Research Article

Identifying comparative opinions in Arabic text in social media using machine learning techniques

Fatmah Rasheed Alharbi[1] · Muhammad Badruddin Khan[1]

Abstract
Social networking sites have become an integral part of everyday life, where people interact, cooperate and quarrel with each other. Social media also encourages them to express their opinions and share their comments about their lives' events or about the product they use. Opinions can be direct without any comparison (I like ABC phone) or they can be comparative (X-phone's camera is better than Y-phone). Comparative opinions are useful in many applications, e.g. marketing intelligence, product benchmarking, and e-commerce. The automatic mining of comparative opinions is an important text mining problem and an area of increasing interest for different languages. This paper focuses on identification of comparative sentence from non-comparative ones in Arabic texts. A corpus was developed consisting of YouTube comments. This paper describes research experiments that aimed to apply data/text mining algorithms, natural language processing and linguistic classification for Arabic comparative text discovery. The results of these experiments along with the analysis are also presented. The results were promising reaching to 91% accuracy for the detection of comparative opinions.

Keywords Text mining · Mining methods and algorithms · Classification · Opinion mining · Machine learning algorithm · Text categorization

1 Introduction

Opinions are important in our life; if we need to make a decision, we often ask our acquaintances to give us their advice or opinions. With the proliferation of technology in every area of our lives, social media—particularly blogs and social networks—has fueled this human nature to know people's opinions about everything. Social media encourages us to express our opinions and share comments about daily life events with others. Individuals as well as organizations and governments want to know about these comments depending on their requirements [1]. Customer opinions were essential to companies even before the advent of the Internet, and many companies use survey forms to evaluate people's opinions about their products [2].

Individuals and organizations are increasingly using social media content for decision making. Potential customers of any product or service access other customers' comments through the internet to gain knowledge of the experiences of other users before buying [3]. Additionally, few organizations continue to conduct surveys about their products and services even though are obtaining reviews about their products and services from their consumers directly via social networking sites [1, 4]. Many organizations are thus developing and improving their business analytics capabilities because potential consumers are using user-generated reviews to know about prior evaluations of products and services

✉ Fatmah Rasheed Alharbi, fatmah.alharbi.87@gmail.com; Muhammad Badruddin Khan, mbkhan@imamu.edu.sa | [1]Information Systems Department, College of Computer and Information Sciences, Al-Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia.

SN Applied Sciences (2019) 1:213 | https://doi.org/10.1007/s42452-019-0183-3

Received: 30 October 2018 / Accepted: 16 January 2019 / Published online: 4 February 2019

by previous consumers and are using the earned knowledge to make decision for their purchases [5].

Opinion Mining (OM), also known as Sentiment Analysis (SA) focuses on development of method that can automatically discover opinionated information and can suggest extent of polarity toward a particular target [1, 6].

Opinions express the opinion holder's view towards an entity and/or its aspects. Sometimes people express their opinion of a product by comparing it with another one. For example, 'X-video game console is better than Y-video game console' is an opinion that compares the two entities, whereas 'X gaming controller is easier than Y' is an example of an opinion that compares two entities based on a single aspect (controller). When opinion holders express their opinion towards an entity in comparison to another, they express a comparative opinion [7].

Social networks are filled with opinions comparing between entities and specifying a certain aspect of comparisons [7]. Additional examples of these types of opinions circulating the social media are 'ABC's tablet is better and cheaper than XYZ', 'I prefer A to B' and 'X cars are more powerful than Y's cars'.

Comparative opinion mining from social networks is useful in many fields, such as business, education, politics and sports [8]. For example, in the business environment, the producer wants to know consumers' opinions about the products and how the product compares with those of its competitors. Availability of such information can be helpful to businesses in boosting their market performance by directing their efforts towards marketing and product benchmarking in successful direction [7, 9]. In education sector, performance of different teachers of the same course can be compared using student's opinion [8]. Governments might want to know the attitude of people towards certain decisions and services to ensure the satisfaction of their people [2], and detecting comparative reviews is important and even sometimes critical for industries [7].

Following the political unrest in the Middle East in 2011, known as the Arabic Spring, there has been an increasing interest in mining opinions written in the Arabic language [10]. Studies of comparative opinion mining in English are extensive and not new, but the area of comparative opinion mining in Arabic language is not yet well established, and is currently in its infancy.

Manual analysis of opinion and reviews, gives better and more accurate understanding of situation, but the problems associated with this method includes more time-consumption, expensiveness and subjectivity due to human involvement. Moreover, the presence of the huge volume of data available on social networks makes the job of manual analysis of data impractical [2].

We used YouTube comments as inputs for our corpus because many comparative opinions in Arabic exist in the form of YouTube comments for different uploaded videos describing features of various products. Additionally, YouTube data are accessible and available for public access through streaming Application Programming Interface (API).

YouTube ranked 2nd globally on the Alexa website of the most visited sites on the internet. YouTube began its journey in 2005 when it was founded in form of simple video sharing website. With passage of time, it transformed itself to largest video-sharing website in the cyber world. After the acquisition by Google, the speed of YouTube's popularity accelerated further and about half a billion unique users visit YouTube in a month [11].

YouTube allows registered users to upload and share video clips on a diverse array of topics and incorporates a growing number of additional features that allow users to interact with the content and other users. Users can review or rate what they have watched and associate comments with videos to express their opinions or respond to the video content [12].

The following are some properties specific to the comments found in social media, particularly YouTube comments:

- Most comments are short.
- The linguistic style is usually informal, with numerous accidental and deliberate errors and grammatical inconsistencies.
- Comments include many abbreviations, idioms and jargon.
- Users do not care about the correct usage of grammar.
- Many comments are unrelated or contain spam.
- Comments are sometimes are unrelated to the video content and are instead used for self-expression, to provide emotional support, to reminisce, to express grief and to give advice [12].

Thus, this paper attempts to understand the techniques that can be used for Arabic comparative opinion mining and to build accurate models that can classify Arabic comparative sentences. The significance of this work is that if sentiment analysis is to be performed to detect sentiments about entities and aspects, first of all it must be identified whether the opinion is comparative in nature or not.

