Towards Arabic Multi-modal Sentiment Analysis

Abdulrahman S Alqarafi[1,3], Ahsan Adeel[1], Mandar Gogate[1], Kia Dashitpour[1], Amir Hussain[1], and Tariq Durrani[2]

[1] CogBID Lab, Dept. of Computing Science and Mathematics
University of Stirling, FK9 4LA Stirling, UK
[2] University of Strathclyde Glasgow, UK
[3] University of Taibah, Madina, Saudi Arabia

Abstract. In everyday life, people use internet to express and share opinions, facts, and sentiments about products and services. In addition, social media applications such as Facebook, Twitter, WhatsApp, Snapchat etc., have become important information sharing platforms. Apart from these, a collection of product reviews, facts, poll information, etc., is a need for every company or organization ranging from start-ups to big firms and governments. Clearly, it is very challenging to analyse such big data to improve products, services, and satisfy customer requirements. Therefore, it is necessary to automate the evaluation process using advanced sentiment analysis techniques. Most of previous works focused on uni-modal sentiment analysis mainly textual model. In this paper, a novel Arabic multimodal dataset is presented and validated using state-of-the-art support vector machine (SVM) based classification method.

Key words: Arabic, Sentiment Analysis, Multi-modal

1 Introduction

Sentiment Analysis (SA) is the automatic extraction of a sentiment or opinion in a content that comes not only in the form of text but visually as well such as videos or acoustic audio for instance [1,2]. The significant number of Smartphones among individuals enable sharing opinions, stories, and reviews through online video sharing platforms such as YouTube, Vine, Snapchat, Facebook etc. [3,4]. This shared Big Data has captured the attention of various organisations, researchers and consumers who are interested in building better viewpoint-mining applications with the aim to aid better decision-making. To date, the field of multimodal sentiment analysis for Arabic language has not received much attention from researchers. Moreover, features extraction from different modalities and their fusion have not been reported in the literature. One of the biggest challenges in multimodal Arabic sentiment analysis is the variability of topics from time to time, as speakers tend to change them abruptly, including state of their opinion. In such scenarios, it is difficult to distinguish and segment the divergent opinions. For instance, a speaker could give additional opinions (positive or negative) for the same utterance at different time[5].

Accepted for publication in *Communications, Signal Processing, and Systems. CSPS 2017.* Lecture Notes in Electrical Engineering, 463. CSPS 2017: Communications, Signal Processing, and Systems, Harbin, China, 14.07.2017-16.07.2017. Harbin, China: Springer Verlag, pp. 2378-2386. The final publication is available at Springer via https://doi.org/10.1007/978-981-10-6571-2_290.

In addition, some speakers reflect their opinion using visual gestures[1]. In this case, when a speaker uses more facial expressions the opinion tends to be obvious in the visual model which needs to be extracted . On the other hand, when vocal expressions are used, the audio data may contain the clues for an opinion. Therefore, a comprehensive novel multimodal framework could result in more consistent sentiment analysis [3]. In this paper, some of the challenges of examining sentiments in online opinion videos for Modern Standard Arabic language are highlighted and addressed, including the process of features extraction from different modalities (e.g. Text and Video) and their integration. In this context, a novel Arabic multimodal Dataset is built and validated using state-of-the-art SVM based classification method, with the aim to detect the polarity from different models for Arabic language.

The rest of the paper is organised as follows: Sections 2 presents the built novel Arabic multimodal dataset. Section 3 explains the experimental setup and finally Section 4 presents the conclusion and future work.

2 Arabic Multi-Modal Dataset (AMMD)

The dataset is compiled from YouTube videos, considering only video-blogging videos. The dataset attempts to include many different meta information about the videos such as audio, visual gestures, transcript, and sentiment analysis annotation, all aligned with each other.

