# Multimodal Sentiment Analysis of Arabic Videos

Hassan Najadat
Department of Computer Information Systems, Jordan University of Science and Technology, Irbid, Jordan
Email: najadat@just.edu.jo

Ftoon Abushaqra
Department of Computer Science, University of Science and Technology, Irbid, Jordan
Email: ftoon.abushaqra93@gmail.com

***Abstract*—A huge number of videos posted every day online, these videos have an enormous of information about people reactions and opinions. Processing this data required an effective method. In this paper we propose a multimodal sentiment analysis classifier, using voice and facial features of the person. In our study, a new dataset for Arabic videos from the YouTube is collected. Several features were extracted including linguistic, audio, and visual data from the videos. The Arabic videos dataset includes a class label either positive, negative, or neutral. We utilized different classifiers including Decision Tree, k Nearest Neighbor (KNN), naive Bayes classifier, Support Vector Machine (SVM), and neural network. Overall accuracy is 76%.**

***Index Terms*—sentiment analysis, multimodal sentiment analysis, Arabic dataset, features extraction**

## I. Introduction

Sentiment analysis is a way to study human behavior and determine peoples' reactions. Basically, it is the process of knowing the feeling of someone depending on what he is writing or saying. Learn such an information helps in decision making at many different fields such as Market, Economic, Business and even Political [1]-[3]. For example [4] provided a twitter sentiment analysis for American elections, the application showed how many positive and negative tweets had written about specific candidates.

Applications for sentiment analysis have become everywhere. For instance, before buying a product, a simple Google product search shows you a brief summary of hundreds of reviews; this will help people making the right decision and let them know how other think about a particular topic.

Different studies focused on text-based sentiment analysis such as [5], where the information is extracted only from a written text. This text could be a Facebook post or comment, a tweet on twitter or a movie reviews. Using the words only to determine the feeling of a person may lead to inefficient result since the context has a great effect on the meaning; for instance, sarcasm or other forms of derisive language are hard to determine.

In the last few years, the social media applications have been increased and developed to contain videos. People now use videos, audio, and images more than ever, for sharing their daily live activities, their mood and their opinions. Analyze this data which contain a different modalities such as sound, video, and text will solve the text sentiment analysis problem and give more accurate results.

In this work, we addressed the task of multimodal sentiment analysis, we extracted several features from linguistic, audio, and visual data from the videos, and we experiment with Arabic videos that have a class label (positive, negative or neutral). Then we apply different classification methods to build our sentiment classifier.

The paper is organized as follows. section 2 reviewed a related work on sentiment analysis for text, audio, images, and the fusion method used to combine the features, section 3 describe the dataset used and the preprocessing phase, section 4 present our framework and described features extraction methods used for sentiment classification. We present our experiments and evaluations in section 5. Finally, we made a conclusion and discussion of the results.

## II. Related Work

Multimodal sentiment analysis application depends on the efficiency of the methods used for features extraction and emotion recognition for different types of data. Furthermore, the fusion method used to combine these features play a central role in increasing the accuracy of the results. In this section, we review the method used for text, audio and images feature extraction in multimodal sentiment analysis previous work. Considering that there is a limited number of such works. We also review different fusion methods that used for combine data from the three different modalities.

### *A. Text Features*

Emotion recognition from text builds either using a rule-based classifier or data-driven methods. For rule-based classifiers, sentiment lexicons must be used to determine the polarity of the text such as [6] which is one of the first widely used lexicons. The lexicons mainly contain a long list of words and its classification label (positive or negative). In multimodal sentiment analysis applications, Rosas *et al.* used a bag-of-words representation to derive unigram counts for his text

Manuscript received January 20, 2018; revised June 22, 2018.

doi: 10.18178/joig.6.1.39-43

sentiment recognition [7]. while the authors of [8] followed the semantic computing paradigm that is introduced in [9] before combine the text features with audios and visual features.

For Arabic text, many recent researches were made in the last few years [10]-[12]. Sentiment analysis for Arabic text has many challenges [13]; data need many preprocessing strategies before build a classifier due to the nature of Arabic language [14].

One of the modern Arabic lexicons was in 2017 by Al-Moslmi *et al.* they introduced Arabic senti-lexicon which contain a list of 3880 positive and negative sentence [15].

