Bimodal Sentiment Analysis Using Textual and Visual Clues

Ahmed Medhat, Mohamed Ashraf Hassan, Mohamed Ahmed Mohamed, Ahmed Samir Ewida, Walid Hamdy

Zewail University of Science and Technology, Egypt

***Abstract*: With the dominance of the internet on all the life areas and the social media channels having an incredible amount of data and information about people either on Facebook, YouTube, etc. More than 10,000 new videos are being posted to those channels everyday. A lot of work has been done to extract information from those videos for many reasons like measurement of customer satisfaction. In this paper, we are working on the fusion of visual and textual sentiment analysis to know the sentimental state of a person, either he is positive or negative or neutral.**

I. INTRODUCTION

Sentiment analysis is the identification of the subjective information in the data. Subjective information defines the attitude of the person in question, the attitude can be emotions, sentiments, opinions, behaviors, and beliefs, about the topic in question. Sentiment analysis determine the polarity of the data, by classifying the data into positive, negative, and neutral classes.

Most of the Sentimental analysis work has been highly going in the direction, of text analysis and natural language processing. while there is a huge rise, in multimedia data in recent years, due to social networks platforms, like facebook, and youtube, where there are millions of hours of videos are uploaded everyday. That leads to the need of doing sentimental analysis, not just for text data, but for other modalities, like visual and audio data.

By integrating sentiment analysis techniques in social robots, it will result in emotionally smarter robots, that are capable of relating to human behavior and emotions. succeeding in that would open a huge window of applications.

there is no much work out there about the multimodal sentiment analysis, and the fusion of extracted information from different modalities. In this paper, we are extracting visual and textual features. we used a dataset developed by Morency et al. (2011). we applied machine learning classifiers, like Naive bayes, support vector machine(SVM), extreme learning machine (ELM).

This paper is structured as follows: Section 2 covers the related work, and the motivation to this paper; Section 3 presents problem formulation and the framework of our approach; Section 4 presents approach we tackled in our work; Section 5 covers performance evaluation and results; Section 6 concludes the paper and mention the future work.

II. RELATED WORK

Sentiment analysis has been done using visual, textual, and audio analysis which is known as multimodal analysis. The process of sentiment analysis can be classified into two main categories, feature extraction for each modality and the fusion of features coming from the two modalities.

video: sentiment analysis using facial expressions

In 1971, a scientist called Ekman (1971) performed a lot of research and studies to detect the emotions and the sentiment state of people through their facial expressions. His research showed that the facial expressions of almost all the people are very similar and express the same emotions. However he used only six basic emotion classes in the studies which are anger, sadness, surprise, fear, disgust, and joy. Such basic affective categories are sufficient to describe most of the emotions expressed by facial expressions. However, this list does not include the emotion expressed through facial expression by a person when he or she shows disrespect to someone; thus, a seventh basic emotion, contempt, was introduced by Matsumoto (1992). Ekman also developed a facial expression coding system (FACS) which express the facial expressions as action units(AU which are specified by the movement of specific face muscles and consists of an AU number, FACS name, and muscular basis. Friesen and Ekman proposed what's called the emotional facial action coding system (EFACS) which defines the sets of action units that participate in building the facial expressions of specific emotions. Later, Bayesian networks, hidden Markov models (HMM) , and artificial neural networks (ANN) have been used in sentiment analysis too (Ueki 1994). Multimodal sentiment analysis had been introduced afterwards in which a fusion between more than one modality is performed.

Text: sentiment analysis using textual data

A lot of research has been done on the text either words, phrases, and paragraphs in order to extract the sentiment state. Textual sentiment analysis has many applications such as studying the opinion of people about a

specific thing or issue; they are positive or negative or neutral. Through the fusion with the audio and visual sentiment analysis, researchers like Poria et al. (2016) and Morency et al. (2011) were able to get more accurate results about the emotional state of the person.

III. PROBLEM FORMULATION AND MODELING

3.1. Problem definition

In this paper we are trying to do multimodal sentiment analysis of youtube datasets. where we depend on two types of data, and fuse them together, visual and textual. we want to classify the polarity of the dataset to either, positive, negative, or neutral.