The work is part of broader research which aims to find opinions about entity and aspect in comparative sentences. Figure 1 suggest where this work stands in broader research. Figure 1 depicts the important tasks or steps needed to be taken for comparative opinion mining. There are three major tasks to deal with comparisons. The first major task is to identify comparative sentences in corpora of opinions. This task is the issue of current work. The second task is about extraction of entities and aspects from

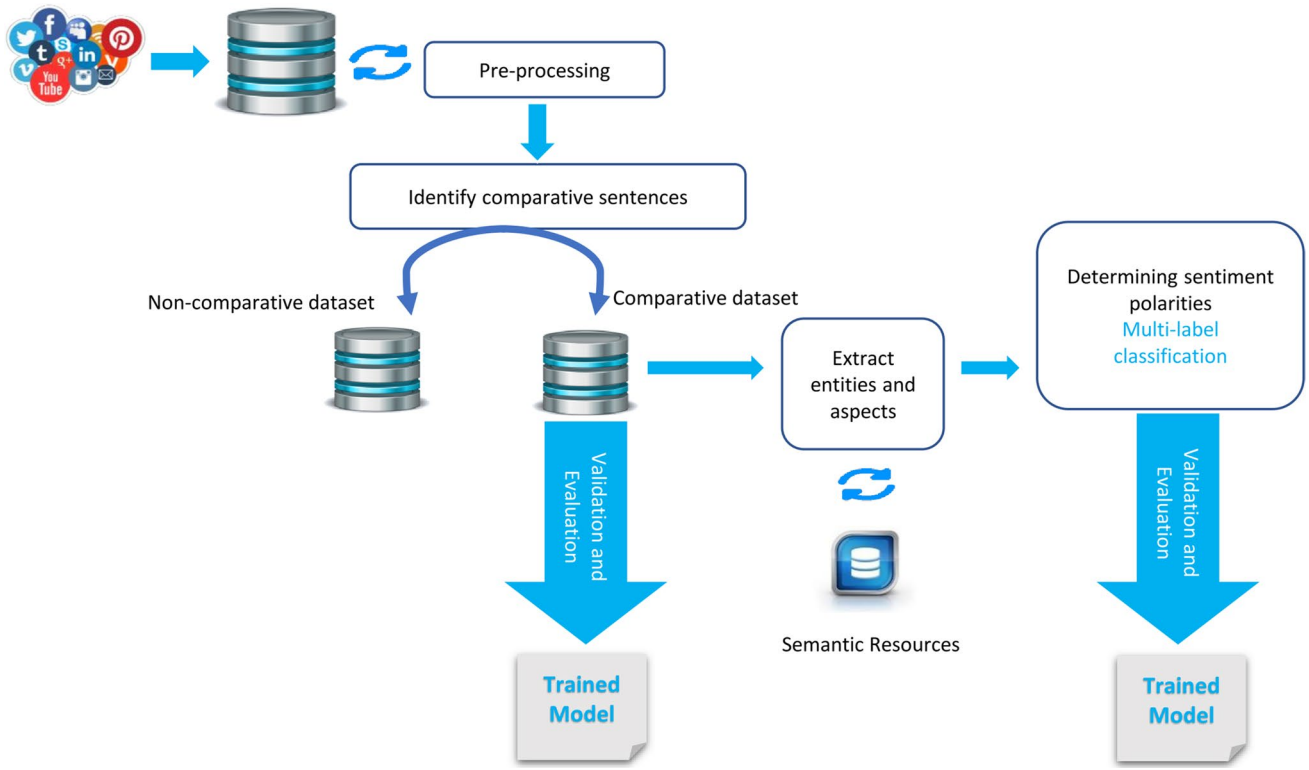

Fig. 1 The pipeline architecture suggest where this work stands in broader research

sentences; and the third task determines sentiment polarities (negative, positive or neutral) for entities or aspects of entities using multi-label classification (MLC).

The rest of the paper is organized as follows. The next section presents the background and discuss some topics related to our work. The existing literature is reviewed in Sect. 3. Section 4 presents the proposed framework and research methodology employed to discover comparative opinions in the developed corpus. In Sect. 5, the details and analysis of the classification results are given. Finally, we conclude this paper with the conclusion and discussion of possible future work.

2 Background

In this section, we discuss different issues that are related to our work.

2.1 Text classification

Text classification "is the assignment of free text documents to one or more predefined categories based on their content" [13]. Generally, building a classification system involves three main phases: data pre-processing, text classification and evaluation.

The data preprocessing phase is performed with the target of transforming the text documents to the state that is appropriate to train the classifier. Using the preprocessed documents, the text classifiers are constructed using training dataset. The performances of the constructed classifiers are evaluated using different evaluation measures, e.g. measures of recall, precision [14].

Saritha and Pateriya [15] studied various methods for comparative sentence identification and classification. They found that few studies had been conducted on comparative sentences, and they used different supervised and unsupervised techniques to identify the comparative sentences and relations.

2.1.1 Linguistic approach

The researchers attempted to categorize different types of comparative sentences based on syntax and semantics. The main concern of Syntax is the issue of structure of language. The logical or grammatical form of sentences come in this category. Semantics, on other hands, deals with the meaning of words and sentences. When researchers study the structure and language of comparative sentences, they explicitly assess the usage of morphemes such as more/-er, less/-er and as to create orders of superiority, inferiority and equality for subject comparison.

2.1.2 Machine learning approaches

Supervised learning algorithms need supervisor for training purpose. The supervisor is provided through data set that has label for every example or record. The labelled data is used by supervised algorithms to create classifiers that create a map from dataset to labels.

2.1.3 Sequential pattern mining (SPM) approach

In order to identify statistically relevant patterns between data samples with same delivery of sequences, SPM technique is used. SPM may be class sequential rule mining or label sequential rule mining. This approach is suitable for structured data.

2.2 Arabic language and challenges

Arabic, the target language of this paper, is the mother tongue of approximately 300 million people in approximately 22 countries [16]; it ranks as the fifth largest natural language among the top 100 used natural languages worldwide [17]. Arabic writing orientation is from right to left. Any Arabic word is combination of any of 28 letters that belongs to the Arabic alphabet set. The 28 letters are extendible to 90 letters due to additional writing shapes, marks and vowels [16].

The Arabic language has two main forms: Standard Arabic and Dialectal (colloquial) Arabic. Standard Arabic includes Classical Arabic (CA) and Modern Standard Arabic (MSA). CA is the historical language used in the Quran and Hadith. MSA is the formal form and is used in books, education, TV, newspapers and formal speeches. However, Arabic speakers use the dialectal form in their daily interactions and when they express their views about different aspects of life on social media [16, 18, 19].

There are many Arabic dialects, but six are dominant namely Egyptian, Gulf, Iraqi, Levantine, Moroccan and Yemini. The dialects are one of the reasons for the introduction of numerous new words into any language particularly stop words [19]. The Colloquial Arabic faces the problem of lack of standardization [18].

Although Arabic is a widely used language, studies of OM have only been conducted in recent years, and this field requires more research due to the unique nature of Arabic language morphological principles [20]. Arabic opinion mining faces many challenges due to the poorness of language resources and to Arabic-specific linguistic features [19].

