Multimodal Sentiment and Gender Classification for Video Logs

Sadam Al-Azani and El-Sayed M. El-Alfy
College of Computer Sciences and Engineering,
King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia

Keywords: Multimodal Recognition, Sentiment Analysis, Opinion Mining, Gender Recognition, Machine Learning.

Abstract: Sentiment analysis has recently attracted an immense attention from the social media research community. Until recently, the focus has been mainly on textual features before new directions are proposed for integration of other modalities. Moreover, combining gender classification with sentiment recognition is a more challenging problem and forms new business models for directed-decision making. This paper explores a sentiment and gender classification system for Arabic speakers using audio, textual and visual modalities. A video corpus is constructed and processed. Different features are extracted for each modality and then evaluated individually and in different combinations using two machine learning classifiers: support vector machines and logistic regression. Promising results are obtained with more than 90% accuracy achieved when using support vector machines with audio-visual or audio-text-visual features.

1 INTRODUCTION

Social media platforms have become a very attractive environment for people to share and express their opinions on all different aspects of life. They use various forms of content including text, audio, and video to expression their attitudes, beliefs and opinions. Flourishing business models with great social impacts have emerged in education, filmmaking, advertisement, marketing, public media, etc. Since the advent of YouTube in 2005, the volume of online videos is dramatically increasing and video blogging is becoming very prevalence with million daily YouTube views. Video blogging is overtaking other forms of blogging due to its visual expression captivation. The advances of smartphones with high quality video cameras have also invigorated people for more user-generated contents. Nowadays, YouTube supports more than 75 languages. Several other online services allowing video sharing include Flickr, Facebook, Instagram, DailyMotion, GoodnessTv, Metacafe, SchoolTube, TeacherTube, and Vimeo.

Analyzing, understanding and evaluating social media content can provide valuable knowledge in different applications and domains such as reviews of products, hotels, hospitals, visited places, and obtained services (Garrido-Moreno et al., 2018; Misirlis and Vlachopoulou, 2018). Sentiment analysis (SA) is defined as the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes in order to understanding the public mode and support decision making and recommendation (Liu, 2012). It incorporates several tasks; among them are subjectivity determination, polarity detection, affect analysis, opinion extraction and summarization, sentiment mining, emotion detection, and review mining in different levels namely, document, sentence and features/aspects (Ravi and Ravi, 2015).

Several approaches and methodologies have been proposed to address sentiment analysis tasks providing many resources and tools. However, most of the techniques in the context of sentiment analysis have focused on textual reviews and blogs. Recently, some researchers have been motivated to explore other communication modalities and combinations of modalities for sentiment analysis (Soleymani et al., 2017). Unimodal analysis and recognition systems suffer from several drawbacks and limitations such as noisy sensor data, non-universality, and lack of individuality. Additionally, each modality has its own challenges. For example, voice-based recognition systems might be affected by different attributes such as low voice quality, background noise, and disposition of voice-recording devices. Text-based recognition systems also suffer from several issues related to morphological analysis, multi-dialectics, ambiguity, temporal dependency, domain dependency,

Al-Azani, S. and El-Alfy, E.
Multimodal Sentiment and Gender Classification for Video Logs.
DOI: 10.5220/0007711409070914
In *Proceedings of the 11th International Conference on Agents and Artificial Intelligence (ICAART 2019)*, pages 907-914
ISBN: 978-989-758-350-6

etc. This is true as well regrading visual modality, which also suffers from illumination conditions, posture, cosmetics, resolution, etc. In consequence, this leads to inaccurate and insufficient representation of discriminating patterns.

Another interesting problem that has a growing research interest is gender recognition with the arising need for personalized, reliable, and secure systems (Alexandre, 2010; Cellerino et al., 2004; Li et al., 2013; Shan, 2012). It has several promising applications, e.g. human-computer interaction, surveillance, computer forensics, demographic studies, and consumer behavior monitoring (Abouelenien et al., 2017). Consequently, many studies have been carried out and several methodologies have been proposed to address this task.

