Phrase-Chunk Level Hierarchical Attention Networks for Arabic Sentiment Analysis

Abdelmawgoud M. Meabed [1†], Sherif Mahdy Abdou[2††] and Mervat Hassan Gheith [3†††]

[†] Faculty of Statistical Studies and Research Cairo University, Cairo, Egypt
[††] Information Technology Department, Faculty of Computers and Information, Cairo University, Cairo, Egypt
[†††] Faculty of Statistical Studies and Research Cairo University, Cairo, Egypt

Abstract
In this work, we have presented ATSA, a hierarchical attention deep learning model for Arabic sentiment analysis. ATSA was proposed by addressing several challenges and limitations that arise when applying the classical models to perform opinion mining in Arabic. Arabic-specific challenges including the morphological complexity and language sparsity were addressed by modeling semantic composition at the Arabic morphological analysis after performing tokenization. ATSA proposed to perform phrase-chunks sentiment embedding to provide a broader set of features that cover syntactic, semantic, and sentiment information. We used phrase structure parser to generate syntactic parse trees that are used as a reference for ATSA. This allowed modeling semantic and sentiment composition following the natural order in which words and phrase-chunks are combined in a sentence. The proposed model was evaluated on three Arabic corpora that correspond to different genres (newswire, online comments, and tweets) and different writing styles (MSA and dialectal Arabic). Experiments showed that each of the proposed contributions in ATSA was able to achieve significant improvement. The combination of all contributions, which makes up for the complete ATSA model, was able to improve the classification accuracy by 3% and 2% on Tweets and Hotel reviews datasets, respectively, compared to the existing models.

Keywords:
Artificial neural networks, Computational linguistics, Machine learning, Natural language processing, Natural languages, Neural networks, Sentiment analysis.

1. Introduction

Social media and bloggers are the most popular sources for sharing and gathering feedback from all different types and levels of users around the globe. Studying ideas and information behind the shared text or comments can monitors the behaviors, evaluation, attitudes, and emotions of such users about specific topic or product. Such type of study is called Sentiment Analysis (SA), also known as Opinion Mining. Sentiment Analysis aims to analyze people's sentiments, opinions, attitudes and emotions [1], towards aspects such as topics, products, individuals, organizations, and services. Many techniques and software tools are being developed to resolve Sentiment Analysis problems.

There are different terms with some different tasks, e.g., sentiment analysis, opinion extraction, opinion mining, subjectivity analysis, sentiment mining, emotion analysis, review mining, effect analysis, etc. However, all of them are commonly studied under the umbrella of opinion mining or sentiment analysis. The term sentiment analysis is commonly used in industries, while both “Sentiment Analysis” and “Opinion Mining” are frequently used in science as well. Despite, they represent the same field of study.

In morphology-rich languages like Arabic, where Arabic is one of the languages having the characteristics that from one root the derivational and inflectional systems can produce a significant number of words. In Arabic, like other Semitic languages, word surface forms may include affixes, concatenated to inflected stems. Detecting idioms or proverbs phrases within the written text or comment in one of these challenges[2]. These types of complexities make the language of Arabic one of the most complex area of research in Sentiment analysis. Moreover, the Arabic writer is used to write their opinions in a different language or different styles. The writers may use English or other language or may write Arabic text in Latin letters (transliteration). Such type of complexities makes the job of sentiment analysis more difficult and requires more considerations to cover such discrepancies of the types and forms of texts written in different sources of the information extracted from social media.

Sentiment Analysis (SA) and opinion mining focus on identifying and evaluating positive and negative views and comments. Opinions may be direct or indirect (explicit/implicit) and comparative. This study aims to determine the sentiment polarity for collected comments or posts from social media using a dataset of Arabic text comment words. Machine learning classification techniques will be used to perform the required classification to identify the polarity of the collected opinions [3]. The writers posted a significant number of reviews and comments on the web providing valuable information to other providers, reviewers, and consumers due to the rapid

Manuscript received September 5, 2023
Manuscript revised September 20, 2023
https://doi.org/10.22937/IJCSNS.2023.23.9.15

development and availability of the web and electronic word-of-mouth interaction.[4]

Sentiment Analysis (SA) provides significant opportunities to develop new applications, mainly due to the fast growth of the availability of the information in many different sources such as blogs, social networks. [1].