While comparative OM is a well-studied problem for English text, not many systems considered extending comparative OM to other languages, such as Arabic. Thus, adapting comparative OM to Arabic text exhibits many challenges, some of which have been discussed in previous papers [8, 18–21]:

- Arabic synonyms are widespread. An Arabic-translated lexicon may have poor coverage due to issue of the morphological and orthographic complexities of Arabic language. The source of derivation of most of Arabic nouns and verbs is a set of 10,000 roots that are cast into stems using templates that may add infixes and double letters, or remove letters. These stems can allow the prefixes or suffixes to be attached. Hence the number of possible surface form of Arabic is very huge and is in order of billions.
- Arabic is morphologically complex compared to the English language.
- Dialect Arabic lacks grammar and rules regarding how to use it, and there are no dictionaries for it.
- Another challenge is common occurrence of broken plurals. They resemble irregular English plural except that the extent of resemblance with their singular form is not close to resemblance level of English irregular plural with their singular form. Due to their lack of conformance with normal morphological rules, existing stemmer are unable to handle them.
- Arabic letters can be written with different shapes according to their position in the word. For example, 'Alif', in which the first letter has four forms (ا, أ, إ, آ) and 'Taa el-marpouta' and 'Haa el-marpouta '(ة, ه).

2.3 Comparative sentences

In this research, we focus on comparative sentences, which are widely used in social media. "A comparative sentence expresses a relation based on the similarities or differences of more than one entity" [1]. Individuals often ask questions such as 'Which one is better: Product A or Product B', and their friends reply to such questions by conducting comparisons, which enable organizations to know their customer's opinions about their products so they can make improvements [9].

Liu [1] divides comparative relations into four main types; the first three types are gradable comparisons and the fourth type is a non-gradable comparison.

1. Non-equal gradable comparisons: This type of comparison expresses an ordering that exist in opinions for some entities with respect to certain aspect. Usually these type of comparisons are presented in type of "Greater or lesser than" form, e.g. 'A laser printer is faster than an inkjet printer'. This type also includes user preferences, e.g. 'I prefer X to Y'. In Arabic, if a comparison source consists of three letters, the word will

be in the following form 'أفعل -> Afael' [8]. For example, 'طول' means 'long' and becomes 'المبنى الثاني هذا المبنى أطول من' ('This building is longer than the other'). If the comparison source contains more than three letters, the sentence will contain one of these words: 'أقل أو أكثر أو أفضل أو أسوأ' (more, less, better or worse) [8]. For example, 'تعلّم' means 'educate' and is used in comparisons such as 'أحمد أفضل تعلما من خالد' ('Ahmad is better educated than Khaled').

2. Equative Comparisons: This type of comparison expresses the equality between two or more entities with regard to shared aspects, e.g. 'The picture quality of X is about the same as that of Y'. In Arabic, many words have meanings similar to equality, such as 'متساويين, متقاربين, نفس بعض, بنفس المستوى', which mean 'equal, the same, similar and same level'. Therefore, 'س وص تقريبا نفس الجودة' means 'X and Y have the same quality'.
3. Superlative Comparisons: Relations of the type 'greater or less than all others' rank one entity above all others, e.g. 'Al-Hilal is the Best'. In Arabic, the phrases 'the best and the worst' are 'الأفضل أو الأسوأ أو الأجمل أو الأقبح', and they have 'Al' prefixes, such as 'الهلال هو الأقوى', which means 'Al-Hilal is the strongest'.
4. Non-gradable Comparisons: These comparisons include relations that compare the aspects of two or more entities but do not grade them. There are three main subtypes of this class:

    - Entity A is similar to or different from entity B with regard to shared aspects, e.g. 'A tastes different from B'.
    - Entity A has aspect a1 and entity B has aspect a2 (a1 and a2 are usually substitutable), e.g. 'Restaurant A has indoor play room for kids, but Restaurant Y has outdoor playground'.
    - Entity A has aspect a1, but entity B does not have aspect a1, e.g. 'Phone-x has earphones, but Phone-y does not'.

In this paper, we cover the first and last type only; the other types will be the subjects of future work.

3 Literature review

In this section, we present the most notable research on identifying and mining comparative opinions in English and Arabic. We also present research on the preprocessing step to solve Arabic language problems and works on YouTube comment mining.

3.1 Comparative opinion mining

Some publications are available regarding comparative opinions of English text. The first widely known paper on comparative OM was presented in 2006 by Jindal and Liu [9], who studied the problem of identifying comparative sentences in English text. The authors developed a binary classifier to classify each sentence into either comparative or non-comparative class.

Later, Jindal and Liu [22] expanded the research on mining comparative sentences. Their corpus was constructed using various items collected from different web sources in order to represent different types of data. The items included reviews from customers, various forum discussions and random news articles. The authors used two new methods to identify comparative sentences and extract comparative relations based on two new types of rules: class sequential rules (CSRs) and label sequential rules (LSRs). The extraction of relations involved extracting the entities and their features (aspects) that were being compared, and comparative keywords.

Xu et al. [23] performed comparative opining classification on data collected from customer reviews of several mobile phones on the Amazon website. The authors employed three domain experts in mobile phones to manually annotate these opinion data. They applied a multi-class SVM-based method to these comparative reviews. The classes were '>', '<', '=' and 'no comparative' relations.

Pereira [24] proposed a genetic algorithm to identify comparative sentences from short sentences based on sequential pattern classification. The authors conducted an experiment using 1000 product reviews from Amazon and 1500 short sentences from Twitter.

Saritha and Pateriya [25] used a rule-based shallow parsing technique to identify comparative sentence from contents that were generated by users.

Some works on mining comparative opinion have been performed in other languages, such as Chinese [26–29], Korean [30, 31], Vietnamese [32] and Indonesian [33].

El-Halees [8] is the first study (to the best of our knowledge) that focused on mining comparative sentences in Arabic. The authors used two approaches to identify comparative text: a linguistic approach and a machine learning approach. The corpus consisted of documents related to opinions expressed in Arabic from three different domains: education, technology and sports. After cleaning and pre-processing the corpus, the researchers used a method that depended on linguistic classification to separate comparative statements from non-comparative ones; they obtained an f-measure of 63.73%. To enhance the f-measure, a combined approach of a linguistic method and three machine-learning methods was used to improve the performance, which obtained an accuracy of 88.87%.

El-Halees [8] made the first step on the long road of mining Arabic comparative opinion. However, our work differs from that of El-Halees because our work focuses on mining comparative opinion. El-Halees' works comprised two steps: First, the sentences were classified as comparative or non-comparative using POS tags and combined using machine learning algorithms (NB, SVM and KNN). Second, El-Halees generated manual rules to distinguish between different types of comparisons. We used comparative analysis to identify comparative opinions between different machine learning algorithms: statistical, rule-based and decision tree, which we combined with POS tags and keywords lists. Regarding the data, we used user-generated text. Our corpus is a collection of opinions obtained from social media (YouTube) in three domains (cars, mobiles and video games), while El-Halees used statements on different posts from three different domains (education, technology and sports).

