Why We Watch the News: A Dataset for Exploring Sentiment in Broadcast Video News

Joseph G. Ellis
Electrical Engineering
Columbia University
New York City, NY
jge2105@columbia.edu

Brendan Jou
Electrical Engineering
Columbia University
New York City, NY
bjou@ee.columbia.edu

Shih-Fu Chang
Electrical Engineering
Columbia University
New York City, NY
sfchang@ee.columbia.edu

ABSTRACT

We present a multimodal sentiment study performed on a novel collection of videos mined from broadcast and cable television news programs. To the best of our knowledge, this is the first dataset released for studying sentiment in the domain of broadcast video news. We describe our algorithm for the processing and creation of person-specific segments from news video, yielding 929 sentence-length videos, and are annotated via Amazon Mechanical Turk. The spoken transcript and the video content itself are each annotated for their expression of positive, negative or neutral sentiment.

Based on these gathered user annotations, we demonstrate for news video the importance of taking into account multimodal information for sentiment prediction, and in particular, challenging previous text-based approaches that rely solely on available transcripts. We show that as much as 21.54% of the sentiment annotations for transcripts differ from their respective sentiment annotations when the video clip itself is presented. We present audio and visual classification baselines over a three-way sentiment prediction of positive, negative and neutral, as well as person-dependent versus person-independent classification influence on performance. Finally, we release the News Rover Sentiment dataset to the greater research community.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding; I.2.7 [**Artificial Intelligence**]: Natural Language Processing–*Discourse*

Keywords

Sentiment Analysis; Video Processing; News Video; Multimodal Processing; Audio Processing; Person Naming

1. INTRODUCTION

Text-based sentiment analysis has become a hot topic in recent years [14], spurred on in particular by the important use cases of automatic political polling, advertisement recommendation, and more. Using sentiment analysis we are able to gauge the public's opinion about topics that would have been impossible to gain widespread opinion data about 10 to 15 years ago cheaply and without interaction. This amount of opinion mining has been made possible by the influx of readily available on-line text data from social media and blog sources such as Twitter, Facebook, and Blogger. However, much of the opinion mining analysis is done in domains that have heavily polarized lexicons and obvious sentiment polarity. For example, a very popular domain for sentiment analysis is movie and product reviews, where the text available is heavily polarized and there is little room for ambiguity. Statements like "I absolutely loved this movie" or "the acting was terrible", have very clear and polarized sentiment that can be attributed to them.

However, in more complicated domains, such as news video transcripts or news articles, the sentiment attached to a statement can be much less obvious. For example, take the statement that has been relevant in the news in the past year, "Russian troops have entered into Crimea". This statement by itself is not polarizing as positive or negative and is in fact quite neutral. However, if it was stated by a U.S. politician it would probably have very negative connotations and if stated by a Russian politician it could have a very positive sentiment associated with it. Therefore, in more complicated domains such as news the text content is often not sufficient to determine the sentiment of a particular statement. For some ambiguous statements it is important to take into account the way that words are spoken (audio) and the gestures and facial expressions (visual) that accompany the sentence to be able to more accurately determine the sentiment of the statement.

Visual and audio elements of a statement can be useful in determining the overall sentiment of a video statement. However, the way that people portray positive, sarcastic, negative, and other feelings can be very different. Many people have difficulty grasping whether someone that they just met was being sarcastic or completely serious with something that he/she had just stated. This is a common phenomenon and happens because people portray emotions in a variety of different ways that are unique to a single person. Therefore, we propose to take advantage of these person-specific actions by focusing on person-specific models that are trained and tested using only videos of a single person. With the advent of video-based social media such as Instagram, Vine, and YouTube finding enough data to build person-specific models is not infeasible.

ICMI'14, November 12–16, 2014, Istanbul, Turkey.

ACM 978-1-4503-2885-2/14/11 ...$15.00.
http://dx.doi.org/10.1145/2663204.2663237.

News provides an interesting opportunity for us to not only study micro-level sentiment trends, but also macro-level trends. By analyzing the sentiment of speakers and the way that topics are covered by different channels, shows, and people, we can address public opinion and media opinion towards topics. Many interesting macro-level experiments can be carried out using sentiment analysis on news broadcasts. For example, we can determine whether the sentiment that a news channel uses in their coverage of a topic effects the companies that choose to advertise on their shows. We could also address regional biases, channel biases, and the changes in how a topic is covered over time.