3.2. Dataset

A. YOUTUBE DATASET

youtube dataset (Morency 2011) consists of 48 videos collected from youtube. the videos aren't concentrating on topic, but rather a wide variety of topics. for example the keywords of the videos include: business, toothpaste, cosmetics, opinion, product review, I like, I hate, war..etc.

demographically, the dataset contains 28 males, 20 females. the ages range from 14-60 years old, with a diverse ethnicity, like: caucasian, Asian, Hispanic, Afro-American. all the videos are in English Language. the videos are in .mp4 format with size 360x480. the videos length is from 2 minutes to 5 minutes.

some preprocessing for the videos had to be done, to overcome some common issues, in youtube videos, like the introductory titles and graphics, and the multiple topics videos. for the introductory parts of the videos, simply 30 seconds of the video was cut out. since most online topics includes many topics, for example, you can easily find a video on youtube, where a lady talks about its vacation in last summer, then switches to how good was her meal this evening, or the movie she watched last week. To overcome this issue, all video sequences are normalised to last 30 seconds.

B. Movie Reviews Dataset

Before applying the machine learning algorithms to the transcripts of youtube dataset, we first tested the algorithms to a well established dataset like Movie Reviews Dataset (Pang 2008). the movie documents were labeled based on their sentimental polarity (positive, or negative). The data includes 2000 review, where 1000 reviews are positive, while the other 1000 are negative. All the reviews were written before 2002. with a cap of 20 reviews per author per category. there is 312 in authors total

3.3. MODELING

Each video is divided into several segments. each video segment is converted into images, according the frame rate of each video. then we compute the final feature vector of each video segment, by extracting facial features from all the images in the segment, and then averaging it. the same thing goes for the textual features, get extracted from video transcription.

Then by fusing the visual and textual features, we form a final vector including information from visual and textual data. Applying a supervised classifier on the fused feature vector, to get the overall polarity of the each segment of the video.

As another way of fusion that can be done, decision-level fusion, which applies classification on visual and textual features individually, then take the classification result as inputs, and output the final sentiment polarity.

3.3.1. Feature Extraction

Feature Extraction is a vital step. Most methods for feature extraction currently is based on unigrams, bigrams, trigrams, and dependency features. And to include linguistic knowledge, some use part-of-speech information. Further, researchers used syntactic information in machine learning models with the help of dependency relations (Agarwal 2016).

there are four types of basic features that could be extracted, unigrams, bigrams, bi-tagged, and dependency parsing. For Unigram, and Bigrams, they are basically bag-of-words features, they got extracted by eliminating the extra spaces and noisy characters between consecutive words. for example, if we have a sentence like " this is a very good book". the unigram features will be "this", "is", "a", "very", "good", and "book". Bigrams will be features consisting of two consecutive words, so it will be like that, " this is", "is a", "a very","very good", "good book".

Bi-tagged features are extracted using part-of-speech patterns. Part-of-speech information is used for the extraction of rich sentiment feature, as it's been found in literature that adj. and adv. have a subjective nature, by extracting two words, such that one of them is adjective or adverb (adj-noun, adv-noun,adj-adj,adv-adv...etc), Also,verbs can have sentiment information (verb-noun, verb-adj, adv-verb,...etc).

Dependency features captures the syntactic relations between words. Dependency parsing tree is used to capture information separated by a long distance. For constructing dependency parsing tree, stanford parser can be used. As an example, for a sentence like " the food looks delicious", part of speech tagging looks like: "the_DT food_NN looks_VBZ delicious_JJ". The dependency relations are: det(food-2, the-1), nsubj(looks-3,food-2),root(Root-0, looks-3), and xcomp(looks-3, delicious-4). so the dependency features are food_the, looks_food, looks_delicious.

IV. PROPOSED APPROACH

Although the one gram tends to overlook many important intuition behind the sentences, as it oversimplifies the hidden opinions behind sentences into a set of positive and negative word, yet still it is considered one of the most ubiquitous features used in Sentiment Analysis due to it's simplicity and acceptable efficiency. For this reason, The one gram is used as our features in the textual sentiment classification.

The features vector shall be constructed using a Bag of Words vector containing the one gram words of each input segment. A simple true value is used if the word is present in the segment instead of using the frequency of occurrence of the word as it proved to be more efficient and robust.

Text preprocessing shall first take place, where lemmatization is applied to the words using NLTK wordnet lemmatizer based on their Part of Speech. Most punctuation is removed, as well as words which occur only once in the training set, so as to minimize the number of distracting features which deteriorates the performance.

