Intelligent Agent for Information Extraction from Arabic Text without Machine Translation

Tarek Helmy[*] Abdirahman Daud

Information and Computer Science Department,
College of Computer Science and Engineering,
King Fahd University of Petroleum and Minerals,
Dhahran 31261, Mail Box#413, Saudi Arabia
{helmy,daud}@kfupm.edu.sa
** On leave from College of Engineering, Department of Computers Engineering and Automatic Control, Tanta University, Egypt*

Abstract. The process of classifying text into two opposing opinions is known as sentiment polarity classification. It has been shown in the literature that this problem cannot reach accuracy higher than 80-85%. This paper shows that a higher accuracy (96%) can be achieved without the need to translate text into English language. More specifically, our case study is: Islamic Hadith Narration. The problem is to tell whether a person is trustworthy or not based on his biographical data. With such high accuracy, the agent can be used to create new books in the area of Hadith automatically instead of manual classification done before. The results of our experiments encourage the use of an intelligent agent for information extraction using supervised learning, domain knowledge and number of natural language processing techniques.

Keywords: Sentiment Analysis, Supervised Learning, Information Extraction, Machine Learning, Natural Language Processing, Arabic, Machine Translation

1 Introduction

Classification is defined as the process of finding a model (or function) that describes and distinguishes data classes or concepts. The goal of classification is to use the model to predict the class of new objects whose class label is unknown [17]. One example of a classification problem is sentiment analysis. Sentiment analysis is defined as the computational treatment of opinion, sentiment, and subjectivity in text or by the use of Natural Language Processing (NLP) techniques [2]. One popular problem of sentiment analysis is Polarity Text Classification (PTC). The problem is defined as follows: given a certain subject, how can we classify opinions (written in text) into positive or negative with regard to that subject? A famous application of PTC is movie reviews where several research papers tried to use supervised machine learning techniques that can tell if a user's review is positive or negative about a movie [4]. Research in PTC could not achieve accuracy more than 85%.

In this paper a new application is introduced in order to find ways to improve PTC and information extraction in general. The new application is Arabic Hadith

Narration (AHN). The problem of AHN is an example of opinion mining (sentiment analysis). This problem has been selected for the purpose of information extraction. In the area of Hadith, many entries are missing for infamous men and it takes a lot of resources to fill these gaps manually. Another motivation for this research is to see how much the properties of a language can have on the accuracy of sentiment analysis? Earlier works where almost all on English language and it has been suggested to translate Arabic text to English first then continue the process [18]. Our intuition is that such translation is not necessary.

The rest of this paper is organized as follows. Section 1 gives backboard information on the AHN process and two classification techniques to make the reader familiar with the rest of the paper. Section 2 summarizes the related work in the area of sentiment analysis. After that, we introduce intelligent AHN in Section 3. Section 4 presents our experiments. Then, the results are analyzed in Section 5. Finally, Section 6 concludes the paper.

1.1 Arabic Hadith Narration

Hadith is tradition relating to the sayings and doings of the Islamic prophet Muhammad and his companions. Hadith had been narrated throughout the centuries. Written in Arabic, it is a sacred source for Islamic wisdom and teachings. Thus, to preserve its authenticity, Islamic scholars have developed a process to save the list of narration that carried Hadith from a generation to generation. This list, composed of number of people, is later reviewed and verified. When a scholar reviews a narrator in the list, he concludes by giving the narrator a label ranging from a trustworthy person to an untrustworthy person. If an untrustworthy person appears in a narration list, that narration becomes invalid and the Hadith is generally not accepted. [1]

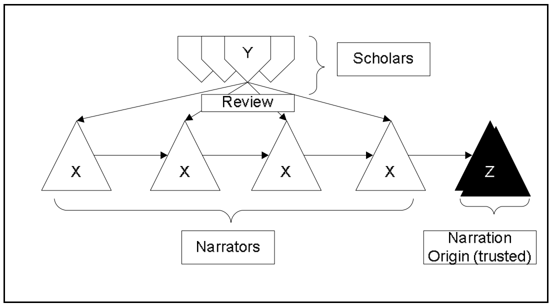

Fig. 1. Process of Hadith Narration

Figure1 illustrates this process. Narrators, labeled as X, are reviewed and criticized by a number of scholars, labeled Y. These Y (scholars) have published books as old as the tenths century. In these books, biographical data about X (narrators) are recorded. Not all X are labeled, only famous ones.

