Exploiting Multimodal Affect and Semantics to Identify Politically Persuasive Web Videos

Behjat Siddiquie
behjat.siddiquie@sri.com
SRI International

Dave Chisholm
davec@cs.columbia.edu
Columbia University

Ajay Divakaran
ajay.divakaran@sri.com
SRI International

ABSTRACT

We introduce the task of automatically classifying politically persuasive web videos and propose a highly effective multimodal approach for this task. We extract audio, visual, and textual features that attempt to capture affect and semantics in the audio-visual content and sentiment in the viewers' comments. We demonstrate that each of the feature modalities can be used to classify politically persuasive content, and that fusing them leads to the best performance. We also perform experiments to examine human accuracy and inter-coder reliability for this task and show that our best automatic classifier slightly outperforms average human performance. Finally we show that politically persuasive videos generate more strongly negative viewer comments than non-persuasive videos and analyze how affective content can be used to predict viewer reactions.

Categories and Subject Descriptors

H.3.3 [[**Information Search and Retrieval**]: Search process; I.2.10 [**Vision and Scene Understanding**]: Video analysis

General Terms

Algorithms; Design; Experimentation

Keywords

Video Classification; Affect Recognition; Audio Concepts; Video Concepts; Sentiment Analysis; Multimodal Fusion

1. INTRODUCTION

In the last few years social media has rapidly emerged as an effective means to disseminate information to a large and geographically diverse audience. Its low barrier of entry allows not just well funded organizations but also individuals to share and propagate their opinions and viewpoints. Multimedia content in particular can both express and evoke strong emotional responses, and there is reason to believe that audio-visual content can affect viewers more strongly than text based content [16]. Viewer reactions are also visible in the form of comments, and so in this way social media allows content creators and uploaders to spread a strong, compelling message and observe its impact.

For these reasons, multimedia content on social networks is particularly suitable to propagate political views, and it is used for public influence, political persuasion and even radicalization [10]. It is thus a potent tool to influence and attract new followers, and there is great interest in many sectors in detecting and assessing politically charged or otherwise persuasive social media content.

In this paper, we propose a solution to detect such politically persuasive videos posted on social media and also develop methods to predict and analyze the sentiment of comment responses. In order to analyze the audio-visual content present in the videos we focus on robust extraction of the semantic and affective information. For the audio domain we detect several grades of speech arousal and related semantic categories such as crowd reaction and music. For the visual domain, we detect visual sentiment and semantic content. Finally we analyze the sentiment of comments associated with a video to determine viewer reaction. Our classification results indicate that politically persuasive videos posted online can be reliably detected based on extracted affective and semantic information. Furthermore, we show that politically persuasive videos often generate more negative reactions among viewers and the overall sentiment of reactions can be predicted with a reasonable degree of success solely from audio-visual features. Figure 1 shows an overview of our approach.

There are two **key contributions** of our work:

- We demonstrate that affective and semantic information extracted from the audio, visual and textual content associated with a video can be used to predict whether the video is of a politically persuasive nature.

- We further show that politically persuasive videos generate more negative reactions in comments, that affective content can indicate how viewers may react, and that viewer reaction to semantically similar videos varies with respect to the videos' affective content.

We would like to emphasize that the goal of our work is to demonstrate the utility of multi-modal affective features to reliably identify politically persuasive media content, rather than to develop sophisticated or novel content understanding techniques. In particular, we show that an approach leveraging these affective and semantic features can generally classify such videos as well or better than humans. We

ICMI 2015, November 9–13, 2015, Seattle, WA, USA.

DOI: http://dx.doi.org/10.1145/2818346.2820732.

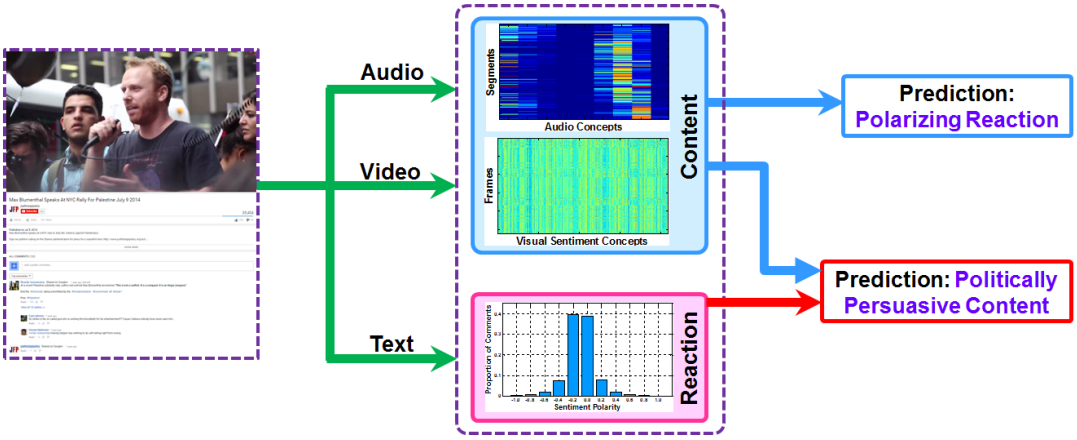

Figure 1: Audio and visual concept scores and features are computed from the audio-visual data of a web video. These scores are used to predict whether the video is politically persuasive and whether the comments in response will be highly polarized. In addition, the actual comments are also used to predict whether a video is politically persuasive, and all three modalities are fused for the best prediction results. In this example our system correctly predicts the video is both politically persuasive and likely to have polarized comments based solely on audio-visual content. (*https://www.youtube.com/watch?v=gDpcsx2CL9E*)

believe that our work opens up an exciting new area of research with a number of interesting applications.