As for visual features, we used two software products to extract basic facial characteristics from each frame which can be used to extrapolate visual features. The first software is called OKAO Vision, it detects 66 points that represents important coordinates of the recognized face's eyes and mouth. This software also gives us indications of smile intensity in each frame, in addition to horizontal and vertical eye gaze angles. The other software is GAVAM, its output per frame is the time of occurrence of the particular frame in addition to displacement and angular displacement around X,Y,Z axes respectively. Visual features were constructed by calculating simple relations between coordinate points extracted by OKAO Vision that resemble important facial expressions related to expressing emotions. GAVAM displacement output was used without changing it. Visual features vector was constructed so that every row represents an utterance as determined in the dataset. To achieve that, we categorized the frames so that frames existing in the same utterance are put together. Selected features vector is represented in Table1.

Table 01: Visual features vector per utterance.

Features	Source	Description (per utterance)
1	GAVAM	The average displacement of the face w.r.t. X-axis
2	GAVAM	The average displacement of the face w.r.t. Y-axis
3	GAVAM	The average displacement of the face w.r.t. Z-axis
4	GAVAM	The average angular displacement of the face w.r.t. X-axis
5	GAVAM	The average angular displacement of the face w.r.t. Y-axis
6	GAVAM	The average angular displacement of the face w.r.t. Z-axis
7	OKAO	Distance between the inner and the outer corner of the right eye
8	OKAO	Distance between the inner and the outer corner of the left eye
9	OKAO	Distance between the upper and the lower corner of the right eye
10	OKAO	Distance between the upper and the lower corner of the left eye
11	OKAO	Distance between the upper and lower outer corner of the mouth
12	OKAO	Distance between the upper and inner outer corner of the mouth
13	OKAO	Distance between the left and the right corner of the mouth
14	OKAO	Distance between right eye and left eye
15	OKAO	Percentage of frames with smile intensity higher than 75
16	OKAO	Percentage of frames with smile intensity higher than 50
17	OKAO	Percentage of frames with both eye gaze angles less than 10

Afterwards, supervised machine learning classifiers were applied to the feature vectors. Three methods were utilized, Naive Bayes, Linear SVM and ELM. Naive Bayes is one of the simplest classification methods. It is based on applying Bayes theorem yet with assuming the independence of features. Linear Support Vector Machines (SVM) is a linear classifier that is based on maximizing the margin between the two classes, which makes it known also as a maximum margin classifier. Extreme learning machine (ELM) is a feedforward neural network mainly for classification and regression with a single layer of hidden node, where the hidden nodes and the connecting inputs are assigned randomly and are also never updated. 10-fold cross validation was carried out on the dataset with 80-85% of the Youtube dataset as training and 15-20% as testing (as the dataset is too small of less than that as training).

V. PERFORMANCE EVALUATION

Python and Matlab were used for writing the code. Natural Language Toolkit (NLTK) package (Bird 2009), which consists of a set of libraries and programs for statistical and symbolic natural language processing, was utilized in processing the text. LIBSVM/LIBLINEAR package (Chang 2011) (as well as SVM in Sklearn which is based on them) were used for applying the SVM classification.

Now we discuss the results obtained when we apply both the Naive Bayes and the SVM on the textual features in both the Movie Reviews dataset and the Youtube dataset. As shown in Table 02, although the Naive Bayes classification method is one of the most simple classification methods, it gave very reliable results in the Movie Reviews dataset. Moreover, if we looked at the most informative features, we find that

magnificent	1 : -1	=	15.0 : 1.0
symbol	1 : -1	=	14.3 : 1.0

outstanding	1 : -1	=	13.6 : 1.0
vulnerable	1 : -1	=	12.3 : 1.0
ludicrous	-1 : 1	=	11.8 : 1.0
uninvolving	-1 : 1	=	11.7 : 1.0
refreshing	1 : -1	=	11.0 : 1.0
nonsense	-1 : 1	=	11.0 : 1.0
plod	-1 : 1	=	10.3 : 1.0
fascination	1 : -1	=	10.3 : 1.0

which seem to be very indicative keywords.

