Sentiment Analysis on Amazon Product Reviews using the Recurrent Neural Network (RNN)

Roobaea Alroobaea
Department of Computer Science
College of Computers and Information Technology
Taif University, P. O. Box 11099, Taif 21944, Saudi Arabia

***Abstract*—In this paper, the problem of sentiment analysis on Amazon products is tackled. In fact, sentiment analysis systems are applied in all business and social fields. This is because the opinions are at the center of all human activities, and they are key influencers of our behaviors. In this study, the recurrent neural network (RNN) model is used to classify the reviews. Three Amazon review datasets were applied to predict the sentiments of the authors. In conclusion, our results are comparable to the best state of the art models with an accuracy of 85%, 70% and 70% for three datasets.**

***Keywords*—Sentiment analysis; natural language processing; deep learning; RNN**

I. Introduction

Sentiment analysis is the area of study that investigates people's opinions and feelings, attitudes, and emotions. It is one of the most active research fields in natural language processing. It is widely studied in data mining filed. Therefore, this research has been extended outside science to cover management and social sciences. The increasing importance of the sentiment analysis overlaps with the evolution of social media, such as chat rooms, blogs, micro-blogs, Twitter [1]. For the first time in human history, there are huge amount of opinions in a digital form for analysis. The sentiment analysis systems are applied in all the business and social fields because of that the opinions are at the center of all human activities. Thus, they are significant influencers of our behaviors. Our beliefs and perceptions of reality as well as the selections we do are conditioned by how others see and value the world? For this reason, when a decision should be made, the opinions of others are often sought. This is true not only for individuals, but also for organizations. On Amazon, Internet users give their opinions on products. However, these reviews vary from a product to another and are used to improve and alert companies that have negative reviews of their products [1]. The aim of this paper is to use the recurrent neural network (RNN) model to classify the reviews of three Amazon review datasets[1] to predict the sentiments of the authors. The rest of the paper is organized as follows. Section two presents a definition of the sentiment analysis. Section three deals with the state-of-the-art. Section four introduces the methodology. Section five gives the results and experiments performed on the Amazon datasets. Finally, the last section concludes the paper.

[1] https://data.world/datafiniti/consumer-reviews-of-amazon-products

II. Contributions

The contributions in this study can be summarized as; 1) Arabic data set are used in this research because few research have been done in this area [12]. 2) We present a method using RNN to predict sentiment of authors. 3) We compare our proposed model to Long short-term memory (LSTM), Gated Recurrent Unit (GRU), and Convolutional Neural Network (CNN).

III. Literature Review

A. Sentiment Analysis

Sentiment analysis, also called the opinion mining, is a sub-field of computer science besides. It is considered as a part of an automatic natural language processing and aims to classify feelings expressed in texts. There are other related names slightly referring to different tasks, such as opinion mining, sentiment mining, subjectivity analysis, effect analysis, and emotion analysis [2].

Furthermore, sentiment analysis is made using one of two fundamental methods, which are rule-based classifiers and machine learning classifiers. The former are rules derived from the linguistic study of a language that are utilized to sentiment analysis. The latter is statistical machine learning algorithms that are used to automatically learn sentiment signals [1][2]. Therefore, there are two terms that should be defined which are sentiment and opinion. In the literature review, there is confusion between these two words. In the Oxford Dictionary, feeling is defined as a point of view or opinion that is held or expressed as an emotion. For the word opinion, it is a belief, or a judgment formed about something, which is not necessarily based on facts or knowledge but on the beliefs or opinions of a group or of the majority of people. Consequently, it can be said that the term of feeling is a person's emotion about something, while the term of opinion represents or shows a person's point of view. According to [3], sentiment analysis, also well-known as opinion mining, is the field of study that analyzes people's opinions, feelings, evaluations, attitudes, and emotions. There are two methods that are used to automatically learn sentiment signals which are supervised and unsupervised learning methods. The former involves the presence of two sets of data which are a training set and a test set. The method is called supervised since the system is trained on a training sub-set, which contains models that have already been processed. The latter (unsupervised), which recommends only one data set, requires the system to autonomously restructure the

information so that the most similar data will be placed in the same group [4].