Combining sentiment with gender recognition can provide more valuable segmentation information that can improve both individual systems. It can overcome the real-world gender bias issue of current sentiment analysis systems (Thelwall, 2018). The proposed method is based on three forms of modalities: text, audio and visual. Incorporating different modalities for a subject can be more reliable and significantly improve the performance since it provides evidences from different perspectives which can overcome the limitations of unimodal systems. Additionally, the proposed system conquers the domain-, topic- and time-independence sentiment analysis issues.

Determining the polarity of user's opinions alongside with gender is very important and have several interesting applications. For example, it is an excellent opportunity for large companies to capitalize on, by extracting user sentiment, suggestions, and complaints on their products from these video reviews. Consequently, they can improve their products/services to meet the customers' expectations. For example, reviews of shaving machines by males are more significant than from females whereas reviews of women-specific products such as make-up products are more appreciated from females than from males. This can also be applicable for governments to explore issues related to citizens according to their genders. Another application is for criminal and forensic investigation where the information obtained can be used as evidences. Amazing applications can be in educational and training systems such as adaptive and interactive educational systems where the content is tailored according to the learners' genders and emotional expressions.

The rest of this paper is organized as follows. Section 2 describes the proposed method. Section 3 presents the experimental work and results. Section 4 concludes the paper.

2 PROPOSED METHOD

Our research aims at exploring a new method for joint classification of sentiments with corresponding gender using multi-modal analysis. Thus, the proposed model analyzes sentiments and recognizes opinions per gender. As a multi-class problem, four classes are defined and detected: negative opinions by females (F_Neg), negative opinions by males (M_Neg), positive opinions by females (F_Pos), and positive opinions by males (M_Pos). The proposed system is evaluated on a video corpus of Arabic speakers.

Figure 1 outlines the general framework of the proposed multimodal analysis. After collecting and preparing the video corpus, various channels are separated and audio is converted to text transcript and some preprocessing steps are conducted. Each audio input is in WAV format, 256bits, 48KHz sampling frequency and a mono channel. Preprocessing operations including normalizing certain Arabic letters written in different forms (e.g. Alefs and Tah Marbotah) are carried out on the associated text transcript. Each video input is resized to 240 × 320 after detecting faces. A feature extractor is developed for each modality. The acoustic feature extractor constructs a feature vector of 68 dimensions for each instance. Moreover, a textual feature extractor is implemented to extract textual features based on Word2Vec word embedding. For each instance, 300 textual features are extracted, i.e. same dimensionality of vectors in Word2Vec pretrained models. The visual feature extraction module extracts 800 features for each input. Various features are then concatenated to form three bimodal systems of audio-text, text-visual and audio-visual modalities, and one trimodal system of audio-text-visual. The generated feature vectors individually and in combinations are used to train two machine learning classifiers: Support Vector Machine (SVM) and Logistic Regression (LR). The following subsections provide more details about the main phases in the system: video corpus preparation, feature extraction and classification.

2.1 Video Corpus Preparation

A video corpus is collected from YouTube. It is composed of 63 opinion videos expressed by 37 males and 26 females. The videos are segmented into 524 opinions distributed as 168 negative opinions expressed by males, 140 positive opinions expressed by males, 82 negative opinions expressed by females, and 134 positive opinions expressed by females. Topics covered in the videos belong to different domains including reviews of products, movies, cultural views,