A study published recently in this field (Arabic comparative mining) [7] focused on extracting the relational elements (entities and aspects) using the CRF algorithm from a dataset of 480 Arabic comparative opinions that they gathered and analysed. The averages of f-measures were 67.27%, 52.81% and 27.8% for entity 1, entity 2 and aspect extraction, respectively.

3.2 Arabic text processing and mining

Some previous works have attempted to solve various processing problems related to the Arabic language, which is morphologically rich. The first systems automatically tokenized text and parts of speech (POS) in Arabic text were discussed in [18]. A Support Vector Machine (SVM)-based approach was developed to automatically tokenize (segmenting of clitics), tag POS and annotate base phrases in Arabic text. The authors took data from the Arabic TreeBank, an MSA corpus containing Agency France Press newswire articles. The corpus comprised 734 news articles (140,000 words corresponding to 168,000 tokens after semi-automatic segmentation) covering topics such as sports, politics and news.

Later, in [34], the authors presented MADAMIRA. It is a system that is constructed for the morphological analysis and disambiguation of Arabic. It combined two valuable tools previously available in Arabic NLP field: MADA and AMIRA.

MADA used ALMOR (an Arabic lexeme-based morphology analyser) for generation of every possible interpretation of each input word. Afterwards, if applied different language models to decide which is the most likely suitable analysis for each word in the light of available context. MADA used tokenizer named TOKAN to tokenize MADA-disambiguated text. The other tool used by MADAMIRA is AMIRA. It is a system for performing number of tasks that are tokenization, POS tagging, Base Phrase chunking (BPC) and Named Entity recognition (NER).

MADAMIRA built up on the two system, was able to provide more robust, portable, extensible and faster implementation.

There are many Arabic papers on text classification using machine learning algorithms, but they concentrate on classifying the text into different categories based on the document content.

Thabtah et al. [35] used text classification and categorization methods to analyze 1562 Arabic documents collected from the Saudi Press Agency. The Arabic documents belonged to 6 categories.

Saad and Ashour [20] discussed Arabic text classification in relation to impact of usage of text preprocessing techniques and different term weighting schemes. The authors performed experiments on an Arabic text dataset collected manually from the Aljazeera news website. The dataset contained 119 text documents belonging to one of three categories (sports, health, computer and communications). They used a C4.5 DT with a tenfold cross-validation.

Alsaleem [14] applied Naïve Bayes (NB) and Support Vector Machine (SVM) algorithms to different Arabic data sets collected from Saudi Newspapers. The experimental results revealed that the SVM algorithm outperformed NB on all measures.

Khorsheed and Al-Thubaity [13] provide results of their experiments using different classification algorithms including C4.5, C5.0, MLP, neutral networks, SVM, NB and KNN algorithms. They discovered that SVM gave the most accurate results.

3.3 YouTube comment mining

Researchers have increasingly been studying YouTube data for number of reasons. Investigation of users' comments and analysis of video popularity using various metrics are interesting areas for research. Madden et al. [12] conducted a content analysis of 66,637 user comments on YouTube videos, and the authors created a classification schema to categorise the types of comments. Their schema revealed 10 broad categories and 58 subcategories that reflect the wide-ranging use of YouTube comments. Additionally, [36] carried out a systematic study of OM from approximately 35,000 YouTube comments by training a set of supervised multiclass classifiers to distinguish between video and product related opinions.

4 Research methodology

This paper discusses comparative opinion discovery, which identifies comparative sentences and non-comparative sentences in the corpora of opinions. Extracting entities and aspects from comparative sentences and then determining sentiment polarities (negative, positive or neutral) for different entities based on different aspects can be performed based on the output of our work.

To evaluate our approach, a set of experiments was designed and conducted. In this section, we describe the experimental design, including the corpus, pre-processing stage and evaluation metrics.

The details of the research design are shown in Fig. 2.

4.1 Data acquisition

This study started by collecting Arabic comments from YouTube. Since there is no publicly available corpus for Arabic comparative opinions, we created our corpus from scratch. To build a corpus of YouTube comments, we focused on a particular set of videos (videos comparing products). YouTube Data API v3 lets users incorporate functions normally executed on the YouTube website into a website or application. The API can retrieve different types of resources such as a videos, playlists, comments or subscriptions. The API also supports methods to list, insert, update or delete many of these resources.

We used CommentThread Resources and List Operations to retrieve a list of comments for different videos containing content and to compare products in different domains (e.g. cars, phones, laptops and video gaming devices).

4.2 Dataset characteristics

This section exhibits several characteristics of the collected Arabic comments.

- Repeated letters were used to stress something, e.g.'أحسسسسن', 'bettttttter'.
- Most of the collected Arabic comments contained spelling mistakes.
- Some of the comments were a mixture of Arabic and English. It is usual to find comments that consist of Arabic and English since most product names are in English, e.g. 'ABC بكثير وبدون مقارنة من أفضل XYZ', which means 'ABC is much better than XYZ, without comparison'.

Some of the comments were removed because they were not suitable for experiment conditions.

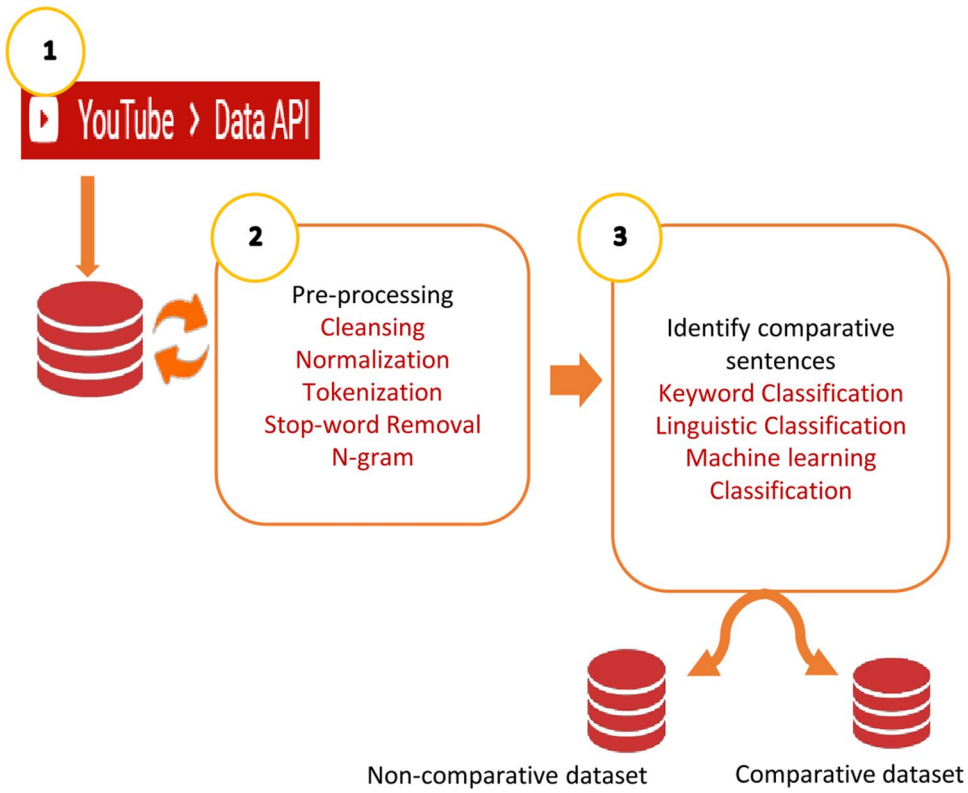

Fig. 2 The work flow summarizing the research work

- Some comments consisted of only one word, e.g. 'هههههه' ('hahaha') or 'س' ('X'), which means the user preferred one product to another without providing a comparison.
- Some comments did not relate to the video subject, e.g. advertisements.
- Some comments were very short and consisted of only 2 words, whereas some were very long with more than 40 words.

