Multimodal Sentiment Analysis in YouTube Videos

Joe Ellis and Jessica Ouyang
Columbia University
jge2105@columbia.edu, ouyangj@cs.columbia.edu

I. Introduction

Sentiment Analysis has been a hot research topic within the past 5-10 years with the advent of social media and vasts amount of available text data on-line. People express their opinions on multiple social media sites such as Facebook, Twitter, Instagram, and Vine multiple times a day, and leave reviews on sites such as Yelp and Amazon. Using this data we have been able to poll public opinion and sentiment about a topic in real-time, and learn about the way that society reacts to a topic in ways never thought possible before the age of social media. This type of sentiment analysis has been used to predict the results of election, stock market prices, natural distasters, and conduct intersting social science experiments.

However, most of this analysis has been completed using text techniques. While on-line text is no doubt growing the rapid rise in social media videos from sites such as Instagram, YouTube, and Vine has risen at an alarming rate in recent times. Videos have been shown to have stronger sentiment than text Need References, but sentiment analysis in videos have not been thoroughly explored throughout the literature. We believe that video is a very promising medium for sentiment analysis given the extra modalities avialable to extract information from such as audio and facial features. These audio and facial features are especially useful in videos that contrain frontal faces and people speaking, which are widely proliferated throughout social media with people posting their video reactions to products, political opinions, news events, and other things of interest. By leveraging the audio, visual, and text components within social media posts we believe we can make a more accurate judgement of the sentiment present within the video content.

We structure the paper as follows: Section 2 will describe our dataset and how we gathered it, Section 3 will describe the features that we extracted from the videos, Section 4 will describe the classification scheme that we used for this project, Section 5 will demonstrate our experimented results, and Section 6 holds the conclusion and discussion of results for this work.

II. DataSet

We have created a datset of videos from the popular "React To" video series on YouTube that were all collected from the Fine Brothers YouTube Channel. Need Ref. Each of the videos that we collected have over 1,000,000 views and contain full frontal faces of people reacting to videos shown to them.

I. Description of Videos

Each of these videos begin with an opening montage and intro music playing for the video and the topic of the video is introduced with an opening slide. Once this is finished there is approxmatley 2-3 minutes of footage of different people reacting to some type of video footage that they are watching on a laptop, and appears in the raw video as a “picture-in-picture” frame in the video. After these 2-3 minutes have passed a question and answer segement that lasts 5-7 minutes begins, where a voice coming from an entity is not on screen asks question to the people who appeared on screen during the opening 2-3 minutes. The people on screen then answer the questions with often times outrageous actions, sarcasm, and other forms of strong reaction. The videos then end with a short 30 second clip advertising the next react to video. We use the entire video in our classificatio scheme, although sections of the video do not have a conclusive sentiment.

II. Data Collection and Processing

We used the open-source community developed executable, youtube-dl ref and interfaced with Google's YouTube API To collect these YouTube videos we developed a command line program to downlaod any video from YouTube and transcode the raw HD video format down to standard definition (640x480). After transcoding we also download the speech transcript generated by Google ASR Need Ref and made available by YouTube. We also download all of the metadata that is present with the YouTube video including how many views, when it was uploaded, the author, if it was every modified, etc. Finally, we also download up to 300 of the most recent comments for each YouTube video, we hoped to be able to use the comments to aid our ability to determine sentiment within the video. However, we found that it was to difficult to determine what portion of the video each comment was referring to, amd therefore were not able to use the comments within the videos.

Finally, because each video was approximately 10 minutes long and a variety of sentiments were expressed we within the video we chose to cut each video into speaker segments, which are sections of the video generally 5-10 seconds long where one speaker is giving an opinion. To cut the long video into these smaller segments we performed speaker diarization on the video using the "SHoUT" open source toolkit need reference. Using this technology we were able to cut the video into smaller segments, which had more coeherent sentiment throughout each specific segment. After speaker diarization and cutting of the videos we were left 1434 video samples that we used for our complete dataset.

III. Feature Extraction

Jessica

I. Text Features

Jessica

II. Visual Feature

We extracted visual features from the face towards the task of building a visual classifier based on extracted facial features. First we detected the faces using the standard and widely implemented Viola-Jones face detector need reference, on frames subsampled every .5 seconds per video and only keeping faces that had a pixel area greater than 10,000 pixels. Once the faces were detected we then found landmark points on the face using the output of structured SVM . The landmark points that we detected were the corners of both the right and left eye, tip of the nose, and the corners of the month. After detecting the points we performed an affine warp on the faces so that each face would be properly aligned, and then extracted SIFT Need Reference features from each of the points. This formed our 896-dimensional feature representatoin for each of the faces that we found within the videos.

III. Audio Features

We also extracted audio features from each of the videos within our dataset, and we extracted a variety of different audio features that targeted different portions of human speech. We extracted energy features, pitch features, and speech-like features from the speech. We cut the audio into smaller windows with 20ms durations and 10ms overlap and then extracted features from each of the windows. We then took found the mean and standard deviation across the features for each window, and used this to constitute our entire feature vector for a video segment. For energy features we extracted the energy below 250 Hz within each window and the total energy within the video. We extracted the fundamental pitch from each window for the total frequency range, and then above and below 1500 Hz, this was shown as a useful feature in audio emotion detection Need Ref. Finally, we extracted 13 MFCCs from each window for speech like features. As stated above, we then found the mean and standard deviation of each feature across all of the windows of a video and then used this as the 72-dimensional feature representation for each of the videos.

IV. Classification

Jessica

V. Experiments

Joe

I. Single Classifier

Joe and Jessica

II. Co-Training Classification

Joe and Jessica

VI. Discussion

Joe or Jessica