Journal of Computer Science

Original Research Paper

Sentiment Analysis of Arabic Reviews for Saudi Hotels Using Unsupervised Machine Learning

[1]Samar Alosaimi, [1]Maram Alharthi, [1]Khloud Alghamdi, [1]Tahani Alsubait and [2]Tahani Alqurashi

[1]College of Computer and Information Systems, Umm Al-Qura University, Makkah, Saudi Arabia
[2]Common First Year Deanship, Umm Al-Qura University, Makkah, Saudi Arabia

Article history
Received: 27-07-2020
Revised: 23-09-2020
Accepted: 24-09-2020

Corresponding Author:
Tahani Alsubait
College of Computer and Information Systems, Umm Al-Qura University, Makkah, Saudi Arabia
Email: tmsubait@uqu.edu.sa

Abstract: Virtual worlds such as social networking sites, blogs and content communities are extremely becoming one of the most powerful sources for news, markets, industries etc. These virtual worlds can be used for many aspects, because they are rich platforms full of feedback, emotions, thoughts and reviews. The main objective of this paper is to cluster Arabic reviews of Saudi hotels for sentiment analysis into positive and negative clusters. We used web scraping to collect Arabic reviews associated only with Saudi hotels, from the tourism website TripAdvisor and obtained in total 4604 Arabic reviews. Then the TF-IDF was applied to extract relevant features. An unsupervised learning approach was applied, in particular K-means and Hierarchical algorithms with two distance metrics: Cosine and Euclidean. Our manual labelled test data shows that the K-means algorithm with cosine distance performed well when applying all of our prepossessing steps. We concluded that the suggested prepossessing steps play a critical role in Arabic language processing and sentiment analysis.

Keywords: Unsupervised, Clustering, Hotel Reviews, Arabic Sentiment Analysis

Introduction

In recent times, we have witnessed a huge flow of data via smart applications, web pages and social media portals. Due to the rapid development of information sharing technologies, people can exchange opinions, feelings and snapshots of situations they have been in, via online outlets. Some of these information sharing outlets, which are heavily used by tourists around the globe, are websites of hotel guests' reviews. It has actually made it easy for tourists to find hotels that they can trust and that meet their expectations by reading previous guest reviews. Simply, the review is a text written by guests to express their feeling/opinion about virtually anything, including, services, room cleanliness and the food offered when they stayed in a particular hotel. Moreover, from an economic perspective, guest reviews have become an important factor affecting hotel bookings/revenues as it has an influence on tourists' decisions. Also, hotel owners can benefit from these reviews, as they can improve their services based on reviewers evaluation. However, it is challenging to make use of online reviews as they are becoming available with huge volumes. In addition, tourists and hotels' owners may find it difficult to understand some reviews and figure out whether they are positive or negative.

To overcome this problem, we present an approach based on unsupervised machine learning methods to discriminate between positive and negative reviews. In particular, we experiment with Arabic language reviews of Saudi hotels which have been gathered for the purposes of this research. From a practical perspective, this work will help hotels assess the quality of the administrative and operational staff and evaluate the satisfaction of customers with the services provided and thus work on improving the quality and upgrading of the hotel services. Also on the client side, it will help in choosing the right hotel and comparing hotels in an easy and quick way. Our work is carried out on Arabic long text reviews with relatively promising results compared to exiting methods applied to short texts. The remainder of this paper is organized as follows: Section 1 encompasses a review of related works that have been applied to process Arabic language reviews. Section 2 presents our proposed methodology. Section 3 shows the results obtained on four versions of our dataset. Finally, section 4 presents concluding remarks and suggests future research lines.

Literature Review

In this section, we shed light on some exiting work in the related literature. For example, (Abuaiadah *et al*., 2017) have worked on sentiment analysis of Arabic tweets using unsupervised methods. They define tweets as short texts. The shortness of the text in this context is a challenge due to the possibility of ambiguity. Also, they have examined text preprocessing and similarity methods for clustering algorithms and examined how they impact the results. They adopted the standard k-means algorithm for sentiment analysis of tweets. They chose a publicly available dataset that contains 2000 tweets. The dataset was labeled as positive or negative, manually. They created four versions of the dataset: (1) Raw (no preprocessing), (2) NoSW (just stop word removal), (3) light10 (use light10 stemmer and stop word removal) and (4) root (use Khoja stemmer and stop word removal). They report that removal of the stop words and applying stemming helped to improve the quality. They chose the number of clusters randomly. The root-based stemming has achieved the highest quality in this study.

Hu *et al*. (2013) focused on emotional signals for sentiment analysis using unsupervised methods. Emotional signals include emoticons and product ratings that are in posts on social media platforms. They tried to investigate if emotional signals help in sentiment analysis. Nowadays, we have big unstructured and unlabeled data, so they proposed to use unsupervised sentiment analysis. Due to that labeling of data is time-consuming; one of the methods of unsupervised sentiment analysis is a lexicon-based method. It is a traditional way but still difficult on this project. Due to character-count limitations in social media posts, especially tweets on Twitter which are limited by 240 letters, the text extracted from these platforms is considered to be short text. In this respect, short text usually lacks more information, creating a challenge for this method. Also, the people in social media are generating and rapidly growing new expressions. They adopted two datasets that are freely available: STS and OMD. They adopted the unigram model to extracting feature weight. They do not apply the stemming or stop word removal. They proposed a novel unsupervised Sentiment Analysis which is Emotional Signals for unsupervised Sentiment Analysis (ESSA). In their experiment, their novel ESSA obtained a higher accuracy on two datasets compared to unsupervised Sentiment Analysis methods.