A sample of two biographical texts for two narrators (one trustworthy and untrustworthy) is given in Figures 2 and 3 taken from [15], [16]. As highlighted, scholars who wrote these samples used a set of words (underlined) to raise or criticize the person in question. Since there are many reviewers, conflicting biographical information are available for the same person in some cases. In total, there are more than 40,000 people whose biographical information is written by more than 1300 scholars [11]. It is very hard if not impossible for a human to summarize or to come into a general conclusion of such huge conflicting information.

إسماعيل بن مجالد صالح .وقال عثمان بن أبي شيبة كان
إسماعيل بن مجالد ثقة وصدوقا وليتني كنت كتبت عنه كان
يحدث عن أبي إسحاق وسماك وبيان وليس به بأس وقال
أحمد بن حنبل ما أراه إلا صدوقا

Ismail bin Mujalid is valid. Othman bin Abi Shayba said Ismail bin Mujalid is trusted and honest I wish I wrote about him, he used to narrate from Abu Ishaq and Samaak and Bayan and there is nothing wrong with him. Ahmed bin Hanbal said “I see him nothing but an honest person”.

Fig. 2. Biographical information for trustworthy person with English translation blow

As mentioned earlier, AHN is very similar to the problem of movie reviews. In this application, biographical data would replace movie reviews. Similarly, the positive and negative labels would be mapped into trustworthy or untrustworthy.

حبة العرني كوفي. حدثنا محمد بن عيسى قال حدثنا عباس
قال سمعت يحيى يقول قد رأى الشعبي رشيدا الهجري وحبة
العرني والاصبغ بن نباتة وليس يسوى هؤلاء كلهم شيئا .
حدثنا محمد قال حدثنا عباس في موضع آخر قال سمعت
يحيى قال حبة العرني لا يكتب حديثه.

Hiba Al-Aerni is Kofi. Muhammad ibn Isa said Abbas told us "I heard Yahya saying, Al-Sha’bi’s have saw Rashid Al Hijri, Hiba Aerni and Asbg bin Nabata and all of them are useless. Mohamed said Abbas told us in another place: I heard Yahya saying the narration of Hiba Al-Aerni should not be written.

Fig. 3. Biographical information for untrustworthy person with English translation below

1.2 Support Vector Machines

Introduced by Cortes and Vapnik [7], Support Vector Machines (SVM) represents the state-of-the-art in the field of machine learning by being one of the most effective classifiers among others in supervised learning [5]. The idea of SVM is to draw a separating hyper plane between data (represented as points). This hyper plane tries to have the minimum amount of error while maximizing the margin between data point and the separating hyper plane as shown in Figure 4.

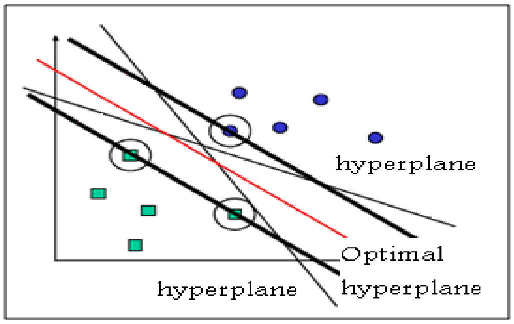

Fig. 4. Data points separation in Support Vector Machine

This classifier works effectively with data that has huge number of features regardless of whether it is linearly separable or not [5]. In the area of sentiment analysis, several papers have shown that SVM is the best classifier for sentiment analysis problem [2], [3], [4].