2. RELATED WORK

There has been some recent work on detecting persuasion. Park et al. collected and annotated a corpus of movie review videos in [31]. From this data, they demonstrated that the verbal and non-verbal behavior of the presenter is predictive of how persuasive they are as well as predictive of the cooperative nature of a dyadic interaction. Our work differs from this in three key ways. First and foremost, while that prior work attempts to predict whether a particular sample will persuade a viewer (e.g. "is this persuasive") we attempt to classify whether a video contains a particular type of persuasive content (e.g. "does this video contain fiery political rhetoric"). Secondly, we rely not only on the affective characteristics of a speaker, but also on the affective characteristics of viewers' reactions as well as semantic content unrelated to a main speaker (e.g. crowd reaction, background music). Finally, we analyze a corpus that is collected from a variety of unscripted situations recorded both indoors and outdoors and whose audio-visual quality ranges from good to poor, whereas the corpus in [31] consists of data captured in controlled environments.

Detection of radical online content is a growing research area due to the interest shown in it by law enforcement agencies; Correa presented a survey of such work in [10]. This detection is typically done in two ways: by analysis of the network structure or by analysis of the content itself. Reid et al. presented an network based approach in [32] that capitalized on the notion that websites that promote the same ideology tend to be interconnected by hyperlinks. Most approaches that analyze content, such as the one presented by Abbasi in [1], tend to focus on the textual and linguistic content of websites rather than audio-visual content of videos. To the best of our knowledge, Fu et al. presented the only published work on detecting extremist videos in [14], and their approach used the meta-data associated with the video rather than the actual audio-visual content. In contrast to these works, our approach is independent of any network structure and exploits the audio-visual content itself as well as comments in reaction to it. Moreover, our target class of politically persuasive videos is broader than Fu et al.'s class of extremist videos.

There has been some work on sentiment analysis of visual content. This includes analysis of image aesthetics [20], analysis of image interestingness and memorability [17] and analysis of image affect and emotion [28]. There has also been work on detecting visual sentiment expressed through human behavior - *e.g.* facial expressions [37] - as well as work on detecting image style [22]. We rely directly on two prior works for visual content analysis. First, we use Jia et al.'s CAFFE implementation from [18] which was trained on concepts presented as part of ImageNet in [12]. Second, we also use the Visual Sentiment Ontology presented in [5], which consists of a large-scale ontology of semantic concepts correlated with strong sentiments.

Detecting affect from human speech has grown in importance because of its applications in enabling realistic human interactions [27, 33]. However, most of the work in this area has been demonstrated on datasets collected in controlled environments with little noise. Similarly detecting semantic concepts in audio has long been an important problem in the audio processing community due to its applications in video retrieval [26, 6, 7]. Our audio concept detection, based on [9] combines the best of both these approaches.

Sentiment analysis and opinion mining from text have long been active research topics with varied applications such as analyzing product reviews, improving question answering systems, and mining political opinions. This area is well covered by Pang and Lee's survey in [30]. Recent work, such as Socher et al.'s approach in [35], has improved the accuracy of text based sentiment analysis significantly. There has also been recent work focusing exclusively on text in social media settings, such as Mohammad et al.'s sentiment analysis of Twitter data in [29]. We employ both of these recent approaches to analyze and gauge user reactions and opinions to videos in the form of comments.

There has also been work on combining information from multiple modalities to better detect affect [21, 3]. The key advantage of multimodal affect detection is that different modalities contain complementary information and therefore fusing the information from them can lead to performance improvements in detecting affect. We also adopt a

multimodal fusion approach to detect persuasive videos, and fuse not only modalities based on audio-visual content (e.g. speech characteristics, crowd response, visual semantics) but also on viewer response (e.g. sentiment of user comments).

3. DATASET AND CLASSIFICATION TASK

For our experiments we use the *Rallying a Crowd* (RAC) dataset, introduced in [9]. The RAC dataset consists of 230 videos from YouTube comprising over 27 hours of content. The dataset contains 132 positive and 98 negative examples of politically persuasive videos. Positive videos contain a speaker vehemently rallying/persuading a crowd for political causes. Negative videos also contain speakers and crowds but lack political themes. Qualitatively, positive examples are often events such as political rallies or protests. Negative examples contain semantically similar content - e.g. speakers addressing crowds - but lack the highly affected, politically persuasive nature of the positive examples, and are often events such as presentations or entertainment events.

Thus the classification task for these videos is neither one of determining speaker valence nor whether a viewer is actually persuaded, but instead is a binary classification problem of determining whether politically charged videos can be distinguished from videos with ostensibly similar audio-visual content but lacking political content and vehemence. The videos in the dataset were recorded under a wide variety of conditions (e.g. outdoor vs. indoor, near vs. far) with various levels of post-production (e.g. professional quality clips vs. unedited amateur footage) and include samples with multiple languages and settings.

