Vol 11 No.2 (2024) 66-78
International Journal of Computers and Information (IJCI)
Available online at https://ijci.journals.ekb.eg/

Analysis and Mining of Arabic Comparative Sentences: A Literature Review

Alaa Sakr[a], Arabi Keshk[a], and Anas Youssef[a],
[a] *Computer Science Department, Faculty of Computers and Information, Menoufia University, Egypt*

Abstract

There are huge Arabic comparative sentences that are generated daily on various social media. These comparative sentences need to be analyzed and mined for different purposes such as service and product reviews. In general, analysis and mining of Arabic text is a big challenge due to the limitations inherited in the Arabic language. Moreover, there are currently no standard datasets for Arabic comparative sentences. This paper provides a background on the different steps required to analyze and mine an Arabic comparative sentence. These steps include the identification of the Arabic comparative sentence, the identification of the sentence type, and finally the extraction of a relation together with a preferred entity. The paper also provides a literature review of current research work applied in this research field. This includes a classification of the various techniques leveraged in this field including three main categories namely: linguistic, machine learning and deep learning approaches. Finally, the paper provides insights on current limitations and future research challenges in this field. To the best of our knowledge, this is the first research paper that provides a dedicated literature review about the analysis and mining of Arabic comparative sentences. This review discusses the specific analysis of Arabic comparative sentences not the general Arabic sentiment analysis. It is noted that this analysis is a subset of the Arabic sentiment analysis field which does not focus on identifying the sentiment of an Arabic sentence, however, it focuses on identifying and analyzing an Arabic comparative sentence and its components.

Keywords: Natural Language Processing; Arabic Text Mining; Comparative Sentence; Type Identification; Relation Extraction.

1. Introduction

Arabic Language has three dialects namely: Modern Standard Arabic (MSA) [1], Colloquial Arabic (CA) [2] and Quranic Arabic (QA). MSA is the Arabic dialect that is the most understandable between many of the Arabic speaking countries. CA represents the spoken language of many Arabic countries; however, this type possesses regional varieties and may even exist in the same country. QA represents that Arabic language written in Quran, the holy book of Islam. The focus of this review is on MSA and QA, however, CA is not considered in this review.

The objectives of this paper are as follows. Firstly, the paper provides a detailed background on the general steps required to analyze and mine Arabic comparative sentences. These steps start with the identification of the Arabic comparative sentence, followed by the identification of the sentence type, and the last two steps are the extraction of a relation together with a preferred entity. Secondly, a classification of the different techniques used in this field is listed and discussed. The different techniques broadly include linguistic, machine learning and deep learning approaches. Finally, the paper presents and provides a detailed discussion of the limitations of current research techniques and future enhancements that can be applied in this field.

The rest of the paper is organized as follows. Background on mining of Arabic comparative sentences is presented and discussed in section 2. Section 3 provides a classification of research techniques for mining of Arabic comparative sentences. Section 4 discusses the limitations that

affect the analysis and mining of Arabic comparative sentences. Section 5 discusses future research challenges. Finally, section 6 summarizes the conclusions.

2. Background

Mining and analysis of Arabic comparative sentences is needed for many purposes such as product and service reviews. This information is very useful for the benefit of many companies so that they can evaluate their products in comparison with other competitors in the market [10]. In general, a comparative sentence is a sentence that describes a similarity or a difference relation that involves one or more entity [3, 4]. For example, the Arabic sentence "أحمد أفضل من عمر في الرياضيات" is considered a comparative sentence while the Arabic sentence "أحمد طالب ممتاز" is not.

Mining and analysis of Arabic comparative sentences include a set of steps as shown in Fig. 1. The figure starts with the identification of a comparative sentence to decide whether it is a comparative sentence or not [3, 5, 6, 9, 10]. This is followed by the identification of the type of comparative sentence for the truly identified comparative sentence. The third step is the extraction of the relation between the different entities in the truly identified comparative sentence. The entities that form any relation are a comparison word, a comparison feature, a first entity, and a second entity. Finally, the preferred entity is extracted. The following subsections will discuss how each of the above-mentioned steps is referred to in the literature.

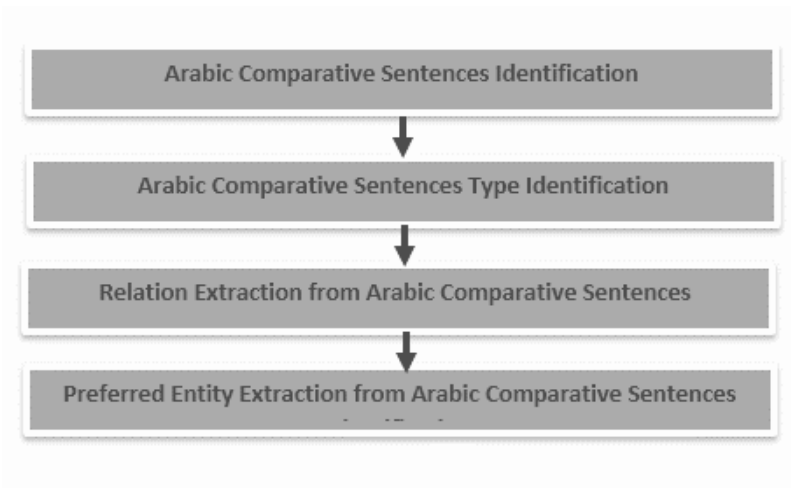

Fig. 1. Analysis and Mining of Arabic Comparative Sentences

2.1 Identification of Arabic Comparative Sentence

Comparative sentence identification means the differentiation between comparative and non-comparative sentences. In [5], the authors used Part of Speech (POS), Support Vector Machine (SVM) [17], Naïve Bayes [18] and K-nearest neighbor [19] approaches for Arabic comparative sentences identification.