The specific contributions of this work are as follows:

- The release of a video dataset in the novel domain of video news, annotated for multimodal sentiment
- A study demonstrating the importance of the audio and visual components of a statement in determining sentiment in news video.
- Baseline audio and visual classifiers and experiments for the dataset are presented.
- Experiments demonstrating improved performance with person-specific audio and visual sentiment classification models compared to global models.

2. RELATED WORK

Text analysis has been used prevalently in social analytics in the past. The authors in [14] present an overview of the field of text-based sentiment analysis up to the state of the art. Recently deep learning techniques have also been applied in the field of text analysis, and have obtained results with high accuracy [18].

There has also been an interest recently in combining multimodal signals (audio, visual, and text) for analysis of sentiment in videos. Emotion recognition by video and audio analysis has been studied extensively, however almost all of this work has been done on heavily controlled datasets that are created for benchmarking performance of algorithms. Popular challenges and datasets have improved performance in controlled scenarios [19, 16]. Recently, however with the advent of video social media sites such as YouTube, Vine, Instragram, and Vimeo, work has begun to perform sentiment analysis "in the wild". This has proven to be much more challenging. While the benchmark datasets have been curated to have very clear sentiment/emotion and little noise, the same can not be said about videos found in social media and television. The field was first pioneered by the authors of [11], where they utilized product reviews in YouTube to show that by using audio and visual features, in conjunction with text features, a higher sentiment classification accuracy can be achieved. Each of the videos within this dataset were labeled on an entire video level, and tri-modal (negative, neutral, positive) sentiment annotation and classification was performed. In [15], sentiment annotation and classification is performed on the utterance level. This is a more natural setting for sentiment analysis, because the sentiment displayed by a person can change between sentences or utterances even if the topic remains the same. Finally, in [22] the more complicated domain of on-line video movie reviews is analyzed using bi-directional long short term memory recurrent neural networks with linguistic, visual and audio features. Recently, a research challenge addressing the recognition of human emotion in the wild was organized, with very promising results [3]. Researchers have also tackled interesting problems such as the "desire to watch again" of a user given their facial expressions in unconstrained on-line content viewing [10]. The engagement of viewers while watching television using their facial expressions [6] and trying to predict movie ratings based on the expressions of viewers throughout a movie [12] have been studied. The authors of [19], state that higher accuracies for visual emotion classification can be obtained if the identity of the person during test time is known beforehand. We will also build off of the ideas in [19], and show the usefulness of person-specific classifiers. We can see that many interesting applications and techniques for emotion and sentiment recognition from video content have been proposed and explored recently.

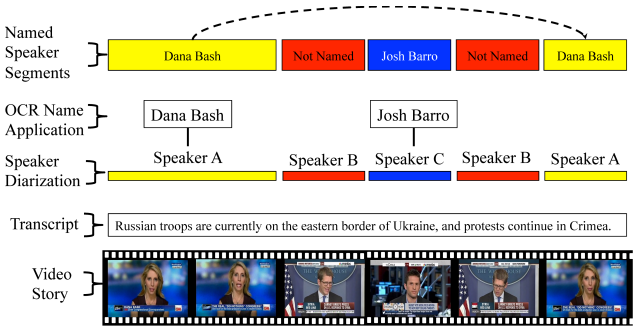

Figure 1: Speaker naming algorithm for video news. Names are detected on-screen using OCR and are used to label speaker segments generated by speaker diarization.

3. NEWS VIDEO SENTIMENT DATASET

In this section we discuss the data collection scheme that was used to create this dataset.

3.1 News Processing Overview

News video is a distinct and interesting domain that is ripe for sentiment mining and analysis due to the variety of opinions that are portrayed within the news. For the past two years, we have been recording up to 12 news programs simultaneously from 100 possible channels and storing and processing the videos. A detailed look at our framework for processing videos in its entirety can be seen here [8].