3.1 Annotation and Human Accuracy

To quantify the difficulty of classifying politically charged content and validate it as a legitimate identifiable class we had four human annotators perform binary labeling of positive and negative RAC videos. This also allows us to compare the performance of our automatic approach to human performance. Ground truth was established by using particular search terms and establish a consensus decision for each video among all the authors. We had four annotators first view a "control" portion of the RAC dataset consisting of 52 positive and 40 negative videos to form an opinion of what constituted politically persuasive content. Each annotator then chose a binary label for each of the remaining 80 positive and 58 negative videos. These 138 videos were the same portion used in the binary classification experiments for automatic classifiers as detailed below.

The four annotators had accuracies of 71.0%, 74.6%, 75.4% and 100.0% ($p < 0.001$ for each) and improvements over chance (I_{oc}) of 31.0%, 39.6%, 41.5% and 100%. I_{oc} represents how many fewer errors an annotator is expected to have versus a chance annotation and is defined as $(A_o - A_e)/(1 - A_e)$, where A_e is expected accuracy by chance. We attribute the perfect performance of the the fourth annotator to the extra diligence and time this annotator took compared to the others, and feel it supports the validity of the ground truth. An inter-coder agreement coefficient of 0.524, indicating "moderate agreement", was measured among all annotators using Fleiss's kappa [13, 24]. Inter-coder agreement between all possible pairs of annotators was also measured using Scott's pi and varied from 0.396 ("fair agreement") to 0.658 ("substantial agreement") [34, 24]. We feel the overall classification success, improvements over chance and annotator agreement validate the class of "politically persuasive" videos as coherent and classifiable.

4. EXPERIMENTS

4.1 Experimental Setup

The same setup involving 138 videos (80 positive, 58 negative) was used for each automatic classification experiment. In order to ensure that any results were not influenced by the larger size of the positive class and to ensure that a random classifier would have a classification accuracy of 50%, we randomly sample 58 positive videos from the group of 80 to create a set of 116 videos with an equal number of positive and negative samples. We then classify these 116 videos using 10-fold cross-validation. This random sampling was performed 25 times for each of the modalities listed below, and final results displayed represent the mean and standard deviation of the accuracy over 25 such randomized runs of 10-fold cross validation. For each modality, an RBF SVM classifier was used to perform the final binary classification task. Details on preprocessing and extraction for each modality follow.

4.2 Classification via Audio Analysis

We first investigated the use of the audio content in the videos to distinguish between persuasive and non-persuasive videos. Politically persuasive videos are often characterized by animated speakers, charged content, crowd response and occasional music. Hence, detecting these audio categories would enable us to identify persuasive videos. In order to do so, we use the audio concept detection approach of [9], which we briefly describe in Section 4.2.1.

4.2.1 Audio Concept Detection

Following the work of [9], we detect the following audio concepts: *Crowd*, *Music + Crowd*, *Music*, *Music + Speech*, *Crowd + Speech*, *Calm Speech*, *Slightly Agitated Speech*, *Agitated Speech*, *Very Agitated Speech*. To do so, we first extract a low level feature vector (either prosody, MFCC, or spectrogram based) for each frame of audio using a sliding window approach. Since a single concept label will very rarely apply to an entire audio track, we apply an algorithm based on the unsupervised Simple Linear Iterative Clustering (SLIC) algorithm for image segmentation [2]. The algorithm was adapted for audio and temporally segments the track into a set of homogenous segments likely to contain a single concept. The approach has been shown to work well for semantic concepts on the CCV dataset [19] as well as for affective concepts on the RAC dataset. The low level features for each segment are then encoded into a bag-of-words representation and used to train a non-linear SVM for concept detection. Note that these concept detectors are trained and tested on a disjoint "control" portion the RAC dataset comprised of 52 positive and 40 negative samples. No video was used to create both the concept detectors and the classifiers.

4.2.2 Binary Classifier Setup

Given a set of n videos $\mathcal{V} = \{V_1, V_2, V_3, \ldots, V_n\}$ and their corresponding binary labels $\{y_1, y_2, y_3, \ldots, y_n\}$ that indicate whether the video contains persuasive content, we trained a politically persuasive versus non-persuasive classifier as follows. For each video V_i, we segment the audio and then compute the concept scores as described above. We denote the audio concept scores for video V_i as O_i, where i denotes the video index. The dimensionality of an audio concept score O_i is $T \times C$, where T is the number of segments (dependent on the length of the video and segmentation behavior) and C is the number of audio concepts (fixed at 9). We quantize O_i by linearly resizing it to $T_{fixed} \times C$, where $T_{fixed} = 100$,

to obtain $\bar{O}_i$. Now corresponding to each video V_i we have a fixed dimensional feature $\bar{O}_i$. We train an RBF SVM on this data for classification.

This process was performed separately for each of the three low level features (prosody, MFCC, or spectrogram) and the results for each are shown below. Additionally, the three feature types were combined using a Multiple Kernel Learning (MKL) approach [4]. The process was also repeated for using three different temporal scales for the SLIC segmentation algorithm, but very little performance difference was observed between them. The results shown below correspond to a medium scale segmentation.