Zhang and Yu (2017) employed Word2Vec tool to obtain one word vector and k-means clustering algorithm to collect similar words to the one cluster. Each cluster was represented by a new text feature vector dimension. But, they encountered a problem in the k-means clustering algorithm which is its sensitivity to the selected number of clusters. They addressed this problem with the ISODATA clustering algorithm. They mentioned having two methods for sentiment analysis: (1) Machine learning method and (2) sentiment dictionary method. In particular, a machine learning method that has two stages. The first stage is a text feature vector, they represented it by the Bag-Of-Words approach. But a drawback of this approach is the high dimensionality, so they worked on reducing the dimensions. The dataset contains 100000 hotel reviews, just taking 4000 samples and manually labeling to 1 as positive and 0 as negative.

Kaur (2018) applied sentimental analysis on reviews of a book. He adopted two machine learning approaches: Supervised and unsupervised. He chose two publicly available datasets collected from GoodReads and Amazon. During preparing the datasets, he removed empty reviews and confused reviews. After that, he selected Naïve Bayes algorithms (NB) and Support Vector Machine algorithm (SVM) for supervised approach and selected Semantic Orientation (SO) - Pointwise Mutual Information (PMI) - Information Retrieval (IR) for unsupervised semantic orientation approach. He concluded that unsupervised techniques obtained better accuracy for long sentences in reviews, whereas supervised techniques had better accuracy for short sentences.

Al-Hadhrami *et al*. (2019) worked on sentiment analysis of tweets, especially English tweets. They adopted supervised and unsupervised approaches. The used algorithms were Support Vector Machine, Random Forest Classification and k-means Clustering. They used a publicly available dataset called Sentiment 140. After that, they applied text preprocessing steps on tweets to remove the noise. They extracted features using uni-grams and bi-gram. They computed Term Frequency-Inverse Document Frequency (TF-IDF) for each type of feature. The SVM algorithm obtained the highest accuracy with Uni-grams features.

Al-Smadi *et al*. (2018) compared between two of the most important supervised machine learning algorithms, namely SVM and RNN, in terms of ability to face the challenges of sentiment analysis in Arabic hotels reviews. One of the most important motivation for this research is that the review may contain more than one feeling, so they used another type of sentiment analysis called Aspect-Based Sentiment Analysis (ABSA) that shows each aspect of the review and the corresponding feelings. They trained the models on an ABSA dataset consisting of 19,226 samples for training and 4802 samples for testing. Sets of morphological, lexical, semantic and syntactic features were extracted to train classifiers to achieve three main targets: Aspect category identification, aspect opinion target expression extraction and aspect sentiment polarity identification. The results

showed the superiority of the SVM model in all tasks in terms of accuracy, but RNN was faster at the implementation time.

After one year, they provided an enhanced approach of their research Al-Smadi *et al*. (2019) using more techniques including Bayes Networks, Naïve Bayes, K-Nearest Neighbor (KNN), Decision Tree and Support-Vector Machine (SVM). The objective behind the selection of these classifiers was their prominent role in the classification of texts. They extracted more features such as Named-Entity Recognition (NER), Part-Of-Speech tagging (POS) and computational morphology to improve the results. All the proposed classifiers outperformed previous works that used the same dataset (SemEval-ABSA16), in addition the best classifier SVM has significantly improved its performance.

Gamal *et al*. (2019) proposed an approach for sentiment analysis in Arabic content where they applied a set of Machine Learning classifiers such as Logistic Regression (LR), Naïve Bayes (NB), Passive Aggressive (PA), Maximum Entropy (ME), Support Vector Machine (SVM), Ridge Regression (RR), Stochastic Gradient Decent (SGD), Multinomial NB (MNB) and Bernoulli Naïve Bayes (BNB) on a dataset they collected containing 151,500 tweets in different Arabic dialects and automatically labeled as negative/positive. The dataset passed through multi stages of preprocessing such as removing noisy data, tokenization, removing diacritics and removing non-Arabic letters. RR and PA achieved best values is 99.96%.

Heikal *et al*. (2018) trained two deep learning models, a CNN and LSTM, with different hyper-parameters on Arabic Sentiment Tweets Dataset (ASTD) to predict sentiment analysis of Arabic tweets. After that, they selected superior models and use them to build an ensemble model. CNN model with fully connected 100 layers gave the best result (64.30%). In addition, the best results for a LSTM model was with dropout rate 0.2 is 64.75%. Ensemble model performed better than the two models when used separately, it achieved 65.05%.