1.3 Bayes Point Machine

Another classifier close to SVM is Bayes Point Machine (BPM) introduced by Herbrich et al. [6]. BPM uses Bayesian inference when drawing a separating line. In Figure5, we can see that BPM tries to take the average line that separates the points while SVM tries to maximize the distance between data classes [8].

Theoretically it has been shown that BPM can have better learning ability than SVM and also in image classification, it has shown that BPM has better accuracy [9], [10].

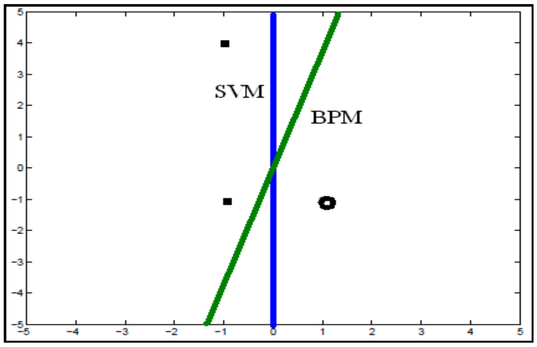

Fig. 5. Difference between SVM and BPM

Our goal of this research is show how an intelligent agent can successfully extract knowledge from Arabic biographical test. In addition, the output of the research will include answers to the following questions:

- Is machine translation needed for sentiment analysis on non-English text?

- Since SVM and BPM are the most popular classifiers, which of these machine learning techniques is best suited for sentiment analysis?

2 Literature Review

Most sentiment analysis research has being going on mostly in English language text [2], [19]. Up to our knowledge, there is no single paper dedicated to sentiment analysis in Arabic Language. The conclusion from these research papers as highlighted by the survey in [2] is that researches in sentiment analysis and more specifically in PTC could not achieve more than 80- 85% accuracy. This accuracy is significantly lower than accuracy achieved in topic text classification [2]. To improve accuracy, several ideas have been proposed as in [4] which implemented NLP techniques along with machine learning. The authors stated that their improvement was marginal and could not add accuracy more than 1% using an SVM classifier.

The field of AHN has not been studied in computer society. However, huge efforts are going on to manually gather and summarize thousands of Arabic books [11]. Up to our knowledge, no work has been suggested to apply machine learning techniques to automate the AHN process. Current scholars of AHN perfume many tasks in order to draw a conclusion about a single narrator in the narration list. The whole process illustrated before in Figure1 is done manually. First, scholars need to search for a single narrator (there are more than 40,000) within books written by over 1300 old scholars in the tenth century. Then, the difference in opinion among old scholars has to be resolved in order to draw a general conclusion about a narrator. Except for searching, the whole process is done manually therefore it is not only prone to subjectivity but it also takes a lot of time. An example is shown blew and in Figure6.

Example 1: In Figure 2, a single biographical entry for Ismail bin Mujalid is shown. Ismail can appear in several narration lists. Therefore, to authenticate every list that has Ismail bin Mujalid, Ismail has to be first classified as a trustworthy or untrustworthy. A scholar will need to gather all the books that mention Ismail (number of books range from tens to hundreds according to the popularity of Ismail). It is common to see that two scholars have different conclusions about Ismail since this is a subjective manual process. According to Hadith scholars, drawing conclusion for all narrators can take up to 10 years with several scholars working together.

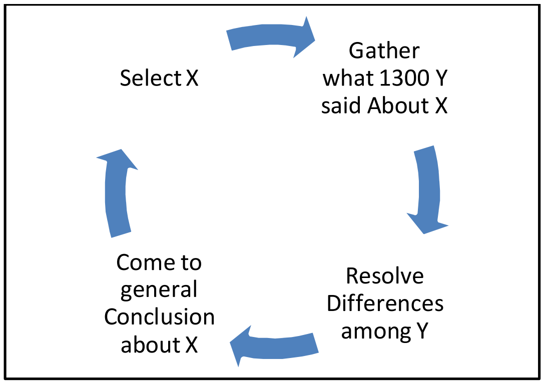
Select X
Gather what 1300 Y said About X
Resolve Differences among Y
Come to general Conclusion about X