2. Background and Related Work

Machine learning techniques and methods have been widely utilized for studying sentiment analysis[5]. The Bag-of-words representation is commonly used for sentiment analysis. With the help of both the availability and ability of machine learning techniques to resolve such intensive data-related tasks, machine learning will assist the process of Sentiment Analysis tasks. In addition, with the aid of feature-selection techniques in machine learning, the machine learning algorithms can reduce the high-dimensional feature space which selects only important features by eliminating the noisy and irrelevant features.

In machine learning, neural networks are models consist of connected layers of computational units, it referred to as neurons [6], Figure 1. Connections to each neuron set to be the weight that is learned through training on example, dataset. Training the network aims to slightly update the network weights to enhance the accuracy of the output.

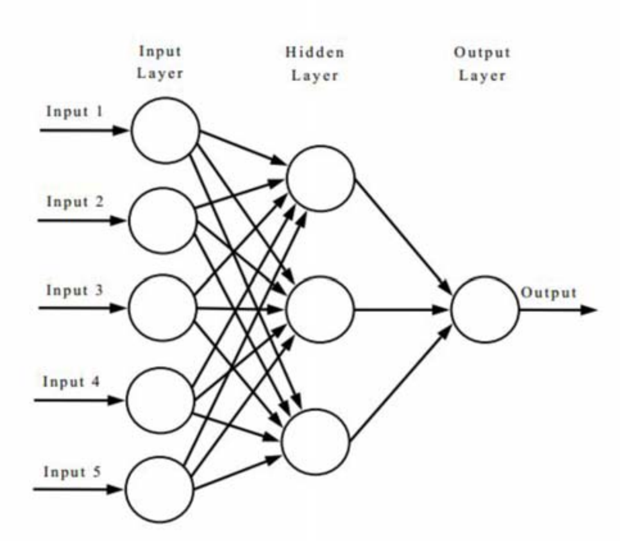

Fig. 1. A simple neural network architecture with five inputs.

The recurrent neural network is that a type of neural networks which connections constitute a directed. This allows neurons cells to store internal state or memory in the previous time step that affects the output of the network at a time step t. Recent improvements in neural networks and their proven performance in multiple applications. But Unfortunately, training neural networks remains to be a difficult problem. A major difficulty in training neural networks is the vanishing gradient problem which is difficulty found in training that causes updates to the network's weights to be very small and can slow down training and reduce performance[6].

With the aid of data science methods; Pre-processing the data can be made easily; where it is the process of data cleaning and preparation of the text for classification. Texts of online blogs contain usually lots of noise and uninformative parts. Using machine learning techniques, we can classify which is useful information along with other features can be extracted like user profile and date posted.

The Sentiment classification techniques can be obviously divided into machine learning approach, lexicon-based approach, and hybrid approach. The Machine Learning Approach (ML) applies the famous ML algorithms and uses linguistic features. The Lexicon-based Approach relies on a sentiment lexicon, a collection of known and precompiled sentiment terms. It is divided into a dictionary-based approach and corpus-based approach which use statistical or semantic methods to find sentiment polarity. The hybrid Approach combines both approaches and is very common with sentiment lexicons playing a key role in many methods. The various approaches and the most popular algorithms of SC are illustrated in Fig. 2 as mentioned before. The text classification methods using the ML approach can be roughly divided into supervised and unsupervised learning methods. The supervised methods make use of many labeled training documents.[7]

The lexicon-based approach depends on finding the opinion lexicon which is used to analyze the text. There are two methods in this approach. The dictionary-based approach which depends on finding opinion seed words, and then searches

the dictionary of their synonyms and antonyms. The corpus-based approach begins with a seed list of opinion words and then finds other opinion words in a large corpus to help in finding opinion words with context-specific orientations. This could be done by using statistical or semantic methods. There is a brief explanation of both approaches' algorithms and related articles in the next subsections.

3. Proposed Arabic Sentiment Analysis Model

In the previous section, related work and background of different approaches for sentiment analysis (SA) have been discussed, and several research limitations have been outlined. Most of the researches been done in this area are based on the contribution of a word in each document or sentence. Since each word or even each chunk in a specific sentence doesn't contribute equally in the overall sentiment polarity of a review. To overcome these limitations, and achieve the objectives of this research, a new framework has been proposed in this chapter. It aims to provide a precise sentiment analysis based on the contribution of each tweet's word.

Researchers are currently focusing on sentence-level sentiment analysis, based on the assumption that each sentence can only contain a single opinion. Actually, sentence-level sentiment analysis has no major difference with document-level analysis[8].