4.2.3 Results

Following the experimental setup detailed in 4.1, we achieve the results shown below in Table 1. The results indicate that using audio features one can perform reasonably well at detecting persuasive content, and that the spectrogram feature performed best among the possible low level feature types.

Audio Concept Features	Classification Accuracy
Prosody	77.82 ± 12.79
MFCC	73.64 ± 12.17
Spectrogram	81.03 ± 11.93
All Features (MKL)	78.65 ± 11.76

Table 1: SVM classification performance using the audio concept detection features. The margins represent the standard deviations across 25 randomized runs.

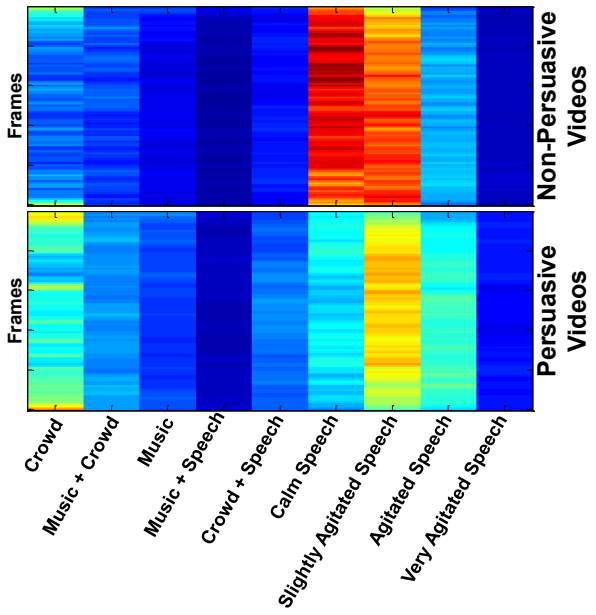

Figure 2: Mean concept detection scores for positive and negative videos. Notice that positive videos tend to have a higher scores for *crowd* and *agitated speech*, while negative videos have higher scores for *calm speech* and *slightly agitated speech*. (Best viewed in color - red/yellow indicate higher scores and blue indicates lower scores.)

4.2.4 Analysis of Results

In order to better understand what the classifiers were learning for distinguishing between the persuasive and non-persuasive videos, we looked at the mean features $\bar{O}_i$ from the persuasive and non-persuasive videos (Figure 2). The figures show that persuasive videos contain a stronger response for *crowd* and *agitated speech*. On the other hand, non-persuasive videos contain a stronger response for *calm speech* and *slightly agitated speech*. Considering the charged content and oratory of many of the persuasive videos, this matched our intuition of how the content would be distributed. Finally, we believe that the concept scores computed from the spectrogram features lead to the best results because they detect the relevant concepts such as *crowd* and *agitated speech* more accurately compared to other features leading to superior results on the classification task.

4.3 Video Analysis

In this subsection, we investigate the use of concepts based on visual content that could be used to differentiate between persuasive and non-persuasive videos. Persuasive videos from the RAC dataset often contain rousing visuals ranging from cheering crowds and heroic figures to images of graphic violence. Overall, the visuals are more striking and extreme than the non-persuasive content, and hence visual features may be useful to identify persuasive content. We primarily focus on deep learning based features that identify semantic concepts and sentiment. Recently deep learning models such as Convolutional Neural Networks (CNNs) have become extremely popular for learning image representations [23]. CNNs, loosely inspired by human vision, are variants of multilayer perceptrons consisting of multiple convolutional and pooling layers followed by fully connected layers [25]. In our work, we implemented a standard network [23] using the popular open source framework CAFFE [18] to extract visual features. We trained two different networks using different datasets. The first network was trained on the ImageNet dataset [12] and the second was trained on the Visual Sentiment Ontology dataset [5]. We next describe each of these processes along with the obtained experimental results.

4.3.1 ImageNet Concepts

In this case, we want to evaluate whether the presence or absence of certain semantic concepts in the video indicates the presence of persuasive content. In order to estimate the presence or absence of concepts in an image we use the CNN trained on the ILSVRC-2012 dataset which is a subset of ImageNet, consisting of around 1.2 million images labeled with 1000 different classes ranging from elephant to space shuttle to stethoscope. The network was trained to maximize the multinomial logistic regression objective for these classes over the training data. We use this 1000 dimensional output as a feature indicating the presence or absence of each class. We also use the outputs of intermediate network layers - which represent more abstract visual features than the final concept outputs and can often provide high classification performance - as features. These three features are referred to as "prob", "fc7" and "fc8" respectively. We extract each of these features from every 30th frame (1 sec.) of the video. As in case of the audio data we now have a vector of dimension F x D (where F is the number of frames sampled) and D is the dimensionality (1000 in case of the prob and 4096 in case of "fc7" and "fc8"). Since F varies based on the length of the video, we linearly resize our feature vector to $F_{fixed} \times D$, where $F_{fixed} = 100$, much like was done for audio in section 4.2.2. Each of the three features were provided individually to an RBF SVM based classifier, and all features were also combined by concatenation. We follow

the same experimental protocol as in the earlier experiments described in section 4.2.3. The results are shown in Table 2.

