The Effect of the Sentence Location on Arabic Sentiment Analysis

Saud S. Alotaibi[1†]
ssotaibi@uqu.edu.sa
Umm Al-Qura University, Information Systems Department, Makkah, Saudi Arabia

Abstract
Rich morphology language such as Arabic needs more investigation and method to improve the sentiment analysis task. Using all document parts in the process of the sentiment analysis may add some unnecessary information to the classifier. Therefore, this paper shows the ongoing work to use sentence location as a feature with Arabic sentiment analysis. Our proposed method employs a supervised sentiment classification method by enriching the feature space model with some information from the document. The experiments and evaluations that were conducted in this work show that our proposed feature in the sentiment analysis for Arabic improves the performance of the classifier compared to the baseline model.

Keywords:
Sentiment Classification, Negation scope, Arabic Natural Language Processing, Arabic Sentiment Sentence Classification, Machine Learning.

1. Introduction

The extraction of sentiment from a text has attracted a considerable amount of attention over the past decade, both in industry and academia. Sentiment analysis attempts to extract the emotions and opinions of individuals from their writing about specific entities. Sentiment Analysis (SA) of Arabic is also still in its early stages, and increased effort and reliability of low-level tools are required in order to build upon this foundation. Many current approaches in Arabic sentiment analysis rely on the bag-of-word (BOW) model to build a feature vector model [1–4]. The other types of features than are added to the baseline model to leverage the performance of the classifier such as Morphological features [3], and stylistic features [5]. Analyzing sentiment in the long text may add some noise and unnecessary information that might influence the actual sentiment. Therefore, in this work, we proposed a new feature that helps to leverage the performance of the Arabic sentiment analysis for long text. Sentiment Analysis (SA) of Arabic is also still in its early stages, and increased effort and reliability of low-level tools are required in order to build upon this foundation. The most common linguistic aspect that affects sentiment analysis is negation. Negation often changes the sentiment and polarity of a sentence. For example, the following two sentences “this is a good movie” and “this is not a good movie” will have the same polarity if the negation item “not” is ignored in the sentiment analysis. The positive sentiment associated with the word “good” is inverted into negative sentiment for the phrase “not good” and may not necessarily be as negative as the sentiment associated with the word “bad”. Therefore, negation items and their scope in the sentence have to be taken into account during sentiment analysis [2]. Determining negation in sentences is not an easy task because of the complexity of the nature of negation. In the Arabic language, the negation words such as “not” and “no” not only show negation in the sentence but other semantic meanings. In addition, the negation could use sometimes to express other meanings or styles such as in questioning and wondering sentences.

The rest of this paper is organized as follows. The next section shows our proposed method follows the experiments and results in the discussion section. The conclusion and future works are displayed in the last section

2. Proposed Method

For document-level classification, we propose that the use of some parts of the documents, especially in the case of the long document, might improve the accuracy result for Arabic sentiment analysis. This intuition comes from the way in which users tend to give their opinions on a particular subject. Users express their feelings either at the beginning or the end of their writing. They almost always place some factual information, sometimes combined with opinions, in the middle of their writing. In order to show this intuition, the next section will show the data that is collected to illustrates this idea.

2.1 Arabic Sentiment Corpus

The authors of this work build their own corpus due to the scarcity of sentiment Arabic corpus. The research corpus was built from two different genres, which include newswire and movie reviews. The news data has been taken from the Sabq[1] website among different domains which are local, sport, economics, technology, and social

[1] http://sabq.org

Manuscript received May 5, 2022
Manuscript revised May 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.5.44

news. The movie reviews[1] were taken from the movie review website and is used in [6]. In total, our corpus contains 384 documents with more than 11269 sentences. There are around 154 positive documents, 52 negative documents, and 32 neutral documents resulting in a total of 268 subjective documents and 116 objective documents. Two Arabic-educated individuals have been chosen to annotate the data, contact the first author for getting this corpus.

2.2 Importance of the Sentence Location

The importance of the sentence location has been investigated in Arabic long text by counting the subjective sentences in different positions in each document. Figure 1 illustrates the number of sentences in each category of subjectivity or polarity in both news and movie reviews domains. In order to do this, each document is divided into five parts. The sentence of each document then is distributed equally to each part. Then, the number of the sentence in each sentiment class is counted in each part of the document.

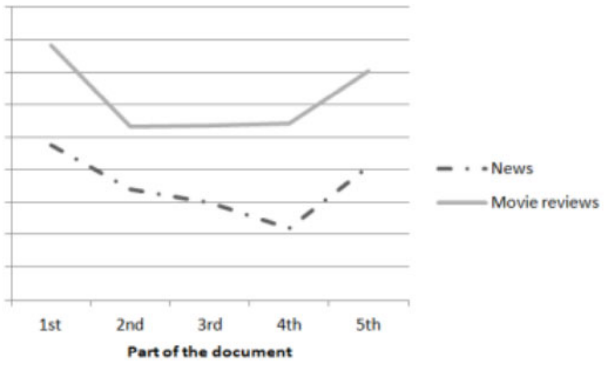

Fig. 1 Number of Opinioned Sentences Depends on its Position in two Different Domains

These figures clearly show that the most "feeling" opinioned sentences occur in either the early or the late positions of the document. Therefore, if the first and the last parts of the documents are only considered while analyzing the sentiment in long Arabic text, this may reduce the noise that could interfere with the actual feeling of the writer. Depending on this observation, we propose a feature that is desired to be included in the document-level sentiment classification in Arabic text.