There exist no established style, template or patter that users need to follow for writing their reviews. Hence, reviews are fully unstructured and for this work, we dealt with the fully unstructured Arabic text.

Labelling: The data set was manually labelled as (comparative, non-comparative). For labelling process, the conditions were determined that must be valid in order to classify an opinion as comparative.

- The comparison type is non-equal gradable, where relations of the type 'greater or less than' express an ordering of entities, or non-gradable.
- A comment must include at least two entities.
- The comment may include aspects (or not).

If the opinion did not follow the above conditions, it was treated as non-comparative. Three Arabic native speakers performed the categorization process. Two labelers categorized the sentences, and the third labeler made decisions about sentences that raised a conflict between the first and second labelers. The labelers were asked to adhere strictly to the abovementioned conditions.

Approximately 43% of the Arabic reviews in the dataset were comparative text, and approximately 57% were non-comparative.

4.3 Pre-processing

Raw data often needs to be pre-processed. Text pre-processing is an important stage in text mining. The following are some popular pre-processing steps:

1. Data Cleansing: Since comments contain several syntactic features that may not be useful for machine learning, the data must be cleaned by removing URLs or website links (http or www). Comments might also have some repeated letters when the user wants to emphasize certain words, and these letters must be removed. Emoticons, special characters and diacritics were also removed.
2. Tokenization: Tokenization breaks a sentence into words, phrases, symbols or other meaningful tokens by removing punctuation marks.
3. Stop word removal: Stop words are common words that do not add meaningful content to a document [37]. For example, (من, إلى, على, أما, و) in English 'from, to, on, as for, and'.
4. Normalization: The letters that have more than one form were normalized into one form. For example, Alef in Arabic has many forms (أ, إ, آ, ا) and was thus normalized to (ا), and Taa Almarbotah (ة, ه) was normalized to (ه).
5. POS tagging: The POS tagging step was performed to identify different POS in the text.

4.4 Classification

After the pre-processing steps and labelling, the classification process was performed using three methods to achieve high accuracy.

1. Linguistic Approach: We used MADAMIRA v2.0 in this research, and the output was saved in XML file format. We used a simple classifier to classify the comments into comparative or non-comparative classes based on adjective comparative words; in Arabic, this is called 'اسم التفضيل' or 'Preference Name' in the POS tag attribute. This method was able to identify direct comparisons only. Some of the comments that were classified as non-comparative were actually comparative; however, because they did not contain adjective comparative words, this method could not identify them. Hence, such opinions require an additional classification in the future.
2. Machine Learning Approach: To overcome the limitations of linguistic classification, we used various well-known supervised learning methods, including the Naïve Bayes statistical classifier, JRip rule-based classifier and C4.5 decision tree, to predict comparative comments given that the trained data contained comparative and non-comparative comments. A word list was generated after processing the text to show the number of occurrences of each word. The high occurrences of words in the comparative category were stored in the keyword list to use later during the keyword classification process.
3. Keywords: This method involved filtering the comments that contained only comparison relations using the keyword strategy. Nitin and Liu [9] used this strategy to filter data and the method worked very well. The authors manually identified a list of 83 keywords and key phrases. However, these keywords achieved high recall but low precision. The authors improved the f-measure using this strategy with machine learning (Naïve Bayes). We manually collected the words with the highest occurrences in each comparative

class from the NB process and we added the most frequent comparative names and adjectives in Arabic to create a list of approximately 30 keywords.

5 Discussion and results

In the experiments, we used RapidMiner software with the text mining extension that includes different tools designed to prepare text documents for mining tasks (tokenization, stop word removal and n-gram). We also used the Weka extension in RapidMiner for the J48 and JRip classifiers. RapidMiner provides an environment in which various machine learning and data mining processes can be performed including a cross-validation process that is used to estimate the performance of several learning operators, such as SVM or NB. The Arabic stop words list included in RapidMiner was also applied to the corpus to remove words without relevant meaning.

Our experiment comprised six runs and two datasets: the original corpus and a corrected copy (corpus+). Corpus+ includes the following changes:

- Corrections of misspelled words that seemed inadvertently misspelled; for example, there is a missing letter in 'س احل من ص', which means 'X is bette than Y'. The word 'احل' ('bette') was meant to be 'احلى' ('better').
- Spaces were added between two or more words connected to each other where there was obviously a space missing between them.
- Some common misspellings were corrected in words commonly misspelled, such as AlHamzah, Alef Almamdodah and Alel Almaqsorah (ء, ا, ى), and incorrect Aldad and Altha characters (ض, ظ) were corrected.

First, we conducted experiments on comparative sentence identification using NB, JRip and J48 with two feature sets: two-grams and remove all stop words; two-grams remove some stop words. Then, we conducted some experiments using POS tags only, keywords only and both. The last experiment combined all methods.

To evaluate our approach, we applied widely used measures in text and opinion mining fields. These measures included Precision, Recall, F-measure, and Accuracy. The overall correctness of the classification is measured by metric of accuracy. Precision is calculated by finding the percentage of predicted document classes that are correctly classified. When the value of precision is higher, we are more confident that what has been predicted by the system, is indeed correct. Recall, on the other hand is the percentage of all documents in a given class that are correctly classified. With higher value of recall, we are sure that system is not missing correct items. The F-measure is a combined metric that takes both precision and recall into consideration as a weighted average of the two [8, 38].

The overall results are given in Fig. 3, which contains the precision, recall and F-score values of all the steps (different techniques). All the results were obtained through tenfold cross validation. We discuss the results below.

The precision, recall and F-score results at each step in the proposed technique are presented in Table 1.