Fig. 6. Manual Hadith Narration Carried by New Scholars

3 Intelligent Agent for AHN

An intelligent agent for AHN can assist scholars and therefore open the door for more application of machine learning in Arabic art, history and culture. In the same time, we would try to suggest improvements that might work in other languages. The first improvement as suggested in [4] is to include NLP techniques while extracting features. Obviously, since Arabic and English have different syntax grammars, we expect the output of their NLP techniques to be different and have different improvement levels on PTC as well. For example, negation in Arabic is some cases is easier to detect compared to English. Instead of only using the adverb "not", there are several adverbs in Arabic language and each of these negation adverbs has a particular pattern based on if negation applies on a verb, noun or an adjective. This can ease the detection task. It also helps the model know when a negative replica of a feature (special terms that assist in classifying text) is needed. This reduces the size of feature lists. Not only negation detection is needed but also irrelevant evaluations in a review (biographical text) have to be removed. Unlike movie reviews, AHN text might contain evaluation for more than one person in the same entry. One entry of biographic information, such as shown in Figure 2, might contain reviews for other people associated with the person in question. This requires the feature extracting module to discard any irrelevant evaluations of a person such as his father, bother or son which appears frequently.

Another problem is most scholars of AHN include other scholars' opinions while writing the biography of a narrator. It is logical to assume that the better a person is, the more frequent positive features (words) will appear in his biography and therefore saving the frequency of features could increase the accuracy of the classifier. This however conflicts with the reported findings in [2] which suggest that appearance, saved as a binary value (exist or does not exist), achieves more accuracy than term frequency. Since this is a new domain, we need to examine if this finding holds true in the case of AHN.

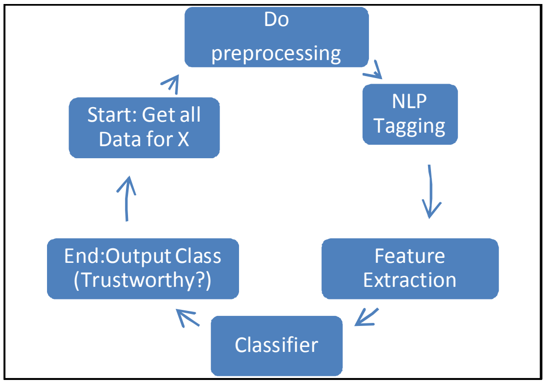

Fig. 7. Automated Process of Hadith Narration

Finally, as we have seen in the literature, BPM and SVM classifiers are closely related and have been compared in several studies [6], [8], [9] and [10]. Therefore it is beneficial to investigate which of these two classifiers can work better in the area of sentiment analysis.

4 System Design, Results & Verification

This section describes the system design for classifying Arabic biographic data into trustworthy or untrustworthy (positive or negative). The system is illustrated in Figure 7. Here we elaborate more on data, features, feature extraction, implementations, workstation environment and success criteria.

4.1 Data

Biographic information is found in either three types of books: books that contain only trust worthy people, those which contain only untrustworthy and those who are mixture of both. We selected two books for our data sampling: 50% of the data came from a book for trustworthy people and the other 50% from a book holding biographical entries for untrustworthy people [15], [16]. This ratio is maintained in the training and testing datasets.

4.2 Features

An expert in the domain of Islamic studies is required to help choose the best features (words) to classify biographical text into trustworthy or untrustworthy. We will use a text book as our domain knowledge expert. This book [12] has listed all the terms usually used by scholars (reviewers) to criticize or upraise a person. The number of these special words is around 160 and they can be considered all to be the feature list. More formally, each feature X_i has a discrete binary value of 0 or 1. If $X_i = 0$, this means the word *i* has not been found. However, these terms need to be refined (using Stop and Stemming techniques) since they are mentioned in the domain knowledge book as phrases (not exactly words). This Stop and Stemming process would be done manually. See the example below for which words the model looks for.