Boudad *et al*. (2018) surveyed the state of the art in Arabic sentiment analysis in previous major works and reported their advantages and disadvantages. They showed that there are three main methods mostly used to sentiment analysis tasks: Unsupervised, supervised and hybrid. However opinion spam detection, opinion holder extraction and Aspect Based Sentiment Analysis are the least analyzed tasks. They reviewed Arabic challenges and found that the main and most influential reason for obtaining an effective sentiment analysis system is trying to overcome challenges in the nature of the Arabic language.

Elnagar *et al*. (2018) used 692586 annotated reviews from BRAD 2.0 which is a large free Arabic dataset that has been used for sentiment analysis to classify book reviews. For the pre-processing steps, extraction and normalization were applied to the raw data. To extract features, the most frequent words were used to build the feature vector. Each review is represented as a feature vector of vocabulary. To verify the proposed dataset, several supervised classifications have been implemented, which are: Naïve Bayes, Decision Tree, Random Forest, XGBoost, Support Vector Machines (SVM). For Unsupervised classifications, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). The results showed high accuracy in both classifications and an increase in accuracy for the non-supervised classification.

The work of (Mataoui *et al*., 2018) describes a new approach for sentiment analysis in Arabic reviews. The approach is based on many steps: Processing, lexical entities disambiguation, separation between aspects, extraction of aspects and grouping of aspects. A database of social content related to hotels and products reviews has been used, which has been collected by ElSahar and El-Beltagy (2015). The Experimental results showed an F-measure of 69.56 for the hotels and 68.29 for the products dataset.

Abd-Elhamid *et al*. (2016) proposed a feature-based sentiment analysis technique for Online Arabic Reviews. Reviews in Arabic were collected from Forums, Facebook, YouTube and google search. Then they were stored in a database table. Cleaning the text is the first step in processing. Then rates organized into positive, negative, or neutral. Finally, before tokenizing, they manually added rates to unannotated reviews. Part of Speech tagging was used to determine if the word being a sentiment or feature. Arabic ToolKit Service (ATKS) has been used for tagging. For sentiment extraction and weight assigning, words with POS tagging to be sentiments word, then the number of those words appearing in the review is calculated. Weights are given for each word based on the frequency at which the word appears in positive, negative and natural reviews. For feature extraction, rules have been used for extracting noun and compound noun. Indeed, no repeated nouns on one review were allowed. The same algorithm has been applied for sentiment weighting. There are five rules for features-sentiments extracting: If Noun followed by Adjective extract both words, If Adjective not associated with noun assign it to the root, if consecutive noun connected with “” (and) assign last sentiment, if feature followed by more than one sentiment assign the average weight of those sentiments, if feature without sentiment assign the average weights within review. For classification unsupervised technique has been applied with (Triple polarity or TP). The evaluation has been performed based on the five mentioned rules.

Ismail *et al*. (2018) used supervised learning for twitter data where text was written in Sudanese Arabic dialect for extracting and analysis purposes. They trained four classifiers which are: Naïve Bayes, SVM, Multinomial Logistic Regression and K-Nearest Neighbor. On 4712 tweets were collected using twitter API. Cross-validation was used for validation. The best accuracy was achieved by KNN (k = 2) and it equals to 92.0 while the highest F1-score (72.0) was achieved by SVM.

Al-Ayyoub *et al*. (2019) provides an overview of the work that has been done in the Sentiment analysis field, the problems they address and the gap that exists. They argue that SA has been widely adopted in the field of English text. While, the Arabic language received little attention. However, there are many problems with SA. One of these problems is to deal with subjectivity. Sentiment analysis considered as subjective/opinionated text. So, any objective text should be excluded. They classified sentiment to two levels: Assigning a single sentiment label to an entire document (document-level SA) or considering each paragraph separately (paragraph-level SA). Other approaches, go into considering each sentence (or even each word) separately. Several attempts have tried to solve this problem by creating Aspect-Based SA (ABSA). The authors also define three sentiment classes: Binary SA (BSA) for Positive and negative, Ternary SA (TSA) include a class for neutral text and Multi-Way SA (MWSA) for positive, positive, neutral, negative and strongly negative. On the other hand, they argue that the domain of the text is also an important issue. What is considered a positive in the domain of sport might not be so in other domains such as Politics or Arts. Finally, Dialectal Arabic (DA), seems to be a common problem in sentiment analysis. The authors also mention some attempts to address the discussed issues. Some approaches have been made such as (Elhawary and Elfeky, 2010). This system consists of an extension of a previous system, designed to determine whether a webpage is in English or not. Another work is (El-Halees, 2014) to study changes in students' opinions between two consecutive semesters. Refaee and Rieser (2014) used two datasets collected at different times. The first one was manually annotated, where each tweet is supplemented with a set of features computed automatically. Sayed *et al*. (2020) used nine supervised machine learning algorithms. Namely, Gradient Boosting, Logistic Regression, Ridge Classifier, Support Vector Machine, Decision Tree, Random Forest, K-Nearest Neighbor KNN, Multi-layer Perceptron (MLP) and Naive Bayes classifiers for Arabic Sentiment Analysis. Their model focused on Arabic language reviews. They also created an Arabic corpus called (RSAC). Furthermore, many preprocessing strategies have been applied such as: Tokenization, normalization, Stop Words Removing and Stemming. For the experimental results, the Ridge Classifier (RC) appears to have the best performance in terms of accuracy, recall, precision, training time and F1-score.