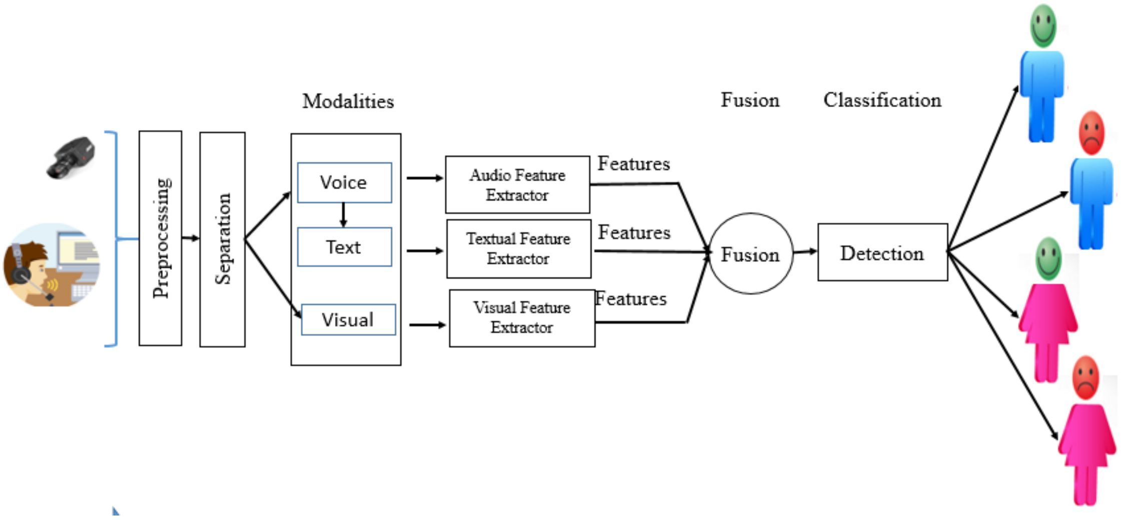

Figure 1: The proposed system pipeline describing the multimodal features combined using machine learning classifiers for classifying sentiment and gender.

etc. They are expressed in six different Arabic dialectics: Egyptian, Levantine, Gulf, Iraqi, Moroccan, and Yemeni using various settings. The collected videos were recorded by users in real environments including houses, studios, offices, cars or outdoors. Users express their opinions in different periods.

2.2 Feature Extraction

In the following subsections, the textual, acoustic and visual feature extraction processes are described in sequence.

2.2.1 Textual Features

For textual features, word embedding based features are employed. Word embedding techniques are recognized as an efficient method for learning high-quality vector representations of words from large amounts of unstructured text data. They refer to the process of mapping words from the vocabulary to real-valued vectors such that elements with similar meaning have a similar representation. Word2Vec word embedding methods (Mikolov et al., 2013a; Mikolov et al., 2013b) efficiently compute word vector representations in a high-dimensional vector space. Word vectors are positioned in the vector space such that words sharing common contexts and having similar semantics are mapped nearby each other. Word2Vec has two neural network architectures: continuous bag-of-words (CBOW) and skip-grams (SG). CBOW and SG have similar algorithms but the former is trained to predict a word given a context whereas the latter is trained to predict a context given a word. Word-embedding based features have been adopted for different Arabic natural language processing tasks and achieved the highest results (Al-Azani and El-Alfy, 2018) comparing to other traditional features such as bag of words.

In this study, a skip-gram model trained using a Twitter dataset with a dimensionality of 300 (Soliman et al., 2017) is used to derive textual features. As illustrated in Figure 2, a feature vector is generated for each sample by averaging the embeddings of that sample (Al-Azani and El-Alfy, 2017).

2.2.2 Acoustic Features

Audio is another important component. The input audio is segmented into short-term frames or windows of length 0.05 millisecond, and each frame is split into sub-frames. For each generated frame, a set of 34 features is computed utilizing pyAudioAnalysis Python package (Giannakopoulos, 2015). These extracted features are: (1) Zero Crossing Rate (ZCR) which is the rate of sign changes of the audio signal during the frame time, (2) Energy which is the sum of squares of the signal values normalized by the frame length, (3) Entropy of sub-frames' normalized energies, which measures the abrupt changes, (4) Spectral Centroid which is the center of gravity of the frame spectrum, (5) Spectral Spread which is the second central moment of the frame spectrum, (6) Spectral Entropy of the normalized spectral energies of a set of sub-frames, (7) Spectral Flux which is the squared difference between the normalized magnitudes of the spectra of two successive frames, (8) Spectral Rolloff which is the frequency below which 90% of the magnitude distribution of the spectrum is concentrated, (9-21) Mel Frequency Cepstral Coeffi-