3.2 Mining Speakers from the News

In this work we create a study dataset with two particular goals. First, the dataset that we create should have an accurate transcript of the words and sentences that are spoken in each video, and this will be automatically created without difficult manual transcribing. Secondly, the dataset would be a diverse representation of the type of people that appear within news programs (politicians, anchors, reporters, etc.), with sufficient spoken sentence videos for each person. After recording the full length programs, removing the commercials, and then segmenting the programs into separate news "stories" [8], we automatically find sections of the

Figure 2: Example Amazon Mechanical Turk Interface. This is an example of the Amazon Mechanical Turk interface used to collect annotations, both the text and video annotation questions and instructions can be seen.

video where particular people of interest are speaking. The speaker naming algorithm presented in this section builds off of the work in [8]. The closed caption transcript associated with each video story accurately transcribes what is stated in the story, but can lag the timing of what is said anywhere between 5-25 seconds. To create an accurate time-aligned transcript of the video we first perform speech recognition on the entire video segment, and then align the associated closed-caption transcript using a modified minimum-edit distance algorithm. This algorithm results in 91.5% precision in the alignment of the transcript to what is said in the sentence level videos used in this experiment. Videos without a properly aligned transcript are removed from the dataset and are not used in the experiments presented later in the paper. Once the time-alignment is completed, we then perform speaker diarization [7] on the audio portion of each video story to determine when each speaker begins and ends speaking throughout the story. To "name" the speech portions we find when a person name appears on-screen, and apply that name as a label to that speech portion. The intuition behind this idea is that when a person is speaking, oftentimes their name will appear in a specific location on the screen (usually in the lower left corner). We first mine the specific "name location" for each separate program by performing OCR on the screen and then comparing the detected OCR text to a list of common first names from the Social Security Name database. Once we find the "name location(s)" on the screen where names most frequently appear for each program we use all of the detected text in this portion as a person name, and apply it as a label to the time-overlapping speech-segment. Once all of the names found are applied to the overlapping speech segments, we find any speaker clusters (the collections of similar speech segments created by speaker diarization) that have only one name applied to them and label these portions of the video according to that name. Finally, we extract the portions of a video that constitute one full sentence within each named speaker portion and add this video to our study. This process is illustrated in Figure 1. We are able to extract from each video specific segments of a person speaking and label them with their appropriate name.

We randomly sampled 198 videos from the dataset to see how many of the videos were named properly and found that 170 out of 198 (85.85%) videos were properly named. The most common error cases occurred during free-flowing debates where it becomes very difficult to detect a change of speaker.

3.3 Dataset Statistics

The videos used for this dataset were recorded and processed between August 13, 2013 and December 25, 2013, and are taken from a large variety of American news programs and channels. A breakdown of the dataset by person, with their occupation and amount of videos within the study can be seen in Table 1. We limit the length of the videos used in the study to be between 4 and 15 seconds long. This is done because it can be difficult to decipher sentiment for very short videos with little speech, and videos longer than 15 seconds could have multiple statements with opposing sentiment.

4. SENTIMENT ANNOTATION

In this section we will discuss our sentiment annotation methodology and how we used Amazon Mechanical Turk to obtain over 3000 unique annotations, and also discuss some interesting statistics of the Turk annotators.

4.1 Annotation Methodology

To obtain sentiment labels for each video we employed a three-way sentiment annotation scheme, where each video was labeled as either positive, negative, or neutral. We utilized Amazon Mechanical Turk to crowd-source the annotation of sentiment in the videos. To prove that the audio and visual components can alter the sentiment implied by a given statement we created an interface that began by showing a sentence to the "turker", a worker on Amazon Mechanical Turk, and asked them to determine the sentiment portrayed by the sentence. Once the turker has labeled the sentence, they are then prompted to watch a video in which the same sentence they previously annotated is spoken by someone in the news and evaluate the sentiment of the video. The turker is told to treat each annotation as a separate entity, and evaluate the sentiment present in the video independently from that of the transcript. We also provided an option to the turker to state that the sentence annotated from the text portion does not match what is said in the video, and we collect these videos and remove them from the overall dataset. An example Amazon Mechanical Turk Human Intelligence Task can be seen in Figure 2.

4.2 Annotation Statistics

We had 72 total annotators provide annotations for our dataset through Mechanical Turk. The number of annotations provided by each user ranged from as little as one annotation, to as many as 343 unique annotations. These annotations were captured over a two week period in April, 2014. We did not collect any personal data from the annotators, as we avoided collecting any data that would make a particular turker identifiable. A histogram breakdown of

Table 1: Dataset statistics by person. The dataset is categorized into occupation, category of news television appearances, and number of videos within each sentiment class.