Kwaik *et al*. (2020) collected and labled 36 k tweets into positive and negative tweets. Besides, 8 k tweets were annotated manually. Also, Distant supervision was applied using emojis as weak labels to annotate the whole dataset. They adopted a method to compare the emoji-based annotation with the human annotation to evaluate the corpus and got an observed agreement of 77.2%. In addition, Sentiment analysis machine learning model was built with unigram features.

In summary, many articles related to sentiment analysis were reviewed in this study. Existing works covered most of the Arabic dialects. Also, they covered many areas, such as news, tweets and reviews.

Methodology

The steps followed in this research are illustrated in Fig.1.

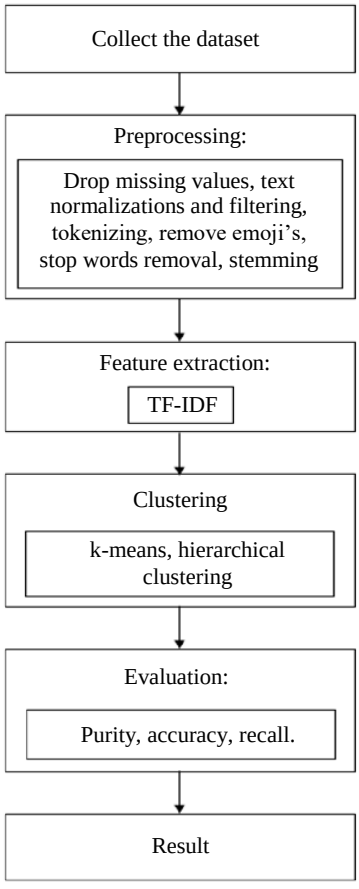

Fig. 1: The system diagram

Dataset

To extract knowledge, first you need to get or collect a dataset. One of the objectives of this research was to collect Arabic reviews of Saudi hotels from TripAdvisor[1] website and save them in a stable format (e.g., CSV). Web scraping was used to collect a dataset that contains 4604 reviews for 121 Saudi hotels. The distribution of collected reviews per Saudi cities is shown in Table 1.

As shown in Fig. 2, our collected dataset has 3 features, which are review number, the date of stay in the hotel and the review text. However, the review text was the only feature that have been used in the sentiment analysis.

For evaluation purposes, we need some labeled data. Therefore, we have manually labeled 30% of the data to positive or negative label by three human annotators. So our test dataset contains an additional feature, which is the cluster label.

Preprocessing

In this stage, we cleaned the unwanted content by performing some preprocessing steps, these are as follows:

1. Drop missing values
2. Text Normalisation and Filtering: It is necessary to clean up the reviews by removing punctuation marks, special characters, non-Arabic characters, dates, time, numbers, links and diacritics, etc.
3. Tokenizing: The purpose of tokenizing is to splitting sentences into tokens or words
4. Remove Emojis: Emoji can be considered as an auxiliary data in the classification of texts into positive and negative, but in this study, we only focus on analysing the written texts so we removed the emoji from our dataset
5. Stop Words Removal: Stop words are known as extremely frequent words, such as (pronouns conjunctions, prepositions and names). Therefore, in this step we removed them.
6. Stemming: We used the light stemming, that make possible to removing prefixes and suffixes.

Regarding these prepossessing steps, we created four versions of our dataset to test our approach, these are: (1) The original raw data, that we do not apply any preprocessing steps to it, (2) With normalization, which we only performed step 2, (3) With remove emojis, stop word and stemming without normalization, (4) With normalization, remove emojis, remove stop words and stemming and we called this version all preprocess.

[1] https://www.tripadvisor.com/

Table 1: The distribution of collected reviews per Saudi cities

City	No. of reviews	No. of hotels
Eastern province	1097	27
Jeddah	736	15
Riyadh	743	23
Makkah	1051	21
Al-Madina	675	23
Other	302	12
Total	4604	121

630	2019 تاريخ الإقامة: فبراير		..ابراج المبربديان هو الفندق عنوان الفخامة و الر
753	2018 تاريخ الإقامة: أبريل		..اشكر الطاقم الاستقبال كان راقين في التعامل ولك
542	2019 تاريخ الإقامة: يوليو		..خدمة سيئة جدا افسدوا علي عبادتي ويومي وادا اطل
535	2016 تاريخ الإقامة: مايو		..فندق سيييئ جداً لا يستحق نجمتين من الاستقبال
852	2017 تاريخ الإقامة: أبريل		..هذا هو المكان المثالي للإقامة إذا كنت تخطط على

Fig. 2: Sample of the collected dataset

Feature Extraction

As it is widely known that machine learning algorithms cannot work with texts as they are, so it must be converted into numerical data that can be easily handle. Therefore, we adopted the most popular approach to extract features from text, which is the Term Frequency-Inverse Document Frequency (TF-IDF) (Salton and Buckley, 1988).