Currently, RNN (Recurrent Neural Network) is considered as a natural generalization of feedforward neural networks of sequences to given sentiment label. Given a series of inputs $(x_1, x_2, x_3, \dots\dots, x_T)$, a standard RNN computes the target label (y) by iterating the following equation:

$$h_t = sigm(w^{hx}x_t + w^{hh}h_{t-1}) \quad (1)$$

$$y = w^{yh}\, h_t \quad (2)$$

The RNN can easily map input data (sentence or tweet) to a single value known as a label or sentimental value of that sequence whenever the mapping between the inputs and the outputs is can be determined[9].

Since words or chunks do not contribute equally in the same sentence where some of them can strongly affect the final sentiment polarity of such tweet, so we use hierarchical attention network concept to recursively traverse the generated parse tree of the sentence to learn the sentence vector based on the importance of each chunk. By using hierarchical attention network, we will be able to extract sequence and arrangement of words which impact the meaning of the tweet/review; similarly, we construct phrase chunks which affects the overall meaning and polarity of the tweet and the representation of those important phrase chunks to build a sentence vector.

3.1 Framework of The Proposed Model

In this section, we present the challenges related to applying the baseline RAE [10] opinion model to Arabic text, in addition to its limitations equally applicable to English. Then, we propose ATSA, which augments the RAE model with the necessary components to address these challenges and constraints.

The proposed framework consists of the six components: base phrase chunking, building TF-IDF matrix, Extract word vectors, calculate work averaging, Vector similarity detection and training classifier.

Figure 2 shows the proposed ATSA framework to address the challenges limitations mentioned in the related work section. We perform morphological tokenization of the input text to overcome the issue of morphological complexity and over-fitting. Then, we proposed a hierarchical neural network architecture to derive embeddings that capture word-level sentiment information[11]. We also propose an unsupervised pretraining block to improve the initialization of both semantic and sentiment embedding models. Finally, we use phrase structure parsers instead of the greedy algorithm to generate grammatically motivated parse trees that are used as a basis for AE recursion.

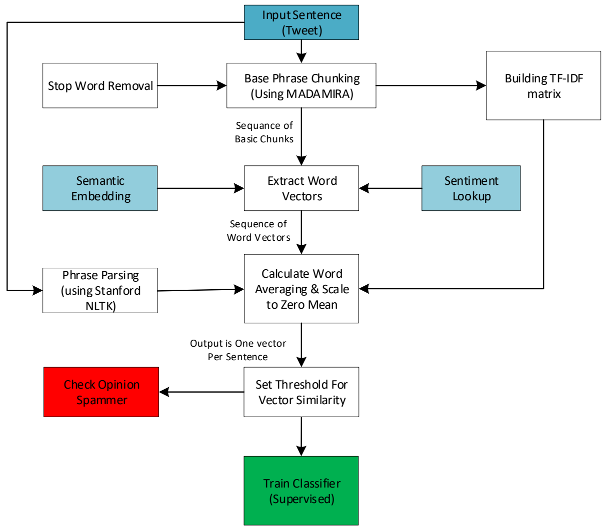

Fig. 2 The framework of the proposed ATSA.

3.2 Handling **Morphological** Features

(i) Word form

Due to morphological richness of Arabic which generated different forms of the same word, in addition to the language vocabulary itself. The resulted word embedding vocabulary can be very huge. To reduce the size of the dimension we utilized GPAM [12] to tokenize the text automatically. We experiment with two different issues:

- Extracting word roots: in the experiment, we analyzed Arabic text to get words' root only. Because the same word structure can be formed from different roots; so many words returned more than one root. In this case, we consider only the output of root if they are unique. Which mean, when the input word returned only one possible root, that root can be considered as a replacement of the original word. Otherwise, we ignore that root and process the word stem.
- Extracting word stems: due to different roots, can be extracted from the same word, we experiment the word stem in case more than one root generated for the input word. In this case, we consider the stem as a replacement for the original word if it is unique.

(ii) Pos Tagging

With POS tagging, base forms of words will be in use, thus the meaningful information about that word can be lost; Thus, we are doing comprehensive POS tagging analysis where we pick only the root of the word in case the morphological analysis determined only one root, in case multiple possible roots appears, we pick only the stem of word. Similarly, if the analysis results multiple possible stems, we pick the word after removing all affixes. Diab [13] used reduced tag set (RTS) and extended reduced tag sets (ERTS). Diab [14] work shows that using ERTS improves performance for higher processing tasks such as Base Phrase Chunking for MSA [15].