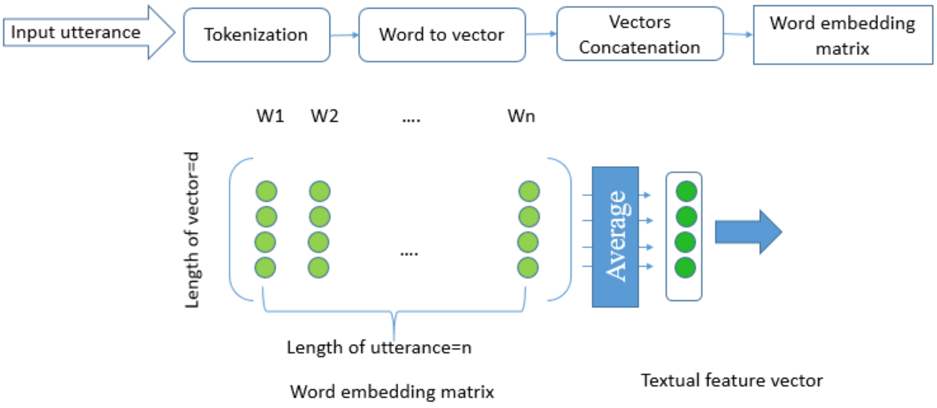

Figure 2: Textual feature extractor.

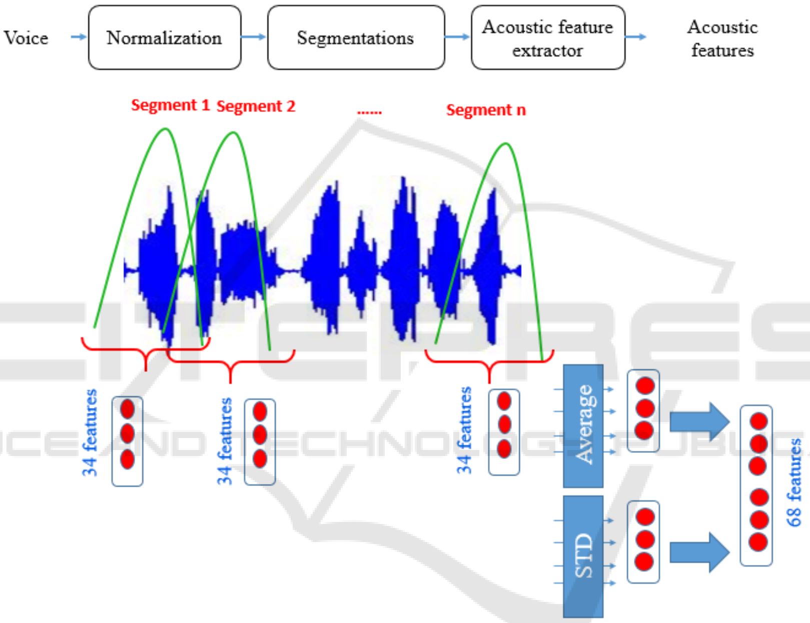

Figure 3: Audio feature extraction.

cients (MFCCs), (22-33) Chroma Vector which is a 12-element histogram representation of the spectral energy, and (34) Chroma Deviation which is the standard deviation of the 12 chroma vector elements. Subsequently, statistics are computed from each audio's segments to represent the whole audio using one descriptor such as the mean and standard deviation in our study. Thus, each input audio is represented by 68 (34 × 2) features (as illustrated in Figure 3).

2.2.3 Visual Features

Figure 4 depicts the general process of visual feature extraction. Facial expressions play very important role in reducing language ambiguity. Recently, it has been shown that face expressions can recognize human's emotional states using a hybrid deep learning model with promising results (Jain et al., 2018). A main step in our study to mine opinions and genders from video is to detect the speaker's face and segment faces from the rest of a given frame. Towards this end, the general frontal face and eye detectors (Viola and Jones, 2004) are utilized. The frontal face detector is based on HAAR features (Viola and Jones, 2001) which combine more complex classifiers in a cascade to detect the face. HAAR feature based face detector is widely used and considered to be very popular with high detection rate (Padilla et al., 2012). Additionally, an eye detector is adopted to locate eye positions which provide significant and useful values to crop and scale the frontal face to 240 × 320 pixels in our case.