Features	Classification Accuracy
prob	68.52 ± 3.62
fc8	67.61 ± 3.97
fc7	69.62 ± 4.48
All Features	71.81 ± 5.37

Table 2: Classification performance using different features extracted from the CNN network trained on ILSVRC-2012 dataset [12] for recognizing object categories. "prob" refers to the final concept output scores, whereas "fc7" and "fc8" are scores for features from intermediate layers of the CNN.

Classification using each of the features significantly outperforms chance ($p < 0.001$ for each), indicating that visual information is indeed useful for detecting persuasive content. We can also see that the features from the "fc7" layer - which contain more abstract, learned features - outperform the features from the "fc8" and "prob" layers indicating the utility of these over the more specific final concepts. Finally, we can also see that combining the features from all of the layers results in a modest improvement in the performance.

4.3.2 *Visual Sentiment Ontology Concepts*

In this case, we want to evaluate whether the presence or absence of certain visual sentiment concepts in a video can provide information on whether the video contains persuasive content. In order to evaluate this we use the Visual Sentiment Ontology dataset [5] which consists of approximately 930k images. This dataset was collected by searching Flickr for Adjective-Noun-Pairs (ANPs) such as "beautiful flower" or "disgusting food". The advantage of using these ANPs is that they relate particular images of sentiment neutral nouns (e.g "flower") to a strong sentiment by adding an adjective (e.g. "beautiful flower"). Thus the concepts capture both semantic and sentiment information. We used the latest version of Visual Sentiment Ontology DeepSentiBank [8] which consists of CNN based concept detectors for 2089 ANPs. These concept detectors are trained using the same deep learning network [23] described above. As in the earlier case, we use the final "prob" outputs as well as the intermediate layer outputs "fc7" and "fc8" as inputs to our classifier. The results are shown in table 3. Here again we can see that the intermediate layers "fc7" and "fc8" perform better than the final outputs ("prob"). The performance of the SentiBank concepts is slightly better than the performance of the ImageNet concepts, which could be due to use of sentiment in the concepts, or just due to the higher number of concepts present in the SentiBank dataset. Finally, we also looked at combining the features from the ImageNet concepts and the SentiBank concepts, but this did not lead to any meaningful increase in performance.

Features	Classification Accuracy
prob	68.46 ± 3.61
fc8	72.40 ± 2.76
fc7	73.95 ± 2.95
All Features	73.13 ± 3.01

Table 3: Classification performance using different features extracted from the CNN network trained on the Visual Sentiment Ontology dataset [5] for recognizing visual sentiment.

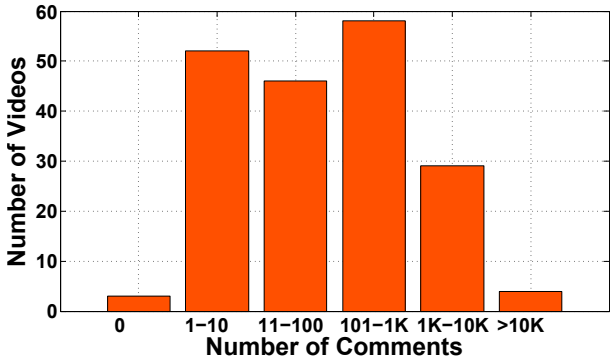

Figure 3: The distribution of the number of comments across the videos in the dataset.

4.3.3 *Analysis of Results*

Ideally, highly scored concepts in positive videos could be used to explain why a video was classified as politically persuasive. In practice we noted a mix of accurate and relevant concepts (e.g. "bad guy" in a violent video) along with concepts that seemed irrelevant. Moreover, we noted that the classification was more accurate using intermediate features than the high level concepts. Thus while the visual concepts are effective for classification, in order to explain classification, a more specific set of visual concepts would need to be used, or detection would need to be improved.

4.4 Text Analysis

We also investigated the use of text associated with the videos, in particular viewer comments. Videos uploaded to YouTube and other video-sharing sites often generate a large number of comments posted by viewers across the world, and many of these comments contain reactions of people to the videos. Intuitively comments generated in response to politically persuasive videos should be of a more polarized nature while other videos generate comments of a more neutral or positive nature. Therefore, exploiting the sentiments contained within these contents should provide us with additional information as to whether the video contains politically persuasive content.

Given a YouTube video we extract all of the associated comments using the YouTube API. The number of comments for to each video varies greatly (Figure 3). Since our videos comprise a geographically diverse range of topics and speakers, many of the comments to some of these videos are in languages other than English. As a pre-processing step we automatically filter out non-English text. In order to do so we use a simple approach of counting the proportion of the words in each comment that are present in a dictionary learnt from a standard English text corpus [36]. Despite its simplicity, we empirically found that this approach worked well for filtering out non-English text in YouTube comments, regardless of the character set of the comments or any spelling or compositional errors. To then detect the sentiment of each comment, we experimented with two different approaches as described below.