(iii) Stop Words

Stop words are those tokens with no important significance in several Natural Language Processing(NLP) tasks. Removing stop words from the proposed model can modify the final lexicon making it a good candidate for analysis.

Since most stop words do not have semantic meaning or sentiment associated, Removing them from the model may improve the lexicon quality.

(iv) Syntactic Parsing

At this stage, we use the Stanford Arabic parser to automatically generate syntactic parse trees, over which the model will be recursively trained. The parser is applied to input text that is morphologically tokenized according to MADAMIRA. Figure 3 shows the resulted parse tree of the tweet with text “جامد جدا و انا بجد بعشق حاجه اسمها مصر” in Arabic; “Very interesting, and I love what so-called Egypt.”

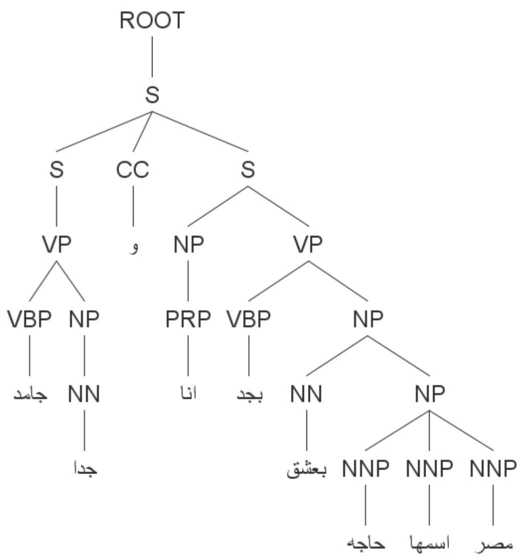

Fig. 3 Sample tweet parse tree using Stanford NLTK package

3.3 Hierarchical Attention Network for Phrase-Chunks

In this work, we consider each tweet as a sentence. A given twee can be combined of one or more chunks. Since words or chunks do not contribute equally in the same sentence where some of them can strongly affect the final sentiment polarity of such tweet, so we use hierarchical attention network concept to recursively traverse the generated parse tree of the sentence to learn the sentence vector based on the importance of each chunk. The Hierarchical Attention Network structure is shown in Fig 4 described by Yang et al. [16]. It consists of multi parts: a word-chunks sequence encoder, a chunk-level attention layer, a sentence encoder and sentence-level attention layer. A detailed description of different components in the following sections.

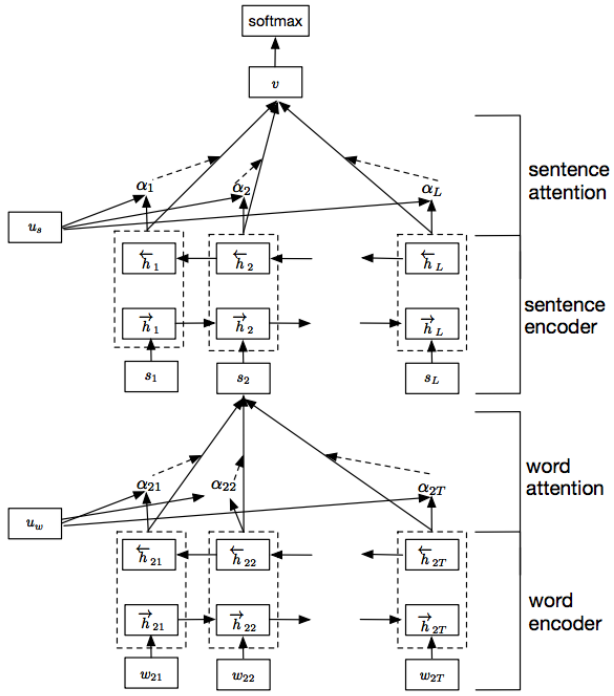

Fig. 4 Hierarchical Attention Network

3.4 GRU-Based Chunks as Sequence Encoding

Using two types of gates to track the state of sequences: the first one is a reset gate r_t and the other one is an update gate z_t. [17] These two together control how information is updated to the n network. state. At time t, the GRU computes the state as:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \qquad (3)$$

This function is a linearly interpolate previous state h_{t-1} and the current new state $\tilde{h}_t$ computed with new sequence information. The gate z_t decides how much past information is kept and how much new information is added. z_t is updated as:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{4}$$

where x_t is the sequence vector at time t. The candidate state $\tilde{h}_t$ is computed in a way like a traditional recurrent neural network (RNN):

$$\tilde{h}_t = tanh(W_z x_t + r_t \odot (U_h h_{t-1}) + b_h) \tag{5}$$

Here r_t is the reset gate which controls how much the past state contributes to the candidate state. If r_t is zero, then it forgets the previous state. The reset gate is updated as follows:

$$r_t = \sigma(W_z x_t + U_h h_{t-1} + b_r) \tag{6}$$

4. Experiments, Results and Discussion

In this section, we evaluated our dataset of manually annotated hotel reviews and tweeter separately against the most knows machine learning models. Then we compare ATA to several models. We focus on binary sentiment classification, where the tweet or review class can be either negative or positive.