Furthermore, there are three analytical approaches which are lexicon-based approach, corpus-based approach, and a Hybrid approach (mixed approach) [4]. Regarding the Lexicon-based approach, it identifies the polarity of a text using two sets of words, the first expresses a positive feeling and the second expresses a negative feeling. Then, the model counts in the text the number of positive words and that of negative words while the sum gives an overall evaluation of the text feeling. Therefore, if the number of positive words outweighs that of the negative ones, the text is considered positive, conversely, the text is considered negative or neutral if the numbers are equal [4].

In terms of Corpus-based approach, the automatic sentiment analysis based on corpora requires the creation of two manually annotated corpora. The first is the learning corpus, which is used to train an automatic system. It includes notes added by human annotators. Thus, a system can perform an analysis on its own while the second uses the test corpus, which is trained to verify the performance of the automatic system. In an ideal scenario, the results of the analysis performed by the automatic system should fully correspond to those of the learning corpus. Therefore, in order to maximize the automatic system's performance, it is important that the learning corpus is representative for the test corpus. [4] gives an example of algorithm like neural network.

In addition, Hybrid approach takes advantage of the previous two methods there are three ways to do it. The first is to use linguistic tools to develop the corpus and then classify the texts using a supervised learning tool. The second way is to use machine learning to build the body of opinion needed for the lexicon-based approach. The third way is the combination of the two previous approaches and the collection of their results [5]. Furthermore, the next section will show the related works on the sentiment analysis.

B. Related Work on Sentiment Analysis

Although the field of sentiment analysis is an emerging field in the natural language processing community, the work done for the Arabic language is still extremely limited [6]. Indeed, most of the research studies have focused on the polarity classification of documents to avoid the excessive cost of sentence annotation. Also, to adopt the machine learning-based approach to escape the excessive cost of creating the lexicon of opinion with good coverage [7]. Thus, the obstacle that slows the advancement of the field of the sentiment analysis for Arabic is the unavailability of resources in terms of annotated corpora and lexicons of opinion.

In this context, [6] are among the first researchers who were interested in constructing an opinion lexicon for the Arabic language. In fact, their work was part of the development of a tool that enables to classify the stakeholders' opinions in the business field. Therefore, the starting number of words to be set were more than 600 positive words, 900 negative words and 100 neutral words. The evaluation tests performed showed good accuracy but a low recall. In addition, [7] proposed the combination approach based on three steps for the classification of Arabic documents according to their polarities. They used an opinion lexicon constructed from the English Lexicon SentiStrength after having done the translation and consulted online dictionaries to enrich the lexicon with the synonyms of the already existing words. However, the size of the lexicon was not mentioned and the accuracy of the classification tool for the lexicon-based step is 48.7%. Also, [8] presented the lexicon "Sifaat", a manually constructed lexicon containing 3325 adjectives categorized in three classes: positive, negative, and neutral. The results of the assessment showed an improvement of 6% in the classification of subjectivity and 40% in the classification of polarity. The authors also proposed to extend the lexicon by translating three English opinion lexicons, namely the SentiWordNet, the Youtube lexicon and the General Inquirer. However, this method focused only on adjectives that have not been evaluated. Furthermore, [9] proposed to create a lexicon by translating 300 words of SentiStrength. Then, enriching it with synonyms and emoticons. The last version of the lexicon included 3479 entries 1262 of which are positive and 2217 are negative. The results of the experiment on the collected corpus reached 59.6% in terms of accuracy. As for [10], they proposed to build an Arabic version of SentiWordNet, which is an opinion lexicon derived from the WordNet database, by going through two stages: updating the base of the WordNet Arabic 2.0 by mapping the WordNet 3.0 for English, and also the base obtained from the SentiWordNet 3.0 for English. In fact, the lexicon coverage assessment reports that 5% of the words in the annotated corpus are not in the lexicon.

In this study the Arabic data set and RNN model will be used to predict sentiments of authors. Next sections will explain the methods uses and demonstrate that our results surpass those obtained by the best deep learning models like LSTM, CNN and GRU.