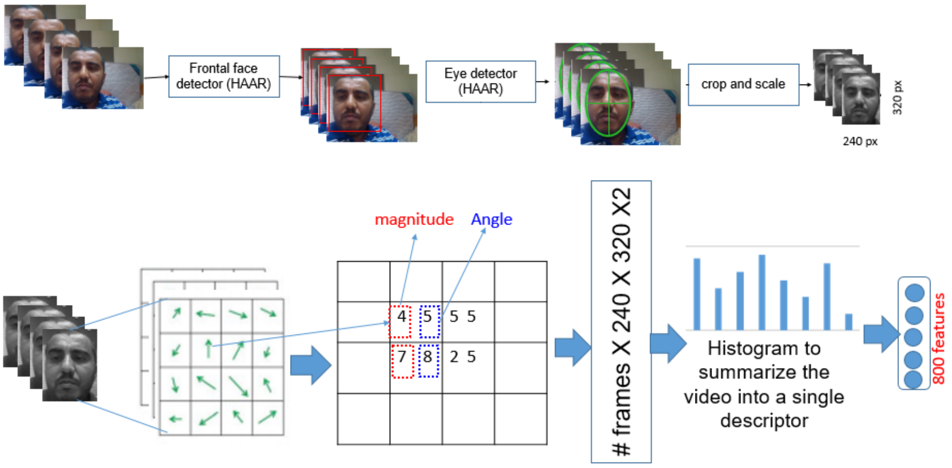

Figure 4: Face localization and histogram of optical flow features extraction.

After segmenting the whole face in successive frames, it is possible to compute the optical flow to capture the evolution of complex motion patterns for the classification of facial expressions. Optical flow is considered to extract the sentiment visual features from the videos processed in the previous step. Optical flow measures the motion relative to an observer between two frames at each point of them. At each point in the image, the magnitude and phase values are obtained which describe the vector representing the motion between the two frames. This results in a descriptor of $NoF \times 240 \times 320 \times 2$ dimensions to represent each video, where NoF refers to the number of frames in a video and 240×320 is the size of one frame. For example, a video of 30 frames is represented by a descriptor of $30 \times 240 \times 320 \times$ dimensions. To describe each video as a single feature vector (descriptor), we need to summarize the generated descriptor of $NoF \times 240 \times 320 \times 2$ dimensions. Towards this end, several statistical methods can be used such as average, standard deviation, min, max, etc. Histogram is considered as a good technique to summarize such descriptors in previous works (Carcagnì et al., 2015; Dalal et al., 2006). In our study, the histogram of the optical flows per video is calculated to summarize the high-dimensional descriptor as a single feature vector.

The scene is split into a grid of $s \times s$ bins, where $s = 10$. The location of each feature is recorded, and the direction of the flow is categorized as one of the eight motions from $\{0, 45, 90, 135, 180, 225, 270, 315, 360\}$. The number of flows belonging to each direction is then counted to end up with $10 \times 10 \times 8$ bins for each frame. Average and max-pooling methods are considered to combine the histograms of various grids in each video.

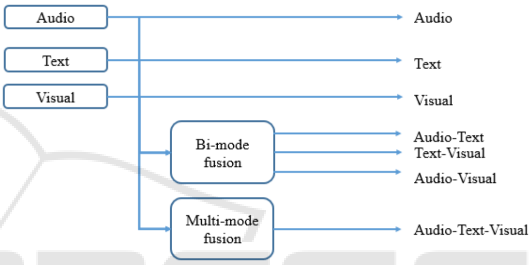

Figure 5: The evaluated models.