TF-IDF

The TF-IDF combines two scores, the term frequency TF, which calculates the frequency of word in each review and the inverse document frequency IDF in order to reduce the weights of words that are repeated frequently and increases the weights of words that are repeated very rarely. Therefore, the TF-IDF is defined as shown in Equation 1:

$$TF - IDF(t,d,D) = TF(t,d) * IDF(t,D) \quad (1)$$

Here, $TF(t, d)$ calculates the number of times word t appears in review d and the IDF is defined as shown in Equation 2, where D is the total number of reviews in the dataset and the $df(t)$ is the number of reviews in which word t appears in D:

$$IDF(t,D) = \log \frac{|D|}{df(t)} \quad (2)$$

Learning

This is an iterative clustering algorithm that aims to partition instances into k clusters in which each observation belongs to the cluster with the nearest mean.

k-means

The k-means algorithm is a famous an unsupervised machine learning algorithm that aims to cluster documents into k clusters in which each document belongs to the cluster with the nearest centroid (MacQueen, 1967). We have initialised the k-means++ to speed up convergence and run 10 times with different centroid and 1000 iterations per run.

After we extracted the numerical features of our dataset, we run the k-means algorithm with two clusters: One for positive and the other is negative. We used the Euclidean distance as well as the cosine distance measures with k-means algorithm (Huang, 2008).

Hierarchical Clustering

The hierarchical clustering like a tree contains a collection of nested clusters (Patel *et al*., 2015). We built a hierarchy clustering with four types of linkage, which are ward, complete, average and single, using a vector matrix of training data.

In order to evaluate the hierarchical clustering result, we used Support Vector Clustering (SVC) model (Ben-Hur *et al*., 2001). So, we train SVC using hierarchy clustering results then we used it to predict testing data.

Evaluation Metrics

In unsupervised learning, we have two different approaches to evaluate the result of a clustering algorithm, which are: (1) Internal validation, if the labeled data is not available and (2) external validation. As we have the label of the test dataset, we applied the external approach.

One of the main external validation methods is the matching set, which includes four measures: Purity, accuracy, recall and f-measure (Palacio-Niño and Berzal, 2019).

Purity

Let us assume that *C* is the cluster result and *k* the number of clusters, which equals to 2. The purity is calculated as the summed number of highest cluster labels in each cluster divided by the total number of reviews in dataset Guerrini *et al*. (2007):

$$P\left(C_r\right)=\frac{1}{n_r}\max_i n_i^r \tag{3}$$

$$Purity\left(C\right)=\frac{1}{N}\sum_{r=1}^{k}P\left(C_r\right) \tag{4}$$

Accuracy

To calculate the Accuracy and the Recall, we must build the contingency, which contains four terms:

- TP: The number of data pairs found in the same cluster, both in our cluster result (C) and in the class label (L)
- FP: The number of data pairs found in the same cluster in C but in different cluster in L
- FN: The number of data pairs found in different clusters in C but in the same cluster in L
- TN: The number of data pairs found in different clusters, both in C and in L

Therefore, the accuracy can be calculated as follows:

$$Accuracy=\frac{TP}{TP+FP} \tag{5}$$

Recall

The Recall can be calculated as follows:

$$Recall=\frac{TP}{TP+FP} \tag{6}$$

Result and Discussion

A partial test set containing 297 reviews for each cluster label was used to achieve a balance evaluation between the two clusters (positive and negative). Figure 3 and 4 show the results of k-means algorithm using the cosine distance and the Euclidean distance respectively. As we can see from these results that the best results in term of recall and accuracy were on our dataset applied to it all preprocessing steps. Whereas, the recall measure of the original dataset and with only normalization dataset was the lowest. Moreover, the result of k-means on the version of dataset when we removed emojis, stopword and applied stemming with Euclidean distance, was below 0.6 using the recall measure.

We run the Hierarchical algorithm using four types of linkage and two types of similarity measures as shown in Table 2. Using ward linkage achieved the best result with Euclidean distance equivalent to 0.56. In the hierarchical algorithm results, shown in Fig. 5, using the recall and accuracy measures achieved well performance with Euclidean distance. On the contrary, the results of hierarchical algorithm with cosine distance shown in Fig. 6.

The recall results with four versions of the dataset of k-means algorithm and hierarchical algorithm as shown in Table 3. The algorithm gave the best recall when applied to the dataset with all the preprocessing steps.

Finally, the k-means algorithm obtains good performance with cosine distance. The second-best performance was obtained when we applied removal emojis, stop word and stemming without normalization

DOI: 10.3844/jcssp.2020.1258.1267

and the best performance when we applied all preprocessing steps.

Table 4 compares our method to a baseline of exiting work. It sheds light on the used datasets, preprocessing steps, algorithms and overall results. It is worth noting that the Arabic language is ambiguous, thus the preprocessing steps are recommended to act on the reduction of the ambiguity.