IV. METHODOLOGY

In 2005, things started to change. In fact, the outlook on the field of artificial intelligence has dramatically changed with the machine learning while the emergence of deep learning, which accounts for the bulk of research conducted by specialists, especially as it intervenes in many areas, such as natural language processing. Deep learning (DL) is a branch of machine learning, where the latter is a branch of artificial intelligence, in which machines can learn by knowledge and gain skills without the human contribution. Therefore, based on artificial neural networks, the algorithms encouraged by the human brain, learn from large amounts of data. Consequently, DL enables models composed of multiple processing layers to understand descriptions of data with multiple levels of abstraction. [11].

A. Deep Learning vs. Machine Learning

Machine learning (ML) is a sub-field of artificial intelligence (AI), which focuses on designing systems that learn or improve performance based on the data. These systems are intended to train a safe set of algorithms for larger amounts of data that can help classify future data. ML is divided into Supervised Machine Learning and Unsupervised Machine Learning. The former is the most generic form of machine learning. It aims to make the algorithm "learn" by comparing

its actual output to the "taught" outputs in order to check for errors and modify the model accordingly. The latter is also known as learning from observations. The learning algorithm in this method finds common points on its own among its input data. On the other hand, in a machine learning process, the algorithm must be instructed on how to make an accurate prediction using more information (for example, by manually performing an extraction of the relevant features) [7]. In fact, Deep Learning (DL) algorithms can learn to make an accurate prediction through their own data processing, thanks to the structure of the artificial neural network. For example, they ignore manual steps and feature extraction, but the modeling process is automatically performed. Another major difference is the fact that deep learning algorithms evolve with the data. Therefore, to succeed in the deep learning application, a large volume of data and large dimension are needed to train the model by adding one or more GPUs (graphics processor) to quickly process the data. This implies that, if these elements are not needed, it is better to use machine learning rather than deep learning.

Indeed, feature extraction is one of the greatest challenges of traditional machine learning models, which has been automated by deep learning models to enable them to achieve a particularly high accuracy rate for computer vision tasks. Therefore, the ability to manage a large number of features makes deep learning powerful when it comes to unstructured data [13]. Nevertheless, DL algorithms can be overkilled for less complex problems because they need access to a large amount of data to be effective. On the other hand, the deep learning algorithms success depends on the availability of more training data. In fact, Google, Facebook, and Amazon have already started using it to analyze their massive amounts of data. In practice, all deep learning algorithms are neural networks, also called ANN. ANNs are information processing models that simulate the functioning of a biological nervous system. They are similar to the way the brain manipulates information at the functional level. Actually, all the neural networks are made up of interconnected neurons that are organized in layers. Then, the details of the neuron and the description of its functioning will be explained in the next part. In fact, what forms the neural networks are the artificial neurons inspired by the real neuron that exists in our brain [13].

On the other hand, the activation function is an essential component of the neural network. This function consists in checking whether the neuron is activated or not by calculating the weighted sum of the inputs and adding the bias. This is a nonlinear transformation of the input value. Moreover, the nonlinearity is particularly important in neural networks however, without the activation function, a neural network simply becomes a linear model [14]. Actually, there are many types of these functions, such as sigmoid function, ReLu function, and Softmax function. The distinct types of neural networks available for use, such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) [14] [15] [16].

B. Steps of our Approach

Deep Learning has been proven to be effective in many complex problems using artificial neural networks to learn and extract patterns and information significant from the data [15]. Therefore, we find many contributions that attempt to adapt this approach as a solution to the problem of sentiment analysis. In this research, the proposed model (RNN) will be used. Also, other types of neural networks will be applied as comparison experiments to the proposed model. Now we are going to present the tasks necessary to conduct this work as showed in Fig. 1.

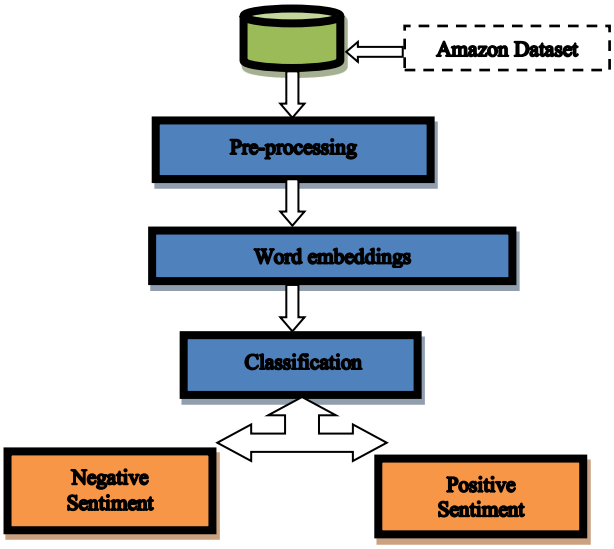

Fig. 1. Architecture of our Method.