### B. Audio Features

Emotion analysis of speech signals aims to identify the emotional states of a person by extracting a set of voice features (e.g., Pause, Duration, Pitch, Intensity, and Loudness). These features help in solving the ambiguity problem that found in text only sentiment analysis.

Speaker-dependent and the speaker-independent are the main approaches used for voice recognition. The first approach depends on knowledge of speaker particular voice, while the second one able to recognize the speech from different users. Despite the speaker-dependent approach gives much better results, it is hard to applied due to the huge number of users speaks every day on the social media [16]. As we note that all of multimodal sentiment analysis researcher [7], [8], [17], [18] used the open source software OpenEAR [19] to extract audio features from an audio track. Then an advanced analysis for these features determined the emotional state of the speakers.

### C. Visual Features

The facial expression defined the unconscious emotion, through subtle movements of facial muscles such as smiling or eyebrow rising. One of the most widely used system for measuring and describing facial behaviors is the Facial Action Coding System (FACS) proposed by Ekman *et al.* [20]. FACS has been used by Datcu and Rothkrantz [21] where the authors provide an automatic emotion recognition using semantic audio-visual data fusion.

Poria *et al.* [8] used the facial recognition software Luxand FSDK along with GAVAM [22] in there multimodal sentiment analysis application. Rosas *et al.* [7] used two series of summary features: Smile duration and Look-away duration.

### D. Multimodal Information Fusion

Fusion is the task of extraction and combination of interrelated information from multiple modalities. Two common fusion strategies used to combine the linguistic-audio-visual recognition: feature-level and decision-level. In feature-level fusion which called an early fusion the combination process applied before performing any classification operations. For instance, the extracted characteristic and features from several modalities are combined into one vector. This vector is classified as one unit.

In decision-level fusion approach each modality is classified independently then we combined the uni-modal results at the end of the process. Many studies favor the decision-level fusion since the errors from different classifiers tend to be uncorrelated.

Rosas *et al.* used Feature-level fusion in his works [7], [18] .while Poria *et al.* [8] carried out both decision and feature-level fusion. Their experimentation showed that decision-level fusion better than feature-level fusion. For our experiments we apply a feature level fusion approach.

## III. DataSet

Videos are known highly available through the internet; therefore, finding a collection dataset for multimodal sentiment analysis is not the hard job. Multimodal sentiment analysis dataset must provide the three different data types: Linguistic, Audio, and Images. These videos could be found on social media websites such as Facebook, Instagram, and Twitter.

Most of the multimodal sentiment analysis studies [7], [8], [17], [23] collected their datasets from YouTube website.

The dataset of 21 Arabic videos from the YouTube were collected. The dataset does not talk about a particular topic. We select the videos that do not have any background sound, and the person must be in front of the camera.

Keywords of the dataset include the following: my opinion (رأي في), my experience (تجربتي مع), and public opinion (رأي الشارع). Finally, the length of the videos varies from 10 to 40 seconds after segmentation process.

All videos are manually segmented to the beginning of opinion utterance. And for each video we manually assign one of three labels: neutral, positive, or negative. The data set contains nine videos labeled as negative, four as neutral, and eight as positive.

## IV. Features Extraction

In our work, many extraction methods are used in our work to extract features from audio and visual data.

### A. Audio

For the audio features extraction, we used the free computer software package, Praat, for the scientific analysis of speech in phonetic. Praat provides many functions including basic speech analysis to graphics and statistics. The source code of Praat is available on the website (www.praat.org), with many help menus and tutorials.

In our experiment, four main temporal and spectral features are automatically extracted from each video:

- Voice energy: The average of the signal energy over the whole audio.
- Voice power: The voice power for the audio.
- Intensity: The average voice intensity over the whole audio
- Pitch: The pitch is a spectral feature represents the variation of voice intonation. We compute the mean of the pitch level for the video

### B. Visual Features

The visual features and facial expression are important to increase the accuracy of sentiment analysis. In our experiment, the visual features extracted automatically from each video, since the video represented by one person who faces the camera. It is easy to extract facial features using any face recognition software. We used Luxand-faceSDK by Luxand Company. FaceSDK used in many applications for different purposes such as face recognition and detecting facial features. Many other software would be good choice such as OpenFace and OpentCV.

In our work, we focused on two main features: smile and eye. We take the average of the percentages of the smile over the video. Also, we take the average of the percentages of eye opening.