2.1 Acquisition Methodology

A total of 40 different videos (spanning two to three minutes in length) were selected from YouTube having 13 different speakers, 10 males and 3 females. The speakers come from 4 Arabic speaking countries, for their speech, covering MSA, the gulf dialect, and the Levantine dialect. Video-blogging or simply vlogs are targeted as these are the types of videos that contain subjective content whether positive, negative or neutral. In these videos, the speakers express their opinions regarding various topics. The topics contain reviews about books, movies, devices, and technologies. The videos were originally retrieved by searching YouTube for a short manually crafted list of keywords/queries. A subset of the retrieved videos is chosen as our dataset depending on their length, clarity and spoken dialect. In these videos, the speakers have used various setups to record their videos and hence, the varying aspects of these setups include the distance from the camera, the background and the lighting conditions. We also discuss some of the challenges we faced while building the dataset. For example, some vloggers prefer using light music as background which could affect automatic acoustic feature extraction. Some noise could result in poor accuracy. In addition, technology video reviews tend to show the device features, so most of the time, the camera focused on the device instead of the speaker, which makes the facial feature extraction difficult. Furthermore, the video has to carry a sentiment to be selected, whether positive, negative or neutral. In addition, we tried to search for videos where the speakers tend to speak the formal language or at least close to it, because most of vloggers and YouTubers speak the dialect instead of the Modern Standard Arabic. There are significant differences between some dialects and the Modern Standard Arabic, that makes the task of building a benchmark dataset challenging. Figure 1 shows some snapshots from the video dataset.

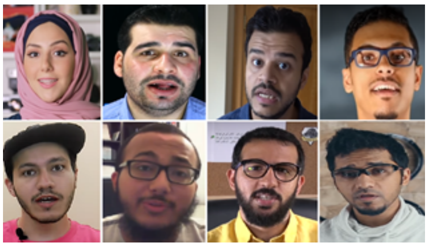

Fig. 1: Snapshots from the video dataset

Some of the keywords used to search for appropriate videos are showed in Table 2,

Keywords in Arabic	**Translation**
هل أشتري نظارة سناب شات؟	**Shall I buy** the Snapchat sunglasses
مراجعة رواية الظل	**Reviewing** the Shadow novel
نظارات الواقع الإفتراضي	**Virtual reality glasses**
أيفون ٧ مميزات وعيوب	IPhone 7,**Pros and Cons**

Table 1: example of keywords search

2.2 Transcription

After collecting the 40 videos, we transcribed all of them manually since there are no attached texts to the videos. Performing manual transcription and segmentation of utterances consume too much effort & time, and hence, it is reliable and results in a better performance. There exist some automatic speech recognition techniques such as Google API, IBM voice recognition etc. However, they are inadequate since they do not fully support the Arabic dialects. Although Google speech recognition API contains all targeted dialects for example Saudi Arabic, Egyptian Arabic, Bahraini Arabic Jordanian Arabic and so on, It's performance is not optimum and needs some improvements which will be addressed in future works.For these reasons, the transcription was conducted manually. At this point in time, we utilized two transcribers. The first transcriber who transcribed all the videos and after that the second transcriber reviewed and evaluated the transcribed content from the first transcriber

2.3 Subjectivity Annotation and Segmentation

Subjectivity annotation is an essential task in the supervised sentiment analysis where the transcribed content is critically evaluated. All the transcribed sentences in the transcript were categorized into either subjective utterance or objective utterance. The subjective utterances are those sentences in which the speaker expresses opinions regarding a product, whereas the objective utterances are those in which the speaker expresses facts. After obtaining the transcription of the videos, we have manually annotated the sentences into subjective and objective. Moreover, the subjective statements were annotated as either positive or negative depending on the orientation of the expressed opinion.. The annotated sentences, thereafter, were used to segment the video contents. The rules we followed during video segmentation based on the rules in [6] [7]:

- Grouping subsequent objective sentences: When the speaker uses successive related objective statements, we group them together into a single segment. For instance, in the following example, the speaker lists three distinct features of a certain product using three objective sentences. In this case, the three sentences are semantically related and hence are grouped in one long segment and given their shared objective label.