- Pre-processing is preparing the dataset to have an excellent quality. This is important tasks that must take place before using a dataset for model training. In order for the models to be professionally trained and to provide the expected results, the data used must be representative, clean, precise, complete, and well labeled. For example, for the prediction the feelings of comments from social networks, the corpus must contain the same type of documents. The preparation of these data is therefore a crucial step [4].

- Word embeddings are the mapping of words to number vectors real in a reduced dimensional space. Word embedding vectors represent words and their contexts; thus, the words with similar meanings (synonyms) or with close semantic relations will have more similar embeddings. In addition, word embeddings should reflect how the words are related to each other. For example, the inscriptions for "man" should be "King" like "woman" is "queen". Since learning word embeddings takes time and the power of calculation, the method started with pre-trained incorporations.

- In the classification step, we apply the RNN model in order to predict the sentiment of the author. We have two output classes: positive sentiment (Happy, in love, smile, etc.) and negative (afraid, not happy, angry).

V. Result and Discussion on Amazon Datasets

For this system to function, the method has to go through the following experiments. The proposed method is validated in two steps, in the first amazon dataset, the corpus is divided into 4,149 test comments, which (which represents 20% of the

dataset) to assess this model and have its precision, and 16,593 training comments, and obtain as precision 75% for the training and 70% for the test. Fig. 2 shows the results.

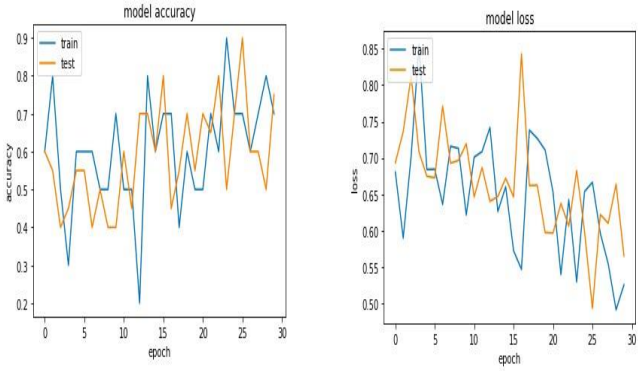

Fig. 2. Precision and Error of this Model for the First Test.

In the second experiment, the data is trained using binary sentiment classification model on a dataset 2 of 66666 items, 33333 positives and 33333 negatives, as shown in the Fig. 4. The best accuracy is obtained with 25 epochs for training and also for test. For the loss, the minimum value is also obtained with 25 epochs for test and 27 for training. Additionally, the corpus is divided into 6667 test comments (which represent 20% of the dataset) to test the model and have its accuracy, and 59999 training comments. Then, we obtained 90% as precision for training and 85% for the test. Fig. 3 shows the results plotted in graphs. For the loss, the best value is obtained with 27 epochs.

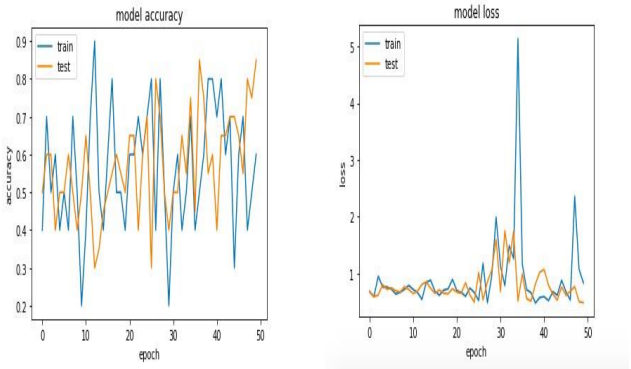

Fig. 3. Precision and Error of the Model for the Second Test.