4.1 Datasets and Experimental Setup

We downloaded over 3.2 million reviews and comments from Twitter, Facebook and the hotel booking website "www.booking.com." The hotel booking reviews collected from hotels located in Egypt, Saudi Arabic, Bahrain and the United Arab Emirates, also, we collected comments on persons from Facebook. The Facebook comments have been selected as comments on artist stars posts. While annotating the posts, we removed comments contains unwanted words. After filtering out the non-Arabic reviews and performing several pre-processing steps to clean up HTML tags and other unwanted content, we built a dataset of 100K Arabic review and comments.

The dataset contains 53,291 hotel reviews and 49934 comments on an event, an artist's post or tweet; Different users submitted that for different hotels in four countries and Comments for various star post. As shown in Fig 6; the total number of tokens on the dataset is 2,181,730 distributed on the positive and negative statement of 103,225 collected statements.

Fig 5. shows the distribution of reviews/comments percentage over the collected dataset. In the followings section, we will demonstrate the properties of both tweets and booking reviews separately.

Table 1. Dataset statistics

Type	Count	Positive	Negative	Tokens
Reviews	53291	43887	9404	1486897
Tweets	49934	24999	24935	694833
Total	**103225**	**68886**	**34339**	**2181730**

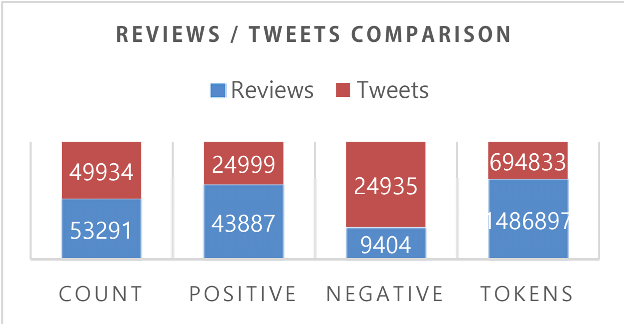

Fig. 5 Reviews/comments percentage over the collected dataset

4.2 Proposed Model Configuration and Training

We tokenize each tweet into words using GPAM [12], then split it into chunks using Stanford's CoreNLP [18]. This training the word vector, wen only pick the words appearing more than 5 times while replacing other words a special token "UNK". We got the phrase chunks embedding by training an unsupervised word2vec model on our dataset splitted as training and validation then use the word embedding to initialize the weights. [19]

Fig. 6 Model Configuration Summary

```
Layer (type)                 Output Shape              Param #
=================================================================
input_2 (InputLayer)         (None, 15, 50)            0

time_distributed_1 (TimeDist (None, 15, 200)           740900

bidirectional_2 (Bidirection (None, 15, 200)           180600

att_layer_2 (AttLayer)       (None, 200)               20200

dense_1 (Dense)              (None, 2)                 402
=================================================================
Total params: 942,102
Trainable params: 942,102
Non-trainable params: 0
```

Having a word vector of 200-dimension, the words having a probability to be closer are similar and get surrounded together. We then reduce them into 2-dimension using a t-SNE tool. Then using Bokeh visualization tool we mapped them directory to 2-dimension plan and interact with it. Fig7 Shows the bokeh chart for top 10,000-word vectors.

Fig. 7 2D-plane visualization of top 10000-word vectors

The hyper parameters of the proposed model are randomly intialized tuned on the validation set. In our experiments, we set the word embedding vector size to to be 200 dimensions and the GRU dimension to be 50, noting that we consider a word as a single unit represeints a phrase-chunk, it can be one or more words in the same node of parse tree of the same statement. In this case a bidirectional GRU-Based chunk as sequence gives us 200 dimensions for word/phrase-chunk annotation. The word/sentence context vectors also have a dimension of 100, initialized at random. For training, we set a mini-batch size of 32 and tweets/reviews of similar length (in terms of the number of words in the tweet/review) are organized to be a batch. Stochastic gradient descent used to train the proposed model. Then we picked the selected learning rates using grid search on the validation set. The proposed model handles each tweet/review equivalent to document in the work of Yang et. al. [16].

```
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 32)                6432

dense_2 (Dense)              (None, 1)                 33
=================================================================
Total params: 6,465
Trainable params: 6,465
Non-trainable params: 0
```