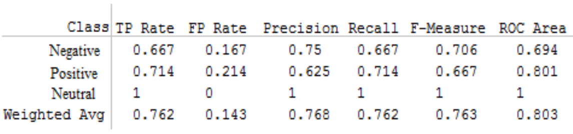
### C. Fusion

As discussed previously, there are two main fusion techniques that have been used: decision-level fusion and feature-level fusion. We used feature-level fusion. We combine the features extracted from visual and audio data together as one vector then we perform the classification methods on the whole vector.

## V. Experiments and Evaluations

In our experiments, we performed five machine learning classifiers; Decision Tree, k nearest neighbor (K-NN), naive Bayes classifier, support vector machine (SVM), and neural network using Weka data mining tool.

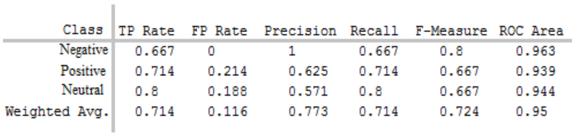
Before applying the classifiers, we increased the dataset using Weka re-sampling option. For the Decision Tree method, we apply J4.8 whose accuracy yielded 76%. It failed to classify 6 out of 21 objects. Fig. 1 shows a detailed accuracy for the decision tree. Decision tree outperforms other classification methods used



```
          Class TP Rate  FP Rate  Precision Recall F-Measure ROC Area
       Negative   0.667    0.167      0.75   0.667    0.706    0.694
       Positive   0.714    0.214      0.625  0.714    0.667    0.801
        Neutral   1        0          1      1        1        1
   Weighted Avg   0.762    0.143      0.768  0.762    0.763    0.803
```

Figure 1. Decision tree classifier evaluation

Using the K-nearest neighbor method with k=3 we got less accuracy than the decision tree. The classifier correctly classified 15 objects; gives about 71% accuracy as shown in Fig. 2.



```
          Class TP Rate  FP Rate  Precision Recall  F-Measure ROC Area
       Negative   0.667  0           1       0.667    0.8      0.963
       Positive   0.714  0.214       0.625   0.714    0.667    0.939
        Neutral   0.8    0.188       0.571   0.8      0.667    0.944
  Weighted Avg.   0.714  0.116       0.773   0.714    0.724    0.95
```

Figure 2. K-nearest neighbors classifier evaluation

Naive Bayes method error rate was 38% with recall and precision values equal to (0.61). Naive Bayes accuracy was 61%. While SVM obtain the lowest performance, it gave an accuracy of 57%. Precision and recall were (0.45) and 0.57 respectively.

We also noted that result for the artificial neural network method was quite similar to the decision tree result. Table I shows the Confusion Matrix for the best three classification methods: Decision Tree, k nearest neighbor, and neural network.

TABLE I. Confusion Matrix for Decision Tree K-NN and Neural Network

| | **Decision Tree** | | | **K Nearest Neighbor** | | | **Artificial Neural Network** | | |
|---|---|---|---|---|---|---|---|---|---|
| **Class Label** | ***Negative*** | ***Positive*** | ***Neutral*** | ***Negative*** | ***Positive*** | ***Neutral*** | ***Negative*** | ***Positive*** | ***Neutral*** |
| **Negative** | 6 | 3 | 0 | 6 | 2 | 1 | 6 | 2 | 1 |
| **Positive** | 2 | 5 | 0 | 0 | 5 | 2 | 2 | 5 | 0 |
| **Neutral** | 0 | 0 | 5 | 0 | 1 | 4 | 0 | 0 | 5 |

In the next experiment, a new attribute, gender, to the dataset was added; since the voice power and energy are different for male than female. From the experiment, the gender attribute did not provide any affect to the result of the classification.

As a final step, we study the effect of combining different data modalities on the sentiment analysis process. We first removed the facial expression features from the dataset and study the result for using a uni-model classification consist of audio only. Then we do the same for the audio features to study the result of sentiment analysis using only visual data.

Tables II, III, and IV provides a summary of the accuracy, recall, and precision values for different types of modalities using the three best classifiers. Fig. 3 shows that k nearest neighbors classifier attained the highest accuracy when audio only used. Also, applying audio only achieved the highest accuracy when the neural networks is used as shown in Fig. 4.