In the third experiment, we trained our binary sentiment classification model on a dataset 3 of 49,870 items, more than 24,900 positives and 24,900 negatives. Moreover, the corpus is divided into 9973 test comments (which represent 20% of the dataset) to evaluate our model and have its accuracy, then 39897 training comments, where results are 90% for training and 70% for accuracy. For the loss, the best results are obtained with 17 epochs for training and 30 epochs for test. Fig. 4 shows the results plotted in the graphs.

In the first validation, the model was applied to the three data sets mentioned above and the evaluation was made in terms of the accuracy and recall metrics, as summarized in the Table I. Accuracy is the number of relevant documents found compared to the total number of documents proposed for a given query. The recall is defined by the number of relevant documents found with regard to the number of relevant documents that the database has [17].

Our decision to choose different datasets based on their size was to see how a DL model works in two situations where we have a large and small amount of data. For the final tests on the datasets, we trained the 30-epoch model. From what we have as results, the effectiveness of the proposed model and the possibility of its use in this area. For second validation, three models were applied on the same data sets as shown in Table II. Our results obtained using RNN are compared with three most known deep learning Models. In fact, our results surpass those obtained by CNN. For LSTM and GRU we did not obtain encouraging result.

In the third validation, as shown in Table III, we compared our results on the dataset amazon reviews2 with Palash experiments. It can be noted that Palash participated on the kaggle3 competition on Amazons' reviews datasets. Palash presents a method using bidirectional LSTM, for the document representation he based on n-grams (a chunk from 2 to 6 grams) to classify sentiments (positive, negative). The best result for his experimentation is obtained using 5-grams. In Comparison with the last study, we obtained the best result for the accuracy with 96%, but we did not surpass those of Palash for the recall.

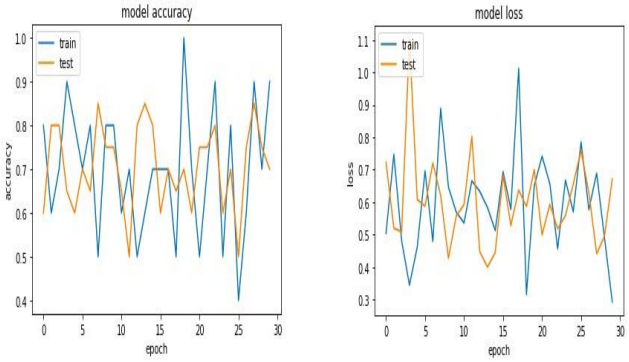

Fig. 4. Accuracy and Loss of our Model for the Third Test.

TABLE I. RESULT ON THREE AMAZON DATASETS

Amazon data	Training	Test	Accuracy
20742	16593	4149	70%
66666	59999	6667	85%
49870	39897	9973	70%

TABLE II. COMPARISON BETWEEN 4 DEEP LEARNING MODELS IN TERMS OF ACCURACY

Deep Learning Model	LSTM	GRU	RNN	CNN
DATA1	76	80	85%	83
DATA2	68	65	70%	70
DATA3	68	64	70%	68

[2] https://jmcauley.ucsd.edu/data/amazon/
[3] www.kaggle.com

TABLE III. COMPARISON RESULTS WITH PALASH4 STUDY USING BIDIRECTIONAL LSTM AMAZON REVIEWS DATASET CONTAINING 400000 REVIEWS

	Bidirectional LSTM (Palash)	**Our Model**
Accuracy	94%	96%
Recall	90%	85%

VI. DISCUSSION

Experiments have shown that our system is efficient compared to other models (GRU, LSTM, etc.). Indeed, for the three datasets our system has obtained the best precision. Nevertheless, for recall we found a worse result than LSTM. Indeed, Lstm used by Spalash gave 90% against our model 85% for the Amazons corpora, our results were less good for data–one and three because the size is minimal, while the results for corpus two were encouraging since we had 66,000 documents.

