (pp. 51-58).

[3] Duwairi, R. M. (2007). Arabic text categorization. *Int. Arab J. Inf. Technol. 4(2)*, 125-132.

[4] Al-Shalabi, R., Kanaan, G., & Gharaibeh, M. (2006). Arabic text categorization using KNN algorithm. *Proceedings of the 4th International Multiconference on Computer Science and Information Technology* (pp. 5-7).

[5] El-Halees, A. (2008). A comparative study on Arabic text classification. *Egyptian Computer Science Journal*, *30(2)*.

[6] Al-Saleem, S. (2010). Associative classification to categorize Arabic data sets. *The International Journal of ACM JORDAN*, *1,* 118-127.

[7] Mesleh, A. A. (2007). Chi square feature extraction based SVMs Arabic language text categorization system. *Journal of Computer Science*, *3(6),* 430-435.

[8] Chantar, H. K., & Corne, D. W. (2011). Feature subset selection for Arabic document categorization using BPSO-KNN. *Proceedings of the 2011 Third World Congress on Nature and Biologically Inspired Computing* (pp. 546-551).

[9] Al-Harbi, S., Almuhareb, A., Al-Thubaity, A., Khorsheed, M., & Al-Rajeh, A. (2008). Automatic arabic text classification, *JADT'08*, 77-83.

[10] Harrag, F., El-Qawasmeh, E., & Pichappan, P. (2009). Improving Arabic text categorization using decision trees. *Networked Digital Technologies*, 110-115.

[11] Khreisat, L. (2006). Arabic text classification using n-gram frequency statistics a comparative study. *Proceedings of the Conference on Data Mining| DMIN*.

[12] Kanaan, G., Al-Shalabi, R., Ghwanmeh, S., &Al - Ma'adeed, H. (2009). A comparison of text — Classification techniques applied to arabic text. *Journal of the American Society for Information Science and Technology*.

[13] Al-Tahrawi, M. M., & Al-Khatib, S. N. (2015). Arabic text classification using polynomial networks. *Journal of King Saud University-Computer and Information Sciences, 27(4)*, 437-449.

[14] Al-Tahrawi, M. M., & Abu, Z. R. (2008). Polynomial networks versus other techniques in text categorization. *International Journal of Pattern Recognition and Artificial Intelligence*, *22(2),* 295-322.

[15] Al-Tahrawi, M. M. (2013). The role of rare terms in enhancing the performance of polynomial networks based text categorization. *Journal of Intelligent Learning Systems and Applications*, *5*, 84-89.

[16] Al-Tahrawi, M. M. (2014). The significance of low frequent terms in text classification. *International Journal of* Intelligent *Systems*, *29(5)*, 389-406.

[17] Al-Tahrawi, M. M. (2015). Class-based aggressive feature selection for polynomial networks text classifiers — An empirical study, *UPB Scientific Bulletin, Series C, 77(2)*, 93-110.

[18] Sawaf, H., Zaplo, J., & Ney, H. (2001). Statistical classification methods for Arabic news articles. *Natural Language Processing in ACL2001*.

[19] El-Halees, A. M. (2007). Arabic text classification using maximum entropy. *Islamic University Journal*. *15(1)*, 157-167.

[20] Said, D., Wanas, N., Darwish, N., & Hegazy, N. (2009). A study of Arabic text preprocessing methods for text categorization. *Proceedings of the 2nd Int. conf. on Arabic Language Resources and Tools, Cairo, Egypt*.

[21] Harrag, F., Al-Salman, A. M. S., & BenMohammed, M. (2010). A comparative study of neural networks architectures on Arabic text categorization using feature extraction. *Proceedings of the 2010 International Conference on Machine and Web Intelligence* (pp. 102-107).

[22] Zahran, B. M., & Kanaan, G. (2009). Text feature selection using particle swarm optimization algorithm