SATSVM: We refer to the approach in [29] as Sentiment Analysis of Tweets using SVMs (SATSVM). SATSVM has been specifically developed for social media data such as tweets and so it is appropriate for our data. The approach relies on extracting a number of features from each comment and training an SVM to classify the comment as having a positive or negative sentiment. We implement a simplified

version of this approach and instead of using the binary SVM outputs, we use the SVM decision scores which roughly indicate the degree of positivity or negativity in the sentiment.

DeepCompositionalModel: We refer to the sentiment detection approach presented in [35] as the DeepCompositionalModel. It uses a Recursive Neural Tensor Network to build a representation of sentences based on their structure and computes sentiment by accounting for how the constituent words compose with each other. Unlike SATSVM, the DeepCompositionalModel splits each comment into its sentences and assigns a separate sentiment score to each sentence. Its output is a 5 dimensional probability vector indicating the probability of the sentence being *Strongly Negative*, *Negative*, *Neutral*, *Positive* or *Strongly Positive*.

4.4.1 Detecting Persuasion based on Sentiment

Given a video V_i and the set of associated comments $\mathcal{C}_i$ consisting of N individual comments $\{c_{i1}, c_{i2}, c_{i3}, \ldots c_{iN}\}$, we run SATSVM on each element of $\mathcal{C}_i$ to get a set of N scores $\{x_{i1}, x_{i2}, x_{i3}, \ldots x_{iN}\}$ normalized within the range $[-1,\ 1]$. We then quantize these scores by binning them into a histogram consisting of eleven equally spaced bins. Using this technique, each video V_i can be represented by a fixed dimensional histogram H_i. We train a RBF SVM using these histogram features for classifiying videos into persuasive versus non-persuasive. The classification results are shown in Table 4.

Similarly, when using the DeepCompositionalModel, we extract the sentiment for each comment $\{c_{i1}, c_{i2}, c_{i3}, \ldots c_{iN}\}$ obtaining $\mathcal{X}_i = \{\mathbf{x_{i1}}, \mathbf{x_{i2}}, \mathbf{x_{i3}}, \ldots \mathbf{x_{iM}}\}$, where $M(> N)$ is the total number of sentences. (Each comment is split into one or more sentences.) Each $\mathbf{x_{ij}}$ is a 5 dimensional probability vector as described above. Each video V_i is now represented by a set of these features $\mathcal{X}_i$. We train an SVM using a pyramid match kernel [15], which has been shown to be very effective for learning with sets of features, for classifying these videos. The results are shown in Table 4. From the results we can see that both SATSVM and DeepCompositionalModel perform similarly and are significantly better than random ($p < 0.001$ for each).

Approach	Classification Accuracy
SATSVM	69.52 ± 4.31
DeepCompositionalModel	69.94 ± 3.98

Table 4: Performance using sentiment features extracted from SATSVM [29] and DeepCompositionalModel [35].

4.4.2 Analysis of Results

We now try and analyze some of the results in an attempt to understand how sentiment analysis helps in detecting persuasive videos. To analyze the results from SATSVM, we looked at the mean normalized histogram of the positive as well as the negative videos. The histograms in Fig. 4 show that politically persuasive videos contain a higher proportion of negative comments than the non-persuasive videos.

We plot similar histograms for the sentiment detection results obtained from DeepCompositionalModel (Fig. 5). Here we can see that persuasive videos have a larger proportion of negative comments compared to non-persuasive videos. Therefore we can see that while the two sentiment extraction approaches result in different distributions, they tend to support the hypothesis that persuasive videos lead to more negative and fewer positive comments when compared to non-persuasive videos and we can use this to help automatically distinguish between politically persuasive and non-persuasive videos.

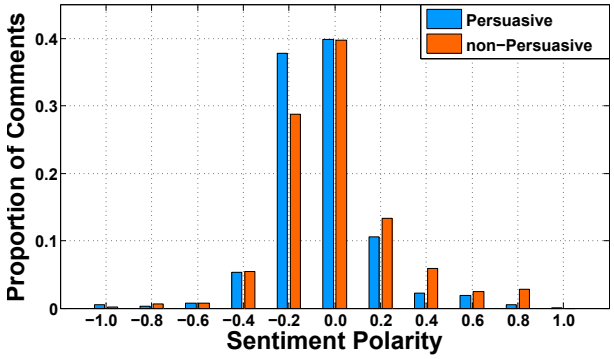

Figure 4: Mean sentiment histograms for the politically persuasive and non-persuasive videos based on SATSVM [29]. Persuasive videos tend to have a higher proportion of negative comments (bin '-0.2').

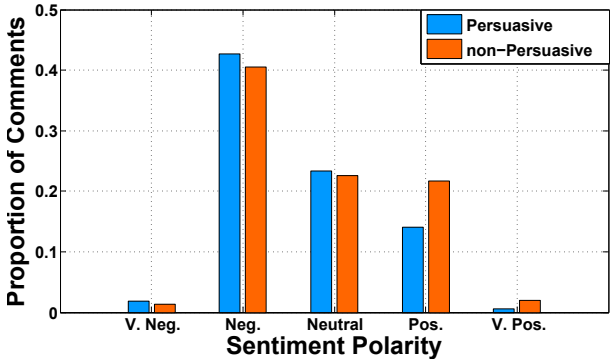

Figure 5: Mean sentiment histograms for the politically persuasive and non-persuasive videos based on the DeepCompositionalModel [35]. Persuasive videos tend to have a lower proportion of positive comments (bin 'Pos.').