Person	Category	Occupation	Negative Videos	Neutral Videos	Positive Videos
Jay Carney	Politician	White House Press Secretary	36 (31.85%)	61 (53.98%)	16 (14.15%)
John Kerry	Politician	U.S. Secretary of State	35 (44.30%)	34 (43.03%)	10 (12.65%)
Goldie Taylor	Pundit/Analyst	Author, Opinion Writer	63 (59.43%)	29 (27.35%)	14 (13.20%)
Josh Barro	Pundit/Analyst	Editor at Business Insider	68 (44.15%)	63 (40.90%)	23 (14.93%)
Dana Bash	Reporter	Capitol Hill Correspondent	36 (23.84%)	100 (66.22%)	15 (9.93%)
Jim Acosta	Reporter	White House Correspondent	27 (23.07%)	79 (67.52%)	11 (9.40%)
Chris Cillizza	Pundit/Analyst	MSNBC Political Analyst	16 (16.84%)	63 (66.31%)	16 (16.84%)
C. Amanpour	Reporter	International Correspondent	22 (19.29%)	72 (63.15%)	20 (17.54%)

the number of annotations provided by the annotators can be seen in Figure 3.

5. ANNOTATION RESULTS

For each video we received annotations from three distinct turkers and used a majority voting scheme to decide the label of the videos. If there was no majority label across the three annotations that particular example is removed from the study due to sentiment ambiguity, approximately 6% of the video annotations fell into this category. In total we collected 929 majority-vote sentiment labels for videos used in this study.

Our study shows that by watching and listening to the statement portrayed by a person the perceived sentiment can change quite frequently compared to the original sentiment perception based solely on the transcript associated with the video. The majority-voted sentiment label changed between the original text annotation and the video annotation in 21.54% of the annotations. Therefore, for many examples in our study text is insufficient to fully understand the sentiment portrayed, and it is necessary to take into account the way that the sentence content is delivered by the speaker. The breakdown of sentiment in the videos by neutral, negative, and positive sentiment categories can be seen in Table 2. The breakdown of sentiment labels remains reasonably consistent throughout the dataset between text and video sentiment even though over 20% of the annotated labels differ between the text and video. This is because there are roughly equal amounts of annotations that change from neutral in the text to positive or negative in the video and vice-versa. It is also interesting to note that most of the differences in sentiment labels between the two modalities consist of one label being neutral and the other label being positive or negative.

When we study the sentiment breakdown of each person's video and group them by occupation we can see that people within the same occupation have a similar percentage of videos in each sentiment category. John Kerry and Jay Carney both portray neutral sentiment in a large portion of their videos, especially Jay Carney who, as the White House Press Secretary, is often tasked with holding news conferences about breaking news events which he portrays a calm and controlled demeanor. John Kerry on the other hand is often criticizing the actions of other countries, which can be seen in his slightly higher proportion of negative videos. Jim Acosta and Dana Bash are both on-site reporters for their respective news channels, and are therefore tasked with giving an un-biased description of the particular news events

Table 2: Percentage of sentiment examples in each class for both transcript and video annotations.

Sentiment	Video	Transcript
Negative	32.61%	34.23%
Neutral	53.92%	53.96%
Positive	13.45%	11.79%

happening that day. This is why they both have around two-thirds of their videos labeled under the neutral sentiment category. Finally, the talk-show political pundits such as Josh Barro and Goldie Taylor are often brought onto television programs to give their opinions and react strongly towards a given topic. This is why we see both exhibiting a relatively large percentage of their videos being deemed positive or negative compared to the other occupations.

6. FEATURE EXTRACTION

We model our baseline feature extraction algorithm off of the baseline feature set in the AVEC 2011 challenge [16], with some slight modifications for our domain. We chose this feature set because the AVEC 2011 challenge is a commonly referenced benchmark dataset and is still actively used for evaluation today.

6.1 Audio Features

Our audio features are extracted independently from each sentence-level video using the openSMILE [5] "emotion base" feature set tuned for emotion recognition. The feature set consists of 988 features, consisting of a variety of low level feature descriptors with functionals applied to them over the entire sentence length video (Mel Frequency Cepstral Coefficients, pitch, etc.).