2.2.4 Fusion and Classification

For the three modalities: Audio, Textual, and Visual, there are four main possibilities to combine them: audio-textual, textual-visual, audio-visual, and audio-textual-visual. In this work, feature-level fusion is carried out by simply concatenating the extracted features of textual, audio and visual modalities. Mathematically, let $T = \{t_1, t_2, ..., t_n\}$, $T \in R^n$, represent the textual feature vector with length n and $A = \{a_1, a_2, ..., a_m\}$, $A \in R^m$, represent acoustic feature vector with size m and $V = \{v_1, v_2, ..., v_o\}$, $V \in R^k$, represent visual feature vector with size k. T,A and V are combined in various ways and their abilities to detect sentiment-gender are evaluated. Two inherent issues arise when fusing features and need to be addressed, namely scaling and the curse of dimensionality. The former arises due to having different scales for features extracted by different methods. Moreover, some features might be redundant or noisy. These two issues are handled through normalization and feature selection. The features are normalized using min-max scheme:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Table 1: Unimodal and multimodal sentiment-gender joint analysis results using a support vector machine.

Modality	*Rec*	*Prc*	F_1	*GM*	*Acc*	*MCC*
Audio	80.73	80.65	80.62	86.62	80.73	0.7375
Text	68.32	67.65	67.83	77.82	68.32	0.5673
Visual	81.68	81.95	81.70	87.09	81.68	0.7505
Audio-Text	83.02	82.94	82.96	88.42	83.02	0.7690
Text-Visual	84.54	84.89	84.59	89.23	84.54	0.7901
Audio-Visual	91.60	91.68	91.61	94.13	91.60	0.8859
Audio-Text-Visual	90.65	90.71	90.63	93.51	90.65	0.8729

Table 2: Unimodal and multimodal sentiment-gender joint analysis results using a logistic regression classifier.

Modality	*Rec*	*Prc*	F_1	*GM*	*Acc*	*MCC*
Audio	75.19	75.05	75.04	82.84	75.19	0.6628
Text	67.18	67.23	67.19	77.23	67.18	0.5542
Visual	83.40	83.59	83.44	88.40	83.40	0.7742
Audio-Text	82.63	82.52	82.54	88.04	82.63	0.7635
Text-Visual	85.11	85.30	85.14	89.68	85.11	0.7977
Audio-Visual	89.50	89.52	89.49	92.71	89.50	0.8572
Audio-Text-Visual	90.27	90.37	90.21	93.21	90.27	0.8677

where $x^{'}$ is the normalized value corresponding to x which falls in the range from x_{min} to x_{max}. For feature reduction, the principal component analysis (PCA) is applied with the criterion to select the number of components such that the amount of variance that needs to be explained is greater than 0.99. Each form of individual and combinations of modalities is evaluated using (1) LibSVM SVM with Linear Kernel, and (2) LR with L2-norm regularization and Liblinear solver.

3 EXPERIMENTS AND RESULTS

Experiments are conducted on the developed video corpus using 10-fold cross validation mode. Sentiment models are built using the extracted features with the aforementioned classifier. The main components in the proposed system including preprocessing module, feature extraction module, fusion module and classification module are developed on Python. Gensim package is applied for textual features extraction, PayAudioAnalysis package is utilized for acoustic features extraction while OpenCV package is implemented for visual features extraction. Scikit-learn package is used for feature reduction and classification.

Precision (*Prc*), Recall (*Rec*), F_1, Geometric mean (*GM*), percentage accuracy (*Acc*), and Matthews Correlation Coefficient (*MCC*) are adopted to evaluate the proposed models. These measure are computed from 10-fold cross-validation mode. By fitting independent models and averaging results over 10 partitions, the variance is reduced, which is preferable over hold-out estimator to get more reliable measures to compare various models when the amount of data is limited.

Since we deal with a multimodal recognition system, several models will be generated from the considered modalities either standalone modalities, bimodalities or trimodalities. In this study, seven main models are generated per each classifier as illustrated in Figure 5. Three models are generated for audio, text and visual modalities. Three other models need to generated for the bimodal approaches of audio-textual, textual-visual, and audio-visual modalities. The last model is for the trimodal case combining audio, text and visual modalities.

Table 1 shows the results of different models using SVM. For the standalone cases, visual modality achieves the highest results with an accuracy of 81.68% which is followed by audio modality with an accuracy of 80.73%. The lowest performance is achieved by the textual modality with an accuracy of 68.32%. Combining audio with textual modalities leads to improving the results of both of them significantly, in all cases. For example, in the best case of bimodality, the accuracy of audio modality and visual modality went up from 80.73% and 81.68%, respectively, to 91.60% in the case of the bimodal audio-visual recognition system. This is true regarding the multimodal recognition system as well; combining the three modalities causes remarkable improvement of the results over single modalities.