The work proposed in [3] differentiates comparative from non-comparative sentences in a set of opinions that were collected from YouTube comments. In this work, comparative sentences were identified using keywords and POS Approaches. The authors employed a set of supervised learning techniques, which included Naïve Bayes classifier [18], JRip rule-based classifier [23, 24] and C4.5 decision tree [20]. Such techniques were able to overcome the limitations of linguistic classification. The work in [6] proposed a deep learning approach based on Probabilistic Neural

Network for the identification of Arabic comparative sentences. The proposed model was applied on two standard datasets namely Corpus and Corpus+.

2.2 Type Identification of Arabic Comparative Sentences

In general, comparative sentence types are classified into four different types [7], namely: non-equal gradable, equative, superlative and non-gradable comparative types [5]. The first three comparative types are categorized as gradable comparatives while the last type is categorized as non-gradable one. Each of the following subsections describes more details on one of the four types.

2.2.1 Non-Equal gradable Type

A non-equal gradable type expresses a relation that involves a comparison between two entities where a non-equality or ordering exists between them. For example, the sentence, "Phone A's battery life is longer than the battery life of Phone B", orders "Phone A" and "Phone B" based on "battery life" which is the shared feature [7].

In Arabic language, the non-gradable comparative keyword will have the form "أفعل" if its verb consists of three letters such as "اللغة العربية أصعب من الانجليزى" [5]. However, if the verb contains more than three letters the comparative statement should contain the word "أقل" or "أكثر" such as "سعيد أكثر اجتهادًا من اخيه". Another example in Arabic language is the sentence " تليفون أ أحسن من تليفون ب فى البطارية". The comparison keywords "أحسن من" refers to ordering of the sentence entities that are "تليفون أ" and "تليفون ب" with respect to their shared feature, i.e., "البطارية".

2.2.2 Equative Type

An equative type expresses an equal relation between two entities with respect to one or more shared features. For example, "Phone1 and Phone 2 are of the same brand" [7]. In Arabic language, the sentence "الجامعتان نفس المستوى فى التعليم" [5], the comparison keyword "نفس" refers to the equation between the sentence entities, that is "الجامعتان" with respect to their shared feature " المستوى فى التعليم".

2.2.3 Superlative Type

A superlative type expresses a superior relation of one entity. This relation ranks one object over all others, for example, "Phone A's battery life is the best" [7]. In Arabic language, it adds "ال" to the comparison keyword such as "الأفضل". For example, "الاهلي المصري الافضل في التاريخ" [5]. The comparison keyword may also be like "أفعل", for example "أحمد أطول طالب في الفصل". In this sentence, the comparison keyword "أطول" ranks one object which is the entity "أحمد" over all others with respect to their shared feature which is the feature "الطول".

2.2.4 Non-Gradable Type

The non-gradable type included sentences which compare features of two or more entities, but do not explicitly grade them. For example, "Computer A and Computer B have different features" [7]. Another example, in Arabic language, "تدريس أستاذ محمد يختلف عن تدريس أستاذ عمر", the comparison keywords "يختلف عن" refers to comparing the two sentence entities "أستاذ محمد" and "أستاذ عمر" with respect to their shared feature "تدريس" but do not explicitly grade them. There are three main sub-types [5] of the non-gradable type which are as follows:

a) The first entity is similar to or different from the second entity with respect to some features. An example is "تدريس أستاذ محمد يختلف عن تدريس أستاذ عمر" where the comparison keyword is "يختلف عن"
b) The first entity has a certain feature, and the second entity has another feature, where the former feature and the latter feature are usually substitutable. An example is " الحاسوب الثابت يستخدم سماعات خارجية أما الحاسوب المتنقل يستخدم سماعات داخلية" where the comparison keyword is "أما".

c) The first entity has a certain feature, where the second entity does not. An example is " الهاتف س يستخدم سماعات أذن و الهاتف ص لا يستخدم" where the compassion keyword is "و".

2.3 Relation Extraction from Arabic Comparative Sentences

A comparative sentence expresses a relation which orders two sets of entities with respect to a common feature. Here, a comparative sentence expresses the comparative relation with a certain relation vector [8]. The relation vector contains the following components: a relation keyword, a feature, the first entity, the second entity, and the comparative relation type. For example, in the comparative sentence "Camera A's optics is better than the optics of Camera B.", the corresponding relation vector is [better, optics, Camera A, Camera B, non-equal gradable].

Extracting a relation vector from an Arabic comparative sentence depends on the Arabic comparative sentence type. The following are examples of different sentences that illustrate the relation extraction from each of the four Arabic comparative sentence types mentioned in the previous section. The work proposed in [9] used Conditional Random Field (CRF) algorithm for relation extraction from Arabic comparative sentences.