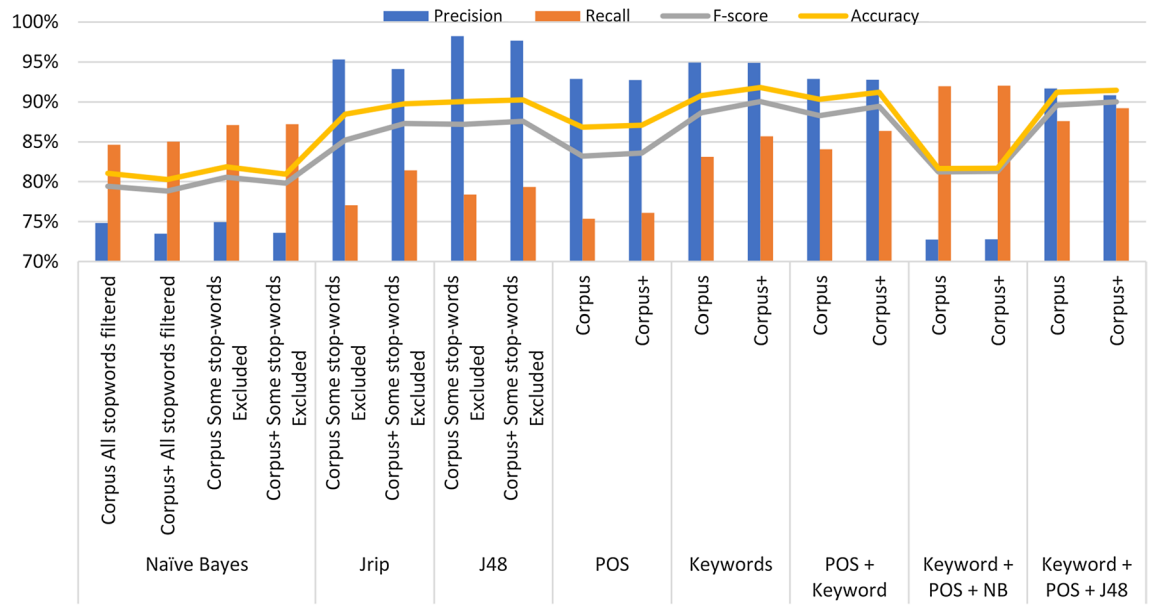

Fig. 3 Results of comparative opinion identification

Table 1 Precision, recall, F-score and accuracy values

Dataset	Approach	Precision (%)	Recall (%)	F-score (%)	Accuracy (%)
Corpus	NB classifier (all stop-words filtered)	74.85	84.64	79.44	81.07
	NB classifier (some stop-words excluded)	74.96	87.11	80.58	81.85
	Jrip (some stop-words excluded)	95.31	77.06	85.22	88.45
	J48 (some stop-words excluded)	**98.22**	78.39	87.19	90.05
	POS classifier	92.87	75.36	83.20	86.84
	Keywords classifier	94.91	83.13	88.63	90.78
	POS + keyword classifier	92.88	84.08	88.26	90.33
	Keyword + POS + NB classifier	72.77	91.94	81.24	81.65
	Keyword + POS + J48 classifier	91.67	87.58	89.58	91.19
Corpus+	NB classifier (all stop-words filtered)	73.52	85.02	78.85	80.29
	NB classifier (some stop-words excluded)	73.60	87.20	79.82	80.95
	Jrip (some stop-words excluded)	94.09	81.42	87.30	89.76
	J48 (some stop-words excluded)	97.67	79.34	87.56	90.25
	POS classifier	92.73	76.11	83.60	87.09
	Keywords classifier	94.86	85.69	**90.04**	**91.81**
	POS + keyword classifier	92.77	86.35	89.44	91.19
	Keyword + POS + NB classifier	72.79	**92.04**	81.29	81.69
	Keyword + POS + J48 classifier	90.83	89.19	90.00	91.44

The bold numbers indicate the highest value for each column

Our results were compared to those of the only study that identified comparative Arabic sentences from non-comparative text [8]. El-Halees [8] used 1048 posts with 435 comparative statements and 613 non-comparative statements. The researchers tested three approaches, using POS tags only, using machine learning algorithms only and then combining these methods. Using POS only, the f-measure average was 63.73%. We achieved better results in this study. El-Halees [8] used a Stanford POS tagger, and we used a MADAMIRA POS tagger. The best result was achieved when El-Halees [8] used SVM with POS tagging because the f-measure was between 87 and 88%. Our result using the J48 classifier is 90% and it is better. Additionally, El-Halees [8] did not use the keyword approach, and we achieved the best result in our work using this approach.

Table 1 and Fig. 3 show that the accuracy in this study ranges from 81 to 91%. The best performance was achieved when we used the keyword only approach, which can be attributed to the fact that most of the keywords that we used were extracted from the same corpus.

In our experiment on a machine with normal computational power, the execution time for each technique to produce the results did not exceed three hours.

In the next few subsections, we will present the analysis of results for the experiments that were performed based on the approaches discussed in this paper.

5.1 NB (Naïve Bayes) classifier

We obtained an accuracy between 80 and 81% using the NB classification technique after applying some pre-processing steps. Some sentences were classified as non-comparative when they were actually comparative for the following reasons:

- Some stop words were filtered from the text during the text pre-processing stage that were important in identifying comparative sentences and for use in comparisons, e.g. 'من, أما, لكن' ('than', 'as' and 'but'). To resolve this issue, we excluded these stop words in another run to compare the performances. The accuracy subsequently increased to 80.95% for corpus + and 81.85% for corpus.
- Variations of the same words exist, i.e. some words in our corpus were written differently but were the same word, and variations occurred because of misspellings or colloquialisms, which made the number of occurrences of such words different. Table 2 shows some examples.

Table 2 Different writing styles of entity names

Words	Translation
كثير, كتير	A lot
اني, انا	I
عندو, عنده	Has
شيء, شي, شيء	Thing
لكن, لاكن	But

- The product about which the comparison was made was written differently in the corpus. Some users wrote their comments in English and others wrote them in Arabic. Some users used both languages and others used abbreviations, which led to a dispersion and division among the occurrences in a class. For example, 'Xbox One' was written in English as (Xbox, Xbox1, Xbox 1 or Xbox one) and in Arabic as (بوكس 1 بوكس, الاكسبوكس, اكسبوكس, اكسبوكس' ون, اكس ون, الاكس, اكس اكس).
- Some comments used brand names while other comments used the name of the product or both to refer to the same thing. For example, 'Galaxy' product name and 'Samsung' brand name which were written in Arabic as 'غالكسي, كالكسي, جلكسي, كلكسي, سامسونج سامسونغ, سمسنك, سامنسغ'.

5.2 JRip rule-based classifier

The results of the classifier based on the RIPPER algorithm are better than those based on the NB method. Notably, the accuracy increased to 88.45% for corpus and 89.76% for corpus+. The number of rules yielded was 15 for corpus and 12 for corpus+. Most rules show that if the sentence includes comparison words (preference name), such as 'اكثر, احلى, اقوى, اطلق, احسن, افضل' ('better', 'stronger', 'more beautiful' and 'more') and conjunctions 'من, اما' ('than' and 'as'), the sentence was classified as comparative.