6.2 Visual Features

Extracting visual features from news video presents some unique challenges not present in other datasets. For example, in news video frames can often be split with multiple faces on screen. However we only want to analyze the sentiment portrayed by the "speaking" face. Another difficult scenario appearing in the news is that a reporter or pundit will be speaking, but on-screen a different video would be shown with a large variety of possible alternate content (on-site scenes, picture of people of interest, etc).

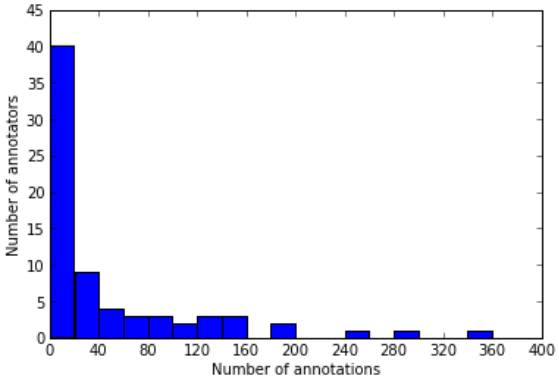

Figure 3: Histogram of annotations per annotator. We can see that many annotators supply few annotations, while approximately 20 provided annotations in large numbers throughout the project.

6.2.1 Visual Speaker Detection

We take inspiration for our visual speaker detection algorithm from [4]. We sample frames in each news video at 10 frames/sec, and on each sub-sampled frame we detect faces using the OpenCV implementation of the Viola-Jones face detector [21]. Using these detected faces we refine the face detections by forming coherent face tracks across the entire video by exploiting the temporal and spatial consistency of the tracks within each successive frame using optical-flow. After this we are left with a consistent set of face images belonging to one identity for the duration of the time that this face appears within the video. For each of the faces within a track we then extract seven landmark points on the detected face [20], and then affine warp and resize the faces to a fixed 200×200 pixel resolution such that the facial landmarks appear at the same point for each face.

Once all of the aligned face tracks are generated from a particular video we find the face track that corresponds to the speaker within the video. We extract the mouth region from each aligned face within a face track and perform a normalized cross-correlation comparison between two mouth patches in successive frames. We then compute the median cross-correlation score between successive faces to obtain a "speaking" score for each face track. We threshold the score for each to determine if it represents a visual speaker. If multiple visual speakers are detected within a video at different times we choose the face that appeared on screen longer as the detected visual speaker. Using this methodology, we have determined that visual speakers appear in 455 out of the 929 videos that appeared in this dataset. The amount of videos in which a visual speaker was detected in each person specific video set is shown in Table 3.

6.2.2 Face Features

Following the AVEC 2011 Challenge [16] we extract dense Local Binary Patterns (LBP) [13] on the 200×200 pixel face images of detected visual speakers. LBP has proven to be a useful feature for a variety of tasks concerning humans faces, such as emotion [17] and person recognition [1].

We first divide each image into 10×10 evenly spaced image sections (20×20 pixel image patches), and extract uniform LBP patches from each section. A histogram of the LBP features are taken over each image patch resulting in 100 59-dimensional LBP histograms for each face image. The histograms are then concatenated together for each image, resulting in a 5900-dimensional feature representation for each face within our dataset. These visual features will be used as the baseline visual feature set for our video classification analysis.

We also present an alternate visual feature representation pipeline based on Block-Based-Bag-of-Words (BBoW) [9]. First, we separate the image into four quadrants, top left, top right, bottom left, and bottom right. Next, within each quadrant we extract LBP histograms in the same manner as described above from each quadrant. Within the LBP features extracted from every image in a particular quadrant we use k-means clustering to generate a codebook, and then quantize the features into a Bag of Words representation for each quadrant. The quantized features from each quadrants are then concatenated together to create the representation for each image. Codebook sizes of 50, 200, and 1000 are used in this implementation, and each codebook is created using the complete representation of features from each quadrant (over 3 million LBP histogram features).

Table 3: Number of videos per person in which a visual speaker is detected.

Person	# of Videos
Jay Carney	44
John Kerry	37
Goldie Taylor	70
Josh Barro	64
Dana Bash	71
Jim Acosta	56
Chris Cillizza	85
Christiane Amanpour	28

7. SENTIMENT CLASSIFICATION

In this section, we explore text, audio, and visual classification of the sentiment of the news videos. We provide results demonstrating the accuracy of global models trained on all of the people in the dataset and person-specific models.