An example of the Arabic non-equal gradable comparative sentence type is " بطارية موبايل سامسونج أحسن من بطارية النوكيا" and its extracted relation vector is ("النوكيا, موبايل سامسونج, بطارية, أحسن", non-equal gradable).

An example of the Arabic equative comparative sentence type is "الجامعتان نفس المستوى فى التعليم" and its extracted relation vector is ("الجامعتان, المستوى, نفس", equative). In this sentence, there is no obvious second entity because "(الجامعتان)" represents both the first and the second entities in the relation vector.

Two examples of the Arabic superlative comparative sentence type are illustrated as follows. The first sentence is "الأهلى الأفضل فى التاريخ" and its extracted relation vector is ("الأهلى, فى التاريخ, الأفضل)", superlative). The second sentence is "رب لا تذرني فردا وأنت خير الوارثين" and its extracted relation vector is ("(أنت, الوارثين, خير)", superlative).

Three examples of the Arabic non-gradable comparative sentence type are illustrated as follows. The first sentence is "تدريس الدكتور رشدى يختلف عن تدريس الدكتور عيسى" and its extracted relation vector is ("(الدكتور عيسى, الدكتور رشدى, تدريس, يختلف عن)", non-gradable). The second sentence is "الكمبيوتر المكتبى يستخدم سماعات خارجية أما اللاب توب يستخدم سماعات داخلية" and its extracted relation vector is ("اللاب توب, الكمبيوتر المكتبى, سماعات, أما", non-gradable). Finally, the third sentence is " جوال أ يستخدم سماعات أذن وجوال ب لا يستخدم" and its extracted relation vector is ("جوال ب, جوال أ, سماعات, و", non-gradable).

2.4 Extraction of Preferred Entity from Arabic Comparative Sentences

As mentioned before, a comparative sentence is used to compare between two or more entities. For example, the sentence, "Phone A is better than Phone B", compares between the two entities "Phone A" and "Phone B". In this sentence, "Phone A" is considered the preferred entity [7]. There are only two comparative sentence types which have a preferred entity. These are non-equal gradable and superlative comparative types.

2.4.1 Non-Equal Type

This section describes preferred entity extraction from Arabic non-equal gradable comparative sentence type. The following are some examples for the extraction of the preferred entity based on the sentiment of the comparison keyword. In the Arabic non-equal gradable comparative sentence type "أحمد أفضل من على فى الدراسة", the comparison keyword "أفضل" has a positive sentiment, so the first entity, i.e., "أحمد", is the preferred entity. In the example "موبايل نوكيا أسوء من سامسونج", the comparison keyword, i.e. "أسوء", has a negative sentiment. Therefore, the second entity, i.e., "(سامسونج)", is the preferred entity.

2.4.2 Superlative Type

This section describes the preferred entity extraction from Arabic superlative comparative sentence type. The following are some examples for the extraction of the preferred entity again based on the sentiment of the comparison keyword. Comparative superlative type sentence has only one entity if the comparison keyword has a positive sentiment. An example of such sentence is "الأهلى الأفضل فى مصر" in which the first entity i.e. "الأهلى", is the preferred entity since the comparison keyword, i.e. "الأفضل", has a positive sentiment. In another sentence like " حسام اسوء طالب", the comparison keyword, i.e., "اسوء" has a negative sentiment, then there is no preferred entity at all.

3. Research Techniques Proposed in the Literature

This section lists and discusses a classification of the different techniques used in the literature to analyze and mine an Arabic comparative sentence. Fig. 2 shows this classification which will be discussed in the following subsections.

3.1 Linguistic-Based Approaches

In the linguistic-based approaches, the Arabic comparative sentence is identified based on linguistic properties [5]. There are two linguistic methods which are keywords or lexicon-based and part of speech (POS). Both approaches will be discussed in the following subsections.

3.1.1 Keywords/Lexicon-Based Approach

Arabic comparative sentence identification using keywords is performed by searching for a comparison keyword in the sentence. If such keyword is present in the sentence, then it is identified as a comparative sentence, otherwise, it is identified as a non-comparative sentence [5]. For example, the sentence, "موبايل سامسونج احسن من نوكيا فى البطارية", is identified as a comparative sentence since there is a comparison keyword which is "أحسن".

The work proposed in [3] identified a set of Arabic comparison keywords which were used to identify an Arabic comparative sentence. The identification results reached 94% precision and 91% accuracy, however, the authors stated that they did not consider the identification of comparative sentences of superlative and non-gradable types.

The work proposed in [10] employed lexicons of comparative keywords and features for only preferred entities identification step. Several sentimental Arabic lexicons were used to form one large lexicon. The lexicons used in this work were ArSenl [12], NileULex [13], and the Bing Liu's lexicon Arabic translation [14,15] which is a translation of the English lexicon presented in [16] and the Arabic hashtag lexicon (dialectal) [14,15]. These lexicons contained Arabic and Egyptian terms with their positive or negative sentiments.