Rule-based classifiers are readable, which can lead to a better understanding of such text. Our evaluation suggests that the rules can provide insights into and a better understanding of relevant text. The result is very promising to do more research on rule-based classification involving comparative OM. However, JRip classification is only feasible when the number of training examples is relatively small. In our experiment on a machine with normal computational power, the execution time to produce the results and rules was approximately three hours.

5.3 J48 DT classifier

This classifier yielded the best results compared to other machine learning algorithms (NB, JRip), and it achieved the highest precision. The accuracy reached 90% for both corpora.

5.4 POS tag classifier (MADAMIRA 2.0)

We obtained an accuracy of 86–87% using this technique. This result shows that these POS tags are good indicators for comparative sentences detection. The reasons why some comparative sentences were misclassified as non-comparative are as follows:

- Some sentences in our corpus did not have comparison terms (comparative adjective), e.g. 'س تتفوق على ص' means 'X outperforms Y'.
- Some comparison terms contained misspellings. Therefore, the POS classifier could not discover them, e.g. 'س سعره اغله من ص' means 'X price is more expensive than Y', or 'س افظل من ص' means 'X is better than Y'.
- Some comparative sentences included no explicit comparison or preference of one product over another, e.g. 'س جيد في الطرق وص ممتاز للبر' means 'X car is good for roads and Y car for desert'. These sentences are the fourth type of non-gradable comparison.

The following are possible reasons as to why non-comparative sentences were classified as comparative:

- Some non-comparative sentences in our corpus had comparison terms (comparative adjectives) that were not used for comparison purposes, e.g. 'الله عزاك أحسن' is a consolation phrase, and the first word literally means 'better'. Another example 'يستخدمون بلايستشن أكثر العرب', which means 'Most Arabs used PlayStation'.

5.5 Keywords classifier

We obtained a high precision of 94% and a high accuracy of 90–91%. In this approach, every sentence that contained words from the keywords list was considered a comparative sentence. This shows that these keywords are very good indicators for comparative sentences detection. The reasons why sentences were classified as non-comparative when they were comparative are as follows:

- Such sentences in our corpus had no explicit comparison or preference of one product over another, e.g. 'قصص خذ ص اذا تبي لعب اون لاين خذ س واذا تبي تلعب نظام' means 'If you like to play online, buy X, if you like to play stories, buy Y'. These sentences belong to the fourth type of non-gradable comparison.
- Misspellings in comparative names, e.g. 'ص س احل من' means 'X is more beautiful than Y' and should be written as 'س احلى من ص'. This problem was solved in corpus+.
- Using stop words in comparisons, e.g. 'استخدامه لكن ص صعب س سهل' means 'X is easy to use but Y is difficult'. These stop words could not be added to the keywords list because they could affect the performance of classifying non-comparative text.
- Using colloquial comparative words that mean another thing in another context, e.g. 'ص س أطلق عزم من' means 'X is better than Y'. The word 'اطلق' here means 'better' or 'faster' but it can also mean 'shoot', 'fire' or 'release' in another context, e.g. 'أحمد أطلق النار على المجرم' means

'Ahamad fires shots at criminals'. However, if 'من' ('than') comes after this phrase, this word is used as a comparison and was added with 'من' ('than') to the keywords. Therefore, when the word is used alone, the sentence is classified as non-comparative to balance the comparative and non-comparative performance.

- Some comparative words may be used as a noun in a comparison or a verb in other sentences, e.g. 'أوضح' ('clearer') could be used as an adjective in a comparison, such as 'الجرافيكس في س أوضح من ص', which means 'X Graphix is clearer than Y', or as a verb, such as 'أوضح المتحدث أن...', which means 'Speaker explained...'. Therefore, if these words are added to the keywords, they could affect the performance of classifying non-comparative text. We solve this issue by adding 'من' ('than') to the keywords as previously discussed in previous reason.

The same reasons noted for the keyword classifier approach could also be responsible for classifying non-comparative sentences as comparative.

5.6 Keywords + POS + NB classifier

The accuracy of combined approaches significantly improved compared to the use of NB alone. We obtained a high recall of 92% and an acceptable precision score of 72%. The keyword classifier is better because it has a higher recall.

5.7 Keywords + POS + J48 classifier

The results of this combination show that it is better than using J48 alone. The accuracy achieved was 91% for both corpora. The best results were observed for corpus; however, for corpus +, a slight difference in performance existed between this combination and the keyword classifier alone.

Finally, we discovered that the corpus with corrected spellings, corpus +, did not achieve the results we expected, despite the time and effort taken to make these corrections. By contrast, the NB classifier achieved better results for corpus compared to corpus + in terms of accuracy. We conclude that correcting misspellings does not improve the performance when the data set is colloquial because most words were colloquial and not corrected. We also recommend not filtering all stop words because some stop words are important in comparative sentences. The decision tree classifier C4.5 (J48 implementation) yielded the best results with 90% accuracy for both corpora compared to the other machine learning algorithms that we used.

6 Conclusion

The Arabic Language is becoming a popular area of research in opinion mining. However little work has been done in the field of comparative opinion mining. Arabic language comparative opinion is a difficult field to work on as Arabic language has more difficulties and problems as compared to other languages that are derived from Latin, because it implies the solving of different types of problems such as the short vowels, Alhamzah, prefixes, suffixes, colloquial, etc. In this work, we have combined many approaches in order to identify Arabic comparative opinions from YouTube comments. We used machine learning algorithms to construct the classifiers that can be used to identify comparative opinions automatically. Experiments with decision tree classifier C4.5 (J48 implementation) yielded the best results. Although obtained results with decision tree classifier were promising, we have shown that combination techniques (J48, keywords and POS) improved the performance that was achieved by J48 only. The keyword classifier is the best for detecting the gradable comparisons but combining different approaches achieved a good performance and balance between the gradable and non-gradable comparison. These results encourage us to continue working along this line. The models developed in this work can be used as input for applications that aim to find opinion about entities and their aspects in comparative sentence. Thus, this paper discusses pioneer work in the field of Arabic comparative opinion mining field and is expected to attract further researches along a similar line.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Human and animal rights This research uses publicly available data and it doesn't involve human participants or animals.