Fig. 8 Model Summary

Additional preprocessing was applied to the Tweets dataset including removing user mentions, re-tweet labels, and URLs, and preprocessing hashtag mentions by removing the hashtag symbol and the "under-scores" connecting between multiple words in a single tag. Hence, these techniques are common practice in the literature [20]. Performance is quantified using accuracy and F1-score averaged over both opinion classes. To ensure the statistical significance of results, the different models are evaluated using 10-fold cross-validation. ATSA is formed of three layers, and both the size of the word embeddings and hidden neurons in each layer are set to 50, which yield the best results in a preliminary experiment on a random fold of the ATSA dataset. For all experiments, we train the word semantic embeddings using an unlabeled dataset of 100,000 reviews extracted from the full hotel reviews and that do not pertain to any evaluation corpus. The reason for using hotel reviews is that it contains a collection of text written in both MSA and DA, and hence its vocabulary is more likely to cover the different evaluation corpora used in this article. The unlabeled dataset is also preprocessed using the above-mentioned steps.

4.3 Results and Evaluation

The Accuracy of the proposed ATSA model is evaluated on large scale sentences classification data sets consists of 103,225 tweets/review collected and annotated manually. These data sets was categorized into two types of samples: hotel reviews and tweets. The dataset was splited into training set of 80% of the data, 10% for validation, and the 10% for testing.

The experimental results show that the proposed moded is proforming significatly better than the most known stright forwart implementations of machine learning algorithms for Arabic sentiment analysis. Visualization ot the model accuacy demonistrated in Figure 9 shwos how much the proposed model is effective in ideniting the important words and phrase-chunks.

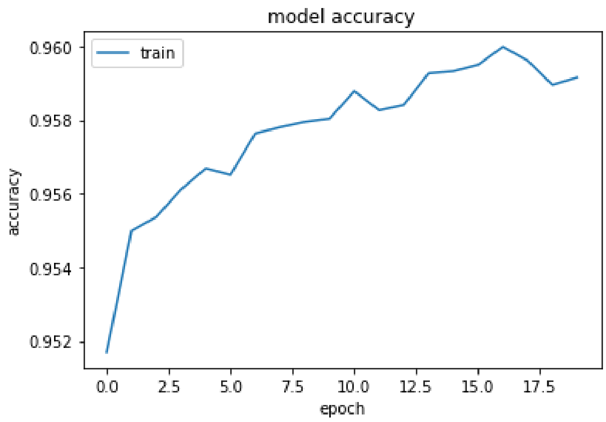

Fig. 9 Training Model Accuracy

The most common measures of classification effectiveness in sentiment classification problems are: accuracy, precision and recall. We used those three measures in addition to the F-

Measure to measure the accuracy of the test data as it consider both the precision and the recall of the test in computing the score.

Table 2 The Impact of Each of The Proposed Contributions

Algorithm	Hotel Reviews		Tweets	
	accuracy	F1-score	accuracy	F1-score
Baseline	74.3	73.5	69.7	61.1
POS Tagging	75.2	73.8	71.1	72.6
Stop Words	79.4	78.3	73.4	73.9
Syntactic Parsing	81.5	82.5	75.6	76.8
HAN	85.6	84.9	79.1	79.3
Combined	88.7	86.3	81.4	79.8

4.4 Results Benchmarking

In this section, we evaluate ATSA against several classification models proposed in the literature. We compare to SVM and NB models trained using BoW features with different choices of preprocessing and feature representation. We train using word, stem, and lemma n-grams (n =1, 2, 3), represented using presence, term frequency (TF), and term-frequency inverse-document frequency (TFiDF) scores. We report the best results achieved using the TFiDF scores. Sentiment Analysis accuracy is measured by general Accuracy, Precision, Recall, and F-measure and cross-validation. In this process as well, we randomly shuffling both positive and negative samples together to make sure that the cross-validation test chunks not fall in a single part either positive or negative. In our comparison we utilized both single and 5-folds cross validation.