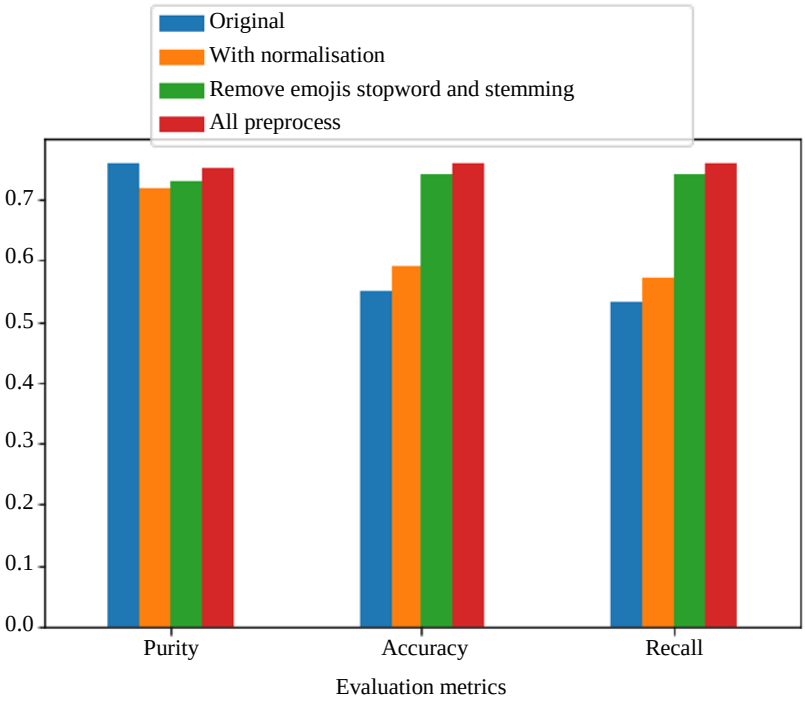

Fig. 3: k-means algorithm with cosine distance

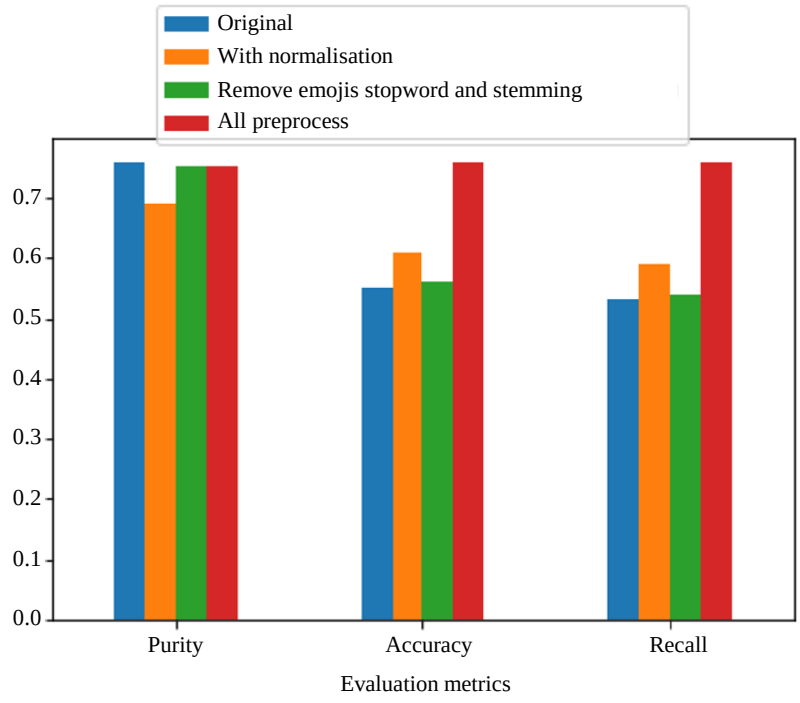

Fig. 4: k-means algorithm with Euclidean distance

DOI: 10.3844/jcssp.2020.1258.1267

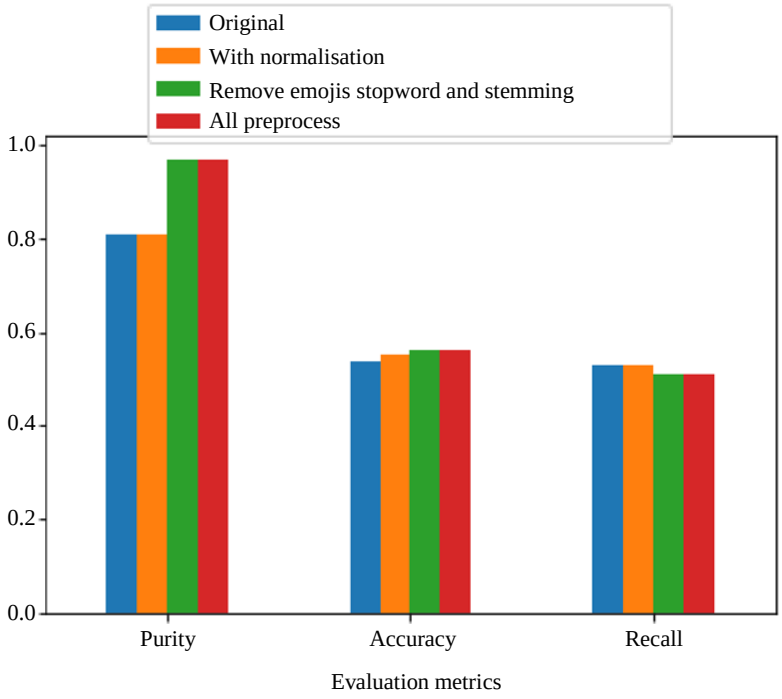

Fig. 5: Hierarchical algorithm with Euclidean distance

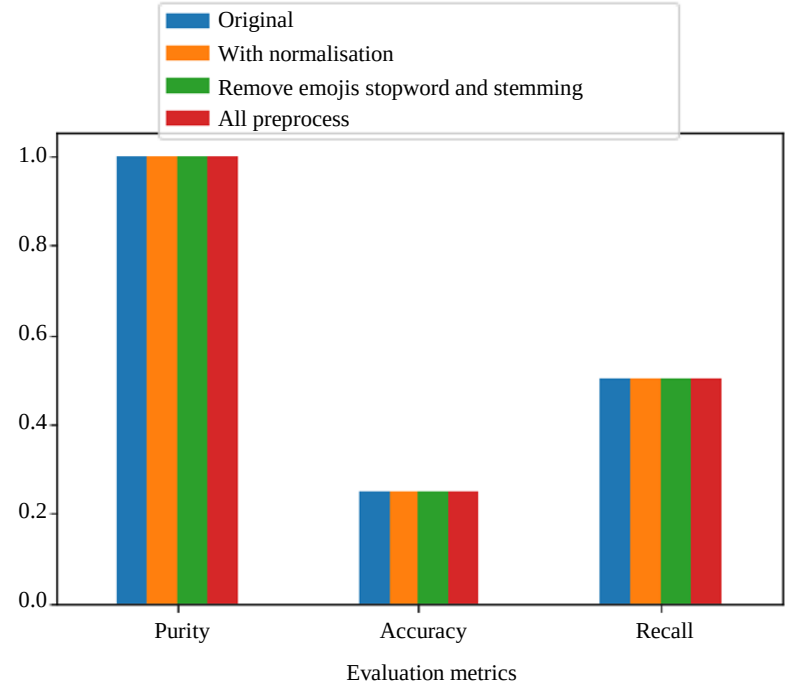

Fig. 6: Hierarchical algorithm with cosine distance

Table 2: Hierarchical algorithm accuracy results with different linkage using TF-IDF matrix

	Ward	Single	Complete	Average
Euclidean distance	0.56	0.25	0.47	0.25
Cosine similarity	0.25	0.25	0.25	

Table 3: Recall clustering results with cosine distance using TF-IDF matrix

Dataset	Version k-means	Hierarchical clustering
Original	0.53	0.50
With normalisation	0.57	0.50
Remove emojis, stopword and stemming	0.74	0.50
All preprocess	0.76	0.50

Table 4: Comparison of Abuaiadah *et al*. (2017) work and our work

	Dataset	Preprocessing	Learning	Rustle
Abuaiadah *et al*. (2017)	They used a freely labeled dataset containing 2000 Arabic tweets.	They applied clustering on four versions: Raw, NoSW, Root and light10.	They use the standard K-means algorithm with five similarity functions: Cosine, Pearson, Jaccard, Euclidean and KLD.	The Root version gives better results with KLD and Jaccard that purity equals 0.69.