7.1 Text Classification

We performed sentiment classification from the transcript annotations obtained from Amazon Mechanical Turk using one of the state-of-the-art deep-learning based text sentiment analysis algorithms [18]. We first perform true-casing on news transcripts, because the news transcripts use only capitalized letters and then use the true-cased transcripts for text sentiment classification. The positive, negative, and neutral classification accuracy of [18] on our corpus is 45.6%. This low level of accuracy shows that news domain transcripts are quite ambiguous with regard to sentiment. There is much room for improvement by incorporating the multimodal information also present within the video.

We believe that the classification accuracy is particularly low due to several reasons. First, it must be stressed that this classification accuracy was achieved using the text system "out-of-the-box" and no domain adaptation to the news

Table 4: Video sentiment classification accuracy.

Feature Set	Accuracy
AVEC Raw LBP	31.47%
BBoW 50-dim codebook	44.41%
BBoW 200-dim codebook	37.94%
BBoW 1000-dim codebook	33.48%

transcripts was done. It is believed that some domain adaptation would improve the results considerably. Second, the grammar in news transcripts is often not perfect, because the transcripts are manually typed during the news program. This could also lead to degraded performance because the sentences are not well formed. Both of these issues are not the focus of this work.

7.2 Global Models

We trained sentiment models using all of the data present in the dataset independent of person identity.

7.2.1 Audio Classification

For audio classification, we used the raw features from openSMILE to train a linear SVM classifier using libSVM [2]. We use four-fold cross-validation to tune the SVM cost parameter, and report here the average accuracy of the four-fold cross-validation. We split each set of person videos into four-folds and use three of the folds across all of the people for training and one for test. We do this so that we know the accuracy of the trained global model on each person, and to insure that training examples exist for each person in each training fold. The four-fold training classification accuracy for audio classification across the entire dataset is 56.57%. random guess model for this classification is 33.33%, and the dominant class guess for this dataset is neutral with 53.92%. The audio classification accuracy of the global model for each particular person in our dataset can be seen in Table 5.

7.2.2 Video Classification

The results in the video classification utilize the 455 videos in which we detected speaking faces. In the same fashion as the audio classification we trained a linear SVM classifier and report the average accuracy over four-fold classification over our four different feature extraction algorithms. The videos are split in the same way as described in Sec. 7.2.1. We label all of the extracted frames within a video with the sentiment label of the video and then train a frame-level classifier for sentiment. We then carry out a majority vote classification over the frames in each video to determine the sentiment on a video-level during test. The results can be seen in Table 4. We can see that our Block-Based Bag of Words feature representation outperforms the LBP standard feature extraction method on this dataset, and performs the best with a small codebook size of 50 codewords per quadrant (200 codewords across the whole face). We believe that the small codebook performs well because we have few label categories (positive, negative, and neutral) and there is high inter-image similarity because all the images are aligned faces. The video classification accuracy of the best performing global model for each particular person in our dataset can be seen in Table 6. The visual model accuracy is not directly comparable to the audio model accuracies, because the visual model was only tested on the subset of videos in which visual speakers were detected.

Table 5: Classification accuracy (%) for audio-based person-specific and global sentiment classifiers.

Person	Global	Person-Specific
Global	56.57	62.56
Jay Carney	58.92	64.17
John Kerry	42.10	48.68
Goldie Taylor	54.80	59.61
Josh Barro	50.00	59.21
Dana Bash	59.45	66.21
Jim Acosta	63.79	67.24
Chris Cillizza	60.86	66.30
C. Amanpour	59.82	65.17

7.3 Person-Specific Models

Although most of the previous work in the field of multimodal sentiment has been done by creating global models that apply to all people, there are some inherent advantages to training classifiers for specific people. People have their own unique ways of displaying anger, happiness, sarcasm, and sentiment in both auditory and visual ways. In the age of social media it is possible to gather large amounts of video for one person, and we show that taking into account person-specific classification can improve or achieve comparable performance with our global models.