The authors of work [10] stated that this is the first work which addresses the problem of sentiment analysis of Arabic comparative opinions. The proposed technique showed an f-measure of correctly identified directions of comparative relations that ranged between 94% and 99%. It limited the human interaction to the initial steps of collecting lexicons and categorizing comparative keywords, which showed a good potential for fully automating the whole identification process. The limitations stated by the authors of this work included the inability to address the implicit features that may exist in the sentiment analysis of Arabic comparative opinions. Moreover, this work did not consider the analysis of the sentiment of comparative opinions with more than one relation and/or with more than two entities.

3.1.2 Part Of Speech

Parts of speech represent words that perform different functions in a sentence to provide the sentence with a proper structure and meaning. Parts of speech words include nouns, pronouns, verbs, adverbs, adjectives, prepositions, conjunctions, and interjections [11]. Using POS approach,

the comparative sentence is identified based on linguistic properties. For example, if the Arabic sentence contains the Adjective Comparative POS tag, such as "أفضل", the Adjective Superlative POS tag, such as "الأفضل", the Adverb Comparative POS tag, such as "في كثير من الاحيان" or the Adverb Superlative POS tag such as, "الاصخب", then it the sentence is expressed as a direct comparison [5]. The downside of this approach is that some comparative sentences do not contain any comparison keywords such as "الكمبيوتر المكتبى يتحمل شغل عن اللاب توب". On the other hand, some sentences have a comparison keyword, but they are not real comparative sentences such as " سيارة أ أعلى من سيارة ب بمتر".

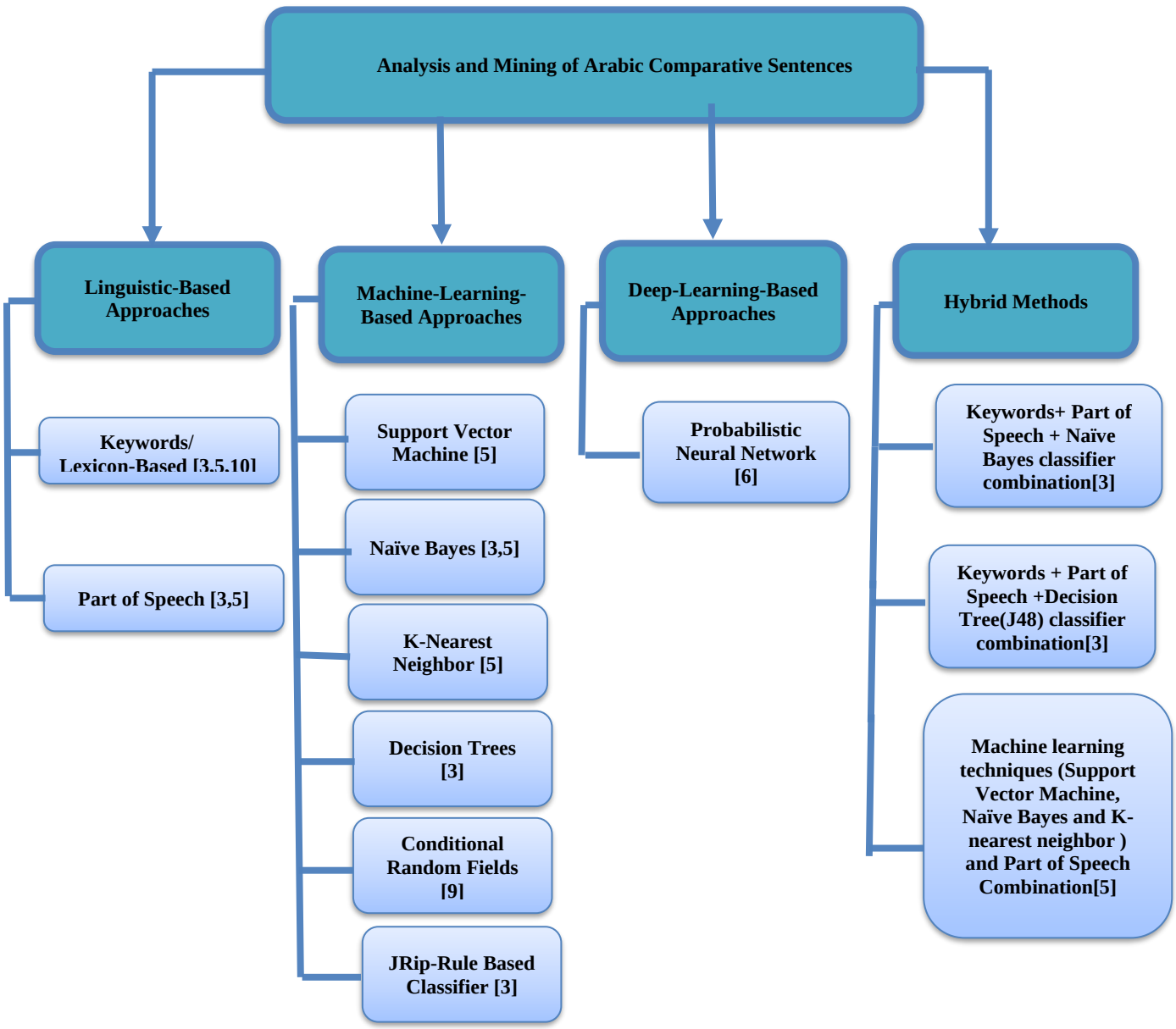

Fig 2: Classification of Research Techniques for Analysis and Mining of Arabic Comparative Sentences

The work proposed in [3] differentiated between comparative and non-comparative sentences in a dataset of sentences that were collected from YouTube comments. The authors used a simple classifier to classify the comparative sentences based on adjective comparison words which in Arabic is called" اسم التفضيل" or 'Preference Name' in the POS tag attribute. The authors used the POS method to comparative sentences identification step only and the accuracy reached 87% using this technique. In the work proposed in [3], the authors mentioned that the methodology used was capable of only identifying direct comparisons. Moreover, there was false identification of some of the comparative sentences because their method did not consider adjective comparative words.