2.3 Classification Process

[1] http://www.filfan.org

The preprocessing phase contains some steps before the text pass to the classifier. These steps include filtering out all rubbish data, normalizing long words that contain some letters redundant, and removing the stop words [7]. The document then is divided into parts, then the beginning and the ending part will be used in our proposed feature. To evaluate our method, experiments were undertaken using a support vector machine classifier (SVM) with a linear kernel with 5-fold cross-validation using the scikit-learn library [8]. As a baseline model, the uni-gram model is applied to all documents. To measure the performance of the classifier, the F1 score is used after computing the precision and recall.

3 Results and Discussion

In this experiment, the investigation of using the proposed feature is carried out. It also shows whether our proposed feature helps the classifier to learn better than using the whole document parts. The performance of the classifier using the sentence location method is shown in table 1. In this table, the bold numbers illustrate the best result that is achieved with different methods either using baseline or without sentence location approach. The NA in this table refers to the classification process is not applicable because there is not any objective document in the movie review domain. It is clear that using the position of the sentence approach works in some cases from the data shown in table 1. With polarity classification, we noticed that this approach plays a central role to increase the accuracy of the SVM. For example, the SVM achieves the best result with 78% in polarity classification with the newswire domain. In addition, the performance of SVM increased by 3% using the position approach in polarity classification in the Movie Reviews area. Moreover, the proposed method also helps in the case of the news field by increasing the result by 4%

Table 1: Result of comparing Baseline with Sentence Location Proposed Feature

		Subjectivity	Polarity
Movie Review	Baseline Model	NA	80%
	Sentence Location Model	NA	**83%**
Newswire	Baseline Model	63%	76%
	Sentence Location Model	**67%**	**78%**

4. Conclusion and Future Work

Our work in this paper shows the concept of using the whole of the document while analyzing sentiment for Arabic text may hurt the performance of the classifier using a machine learning algorithm. This work describes and comes up with the proposing investigating a new proposed feature that includes some of the document sentences to the process of the sentiment classification.

The results achieved by this paper demonstrate the potential gain obtained by our proposed feature. The performance increased by 3% during the process of classification. This improvement encourages to continue investigating this method in the future. One of possible direction is to use this proposed method with the medium Arabic text or another language such as English to investigate its effect to the process of sentiment analysis.

References

[1] Abdul-Mageed, M., Diab, M.T., Korayem, M.: Subjectivity and Sentiment Analysis of Modern Standard Arabic. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2. HLT '11, Stroudsburg, PA, USA, Association for Computational Linguistics (2011) 587–591

[2] Abdul-Mageed, M., Diab, M.: AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis. Proceedings of LREC, Istanbul, Turkey (2012) Pages 19–28

[3] Abdul-Mageed, M., K¨ubler, S., Diab, M.: SAMAR: A System for Subjectivity and Sentiment Analysis of Arabic Social Media. WASSA 2012 (2012)

[4] El-Halees, A.: Arabic opinion mining using combined classification approach. In: Proceeding The International Arab Conference On Information Technology, Azrqa, Jordan. (2011)

[5] Abbasi, A., Chen, H., Salem, A.: Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums. ACM Trans. Inf. Syst. 26 (2008) 12:1–12:34

[6] Farra, N., Challita, E., Assi, R.A., Hajj, H.: Sentence-Level and Document-Level Sentiment Mining for Arabic Texts. In: Data Mining Workshops (ICDMW), 2010 IEEE International Conference on. (2010) 1114 –1119

[7] El-Khair, I.A.: Effects of stop words elimination for arabic information retrieval: a comparative study. International Journal of Computing & Information Sciences 4 (2006) 119–133

[8] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12 (2011) 2825–2830.

Dr. Saud S. Alotaibi is an assistant professor of Computer Science at the Umm Al-Qura University, Makkah, Saudi Arabia. His current research interests include AI, Machine learning, Natural language processing, Neural computing IoT, Knowledge representation, Smart cities, wireless, and sensors. He received his Bachelor of Computer Science degree from King Abdul Aziz University, in 2000. Dr. Saud started his career as Assistant Lecturer in July 2001 at Umm Al-Qura University, Makkah, Saudi Arabia. He then earned his master's degree in computer science from King Fahd University, Dhahran, in May 2008. After that, he worked as a Deputy of the IT-Center for E-Government and Application Services, in January 2009, at Umm Al-Qura University. Under Dr. Charles Anderson's supervision, Saud completed his Ph.D. degree in Computer Science from Colorado State University in Fort Collins, US, in August 2015. From 2015 to 2018, Dr. Saud worked with the Deanship of Information technology to improve the IT services that are provided to the Umm Al-Qura University. Nowadays, Dr. Saud is working right now with the Computer and Information College as a vice dean for academic affairs.