References

1. Liu B (2011) Opinion mining and sentiment analysis. Web data mining, pp 459–526
2. Al-Kabi M, Gigieh A, Alsmadi I, Wahsheh H, Haidar M (2014) Opinion mining and analysis for Arabic language. Int J Adv Comput Sci Appl 5:181–195
3. Martínez-Cámara E, Martín-Valdivia M, Ureña-López L, Montejo-Ráez A (2012) Sentiment analysis in Twitter. Nat Lang Eng. https://doi.org/10.1017/S1351324912000332
4. Basari A, Hussin B, Ananta I, Zeniarja J (2013) Opinion mining of movie review using hybrid method of support vector

machine and particle swarm optimization. Proc Eng. https://doi.org/10.1016/j.proeng.2013.02.059
5. He W, Shen J, Tian X, Li Y, Akula V, Yan G, Tao R (2015) Gaining competitive intelligence from social media data. Ind Manag Data Syst. https://doi.org/10.1108/IMDS-03-2015-0098
6. Varathan K, Giachanou A, Crestani F (2016) Comparative opinion mining: a review. J Assoc Inf Sci Technol. https://doi.org/10.1002/asi.23716
7. Defrawi M, Salah M, AbdAlAziz A, Eldin S (2017) Comparative relation extraction from Arabic opinions. Int J Comput Sci Inf Secur 15:230–235
8. El-Halees A (2012) Opinion mining from Arabic comparative sentences. In: The 13th international Arab conference on information technology ACIT, pp 265–271
9. Jindal N, Liu B (2006) Identifying comparative sentences in text documents. In: Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, pp 244–251
10. Refaee E, Rieser V (2014) An Arabic Twitter corpus for subjectivity and sentiment analysis. In: LREC International Conference on Language Resources and Evaluation, pp 2268–2273
11. Severyn A, Moschitti A, Uryupina O, Plank B, Filippova K (2014) Opinion mining on YouTube. In: Proceedings of the 52nd annual meeting of the association for computational linguistics, vol 1, pp 1252–1261
12. Madden A, Ruthven I, McMenemy D (2013) A classification scheme for content analyses of YouTube video comments. J Doc. https://doi.org/10.1108/JD-06-2012-0078
13. Khorsheed M, Al-Thubaity A (2013) Comparative evaluation of text classification techniques using a large diverse Arabic dataset. Lang Resour Eval. https://doi.org/10.1007/s10579-013-9221-8
14. Alsaleem S (2011) Automated Arabic text categorization using SVM and NB. Int Arab J e-Technol 2:124–128
15. Saritha S, Pateriya R (2014) Methods for identifying comparative sentences. Int J Comput Appl 108:23–26
16. Ibrahim H, Abdou S, Gheith M (2015) Sentiment analysis for modern standard Arabic and colloquial. Int J Nat Lang Comput. https://doi.org/10.5121/ijnlc.2015.4207
17. Al-Kabi M, Al-Qudah N, Alsmadi I, Dabour M, Wahsheh H (2013) Arabic/English sentiment analysis: an empirical study. In: The fourth international conference on information and communication systems, pp 23–25
18. Diab M, Hacioglu K, Jurafsky D (2004) Automatic tagging of Arabic text: from raw text to base phrase chunks. In: Proceedings of HLT-NAACL 2004: short papers, pp 149–152
19. Mourad A, Darwish K (2013) Subjectivity and sentiment analysis of modern standard Arabic and Arabic microblogs. In: Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis, pp 55–64
20. Saad M, Ashour W (2010) Arabic text classification using decision trees. In: Proceedings of the 12th international workshop on computer science and information technologies CSIT, vol 2, pp 75–79
21. Mobarz H, Rashown M, Farag I (2014) Using automated lexical resources in Arabic sentence subjectivity. Int J Artif Intell Appl 5:1
22. Jindal N, Liu B (2006) Mining comparative sentences and relations. AAAI 22:1331–1336
23. Xu K, Wang W, Ren J, Xu J, Liu L, Liao S (2011) Classifying consumer comparison opinions to uncover product strengths and weaknesses. Int J Intell Inf Technol. https://doi.org/10.4018/jiit.2011010101
24. Pereira F (2015) Mining comparative sentences from social media text. In: Second workshop on interactions between data mining and natural language processing co-located with European conference on machine learning and principles and practice of knowledge discovery in databases, pp 41–48
25. Saritha S, Pateriya R (2016) Rule-based shallow parsing to identify comparative sentences from text documents. In: Emerging research in computing, information, communication and applications. https://doi.org/10.1007/978-981-10-0287-8_33
26. Hou F, Li G (2008) Mining Chinese comparative sentences by semantic role labeling. In: Machine learning and cybernetics international conference. https://doi.org/10.1109/ICMLC.2008.4620840
27. Gao S, Wang H, Song Y, Lu T (2016) Mining comparison opinions from Chinese online reviews for restaurant competitive analysis. In: WHICEB 2016 proceedings
28. Shi L, Li S, Jiang P, Liu H (2016) Improving comparative sentence extraction of chinese product reviews by sentiment analysis. J Eng Sci Technol Rev 9:149–156
29. Wang H, Gao S, Yin P, Liu J (2017) Competitiveness analysis through comparative relation mining: evidence from restaurants' online reviews. Ind Manag Data Syst. https://doi.org/10.1108/IMDS-07-2016-0284
30. Yang S, Ko Y (2009) Extracting comparative sentences from Korean text documents using comparative lexical patterns and machine learning techniques. In: Proceedings of the acl-ijcnlp 2009 conference short papers, pp 153–156
31. Yang S, Ko Y (2011) Extracting comparative entities and predicates from texts using comparative type classification. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, pp 1636–1644
32. Bach N, Van P, Tai N, Phuong T (2015) Mining Vietnamese comparative sentences for sentiment analysis. In: Knowledge and systems engineering (KSE) 2016 eighth international conference. https://doi.org/10.1109/KSE.2015.36
33. Saelan A, Purwarianti A, Widyantoro D (2017) Question analysis for Indonesian comparative question. J Phys: Conf Ser. https://doi.org/10.1088/1742-6596/801/1/012077
34. Pasha A, Al-Badrashiny M, Diab M, El Kholy A, Eskander R, Habash N, Pooleery M, Rambow O, Roth R (2014) MADAMIRA: a fast, comprehensive tool for morphological analysis and disambiguation of Arabic. Proc Ninth Int Conf Lang Resour Eval LREC 14:1094–1101
35. Thabtah F, Eljinini M, Zamzeer M, Hadi W (2009) Naïve Bayesian based on Chi square to categorize Arabic data. In: Proceedings of The 11th international business information management association conference (IBIMA) conference on innovation and knowledge management in twin track economies, pp 4–6
36. Severyn A, Moschitti A, Uryupina O, Plank B, Filippova K (2014) Opinion mining on YouTube. In: Proceedings of the 52nd annual meeting of the association for computational linguistics, pp 1252–1261
37. Solka J (2008) Text data mining: theory and methods. Statist Surv. https://doi.org/10.1214/07-SS016
38. Maynard D, Peters W, Li Y (2006) Metrics for evaluation of ontology-based information extraction. CEUR Workshop Proc 179:1–8