Example 2: In Figure 2, the underlined words (tested, valid, honest, nothing wrong, useless, should not be written from) are marked as features. These features including other tens of words have been specified by [12] who acts as our domain expert.

4.3 Feature extraction Module

Feature extraction module first applies Parts Of Speech (POS) labeling on the feature lists to know whether a negative replica is required or not. For example, the feature Valid can appear without negation or with negation. As an output, the module processes the biographical Arabic text and feature list to output a matrix. This matrix contains data label (trustworthy or untrustworthy), and a string of integers that indicate the count of each feature in every biography entry. In the process of extracting features, the module will perform the specified techniques in Section 3. The example below illustrates in more detail.

Example 3: In Figure 2, the Arabic text can be represented as one raw similar to the following (1:101013) where the first bit indicates A is trustworthy and the following numbers mark the count of certain words (features) in A. Similarly B can be represented as (0:012120) which means B is untrustworthy.

4.4 Classifier and Post-Processing

After features are extracted, it is represented in a matrix form. The rows mark data points whereas columns stand for feature numbers. This matrix is then fed into SVM and BPM classifiers. The experiment is repeated with fixed portion of the matrix to tune for best attributes for each classifier. Then, each classifier is trained and tested for 10 epochs. Finally, the success criteria are calculated for SVM and BPM. This procedure is for testing and training the classifiers. However in practice, once the classifiers are trained, they are used directly.

Post-processing is done in case a narrator X has several biographical entries, each entry is classified separately. Then, the average output marks the final class of X. Optionally the user can give higher weight to preferred scholars/biographical sources.

4.5 Environment

The BPM classifier used is implemented using Microsoft Infer.Net framework [13]. The SVM classifier is using SVM.NET which is a Microsoft .NET library for LIBSVM [14]. The experiments were carried out on AMD workstation with CPU speed 2 GHz and 2.5 GB of RAM running Windows XP.

4.6 Success criteria

The success criteria are expressed in terms of percentages of Accuracy, Positive Perdition rate and Negative Perdition rate. Let TP, FP, TN and FN stand for: True

	Data size	Training Time	Accuracy	Positive Prediction Rate	Negative Prediction Rate
BPM	526	~ 2 min	51.83 %	6.3%	98%
BPM+ NLP	526	~ 2 min	57.6 %	6.7%	99.3%
BPM +NLP+ Term Presence	526	~ 2 min	59.7%	7.1%	100%
SVM	526	~ 40 sec	95.91%	96.5%	95%

Table 1. Results Using SVM and BPM Classifiers

Positive, False Positive, True Negative and False Negative respectfully. Positive Predation value (or rate) and Negative Prediction rate are calculated as follows:

$$\text{Positive Predation rate} = \frac{TP}{TP+FP}$$

$$\text{Negative Predation rate} = \frac{TN}{TN+FN}$$

Accuracy is the ratio of sum of all data points classified correctly over the number of all data points.

$$\text{Accuracy} = \frac{TP+TN}{All\ Points}$$

5 Results

The Experiments were carried out on 526 biographical information entries which are equal to 526 individuals from two sources [15], [16] as explained in Section 4. The results are shown in Table 1 for BPM, BPM with NLP techniques, BPM+NLP with term presence instead of term frequency and finally with SVM. It can be seen from Table 1, NLP techniques and term presence have improved the accuracy of BPM classifier for about 6% and 1.4% respectfully. This improvement is only visible to BPM while the SVM classifier has not experienced any changes with these modifications and maintained the same high accuracy. The time efficiency of SVM is also higher than that of BPM. This high accuracy proves that there is no need for machine translation for non-English language.

An important point to notice is the positive effect of NLP tasks on BPM which is more obvious than SVM. This suggests that it is possible, with more NLP processing, that BPM or any other classifier can excel SVM.