3.2 Machine Learning Methods

Machine learning methods were used to mitigate the limitations of linguistic-based methods in comparative sentence identification. There are several machine learning methods that were used in the analysis and mining of Arabic comparative sentences. Examples of such methods include Support Vector Machine (SVM) [17], Naïve Bayes [18], K-Nearest Neighbor [19], Decision Tree [20], Conditional Random Fields [21, 22], and JRip Rule−Based classifier [23, 24] algorithms.

3.2.1 Support Vector Machine

SVM is a machine learning method which is used to separate a set of positive samples from a set of negative samples with a maximum separation margin [17]. The work proposed in [5] used SVM for Arabic comparative sentence identification. In this work, reviews were classified according to their positions with respect to a separation margin. The results showed an f-measure which reached more than 80%. The work proposed in [5] did not explain the details of using the SVM technique in their proposed approach and did not provide working examples that describe the operation of SVM on Arabic comparative sentences.

3.2.2 Naïve Bayes

Naïve Bayes classification method was used in classification because of their simplicity and computational efficiency [18]. This method was used to categorize a document based on the existence of some words in the document with certain frequency [18]. The work proposed in [3, 5] used a Naïve Bayes classifier for Arabic comparative sentence identification and the accuracy was between 80% and 83%.

The authors of the work proposed in [3] mentioned that some words were deleted from the text during the text pre-processing step. These words were found to be important in identifying comparative sentences. This pre-processing step affected the accuracy of the obtained results. The work proposed in [5] did not explain the details of using Naïve Bayes technique in their proposed approach. Moreover, they did not provide working examples that describe the operation of the Naïve Bayes classifier on Arabic comparative sentences.

3.2.3 K-Nearest Neighbor

K-Nearest Neighbor is another method used for document classification [19]. In the training step, documents were formulated in the form of a vector representation. To classify a new document, its vector is compared with each corresponding vector in the training set. Afterwards, the sentence k-nearest neighbor was determined based on the *Euclidean* distance similarity. The work proposed in [5] used K-Nearest Neighbor classifier for Arabic comparative sentence identification and obtained the best f-measure 86.63% among all other classifiers used in the authors' study. The work proposed in [5] did not provide the details of using the K-Nearest Neighbor technique in their proposed approach. They also did not provide any Arabic comparative sentences examples that illustrate the usage of the K-Nearest Neighbor technique.

3.2.4 Decision Trees

A decision tree is a graph used to take decisions [20], where a specific feature of a feature vector is investigated in each branch of this graph. If the value of the investigated feature is found to be lower than a certain threshold, then the left branch is explored; otherwise, the right branch is explored. The decision is made on the final classification when the leaf node is encountered. The work proposed in [3] employed the J48 classifier to identify comparative sentences from Corpus and Corpus+ datasets. The J48 classifier achieved 90% accuracy for both datasets. The work proposed in [3] only applied one step of Arabic comparative sentence analysis and mining which is the Arabic comparative sentence identification. The authors of the work proposed in [3] did not explain the details of using the decision tree technique in their proposed approach and did not illustrate the application of this technique on sample Arabic comparative sentences.

3.2.5 Conditional Random Fields

Conditional Random Fields (CRFs) is a probabilistic model for predicting sequences based on contextual information [21, 22]. CRFs is a graphical model that defines a log-linear distribution over label sequences provided an observation sequence. It is a discriminative model which models the conditional probability of labels provided the observations. CRFs model has applications in natural language processing, computer vision, and bioinformatics [22]. The work proposed in [9] used CRFs model for relation extraction from Arabic comparative sentences. The evaluation results of this work achieved an accuracy of 86.3%. The work proposed in [9] applied CRFs model on only two types of comparative sentences which are non-equal gradable and superlative types. This work did not consider the other two comparative types which are equative and non-gradable types.

3.2.6 JRip rule‑based classifier

The work proposed in [3] employed JRip rule‑based classifier [23, 24] for Arabic comparative sentence identification. The evaluation results obtained by using this classifier with the RIPPER algorithm were better than the results obtained using the Naive Bayes classifier. A set of rules were applied for both Corpus and Corpus+ datasets. The applied rules were used to decide, if the Arabic sentence includes comparison keywords such as "أفضل, احسن, اقوى" and conjunctions such as " من ,اما", then the sentence was classified as a comparative sentence. The obtained accuracy was 88.45% and 89.76% for the Corpus dataset and the Corpus+ dataset, respectively.