Our work	We collected a dataset containing 4604 hotel reviews and labeled 0.25 from that dataset into positive and negative.	We have four versions: (1) original, (2) with normalization, (3) With remove emojis and stop word, stemming without normalization and (4) With normalization, remove emojis and stop words, stemming.	We applied k-means and Hierarchical clustering with cosine distance measures and Euclidean.	The k-means algorithm obtains good performance that purity equals 0.75 in the fourth version.

Conclusion and Future Work

In this study, we used unsupervised machine learning to detect and cluster sentiments in Saudi hotels reviews. To this end, we have collected about 4604 reviews. We then manually labeled 1382 reviews. Also, we used clustering, features and preprocessing strategies to find the best models to predict the sentiment label. The results showed that k-means clustering achieved the best accuracy.

Author's Contributions

All authors have equally contributed to the final version of the manuscript.

Ethics

This article is original and contains unpublished material.

References

Abd-Elhamid, L., Elzanfaly, D., & Eldin, A. S. (2016, December). Feature-based sentiment analysis in online Arabic reviews. In 2016 11th International Conference on Computer Engineering & Systems (ICCES) (pp. 260-265). IEEE.

Abuaiadah, D., Rajendran, D., & Jarrar, M. (2017, October). Clustering Arabic tweets for sentiment analysis. In 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA) (pp. 449-456). IEEE.