4.5 Multimodal Fusion

We next looked at fusing the information from the audio, visual and text modalities. We believe that these modalities contain complementary information and therefore fusing them should boost the overall classification performance. For fusion we considered three different fusion strategies - *Early Fusion*, *Simple Late Fusion* and *Learning based Late Fusion*. For the purpose of fusion, we use the spectrogram features for audio (Table 1), the "fc7" features from the sentiment ontology for video (Table 3) and the SATSVM features for text (Table 4). In case of *Early Fusion* we simply concatenate the features from all of the modalities and train a RBF SVM for classification. In *Simple Late Fusion*, we add up the decision scores obtained from each modality to arrive at a composite decision score to perform classification. For *Learning based Late Fusion*, we train a logistic regression based fusion that combines the decision scores from each modality in a weighted manner. The results are shown in Table 5 and they demonstrate that fusion tends to improve classification results over the individual modalities and late fusion is more effective than early fusion. Furthermore, *Learning based Late Fusion* leads to further performance gains over *Simple Late Fusion*.

4.5.1 Significance of Multimodal Approach

Modality	Classification Accuracy
Audio	81.03 ± 11.93
Video	73.13 ± 3.01
Text	69.52 ± 2.31
Early Fusion	81.37 ± 1.47
Simple Late Fusion	83.32 ± 3.40
Learning based Late Fusion	**85.09 ± 1.83**

Table 5: Classification performance on the RAC dataset using different fusion techniques.

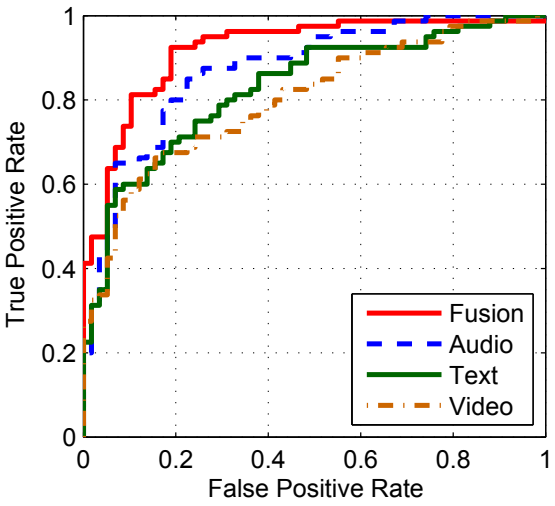

Figure 6: ROC curves for each modality.

In order to further compare the performance of different modalities and establish the significance of a multimodal approach over individual modalities, ROC curves and corresponding statistics were computed for each modality as well as learning based late fusion. The ROC curves for each modality are shown in Figure 6 and were derived from the mean rankings of samples over multiple randomized experiments as noted in Section 4.1. For each ROC curve the area under the curve (AUC) was computed, and for each possible pair of ROC curves the p-value was computed for the difference between them. Since our AUCs values are computed on the same underlying set of samples, we chose the method of [11] to compute p-values, as it accounts for the correlated nature of data generated from pairs of tests performed on the same samples. A significance threshold of 95% ($p < 0.05$) for this null-hypothesis test was chosen to determine if a performance gain was significant. Using this criteria, the performance gain of the learning based late fusion approach over every individual modality was significant; the corresponding p-values are shown in Figure 6. However, the performance differences between individual modalities all failed to meet this threshold. These statistics strongly suggest that fusion of multiple modalities leads to a statistically significant improvement in results.

Additionally, the I_{oc} statistic (as defined in 3.1) for typical performance of each modality was computed. We note that for this statistic, the fusion approach outperformed 3 of the 4 human annotators.

4.6 Predicting Viewer Response

Modality	AUC	p-Value	I_{oc}
Audio	0.873	$p = 0.02$	62.1%
Video	0.801	$p < 0.01$	46.3%
Text	0.828	$p = 0.03$	39.0%
Fusion	0.919	(n/a)	70.2%

Table 6: AUC values, p-values for the difference versus fusion AUC, and improvement-over-chance statistics

Finally, we investigated the prediction of viewer response to a video. Our goal here is to see whether given a video's audio-visual content, can we predict the sentiment polarity of the comments posted in response to it. In order to do so, we first clustered the test videos based on their sentiment histograms H_i (subsection 4.4), computed using SATSVM [29], in an unsupervised manner. We set the number of clusters to two, partitioning the set of test videos into two clusters that roughly correspond to videos that generated a positive response and videos that generated a negative response. Also note that while these clusters roughly map to the persuasive and non-persuasive classes, the correspondence is not exact. We treat this as a supervised classification problem, using the cluster indices as the class labels, which correspond to videos generating a positive and negative response. As features, we use the spectrogram features for audio (Table 1) and the "fc7" features from the sentiment ontology for video (Table 3). We train non-linear SVMs for classification based on unimodal features and a logistic-regression based late-fusion for multimodal fusion. The results are shown in Table 7. The results show that we can predict the viewer response in advance based on just the extracted audio-visual content with a reasonable degree of accuracy (random accuracy is 50%). Furthermore, fusing the audio and visual modalities leads to an increase in performance.