1. *World Applied Sciences Journal*.
[23] Alsaleem, S. (2011). Automated Arabic text categorization using SVM and NB. *Int. Arab J. e-Technol, 2(2)*, 124-128.
[24] Khorsheed, M. S., & Al-Thubaity, A. O. (2013). Comparative evaluation of text classification techniques using a large diverse Arabic dataset. *Language Resources and Evaluation*, *47(2)*, 513-538.
[25] Belkebir, R., & Guessoum, A. (2013). A hybrid BSO-CHI2-SVM approach to Arabic text categorization. In *Proceedings of the 2013 ACS International Conference on IEEE Computer Systems and Applications*.
[26] Sharef, B. T., Omar, N., & Sharef, Z. T. (2014). An automated Arabic text categorization based on the frequency ratio accumulation. *Int. Arab J. Inf. Technol, 11(2)*, 213-221.
[27] Ababneh, J., Almomani, O., Hadi, W., El-Omari, N. K. T., &Al-Ibrahim, A. (2014). Vector space models to classify Arabic text. *International Journal of Computer Trends and Technology (IJCTT)*, *7(4)*, 219-223.
[28] Al-Tahrawi, M. M. (2015). Arabic text categorization using logistic regression. *International Journal of Intelligent Systems and Applications*, 71-78.
[29] Keinosuke, F. (1990). *Introduction to statistical pattern recognition.* Academic Press, *Boston*.
[30] Liu, C.-L. (2006). Polynomial network classifier with discriminative feature extraction. *Proceedings of the 2006 Joint IAPR International Conference on Structural, Syntactic, and Statistical Pattern Recognition* (pp. 732-740).
[31] Campbell, W., Assaleh, K., & Broun, C. (2001). A novel algorithm for training polynomial networks. *Proceedings of the International NAISO Symposium on Information Science Innovations ISI*'2001.
[32] Assaleh, K., & Al-Rousan, M. (2005). A new method for Arabic sign language recognition. *Hindawi Publishing Corporation.*
[33] Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.
[34] Zheng, Z., Wu, X., & Srihari, R. (2004). Feature selection for text categorization on imbalanced data. *ACM Sigkdd Explorations Newsletter*, *6(1)*, pp. 80-89.
[35] Domingos, P., & Pazzani, M. (1997). On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*.
[36] Lewis, D. D., & Ringuette, M. (1994). A comparison of two learning algorithms for text categorization. In *Proceedings of the Third annual Symposium on Document Analysis and Information Retrieval*.
[37] Kim, S.-B., Seo, H.-C., & Rim, H.-C. (2003). Poisson naive bayes for text classification with feature weighting. Proceedings *of the Association for Computational Linguistics Sixth International Workshop on Information Retrieval with Asian languages.*
[38] Lee, Y. (1991). Handwritten digit recognition using K nearest-neighbor, radial-basis function, and backpropagation neural networks. *Neural computation*, *3(3),* 440-449.
[39] Wettschereck, D., & Dietterich, T. (1991). Improving the performance of radial basis function networks by learning center locations. *Advances in Neural Information Processing Systems, 4.*
[40] Vapnik, V. (2013). *The Nature of Statistical Learning Theory*, Springer Science & Business Media.
[41] Joachims, T. (1998). *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*. Springer.
[42] Joachims, T. (2002). *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms.* Kluwer Academic Publishers.
[43] Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and* Knowledge *Discovery*..
[44] Vapnik, V. N., & Vapnik, V. (1998). *Statistical Learning Theory*. Wiley, New York.
[45] Hoi, S. C., Jin, R., & Lyu, M. R. (2006). Large-scale text categorization by batch mode active learning.

Proceedings of the 15th ACM international conference on World Wide Web.

[46] Zhang, J., Jin, R., Yang, Y., & Hauptmann, A. G. (2003). Modified logistic regression: An approximation to SVM and its applications in large-scale text categorization. *Proceedings of the 20th Int. Conf. Machine Learning (ICML)*.

[47] Komarek, P., & Moore, A. (2005). Making logistic regression a core data mining tool: A practical investigation of accuracy, speed, and simplicity, Carnegie Mellon University, Pittsburgh, PA.

[48] Alj-news. Retrieved, from http://filebox.vt.edu/users/dsaid/Alj-News.tar.gz

[49] Samir, A., Ata, W., & Darwish, N. (2005). A new technique for automatic text categorization for Arabic documents. *Proceedings of the 5th IBIMA Conference the Internet and Information Technology in Modern Organizations*.

[50] Khoja, S., & Garside, R. (1999). Stemming arabic text. *Lancaster*. UK, Computing Department, Lancaster University.

[51] Arabic stemmer. Retrieved, from http://zeus.cs.pacificu.edu/shereen/ArabicStemmerCode.zip

[52] Brank, J., Grobelnik, M., Milic-Frayling, N., & Mladenic, D. (2002). Interaction of feature selection methods and linear classification models.

[53] Bekkerman, R. (2003). Distributional clustering of words for text categorization. Citeseer.

[54] Rogati, M., & Yang, Y. (2002). High-performing feature selection for text classification. *Proceedings of the Eleventh International Conference on Information and Knowledge Management*.

[55] Thabtah, F., Eljinini, M., Zamzeer, M., & Hadi, W. (2009). Naïve Bayesian based on chi square to categorize Arabic data. *Proceedings of the 11th International Business Information Management Association Conference (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies, Cairo*.

[56] [Fodil, L., Sayoud, H., & Ouamour, S. (2014). Theme classification of Arabic text: A statistical approach. *Terminology and Knowledge Engineering*.

[57] Eldin, S. (2007). Development of a computer-based Arabic lexicon. *Int. Symposium on Computers & Arabic Language.*

[58] Eldos, T. (2002). *Arabic Text Data Mining: A Root Extractor for Dimensionality Reduction*. ACTA Press, A Scientific and Technical Publishing Company.

[59] Debole, F., & Sebastiani, F. (2005). An analysis of the relative hardness of Reuters — 21578 subsets. *Journal of the* American *Society for Information Science and Technology*.

[60] Van, R. C. J. (1979). *Information Retrieval*. London.

Mayy M. Al-Tahrawi obtained her B.Sc. degree in computer science from Kuwait University in 1986, a M.Sc. also in computer science (compiler generators) from Kuwait University in 1988, and a Ph.D. in computer information systems from the Arab Academy for Banking and Financial Sciences, Jordan in 2006. She has worked as a lecturer in several universities in Jordan and Kuwait since 1988. Currently, she is an associate professor and the head of Computer Science Department at Al-Ahliyya Amman University, Amman, Jordan. Her research interests are in machine learning, pattern recognition, feature selection, text categorization, spam filtering and information retrieval.