Table 3: Hotel Review Analysis Result

Model		Accuracy	Precision	Recall	F-measure
Single fold result	Naive Bayes	0.83	0.83	0.82	0.83
	Maximum Entropy	0.78	0.80	0.76	0.77
	SVM	0.82	0.82	0.82	0.82
N-fold cross validation result	Naive Bayes	0.85	0.84	0.84	0.84
	Maximum Entropy	0.80	0.81	0.79	0.79
	SVM	0.83	0.82	0.82	0.82
ATSA	**HAN**	**0.88**	**0.86**	**0.88**	**0.87**

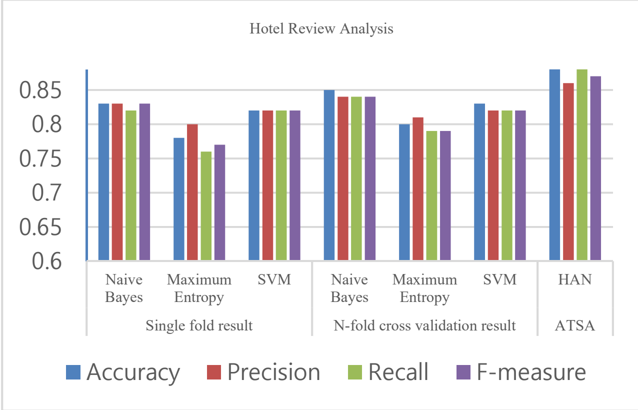

Fig. 10 Hotel Review Analysis Results

In the other hand we applied the same models into tweets dataset, the results show higher accuracies due to the length and number of chunks in each tweet.

Table 4. Tweets analysis results

Model		Accuracy	Precision	Recall	F-measure
Single fold result	Naive Bayes	0.80	0.85	0.79	0.79
	Maximum Entropy	0.78	0.84	0.78	0.77
	SVM	0.90	0.90	0.90	0.90
N-fold cross validation result	Naive Bayes	0.96	0.96	0.96	0.96
	Maximum Entropy	0.95	0.95	0.95	0.95
	SVM	0.97	0.97	0.97	0.97
ATSA	**HAN**	**0.99**	**0.98**	**0.99**	**0.99**

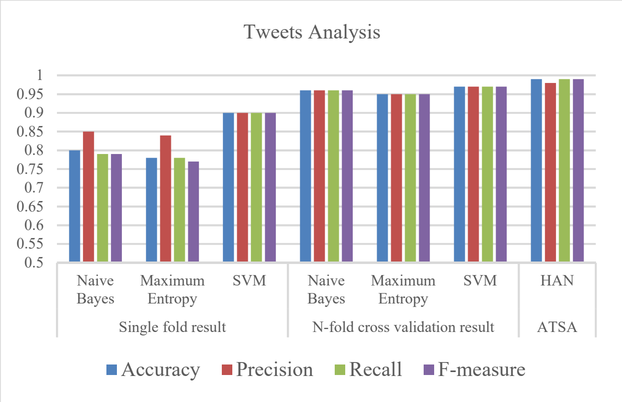

Fig. 11 Tweets Analysis Results

The ATSA model outperforms all the other models by a margin of around 3% for hotel review and 2% of tweets on the same datasets. As pointed out earlier, in this model, semantic context and the parsing order of phrase-chunks are considered. In the same time, no lexicon is used, and no special features are used, but only raw words as input. Table III and Table IV show the result of benchmarking the Hierarchical attention deep learning models proposed against other models in literature, like linear SVM applied to the same dataset which we created and annotated which represent the state-of-the-art results the dataset in Arabic sentiment classification. ATSA outperformed the classic implementations of machine learning models by around 3%.

5. Conclusions

In this work, we have presented ATSA, a hierarchical attention deep learning model for Arabic sentiment analysis. ATSA was proposed by addressing several challenges and limitations that arise when applying the classical models to perform opinion mining in Arabic. Arabic-specific challenges including the morphological complexity and language sparsity were addressed by modeling semantic composition at the Arabic morphological analysis after performing tokenization. We also proposed to perform phrase-chunks sentiment embedding to provide a broader set of features that cover syntactic, semantic, and sentiment information. At last, we used phrase structure parser to generate syntactic parse trees that are used as a reference for ATSA. This allowed modeling semantic and sentiment composition following the natural order in which words and phrase-chunks are combined in a sentence.