Modality	Classification Accuracy
Audio	61.97 ± 7.26
Video	61.67 ± 8.83
Learning based Late Fusion	**64.69 ± 4.63**

Table 7: Classification performance for predicting the viewer response on the RAC dataset.

5. CONCLUSION

We have demonstrated that affective and semantic audio and visual concepts as well as sentiment measures on viewer comments are effective at predicting whether a video contains politically persuasive content. Notably, the best automatic classification approach generally outperforms human annotators. Individually, audio concepts are the best predictors, while a fusion of these modalities produces the best results, indicating that each contains some complementary information to the others. Both visual and audio features are also predictive of negatively polarized comments, allowing us to potentially predict the viewer response to a video before comments are left. There are several possible areas for future work. For instance, an attempt to try to identify new videos with politically charged content as they are posted could be based on this approach, or a similarly tailored approach could be applied to other specific types of content besides politically persuasive videos.[1]

[1]This material is based upon work sponsored by the Defense Advanced Projects Agency under the U.S. Army Research

6. REFERENCES

[1] A. Abbasi. Affect intensity analysis of dark web forums. *Intelligence and Security Informatics, IEEE*, 2007.

[2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. In *IEEE PAMI*, 2012.

[3] M. R. Amer, B. Siddiquie, S. Khan, A. Divakaran, and H. Sawhney. Multimodal fusion using dynamic hybrid models. *WACV*, 2014.

[4] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. *ICML*, 2004.

[5] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM*, 2013.

[6] S.-F. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A. C. Loui, and J. Luo. Large-scale multimodal semantic concept detection for consumer video. *ACM MIR*, 2007.

[7] S. Chaudhuri, M. Harvilla, and B. Raj. Unsupervised learning of acoustic unit descriptors for audio content representation and classification. In *INTERSPEECH*, 2011.

[8] T. Chen, D. Borth, T. Darrell, and S. Chang. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv*, 2014.

[9] D. Chisholm, B. Siddiquie, A. Divakaran, and E. Shriberg. Audio-based affect detection in web videos. In *ICME*, 2015.

[10] D. Correa. Solutions to detect and analyze online radicalization.

[11] E. DeLong, D. DeLong, and D. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, pages 837–845, 1988.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[13] J. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, pages 378–382, 1971.

[14] T. Fu, C.-N. Huang, and H. Chen. Identification of extremist videos in online video sharing sites. *Intelligence and Security Informatics, IEEE*, 2009.

[15] K. Grauman and T. Darrell. The pyramid match kernel: Efficient learning with sets of features. *JMLR*, 2007.

[16] J. Hinojosa, L. Carretie, M. Valcarel, C. Mendez-Bertolo, and M. Pozo. Electrophysiological differences in the processing of affective information in words and pictures. *Cognitive, Affective, and Behavioral Neuroscience*, 9 2009.

[17] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? *CVPR*, 2011.

Office Contract Number W911NF-12-C-0028, through IBM Corporation subcontract 4914004308. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the position or policy of the U.S. Army Research Office, DARPA, DoD and IBM Corporation and no official endorsement should be inferred.

[18] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv*, 2014.

[19] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. *ICMR*, 2011.

[20] D. Joshi, R. Datta, E. Fedorovskaya, Q. Luong, J. Wang, J. Li, and J. Luo. Aesthetics and emotions in images. *Signal Processing Magazine*, 2011.

[21] A. Kapoor and R. W. Picard. Multimodal affect recognition in learning environments. *MULTIMEDIA*, 2005.

[22] S. Karayev and et al. Recognizing image style. In *BMVC*, 2014.

[23] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[24] J. Landis and G. Koch. The measurement of observer agreement for categorical data. *Biometrics. 33, 1*, pages 159–174, 1977.

[25] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time series. In *The handbook of brain theory and neural networks*, 1995.

[26] K. Lee and D. P. W. Ellis. Audio-based semantic concept classification for consumer video. *IEEE TASLP*, 2010.

[27] D. Littman and K. Forbes. Recognizing emotions from student speech in tutoring dialogues. *ASRU*, 2003.

[28] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. *ACMMM*, 2010.

[29] S. Mohammad, S. Kiritchenko, and X. Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *SemEval*, 2013.

[30] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2008.

[31] S. Park, H. S. Shim, M. Chatterjee, K. Sagae, and L.-P. Morency. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. *ICMI*, 2014.

[32] E. Reid, J. Qin, Y. Zhou, G. Lai, M. Sageman, G. Weimann, and H. Chen. Collecting and analyzing the presence of terrorists on the web: A case study of jihad websites. *Intelligence and Security Informatics*, 2005.

[33] B. Schuller and et. al. Avec 2011 -the first international audio visual emotion challenge. *ACII*, 2011.

[34] W. Scott. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 19*, pages 321–325, 1955.

[35] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. *EMNLP*, 2013.

[36] E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Text genre detection using common word frequencies. In *18th Conference on Computational Linguistics*, 2000.

[37] Y.-L. Tian, T. Kanade, and J. F. Cohn. Facial expression analysis. *Handbook of face recognition*, 2005.