Objective sentences	**Translation**
موجود عندكم الذاكرة ثلاثة جيجا بايت	**There is a 3 GB RAM**
سعة هذي الذاكرة حقته سعة تخزينية ٢٣ قيقا بايت	**The capacity of RAM is 32 2GB**
وكرت ذاكرة يدعم الى ٨٢١ قيقا بايت	**The memory card support to 128 GB**

Table 2: Example of objective sentences

- Segmentation of subjective sentences: A sentence in which the speaker expresses his/her opinion i.e. whether positive/negative or conveys people opinion is segmented.
- Change of scenes: Whenever the video moves from one scene to another, we segment the utterance even if we split semantically related objective sentences. However, if the scene changes within the same statement, we do not split the utterance.

We requested trained annotators to annotate the whole dataset and review the results. The dataset contains 40 different videos. There are 830 utterances after segmentation process. The total number of the objective segments is 467 while the subjective segments are 363. The number of positive examples is 269, while the negative examples are 94. In addition, the dataset contains a total of 10,903 words. The Table 3 shows the statistics of the dataset.

Total no of Videos	40
Total no of segmentation	830
Total no of objective segments	467
Total no of Subjective segments	363
Total no of positive segments	269
Total no of negative segments	94
Total no of words	10903

Table 3: Dataset details

3 Experimental Setup

In multimodal sentiment analysis, not only but gestures and facial expressions can also indicate an opinion. Analysing sentiment from text, audio and video could lead to better accuracy in results. This experiment followed some of the procedures as mentioned in [6]. There are four different experiments conducted in this paper.

- Verbal experiment (Text based): classifying video segments using only transcript text.
- Visual: classifying video segments using only visual gestures.
- Classifying video segments using the concatenation of both verbal and visual predictors.
- Classifying video segments using joint features of both verbal and visual predictors.

3.1 Text (Verbal)

The verbal feature vector is extracted by converting the utterance into BoW representation:

- All distinct words in the corpus are gathered first to form a dictionary.
- Words are normalized by removing punctuations and converting all numbers to a single word "number".
- Words are stemmed so that we do not differentiate between multiple forms of the same lexical word (a man, the man)".
- The dictionary is filtered by removing frequent words (that occurred in at least half the videos), stop-words, and rare words that occurred less than k times. We empirically found that k=5 gives the better balance between vocabulary size (number of features) and accuracy.
- The final dictionary contains 604 normalized stemmed words.

3.2 Video (Visual)

In this work, we included some facial gestures by first extracting them and then annotating these features. We extracted the visual features manually for each segment. Four gestures and expressions were considered for manual annotation: smile, frown, head nod, and head shake. From all utterances, we annotated all the utterances. Two experts were asked to annotate the facial gestures. They were asked to watch the videos and observe for any of these four expressions mentioned above.

3.3 Verbal + Visual

"Verbal + Visual" input is the result of concatenating BoW and visual vectors for each utterance.

Experiment	Verbal	Visual	Verbal + Visual	Multimodal
Best Results Linear	58.21	60.29	63.16	65.59
Value of C Linear	0.01	1	0.1	0.01
Best Results RBF	57.41	59.33	60.29	63.2
Value of C	10	100	100	10

Table 4: Results for Subjectivity Classification models

3.4 Multimodal

The input for multimodal experiments is a joint representation of words and gestures of the same segment. We construct the feature vector as described in [6]. For each word, wi from the dictionary and gesture, gk from the four visual gestures, we add the two pairs of word-gesture tuple: (wi ,∼gk), (wi , gk) to the feature vector of the segment. The first pair indicates the co-existence of both word wi and visual expression gk in the video segment, while the second pair indicates the existence of the word without the gesture. For all types of experiments, we partitioned our data into two parts. The first quarter of the data is set aside for testing purposes. We call this "held-out dataset". The rest of the data is used for training and validation using a 4-fold cross-validation. Cross-validation was applied in particular to choose the hyper-parameters of the classifier, and to ensure that the model results are generalizable. All models were trained using C-SVM model. We only switched the value of C from the set 0.001, 0.01, 0.1, 1.0, 10, 100, 1000, and the type of kernel from the set linear, RBF. Generally speaking, cross-validation results preferred low values of C for linear kernel SVM and high values for RBF kernel SVM. This can be depicted from Table 4, Table 5 and Table 6.