Al-Ayyoub, M., Khamaiseh, A. A., Jararweh, Y., & Al-Kabi, M. N. (2019). A comprehensive survey of arabic sentiment analysis. Information processing & management, 56(2), 320-342.

Al-Hadhrami, S., Al-Fassam, N., & Benhidour, H. (2019, May). Sentiment Analysis Of English Tweets: A Comparative Study of Supervised and Unsupervised Approaches. In 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS) (pp. 1-5). IEEE.

Al-Smadi, M., Al-Ayyoub, M., Jararweh, Y., & Qawasmeh, O. (2019). Enhancing aspect-based sentiment analysis of Arabic hotels' reviews using morphological, syntactic and semantic features. Information Processing & Management, 56(2), 308-319.

Al-Smadi, M., Qawasmeh, O., Al-Ayyoub, M., Jararweh, Y., & Gupta, B. (2018). Deep recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews. Journal of computational science, 27, 386-393.

Ben-Hur, A., Horn, D., Siegelmann, H. T., & Vapnik, V. (2001). Support vector clustering. Journal of machine learning research, 2(Dec), 125-137.

Boudad, N., Faizi, R., Thami, R. O. H., & Chiheb, R. (2018). Sentiment analysis in Arabic: A review of the literature. Ain Shams Engineering Journal, 9(4), 2479-2490.

El-Halees, A. M. (2014). Mining changes of opinions expressed by students to improve course evaluation. Mining Changes of Opinions Expressed by Students to Improve Course Evaluation.

Elhawary, M., & Elfeky, M. (2010, December). Mining Arabic business reviews. In 2010 IEEE international conference on data mining workshops (pp. 1108-1113). IEEE.

Elnagar, A., Lulu, L., & Einea, O. (2018). An annotated huge dataset for standard and colloquial arabic reviews for subjective sentiment analysis. Procedia computer science, 142, 182-189.

ElSahar, H., & El-Beltagy, S. R. (2015). Building large arabic multi-domain resources for sentiment analysis. In Gelbukh, A., editor, Computational Linguistics and Intelligent Text Processing, pages 23–34, Cham. Springer International Publishing.

Gamal, D., Alfonse, M., El-Horbaty, E. S. M., & Salem, A. B. M. (2019). Implementation of Machine Learning Algorithms in Arabic Sentiment Analysis Using N-Gram Features. Procedia Computer Science, 154, 332-340.

Guerrini, G., Mesiti, M., & Sanz, I. (2007). An Overviewof Similarity Measures for Clustering XML Documents. In Web Data Management Practices: Emerging Techniques and Technologies (pp. 56-78). IGI Global.

Heikal, M., Torki, M., & El-Makky, N. (2018). Sentiment analysis of Arabic Tweets using deep learning. Procedia Computer Science, 142, 114-122.

Hu, X., Tang, J., Gao, H., & Liu, H. (2013, May). Unsupervised sentiment analysis with emotional signals. In Proceedings of the 22nd international conference on World Wide Web (pp. 607-618).

Huang, A. (2008, April). Similarity measures for text document clustering. In Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand (Vol. 4, pp. 9-56).

Ismail, R., Omer, M., Tabir, M., Mahadi, N., & Amin, I. (2018, August). Sentiment analysis for arabic dialect using supervised learning. In 2018 International Conference on Computer, Control, Electrical and Electronics Engineering (ICCCEEE) (pp. 1-6). IEEE.

Kaur, V. D. (2018). sentimental analysis of Book Reviews using Unsupervised Semantic Orientation and Supervised Machine Learning Approaches. In Second International Conference on Green Computing and Internet of Things (ICGCIoT).

Kwaik, K. A., Chatzikyriakidis, S., Dobnik, S., Saad, M., & Johansson, R. (2020, May). An Arabic Tweets Sentiment Analysis Dataset (ATSAD) using Distant Supervision and Self Training. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection (pp. 1-8).

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

Mataoui, M. H., Hacine, T. E. B., Tellache, I., Bakhtouchi, A., & Zelmati, O. (2018, April). A new syntax-based aspect detection approach for sentiment analysis in Arabic reviews. In 2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP) (pp. 1-6). IEEE.

Palacio-Niño, J. O., & Berzal, F. (2019). Evaluation metrics for unsupervised learning algorithms. arXiv preprint arXiv:1905.05667.

Patel, S., Sihmar, S., & Jatain, A. (2015, March). A study of hierarchical clustering algorithms. In 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 537-541). IEEE.

Refaee, E., & Rieser, V. (2014, May). An arabic twitter corpus for subjectivity and sentiment analysis. In LREC (pp. 2268-2273).

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5), 513-523.

Sayed, A. A., Elgeldawi, E., Zaki, A. M., & Galal, A. R. (2020, February). Sentiment Analysis for Arabic Reviews using Machine Learning Classification Algorithms. In 2020 International Conference on Innovative Trends in Communication and Computer Engineering (ITCE) (pp. 56-63). IEEE.

Zhang, X., & Yu, Q. (2017, September). Hotel reviews sentiment analysis based on word vector clustering. In 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA) (pp. 260-264). IEEE.