The proposed model was evaluated on three Arabic corpora that correspond to different genres (newswire, online comments, and tweets) and different writing styles (MSA and dialectal Arabic). Experiments showed that each of the proposed contributions in ATSA was able to achieve significant improvement. The combination of all contributions, which makes up for the complete ATSA model, was able to improve the classification accuracy by 3% and 2% on Hotel reviews and Tweets datasets, respectively, compared to the existing models. Furthermore, ATSA outperformed several basic approaches by 3% and 2% on the same datasets. These results indicate the ability of ATSA to perform accurate opinion classification when applied to a complex language such as Arabic that lacks large-scale opinion lexical resources.

References

[1] Serrano-Guerrero, J., et al., *Sentiment analysis: A review and comparative analysis of web services.* Information Sciences, 2015. **Vol. 311**: p. pp. 18-38.

[2] Ibrahim, H., S. Abdou, and M. Gheith, *Idioms-Proverbs Lexicon for Modern Standard Arabic and Colloquial Sentiment Analysis.* arXiv preprint arXiv:1506.01906, 2015.

[3] Gonçalves, P., et al. *On the combination of off-the-shelf sentiment analysis methods.* in *Proceedings of the 31st Annual ACM Symposium on Applied Computing*. 2016. ACM.

[4] Manek, A.S., et al., *Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier.* World Wide Web, 2016: p. pp. 1-20.

[5] Agarwal, B. and N. Mittal, *Machine Learning Approach for Sentiment Analysis*, in *Prominent Feature Extraction for Sentiment Analysis*. 2016, Springer. p. pp. 21-45.

[6] Al-Mansouri, E., *Using Artificial Neural Networks and Sentiment Analysis to Predict Upward Movements in Stock Price*. 2016, Worcester Polytechnic Institute.

[7] Medhat, W., A. Hassan, and H. Korashy, *Sentiment analysis algorithms and applications: A survey.* Ain Shams Engineering Journal, 2014. **Vol. 5**(Issue 4): p. pp. 1093-1113.

[8] Qin, Z., *A framework and practical implementation for sentiment analysis and aspect exploration*, in *Management Sciences and Marketing (MSM) Division*. 2017, University of Manchester: Faculty of Humanities.

[9] Sutskever, I., O. Vinyals, and Q.V. Le. *Sequence to sequence learning with neural networks.* in *Advances in neural information processing systems*. 2014.

[10] Socher, R., *Recursive deep learning for natural language processing and computer vision.* 2014, Citeseer: Stanford University.

[11] Al-Sallab, A., et al., *Aroma: A recursive deep learning model for opinion mining in arabic as a low resource language.* ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 2017. **16**(4): p. 25.

[12] Maabid, A., T. Elghazaly, and M. Ghaith, *An Enhanced Rule Based Arabic Morphological Analyzer Based on Proposed Assessment Criteria*, in *The Sixth International Conference on Swarm Intelligence and the Second BRICS Congress on Computational Intelligence (ICSI-CCI'2015) (IEEE Conference #36364).* 2015, Springer Lecture Notes in Computer Science series: Beijing, China.

[13] Diab, M. *Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking*. in *2nd International Conference on Arabic Language Resources and Tools*. 2009.

[14] Diab, M.T. *Improved Arabic base phrase chunking with a new enriched POS tag set*. in *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*. 2007. Association for Computational Linguistics.

[15] Abdul-Mageed, M., S. Kübler, and M. Diab. *Samar: A system for subjectivity and sentiment analysis of arabic social media*. in *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*. 2012. Association for Computational Linguistics.

[16] Yang, Z., et al. *Hierarchical attention networks for document classification*. in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2016.

[17] Bahdanau, D., K. Cho, and Y. Bengio, *Neural machine translation by jointly learning to align and translate.* arXiv preprint arXiv:1409.0473, 2014.

[18] Manning, C., et al. *The Stanford CoreNLP natural language processing toolkit*. in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 2014.

[19] Mikolov, T., et al. *Distributed representations of words and phrases and their compositionality*. in *Advances in neural information processing systems*. 2013.

[20] Zhang, L., et al., *Combining lexicon-based and learning-based methods for Twitter sentiment analysis.* HP Laboratories, Technical Report HPL-2011, 2011. **89**.

Abdelmawgoud M Meabed received the B.S. degree in specific education, computers department from Cairo University, Egypt, in 2001 and the M.S. degree in Computers & Information Sciences from Cairo University, Cairo, Egypt, in 2014. He is currently pursuing the Ph.D. degree in Computer Science at Cairo University, Cairo, Egypt.

His research interest includes the development Natural language understanding & processing of Standard and Colloquial Arabic, in 2014, he Implemented a complete full-fledge a General-purpose Arabic morphological analyzer and Generator, (GPAM) c.