On the other hand, when Naive Bayes was applied to the textual data of the Youtube dataset, the results were not as satisfying. This becomes more obvious when we look at the most informative words:

up	0 : -1	=	7.2 : 1.0
me	-1 : 1	=	7.1 : 1.0
love	1 : -1	=	5.9 : 1.0
down	1 : -1	=	5.1 : 1.0
some	-1 : 0	=	4.4 : 1.0
then	0 : 1	=	4.2 : 1.0
come	0 : 1	=	4.2 : 1.0
he	-1 : 1	=	3.9 : 1.0
way	1 : 0	=	3.8 : 1.0
good	1 : 0	=	3.8 : 1.0

This comes to many reasons, the first reason is that the number of samples is much lower than the number of features. Moreover, the number of samples is in itself very small to make an effective sample, As it consists of 47 videos which each video averaging to around 6 sentences, which makes it very unlikely to create an inclusive dictionary that could be used to train a representative Bag of Words. This problem may be overcome through the use of a lexicon based unsupervised learning using a sentiment dictionary, but here comes a problem that needs to be taken into account, which is that the data is transcribed from real talks on Youtube and hence there are many informal words and words not taking their proper shape. One last issue is that the Youtube dataset contains subjective sentences as well as many neutral sentences, this issue was treated here through treating the neutral sentences as a third class and doing classification among the three classes of negative, positive and neutral. However, it may be better if it was a two step classification where the first step deals with whether it is opinionated or not, and the second step deals with whether opinionated parts are positive or negative.

SVM was then applied to both datasets, and apparently superior results were obtained as SVM is considered one of the state of the art methods in textual sentiment analysis. However, in the case of the Youtube database, the same problems mentioned above caused the still not very impressive results.

Dataset	Naive Bayes		SVM	
	Training Accuracy	Test Accuracy	Training Accuracy	Test Accuracy
Movie Reviews	97.73%	70.6%	100%	83.2%
Youtube	93.28%	57.14%	94.54%	52.38%

Table 02

We discuss now the experimental results on the visual features vector of the YouTube dataset. Two supervised classifiers, SVM and ELM, were used on the visual feature vector. The best accuracy was achieved by the SVM classifier.

Classifier	Parameter	Mean Train Accuracy	Mean Test Accuracy
ELM	n = 100000 (Hidden neurons)	100%	42.2%
SVM	c = 1000 (cost)	97.4%	45.4%

Table 03

It is recognizable that training accuracy is significantly lower than literature. This could be due to numerous reasons. It needs furthermore debugging to find the source of low accuracy, but the following are some of the possible reasons.

- Visual features extracted by software have noise or low confidence level.
- Model parameters need further tuning.
- Utterance-based classification makes the dataset smaller than needed to make an acceptable model.

VI. CONCLUSION

In this paper we presented a bimodal sentiment analysis study to improve the classification of opinion polarities by utilizing both visual and textual features. In

future work, we plan to expand research area in multi-directions.

First, we concluded that part of the error is due to the relative small size and irregularities of the dataset. So, a new larger and more inclusive dataset needs to be developed. Another improvement is to use Luxand FSDK commercial software in facial features extraction as it showed to provide better results in literature. Including audio features is also expected to increase the accuracy of our model as proposed by literature, as intonations are a great factor in showing the sentiment.

Moreover, The usage of concept based Bag of Concepts instead of the Bag of Words shows the potential to provide superior results, especially that it tries to capture the hidden sentiments within the sentences.

All previous suggestions are aimed to improve the performance of the model. Hopefully, after reaching a sufficient level of accuracy, construction of a real-time product that utilizes automatic transcription and real-time feature extraction and classification can be achieved. It then can be mounted on a mobile Robot to perform real-time opinion collection and analysis in public places.

Bibliography

B. Agarwal, *Prominent feature extraction for sentiment analysis*. Place of publication not identified: Springer, 2016.

S. Bird, E. Klein, and E. Loper, *Natural language processing with Python*. Beijing: O'Reilly, 2009.

P. Ekman, *Universals and cultural differences in facial expressions of emotion*. Lincoln: University of Nebraska Press, 1971.

D. Matsumoto, “More evidence for the universality of a contempt expression,” *Motivation and Emotion*, vol. 16, no. 4, pp. 363–368, 1992.

L.-P. Morency, R. Mihalcea, and P. Doshi, “Towards multimodal sentiment analysis,” *Proceedings of the 13th international conference on multimodal interfaces - ICMI '11*, 2011.

B. Pang and L. J. Lee, *Opinion mining and sentiment analysis*. Hanover, MA: Now Publishers, 2008.

B. Pang and L. Lee, “A sentimental education,” *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04*, 2004.

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, pp. 50–59, 2016.

N. Ueki, S. Morishima, H. Yamada and H. Harashima, “Expression analysis/synthesis system based on emotion space constructed by multilayered neural network” Systems and Computers in Japan, vol. 25, no. 13, pp. 95–107, 1994.

C.-C. Chang and C.-J. Lin. LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011.