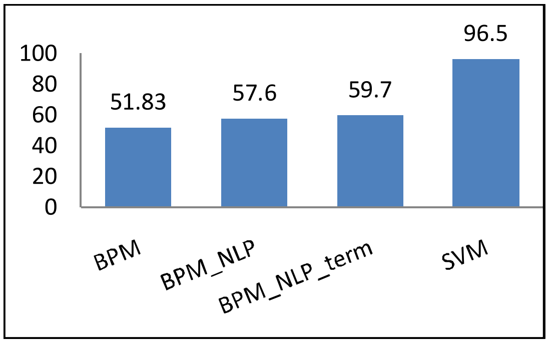

Fig. 8. Difference in Accuracy between SVM and BPM\

6 Conclusion

This work has showed how an intelligent agent can perform near accurate information extraction. Therefore, this paper encourages the use of machine learning without machine translation information extraction. The paper confirms the results found in literature that SVM is a more practical classifier than BPM. However, it can be seen that the accuracy of BPM can be increased using NLP techniques significantly along with term presence instead of term frequency. More notably, this is the first time we get accuracy results in sentiment analysis similar if not higher than other textual classification (95%-96% using SVM). A future work could be to apply sentiment analysis on other languages and domains.

Acknowledgments

The authors would like to thank King Abdulaziz City for Science and Technology (KACST) Project No: GSP-18-106, for the financial support. Also we would like to thank King Fahd University of Petroleum & Minerals for providing the support and utilized computing facilities.

References

1. Kathir I.O, an Abbreviation for Hadith studies (1339)
2. B. Pang and L. Lee.: Opinion Mining and Sentiment Analysis. Now Publishers Inc, July (2008)
3. B. Pang, L. Lee, and S. Vaithyanathan: Thumbs up? Sentiment classification using machine learning techniques, In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 79–86, (2002)
4. A. Kennedy and D. Inkpen: Sentiment classification of movie reviews using contextual valence shifters , Computational Intelligence, vol. 22, pp. 110–125, (2006)

5. S. Kotsiantis: Supervised Machine Learning: A Review of Classification Techniques, Informatica Journal, pp. 249-268. (2007)
6. R. Herbrich, T. Herbrich, T. Graepel, and C. Campbell: Bayes Point Machines, Journal of Machine Learning Research, pp.245-279, (2001)
7. C. Cortes, and V. Vapnik: Support Vector Networks, Machine Learning, pp.273-297 (1995)
8. A. Kapoor, an Analytical Comparison between Bayes Point Machines and Support Vector Machines, (2002), retrieved from: http://research.microsoft.com/en-us/um/people/akapoor/papers/kapoor_bpm.pdf.
9. G. Wu, E. Chang, L. Chung-Sheng: BPMs versus SVMs for image classification. In: IEEE International Conference on Multimedia and Expo, pp. 505-508.(2002)
10. W. Cao, S. Meng: Image classification based on Bayes point machines. In: IEEE International Workshop on Imaging Systems and Techniques,, pp.164-167, (2009).
11. Jaamia' Al Sunna. (2009) Egypt: http://arabia-it.com/esunna.aspx
12. Al-Dahabi M.A. :Words of Jarh and Ta'dil, (1330 AD)
13. T. Minka, J. Winn, J. Guiver, and A. Kannan: Infer.NET 2.3. Microsoft Research Cambridge, (2009) http://research.microsoft.com/infernet
14. C. C. Chang and C.J. Lin, LIBSVM: a library for support vector machines, (2001) http://www.csie.ntu.edu.tw/~cjlin/libsvm.
15. O.A. Shahin: History of Trustworthy Names, (995)
16. A.H. Okeyli,: The Great Book of the Untrustworthy, (933)
17. J. Han, and M. Kamber: Data Mining: Concepts and Techniques, Morgan Kaufmann, San. Francisco, (2000)
18. Carmen et al.: Multilingual subjectivity analysis using machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp.127-135 Honolulu, Hawaii (2009)
19. Tarek Helmy, Ali Bahrani and Jeffery Bradshaw: Agent-Oriented Service Model for Personal Information Manager, Lecture Notes in Computer Science (LNCS-Springer-Verlag), Volume 5907, pp.24-40, (2009)