The work proposed in [3] did not apply the JRip classifier on equative and superlative comparative types. Moreover, the authors mentioned that the JRip classification is only feasible when the number of training examples is relatively small. They also showed that in their experiment on a machine with normal computational power, the execution time to produce the results and rules was approximately three hours. Such computational time will not be appropriate for real time applications that require a deadline of a few seconds or minutes.

3.3 Deep Learning Approaches

The Probabilistic Neural Network (PNN) is a recent deep learning approach employed in the field of analyzing Arabic comparative sentences. PNN is implemented by employing an exponential function instead of the sigmoid activation function which is usually used in neural networks [25, 26]. The PNN provides a hybrid method of Bayes theorem of conditional probability and Parden's method that estimates random variables probability density functions. The recent work in [6] proposed a PNN-based model for the identification of Arabic comparative sentences. This work applied the proposed model to two standard datasets which are Corpus and Corpus+. The applied model obtained an average accuracy of 98.50%. The authors of the work proposed in [6] did not provide a detailed description of the proposed approach using working examples of Arabic comparative sentences.

3.4 Hybrid Methods

Hybrid methods used a combination of linguistic and machine learning methods to enhance the performance and efficiency of analysis and mining of Arabic comparative sentences. This combination enhanced the evaluation results than those obtained using only linguistic or machine learning methods. The work proposed in [3, 5] used linguistic approaches, machine learning approaches and a combination between them. The work in [5] obtained an f-measure of 88.87% using a combination of SVM and POS for the identification of Arabic comparative sentences. The work in [3] also used a combination between linguistic approaches and machine learning approaches for the identification of Arabic comparative sentences. After Applying keywords, POS, and decision tree classifier combination in [3], the authors concluded that correcting misspellings did not improve the performance when the dataset is colloquial because most words were colloquial and cannot be corrected which is not the case when the dataset is based on modern standard Arabic. Authors also recommend not filtering all stop words because some stop words can be important which will affect the accuracy of identifying comparative sentences.

The work proposed in [5] applied two steps of Arabic comparative sentences mining which were comparative sentence identification and comparative sentence type identification. For the second step, the set of generated rules were applied for only three types of Arabic comparative sentences. These types included non-equal gradable. equative and superlative comparative sentence types but the fourth type, i.e., non-gradable comparative type, was not included. The authors of the work proposed in [5] mentioned that, in the case of using KNN and POS combination, the performance decreased when compared to only using the KNN classifier.

4. Limitations of Current Techniques

Arabic language has several inherent challenges and limitations that affect the analysis and mining of Arabic comparative sentences. Some of these limitations are discussed in this section. The Arabic word may have more than one meaning if the diacritical marks, i.e., "علامات التشكيل" are applied. Arabic characters may add to words such as prepositions, i.e., حروف الجر, such as(ل-ب) in (أحسن-لأحسن). This complicates the detection of comparison keywords. The comparison keywords may be falsely written such as the comparison keyword "أرحم" may be written as "ارحم" and be a identified as command verb not a comparative keyword.

The Arabic word "لكن" converts the sentiment of the sentence or provides two different sentiments in the same sentence. For instance, in the sentence, " سامسونج احسن من اوبو لكن اوبو احسن في الكاميرا", there are two comparative sentences "موبايل سامسونج احسن من اوبو" and "اوبو احسن فى الكاميرا" are combined using the Arabic word "لكن".

The exclamation mark used in the Arabic sentence such as "ما أسرع النزول!" can be falsely identified as a comparative sentence because the word "أسرع". If the presence of exclamation symbol at the end of the sentence is taken into consideration, this limitation can be solved.

There will be a misidentification between the non-equal gradable and superlative sentence comparative types since the comparison keywords is the same such as "أسرع" and ""أحسن. The difference between the two types is that non-equal gradable type uses the word "من" after the comparison keyword such as أحسن من. However, the word "من" is not used in the superlative comparative sentence type. If this issue is not considered, this sentence can be falsely identified as superlative comparative type.

In the non-gradable comparative sentence type, the keyword "أما" can be wrote as "اما" which have the same meaning. Also, it may have the meaning of a mother, i.e., "أُماً", which can be falsely identified as a non-gradable comparative sentence type keyword. The comparative keywords of the equative sentence type, "نفس" and "نفسه" may not be a comparison keyword because it may have a meaning of a person or spirit. The different meaning of Arabic comparison keywords may lead to false identification of comparative and non-comparative sentences.

Relation extraction from Arabic comparative sentence has some limitations. For example, the first and second entities may be more than one word where a part of the first entity or the second entity may be a comparison keyword. This leads to false identification and extraction of comparative sentence and relation elements. For example, consider the sentence "أغنية أجمل سنين عمرنا". In this sentence, word "أجمل" is not a comparison keyword. Also, the first entity and the second entity may be a number or a number and text such as "أفضل الطلاب هم ١٠ فقط".

The preferred entity extraction based on comparison keyword and features together has some limitations in the superlative comparative sentence type. For example, consider the two comparative sentences "الفندق الأكثر جمالا" and "الفندق الأكثر قبحا". In the former comparative sentence, the comparison keyword "الأكثر" has a positive sentiment while the feature "جمالا" is a sentimentally positive word so the first entity ""الفندق should be the preferred entity. However, in the latter sentence, the feature "قبحا" is a sentimentally negative word, so the first entity " الفندق" is not the preferred entity. Also, the Arabic negation words such as "ليست" converse the sentiment of the sentence such as "ليست أحسن الكتب".