TABLE II. Sentiment Classification Performance for Different Models Using Decision Tree

| Modality | Accuracy | Precision | Recall |
|---|---|---|---|
| Audio only | 80% | 0.83 | 0.81 |
| Visual only | 76% | 0.76 | 0.76 |
| Audio- Visual | 76% | 0.76 | 0.76 |

TABLE III. Sentiment Classification Performance for Different Models Using K-nearest Neighbors Algorithm

| Modality | Accuracy | Precision | Recall |
|---|---|---|---|
| Audio only | 71% | 0.81 | 0.71 |
| Visual only | 66% | 0.79 | 0.66 |
| Audio- Visual | 71% | 0.77 | 0.71 |

TABLE IV. Sentiment Classification Performance for Different Models Using Neural Network

| Modality | Accuracy | Precision | Recall |
|---|---|---|---|
| Audio only | 85% | 0.91 | 0.85 |
| Visual only | 47% | 0.46 | 0.47 |
| Audio- Visual | 76% | 0.75 | 0.76 |

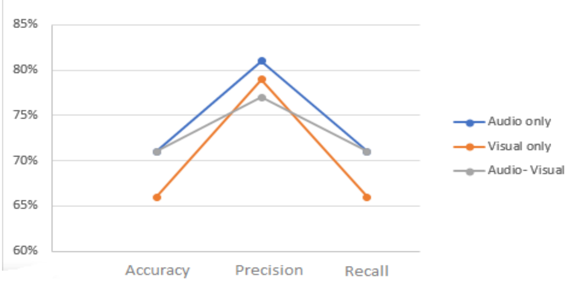


Figure 3. Sentiment classification performance for different models using K-nearest neighbor's algorithm

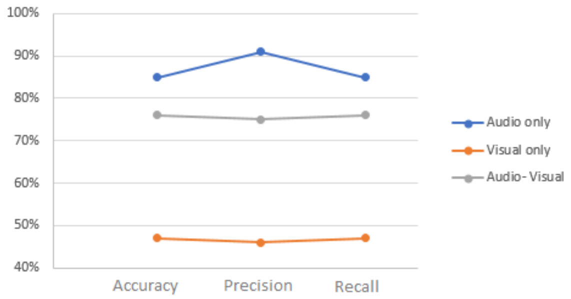


Figure 4. Sentiment classification performance for different models using neural network

## VI. Results and Discussion

As shown in Tables II, III and IV; the result obtained with one model and two modalities at a time are not as expected. The results for audio-only modal classifier yielded better than two modalities classifiers, these results caused by the lack of enough extracted features. Two facial features including smile and eye were not enough to detect the emotional state of the person. As the videos indicate many persons have a positive opinion about a particular topic but they express their opinion without smiling. Although the overall results were promising, extracted more features from both visual and audio data may increase the performance of the classifier.

## VII. Conclusion and Future Work

In this paper, the multimodal sentiment analysis is addressed for Arabic videos with a discussion of the features extraction technique used for different data types such as text, audio, and visual. The fusion method used in multimodal sentiment analysis was reviewed. Five different classification methods including Decision Tree, k-nearest neighbor, naive Bayes classifier, support vector machine and neural network were employed.

As a future work, we will collect more dataset as we will extract more features for all modalities. And we aim to recognize more emotions for a person rather than the polarity of his opinion.

## References

[1] M. S. Vohra and J. Teraiya, “Applications and challenges for sentiment analysis: A Survey,” *International Journal of Engineering Research and Technology*, vol. 2, no. 2, 2013.

[2] M. Hu and B. Liu, “Mining and summarizing customer reviews,” in *Proc. Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2004, pp. 168–177.

[3] A. Levenberg, S. Pulman, K. Moilanen, E. Simpson, and S. Roberts, “Predicting economic indicators from web text using sentiment composition,” *Int. J. Comput. Commun. Eng.*, vol. 3, no. 2, pp. 109–115, 2014.

[4] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, “A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle,” in *Proc. ACL 2012 System Demonstrations, Stroudsburg*, PA, USA, 2012, pp. 115–120.

[5] N. Nehra, “A survey on sentiment analysis of movie reviews,” *Int. J. Innov. Res. Technol.*, vol. 2, no. 7.

[6] P. J. Stone and E. B. Hunt, “A computer approach to content analysis: Studies using the general inquirer system,” in *Proc. Spring Joint Computer Conference*, New York, NY, USA, 1963, pp. 241–256.