VII. CONCLUSION

In this paper, we aimed to contribute to the sentiment analysis field by presenting the tools and Arabic datasets, as well as the steps we have followed to obtain the appropriate results for our model. In fact, our sentiment classification model helped us to correctly categorize more than 85% of all the 7480 test items, which means that our system has correctly categorized more than 5984 items. Moreover, the experiments proved that as the size of the training data increases, the accuracy rate increases, which implies that the percentage of the test dataset is as important as it was in the first experiments. Therefore, our model performs well if we use it for the sentiment analysis problem, where the amount of learning data plays a significant role. The proposed system succeeded in obtaining an accuracy of 85%. As Future work, we will depend on multilingual corpora to predict sentiments. Also, we will use reviews about Covid-19 in KSA twitter to predict emotions of people in period of virus vaccination.

REFERENCES

[1] Abbasi, H. Chen, A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums," ACM Transactions on Information Systems. 008 Jun 20;26(3):1-34.

[2] B. Liu, "Sentiment analysis and subjectivity," In: Handbook of Natural Language Processing, 12. 2010 Feb;2(2010):627-66.

[3] J. Waters, J. Lester, "The Everything Guide to Social Media: All you need to know about participating in today's most popular online communities". Simon and Schuster; 2010 Oct 18.

[4] V. Hee Cynthia," L'analyse des sentiments appliquée sur des tweets politiques : une étude de corpus", Faculté associée de linguistique appliquée Université Bruxelles Belgique, 2013.

[5] D. Poirier, F. Fessant, C. D.e. Bothorel, E.G, Neef, M. Boullé, "Approches statistique et linguistique pour la classification de textes d'opinion portant sur les films". Revue des Nouvelles Technologies de l'Information. 2009 Nov 1: Pages-147.

[6] M. Elhawary and M. Elfeky. "Mining Arabic Business Reviews". In Proceedings of International Conference on Data Mining Workshops (ICDMW), pages 1108–1113. IEEE, 2010.

[7] A. El-Halees. "Arabic Opinion Mining Using Combined Classification Approach". In Proceedings of the International Arab Conference on Information Technology (ACIT), 2011.

[8] M. Abdul-Mageed and M. Diab. "Toward building a large-scale Arabic sentiment lexicon". In Proceedings of the 6th International Global WordNet Conference, Matsue, Japan, 2012.

[9] A. Nawaf, , N. Abdulla, A. Ahmed, A. Mohammed and M. Al-Ayyoub: "Arabic Sentiment Analysis: Lexicon-based and Corpus-based", 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), 2013.

[10] S. Alhazmi, W. Black, and J. McNaught, "Arabic SentiWordNet in Relation to SentiWordNet 3.0", International Journal of Computational Linguistics. 2013 et 4(1):1-11.

[11] Patterson J, Gibson A. Deep learning: A practitioner's approach. " O'Reilly Media, Inc."; 2017 Jul 28.

[12] H. Alamro, M. Alshehri, B. Alharbi, Z. Khayyat, M. Kalkatawi, II. Jaber, X. Zhang , "OVERVIEW OF THE ARABIC SENTIMENT ANALYSIS 2021 COMPETITION AT KAUST", 2021.

[13] I. Vasilev, D. Slater, G. Spacagna, P. Roelants, V. Zocca, " Python Deep Learning: Exploring deep learning techniques and neural network architectures with Pytorch" , Keras, and TensorFlow. Packt Publishing Ltd; 2019 Jan 16.

[14] I. Goodfellow, Y. Bengio, A. Courville "Convolutional networks". InDeep learning 2016 Nov (Vol. 2016, pp. 330-372). Cambridge, MA, USA: MIT Press.

[15] N. Buduma N. Locascio," Fundamentals of deep learning: Designing next-generation machine intelligence algorithms. " O'Reilly Media, Inc."; 2017 May 25.

[16] A. Almulihi, F. Alharithi, S. Mechti, R. Alroobaea, S. Rubaiee, "A Software for Thorax Images Analysis Based on Deep Learning". International Journal of Open Source Software and Processes (IJOSSP). 2021 Jan 1;12(1):60-71.

[17] D. Kumar, U. Bansal, R. Alroobaea, A. Baqasah, M. Hedabou. An Artificial Intelligence Approach for Expurgating Edible and Non-Edible Items. Frontiers in Public Health. 2021;9.

[4]Graduate Researcher at Pukyong National University Busan, Busan, South Korea