- For linear kernel experiments, 1.0. 0.1 and 0.01 are good values for C.

- For RBF kernel, the experiments chose C to be as high as 10, or even 100. Increasing the value of C over 100 can give better results for some folds, but leads to a less generalizable model, i.e, a model that is susceptible to overfitting.

Experiment	Verbal	Visual	Verbal + Visual	Multimodal
Best Results Linear	72.83	59.78	71.74	76.09
Value of C Linear	0.1	1	1	0.1
Best Results RBF	70.65	61.96	71.74	76.09
Value of C	100	10	100	100

Table 5: Results for Polarity Classification models

For each of the four feature representations, we trained three different models to address three different tasks:

- Subjectivity classification model (classify subjective vs. objective).
- Polarity classification model (classify positive vs. negative after filtering out all objective examples).
- A general sentiment analysis model that classifies input as negative, zero or positive.

The hyper-parameters for each model were selected using the 4-fold cross-validation described above. The resulting model was then applied to the “held-out” dataset.

Table 4 shows the results for subjectivity classification models. Table 5 shows polarity classification results, and finally, Table 6 presents the general sentiment analysis experiment results . For each of the three tables, we put three feature representations on columns and the tested hyper-parameters on rows.

Experiment	Verbal	Visaul	Verbal + Visual	Multimodal
Best Results Linear	48.8	58.85	58.55	53.11
Value of C Linear	1	1	0.01	0.1
Best Results RBF	48.32	57.89	54.06	58.42
Value of C	100	10	100	100

Table 6: Results for general Sentiment Analysis models

It can be shown from the tables that fusing different features (utterance, visual) for different opinion mining tasks (such as subjectivity analysis and polarity classification) outperforms using the utterance features only in terms of results. Using our dataset, we could improve the subjectivity identification from 58.21% to 65.59%, the binary sentiment polarity classification from 72.83% to 76.09%, and we even obtained a better margin for the three polarities sentiment classification where we go up around 10 percent from 48.8% to 58.42%. In addition, it can be depicted from tables 4 and 5 is that using verbal features alone surpasses using visual features alone in the polarity classification, while the visual features are more useful when it comes to subjectivity detection. In other word, we can conclude that people tend to change their facial expressions when they start talking about their opinions and subjective states, whether positive or negative.

4 Conclusion & Future Work

In this paper, a novel Arabic multimodal sentiment analysis dataset is presented, which is publicly available for future researchers. For preliminary sentiments classification analysis, state-of-the-art SVM based classification models are built. The classification results have validate the extracted dataset and shows the feasibility and advantage of using multimodal data over textual only data for sentiment analysis task. The initial built dataset contains only textual and video models. In future, we intend to further expand our dataset to 100 videos and include audio features to build a real multimodal dataset covering all possible modalities.

5 Acknowledgements

The authors are grateful to the anonymous reviewers for their insightful comments and suggestions which helped improved the quality of the paper. This work was part-supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant no. EP/M026981/1 (AV-COHGEAR)

References

1. Morency, L.P., Mihalcea, R., Doshi, P.: Towards multimodal sentiment analysis: Harvesting opinions from the web. In: Proceedings of the 13th international conference on multimodal interfaces, ACM (2011) 169–176
2. Wöllmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., Morency, L.P.: Youtube movie reviews: Sentiment analysis in an audio-visual context. IEEE Intelligent Systems **28**(3) (2013) 46–53
3. Poria, S., Cambria, E., Howard, N., Huang, G.B., Hussain, A.: Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Neurocomputing **174** (2016) 50–59
4. Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion **37** (2017) 98–125
5. Rosas, V.P., Mihalcea, R., Morency, L.P.: Multimodal sentiment analysis of spanish online videos. IEEE Intelligent Systems **28**(3) (2013) 38–45
6. Zadeh, A., Zellers, R., Pincus, E., Morency, L.P.: Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv preprint arXiv:1606.06259 (2016)
7. Wiebe, J., Wilson, T., Cardie, C.: Annotating expressions of opinions and emotions in language. Language resources and evaluation **39**(2) (2005) 165–210