In the case of the LR classifier and standalone modalities, visual modality again reports the highest results with an accuracy of 83.40% which is followed by audio modality that achieves an accuracy of 75.19%. The least results are obtained when using textual modality with an accuracy of 67.18%. Combining the standalone modalities leads to improve the results significantly in all cases using bimodal fusion and trimodal fusion. The highest results of LR are obtained from the combination of the three modalities with an accuracy of 90.27%.

Overall, the highest accuracy of 91.6% is obtained when using audio-visual support vector recognition system.

4 CONCLUSIONS

Joint recognition of gender and sentiment polarity is addressed in this paper as a multi-class classification problem. A video corpus of Arabic speakers is collected and processed. Two machine learning classifiers are evaluated using various modalities. Features are extracted and evaluated individually and after fusion. The experimental work using 10-fold cross validation showed that significant improvements can be achieved when combining modalities and using a support vector machine classifier. As future work, we suggest exploring parameter optimization and deep learning approaches to further improve the results.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the support provided by the Deanship of Scientific Research at King Fahd University of Petroleum & Minerals (KFUPM) during this work.

REFERENCES

Abouelenien, M., Pérez-Rosas, V., Mihalcea, R., and Burzo, M. (2017). Multimodal gender detection. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 302–311.

Al-Azani, S. and El-Alfy, E.-S. M. (2017). Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short arabic text. *Procedia Computer Science*, 109:359–366.

Al-Azani, S. and El-Alfy, E.-S. M. (2018). Combining emojis with arabic textual features for sentiment classification. In *9th IEEE International Conference on Information and Communication Systems (ICICS)*, pages 139–144.

Alexandre, L. A. (2010). Gender recognition: A multiscale decision fusion approach. *Pattern recognition letters*, 31(11):1422–1427.

Carcagnì, P., Coco, M., Leo, M., and Distante, C. (2015). Facial expression recognition and histograms of oriented gradients: a comprehensive study. *SpringerPlus*, 4(1):645.

Cellerino, A., Borghetti, D., and Sartucci, F. (2004). Sex differences in face gender recognition in humans. *Brain research bulletin*, 63(6):443–449.

Dalal, N., Triggs, B., and Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. In *European conference on computer vision*, pages 428–441. Springer.

Garrido-Moreno, A., Lockett, N., and García-Morales, V. (2018). Social media use and customer engagement. In *Encyclopedia of Information Science and Technology, Fourth Edition*, pages 5775–5785. IGI Global.

Giannakopoulos, T. (2015). pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one*, 10(12).

Jain, N., Kumar, S., Kumar, A., Shamsolmoali, P., and Zareapoor, M. (2018). Hybrid deep neural networks for face emotion recognition. *Pattern Recognition Letters*.

Li, M., Han, K. J., and Narayanan, S. (2013). Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. *Computer Speech & Language*, 27(1):151–167.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at International Conference on Learning Representations*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Misirlis, N. and Vlachopoulou, M. (2018). Social media metrics and analytics in marketing–s3m: A mapping literature review. *International Journal of Information Management*, 38(1):270–276.

Padilla, R., Costa Filho, C., and Costa, M. (2012). Evaluation of haar cascade classifiers designed for face detection. *World Academy of Science, Engineering and Technology*, 64:362–365.

Ravi, K. and Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, 89:14–46.

Shan, C. (2012). Learning local binary patterns for gender classification on real-world face images. *Pattern recognition letters*, 33(4):431–437.

Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.-F., and Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14.

Soliman, A. B., Eissa, K., and El-Beltagy, S. R. (2017). Aravec: A set of arabic word embedding models for use in arabic nlp. In *Proceedings of the 3rd International Conference on Arabic Computational Linguistics (ACLing 2017)*, volume 117, pages 256–265.

Thelwall, M. (2018). Gender bias in sentiment analysis. *Online Information Review*, 42(1):45–57.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–I.

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154.