The dataset was separated into smaller single-person video sets based on the names applied during the named multimodal speaker diarization algorithm described in Sec. 3.2. We randomly sampled videos from our dataset and found that our naming algorithm has an accuracy of 85.85%. The incorrectly named videos within each person-specific video set will alter slightly the learned classifiers and results, but since these outliers are a small minority we do not expect the effect on results to be drastic. This also demonstrates the feasibility of automatically mining person video data and building person-specific classifiers. The entire processing pipeline from raw-video content to predicted sentence-level video sentiment can be seen in Figure 4.

7.3.1 Audio Classification

The audio classification scheme described in Sec. 7.2.1 followed in exactly the same way, but person-specific models are trained and tested only using the videos that are labeled as one particular person within our dataset. The results can be seen in Table 5.

We can see that building the person-specific classifiers improves the results of classification in every case. We postulate that this occurs for two reasons. First, we are able to take into account person-specific quirks in the audio segment that within a person can be descriptive for sentiment, but are lost when compared across the whole dataset. Second, due to the imbalance of sentiment label classes within each person we are able to take into account the likelihood of a person demonstrating positive, negative, or neutral sentiment when only using their videos as training data compared to the global models. Based on the results we can see that building person-specific sentiment models can achieve similar or improved results over the global model with less data and therefore faster training time.

Table 6: Classification accuracy (%) for visual-based person-specific and global sentiment classifiers.

Person	Global	Person-Specific
Global	44.41	56.85
Jay Carney	36.36	34.09
John Kerry	50.00	52.77
Goldie Taylor	55.88	42.64
Josh Barro	51.56	59.37
Dana Bash	35.29	48.52
Jim Acosta	28.57	66.07
Chris Cillizza	55.95	67.85
C. Amanpour	25.00	85.71

It is also interesting to note that some person-specific models produce the same classification accuracy as the dominant label classification strategy within person. For example, the audio person-specific classifier for reporters, such as Dana Bash or Jim Acosta that has the highest accuracy across the cross-validation folds is the one that classifies every video as "neutral". This shows that it is quite difficult to learn sentiment models for on-site reporters/correspondents, something that is very intuitive because they are expected to give an un-biased description of news events.

7.3.2 *Video Classification*

The video classification scheme described in Sec. 7.2.2 is replicated in exactly the same way, but models are trained and tested using only the person-specific data. We use the best performing model, which was our block-based bag of words with a codebook size of 50. The results can be seen in Table 6. We can see that the results improve in most cases, and in some cases greatly using the person-specific models. Due to the fact that each frame extracted from the training set videos is used for building the classifiers, the visual classifier is trained with a larger set of data making it more robust for the person-specific classifiers as compared to the person-specific audio classifiers. Due to the majority voting scheme within video tracks, this method is also robust to spurious classifications, because the results of the classification on each frame are effectively averaged out over the entire duration of the video. The visual model performance is not directly comparable to the audio model performance, because the visual model was only tested on the subset of videos in which visual speakers were detected.

8. CONCLUSIONS

In this work we describe an algorithm for extracting named speech segments from broadcast news videos, which are then used in a framework for joint sentiment annotation of text and video content using Amazon Mechanical Turk. Using this data we have performed a study detailing the added value of seeing and hearing a statement delivered when determining sentiment in comparison to simply reading the sentence. In particular we show that 21.54% of annotations differ when users are given only the text content compared to the full video. These results imply that when dealing with complicated domains for sentiment analysis such as spoken news video, it is important to take into account the video content and not focus solely on text.

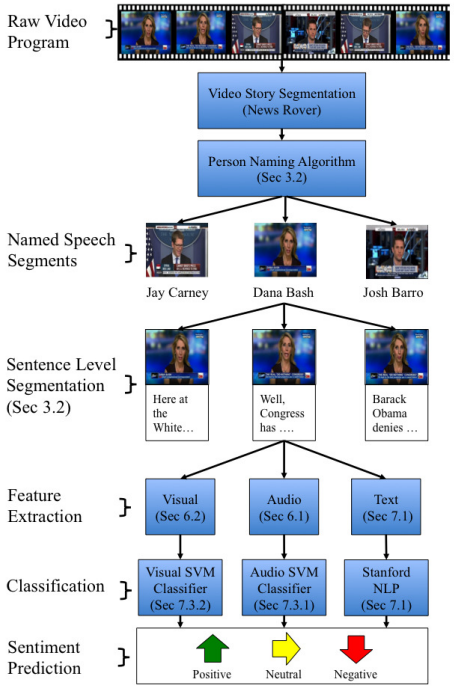

Figure 4: Full News Rover sentiment processing pipeline. Here we represent the ingestion of content into our system from a raw-video program to predicted sentence-level sentiment.