[7] V. P. Rosas, R. Mihalcea, and L. P. Morency, “Multimodal sentiment analysis of spanish online videos,” *IEEE Intell. Syst.*, vol. 28, no. 3, pp. 38–45, May 2013.

[8] S. Poria, E. Cambria, N. Howard, G. B. Huang, and A. Hussain, “Fusing audio, visual and textual clues for sentiment analysis from multimodal content,” *Neurocomputing*, vol. 174, part A, pp. 50–59, Jan. 2016.

[9] E. Cambria, A. Hussain, C. Havasi, and C. Eckl, “Sentic computing: Exploitation of common sense for the development of emotion-sensitive systems,” in *Development of Multimodal Interfaces: Active Listening and Synchrony*, A. Esposito, N. Campbell, C. Vogel, A. Hussain, and A. Nijholt, Eds., Springer Berlin Heidelberg, 2010, pp. 148–156.

[10] N. A. Abdulla, N. A. Ahmed, M. A. Shehab, and M. Al-Ayyoub, “Arabic sentiment analysis: Lexicon-based and corpus-based,” in *Proc. IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies*, 2013, pp. 1–6.

[11] M. Abdul-Mageed, M. Diab, and S. Kübler, “SAMAR: Subjectivity and sentiment analysis for Arabic social media,” *Comput. Speech Lang.*, vol. 28, no. 1, pp. 20–37, Jan. 2014.

[12] A. Shoukry and A. Rafea, “Sentence-level Arabic sentiment analysis,” in *Proc. International Conference on Collaboration Technologies and Systems*, 2012, pp. 546–550.

[13] A. Assiri, A. Emam, and H. Aldossari, “Arabic sentiment analysis: A survey,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 6, no. 12, 2015.

[14] R. Duwairi and M. El-Orfali, “A study of the effects of preprocessing strategies on sentiment analysis for Arabic text,” *J. Inf. Sci.*, vol. 40, no. 4, pp. 501–513, Aug. 2014.

[15] T. Al-Moslmi, M. Albared, A. Al-Shabi, N. Omar, and S. Abdullah, “Arabic senti-lexicon: Constructing publicly available language resources for Arabic sentiment analysis,” *J. Inf. Sci.*, Feb. 2017.

[16] H. Atassi and A. Esposito, “A speaker independent approach to the classification of emotional vocal expressions,” in *Proc. 20th IEEE International Conference on Tools with Artificial Intelligence*, 2008, vol. 2, pp. 147–152.

[17] L. P. Morency, R. Mihalcea, and P. Doshi, “Towards multimodal sentiment analysis: Harvesting opinions from the web,” in *Proc. 13th International Conference on Multimodal Interfaces*, New York, NY, USA, 2011, pp. 169–176.

[18] V. Pérez-Rosas, R. Mihalcea, and L. P. Morency, “Utterance-level multimodal sentiment analysis,” presented at the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 2013.

[19] F. Eyben, M. Wöllmer, and B. Schuller, “OpenEAR #x2014; Introducing the munich open-source emotion and affect recognition toolkit,” in *Proc. 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1–6.

[20] D. Datcu and L. J. M. Rothkrantz, “Semantic audio-visual data fusion for automatic emotion recognition,” in *Proc. Euromedia*, 2008, pp. 1–6.

[21] J. M. Saragih, S. Lucey, and J. F. Cohn, “Face alignment through subspace constrained mean-shifts,” in *Proc. IEEE 12th International Conference on Computer Vision*, pp. 1034–1041, 2009.

[22] P. Boersma, “Praat, a system for doing phonetics by computer,” *Glot Int.*, vol. 5, 2002.

[23] P Ekman and H. Oster, “Facial expressions of emotion,” *Annu. Rev. Psychol.*, vol. 30, no. 1, pp. 527–554, 1979.

**Hassan Najadat** is an associate professor of computer science at Jordan University of Science and Technology. He earned his Ph.D. in computer science from North Dakota State University, USA. His research interests include data mining, artificial intelligence and data envelopment analysis.

**Ms. Ftoon Abushaqra** holds a bachelor degree in Computer Information Systems from Jordan University of Science and Technology, Jordan. She is currently pursuing Master in Computer Science at the same university. She is interested in data mining filed and its application.