There are publicly available datasets that can be used to evaluate research techniques in the general field of Arabic sentiment analysis. However, there is still a limitation with the public availability of standard datasets that can specifically be used to develop and evaluate new research techniques that serve in the field of mining of Arabic comparative sentences. Most of the previous studies were used to develop their own dataset to evaluate their proposed approach. For example, the authors in [3] collected Arabic text comments from YouTube and mentioned that that the reason was that there was no publicly available corpus for Arabic comparative opinions. Another example, the authors in [5] manually collected documents of Arabic opinions in three different domains, i.e., education, technology, and sports, and classified them into comparative and non-comparative sentences. The work proposed in [6] employed Corpus and Corpus+ datasets from Arabic social media content. The authors in [9] collected 480 comparative opinions expressed in Egyptian and Khaliji dialects and in Modern standard Arabic. Finaly, the authors in [10] collected 830 comparative opinions expressed in Egyptian dialect and Modern standard Arabic. This dataset was collected from some other Arabic datasets, Facebook, Twitter, and public blogs.

5. Future Research Challenges

Future research work needs to address all steps of analysis and mining of Arabic comparative sentences. Specifically, this work needs to improve the accuracy of each of the outlined steps and needs to solve the above-mentioned limitations and challenges of the Arabic language. Future work can employ other linguistic features than POS which include word semantics. Additionally, future work can leverage deep learning models. These models can be trained with more features and comparison keywords that address the above-mentioned limitations. Also, the various cases of adding characters to the comparison keywords such as prepositions for true identification need to be handled. Moreover, comparative, and non-comparative datasets of larger sizes need to be used in the evaluation of the various steps of analyzing and mining Arabic comparative sentences. The Arabic comparative datasets and Arabic lexicons need to be continuously updated with more comparison keywords and features that can fit in different domains.

To the best of our knowledge, there is no current research work on the non-gradable comparative sentence type. The limitations of implicit features which are not directly stated in the text but inferred from the text, need to be addressed in future research work. Finally, the analysis and mining of Arabic comparative sentences which contain more than one relation, and more than two entities needs to be covered.

6. Conclusions

There are few research works on analysis and mining of Arabic comparative sentences. This paper provided a literature review that included a classification of the different techniques that

addressed the various steps employed in this field. These techniques employed linguistic approaches, machine learning approaches, a combination between linguistic approaches and machine learning approaches and deep learning models. The current research work which combined linguistic with machine learning approaches achieved higher accuracy results than all other employed techniques. This paper also provided a discussion of the limitations and challenges that inherently exist in the Arabic language which affect the analysis and mining of Arabic comparative sentences. Moreover, the paper summarized the insights that provide researchers with directions that can guide future research work in this field. To the best of our knowledge, this literature review is the work that addresses the current research work employed in this field.

References

[1] Sawi, Laila, and Iman Saad. "Al-Murshid". *American University in Cairo Press*, 2012.

[2] DoniaGamal, Marco Alfonse, El-Sayed M. El-Horbaty, and A. B. Salem. "Opinion mining for Arabic dialects on twitter." *Egyptian Computer Science Journal* 42.4 (2018).

[3] Alharbi, F.R. and Khan, M.B., 2019."Identifying comparative opinions in Arabic text in social media using machine learning techniques". *SN Applied Sciences*, 1(3), p.213.

[4] Liu, B. and Liu, B., 2011. "Opinion mining and sentiment analysis". *Web data mining: exploring hyperlinks, contents, and usage data*, pp.459-526.

[5] El-Halees A (2012) "Opinion mining from Arabic comparative sentences".*In: 13th International Arab conference on information technology ACIT ASSESSMENT*, pp 265–271.

[6] Alotaibi N., Al-onazi B.B., Nour M.K., Mohamed A., Motwakel A., Mohammed G.P., Yaseen I., Rizwanullah M. "Political optimizer with probabilistic neural network-based Arabic comparative opinion mining". *Intelligent Automation & Soft Computing, 2023, vol. 36, no. 3, pp. 3121–3137.*

[7] Ganapathibhotla, Murthy, and Bing Liu. "Mining opinions in comparative sentences." In *Proceedings of the 22nd international conference on computational linguistics (Coling 2008)*, pp. 241-248. 2008.

[8] Jindal, Nitin, and Bing Liu. "Mining comparative sentences and relations." In *Aaai*, vol. 22, no. 13311336, p. 9. 2006.

[9] El Defrawi, Mai, Marwa Salah, Ahmed Abd Al-Aziz, and Ahmed Sharaf Eldin. "Comparative relation extraction from Arabic opinions." *Int J Comput Sci Inf Secur* 15, no. 10 (2017).

[10] Eldefrawi, M.M., Elzanfaly, D.S., Farhan, M.S. and Eldin, A.S., 2019. "Sentiment analysis of Arabic comparative opinions". *SN Applied Sciences*, *1*, pp.1-11.

[11] Martinez, Angel R. "Part of speech tagging." *Wiley Interdisciplinary Reviews: Computational Statistics* 4.1 (2012): 107-113.