The News Rover Sentiment dataset has been presented and baseline classification algorithms benchmarked on this dataset. We think this dataset presents a rich opportunity for the research community to tackle difficult video-based sentiment analysis questions "in the wild". The News Rover Sentiment dataset described in this paper is available at `www.ee.columbia.edu/dvmm/newsrover/sentimentdataset`. Because it was automatically created using the algorithms for naming and visual speaker detection described in this work it contains some noise. Therefore, we have manually inspected the dataset and provide a "clean" dataset with videos of only properly named visual speakers as well. The same classification algorithms are performed on this dataset, and results on the "clean" dataset can be seen on the provided website, and can be used for benchmarking against.

9. FUTURE WORK

We hope to expand this work by fusing the multimodal features for sentiment classification on this dataset, and transferring current state-of-the-art techniques to this domain. We believe that gains can be made in text-based sentiment classification on the news transcripts by developing techniques that are robust to errors in grammar and are more domain specific. In the future we will pursue a fusion method for the person-specific and global models to leverage the "big-data" advantages of a global model, while still learning person-specific features. Finally, once reliable sentiment classifiers are built then macro-level experiments exploring news program trends will be carried out.

10. REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006.

[2] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2011.

[3] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge (emotiw) challenge and workshop summary. In *International Conference on Multimodal Interaction*, 2013.

[4] M. Everingham, J. Sivic, and A. Zisserman. Taking the bite out of automatic naming of characters in TV video. *Image and Vision Computing*, 2009.

[5] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: The munich versatile and fast open-source audio feature extractor. In *ACM Multimedia*, 2010.

[6] J. Hernandez, Z. Liu, G. Hulten, D. Debarr, K. Krum, and Z. Zhang. Measuring the engagement level of tv viewers. In *IEEE Automatic Face and Gesture Recognition*, 2013.

[7] M. Huijbregts. *Segmentation, Diarization, and Speech Transcription: Suprise Data Unraveled*. PhD thesis, University of Twente, 2008.

[8] B. Jou*, H. Li*, J. G. Ellis*, D. Morozoff, and S.-F. Chang. Structured exploration of who, what, when, and where in heterogeneous multimedia news sources. In *ACM Multimedia*, 2013.

[9] Z. Li, J. Imai, and M. Kaneko. Robust face recognition using block-based bag of words. In *International Conference on Pattern Recognition*, 2010.

[10] D. McDuff, R. E. Kaliouby, D. Demirdjian, and R. W. Picard. Predicting online media effectiveness based on smile responses gathered over the internet. In *IEEE Automatic Face and Gesture Recognition*, 2013.

[11] L.-P. Morency, R. Mihalcea, and P. Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *International Conference on Multimodal Interaction*, 2011.

[12] R. Navarathna, P. Lucey, P. Carr, E. Carter, S. Sridharan, and I. Matthews. Predicting movie ratings from audience behaviors. In *IEEE Winter Conference on Applications of Computer Vision*, 2014.

[13] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002.

[14] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), 2008.

[15] V. Perez-Rosas, R. Mihalcea, and L.-P. Morency. Utterance-level multimodal sentiment analysis. In *Association for Computational Linguistics*, 2013.

[16] B. Schuller, M. F. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. Avec 2011-the first international audio/visual emotion challenge. In *Affective Computing and Intelligent Interaction*, 2011.

[17] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 2009.

[18] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Empirical Methods in Natural Language Processing*, 2013.

[19] U. Tariq, K.-H. Lin, Z. Li, X. Zhou, Z. Wang, V. Le, T. Huang, X. Lv, and T. Han. Recognizing emotions from an ensemble of features. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2012.

[20] M. Uřičář, V. Franc, and V. Hlaváč. Detector of facial landmarks learned by the structured output SVM. In *International Conference on Computer Vision Theory and Applications*, 2012.

[21] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 2004.

[22] M. Wollmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, and L.-P. Morency. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 2013.