[12] Badaro, G., Baly, R., Hajj, H., Habash, N. and El-Hajj, W., 2014, October. "A large scale Arabic sentiment lexicon for Arabic opinion mining". In *Proceedings of the EMNLP 2014 workshop on arabic natural language processing (ANLP)* (pp. 165-173).

[13] El-Beltagy, Samhaa R. "Nileulex: "A phrase and word level sentiment lexicon for egyptian and modern standard arabic." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 2900-2905. 2016.

[14] Mohammad, S.M., Salameh, M. and Kiritchenko, S., 2016. "How translation alters sentiment". *Journal of Artificial Intelligence Research*, *55*, pp.95-130.

[15] Salameh, M., Mohammad, S. and Kiritchenko, S., 2015. "Sentiment after translation: A case-study on arabic social media posts". In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 767-777).

[16] Hu, M. and Liu, B., 2004, August. "Mining and summarizing customer reviews". In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177).

[17] Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995, pp. 273-97.

[18] Du, R., Safavi-Naini, R. and Susilo, W., 2003, October. "Web filtering using text classification". In *The 11th IEEE International Conference on Networks, 2003. ICON2003.* (pp. 325-330).

[19] Dasarathy, Belur V. "Nearest neighbor (NN) norms: NN pattern classification techniques." *IEEE Computer Society Tutorial* (1991).

[20] Burkov, A. (2019). "The hundred-page machine learning book" (Vol. 1, p. 32). *Quebec City, QC, Canada: Andriy Burkov.*

[21] Wallach, H.M., 2004. "Conditional random fields: An introduction". *University of Pennsylvania CIS Technical Report MS-CIS-04-21*, *24*,pp.33-42.

[22] Sutton, C. and McCallum, A., 2012. "An introduction to conditional random fields". *Foundations and Trends® in Machine Learning*, *4*(4), pp.267-373.

[23] Cohen, W.W., 1995. "Fast effective rule induction". In *Machine learning proceedings 1995* (pp. 115-123). Morgan Kaufmann.

[24] Shahzad, W., Asad, S. and Khan, M.A., 2013. "Feature subset selection using association rule mining and JRip classifier". *International Journal of Physical Sciences*, *8*(18), pp.885-896.

[25] Hajmeer, M. and Basheer, I., 2002. "A probabilistic neural network approach for modeling and classification of bacterial growth/no-growth data". *Journal of microbiological methods*, *51*(2), pp.217-226.

[26] Specht, D.F., 1990. "Probabilistic neural networks". *Neural networks*, *3*(1), pp.109-118.

المجلد (١١) العدد (٢) (السنة ٢٠٢٤)

المجلة الدولية للحاسبات والمعلومات

متاح على الإنترنت على الرابط : https://ijci.journals.ekb.eg/

تحليل وتعدين جمل المقارنة باللغة العربية: استعراض أدبي

آلاء صقر – عربي كشك – أنس يوسف

قسم علوم الحاسب– كلية الحاسبات والمعلومات – جامعة المنوفية

الايميل الرسمي: anas.youssef@ci.menofia.edu.eg

ملخص البحث

هناك كم هائل من جمل المقارنة في اللغة العربية التي يتم كتابتها يوميا على وسائل التواصل الاجتماعي المختلفة. تحتاج هذه الجمل إلى التحليل والتنقيب لأغراض مختلفة مثل تقييمات الخدمات والمنتجات. بشكل عام، يواجه تحليل النص العربي والتنقيب عنه تحديًا كبيرًا بسبب القيود الموروثة في اللغة العربية. علاوة على ذلك، لا توجد حاليًا قواعد بيانات قياسية لجمل المقارنة باللغة العربية. تقوم هذه الورقة البحثية بتوفير خلفية عن الخطوات المختلفة المطلوبة لتحليل واستخراج جملة مقارنة باللغة العربية. وتشمل هذه الخطوات تحديد ان كانت الجملة العربية هي جملة مقارنة ام لا، وكذلك تحديد النوع الدقيق لجملة المقارنة، وأخيراً استخراج مكونات العلاقة من جملة المقارنة وتشمل العنصر المفضل في الجملة. توفر الورقة البحثية أيضًا استعراض أدبي للأعمال البحثية الحالية التي تم طرحها في هذا المجال. يتضمن ذلك تصنيفًا للتقنيات المختلفة المستخدمة في هذا المجال بما في ذلك ثلاث فئات رئيسية وهي: أساليب لغوية وأساليب التعلم الآلي وأساليب التعلم العميق. وأخيرا، تقدم الورقة البحثية رؤى حول القيود الحالية وتحديات البحث المستقبلية في هذا المجال. على حد علمنا، هذه هي الورقة البحثية الأولى التي تقدم استعراض أدبي للمراجع التي تقوم بتحليل واستخراج الجمل المقارنة باللغة العربية. يناقش هذا الاستعراض التحليل المحدد لجمل المقارنة باللغة العربية وليس تحليل الميل للجملة العربية بشكل عام. ويلاحظ أن هذا التحليل هو فرع من فروع تحليل الميل في الجملة العربية الذي لا يركز على تحديد الميل للجملة العربية، ولكنه يركز على تحديد وتحليل جمل المقارنة باللغة العربية ومكونات تلك الجمل.