

Sentiment in Speech

Ahmad Elshenawy
Steele Carter

May 13, 2014

Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web

What can a video review tell us that a written review can't?

- By analyzing not only the words people say, but *how* they say them, can we better classify sentiment expressions?



The trouble with director Marc Webb's disappointing sequel is it wants to have it both ways: to take seriously human connection and loss and also spin the spectacular and pulpy... Spider-Man 2 never locates that sweet spot.

[Full Review](#)

May 2, 2014



Lisa Kennedy

Denver Post

★ Top Critic

Prior Work

For Trimodal (textual, audio **and** video) not much, really...

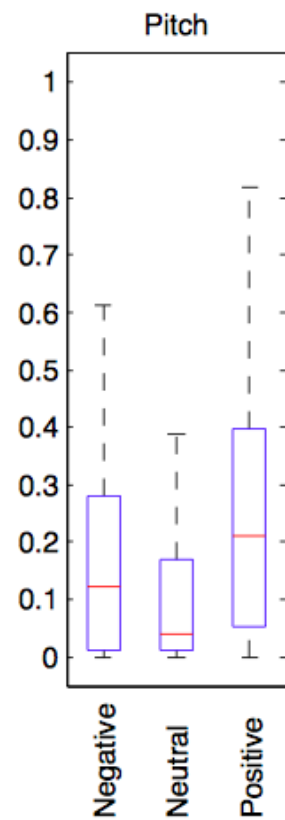
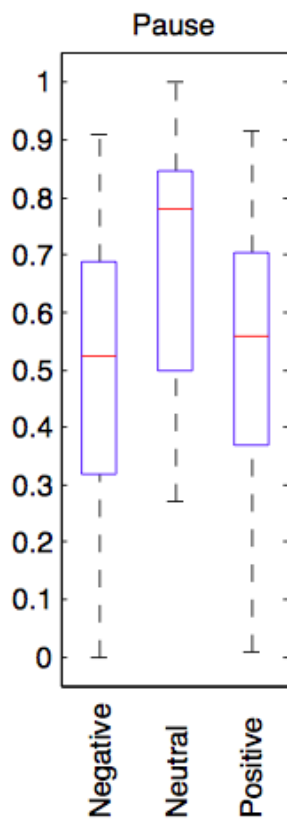
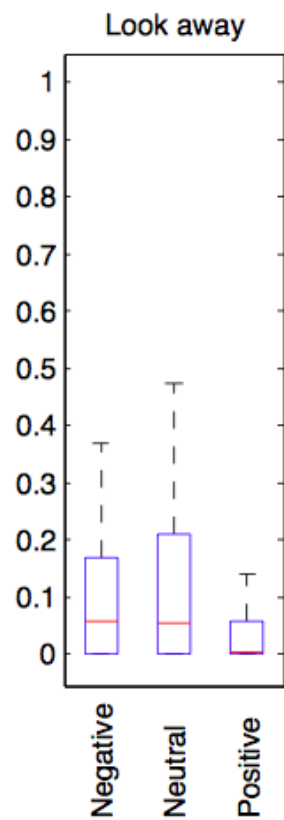
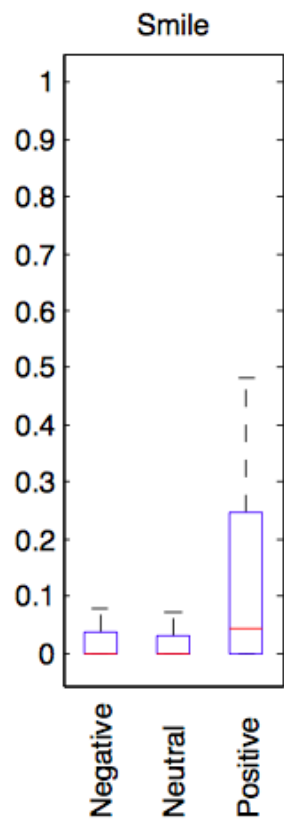
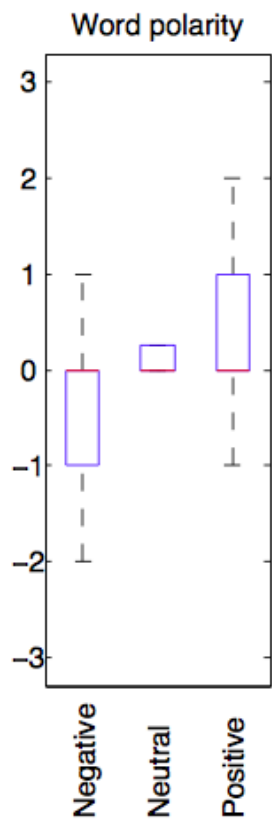
- As we have seen, a plethora of work has already been done on analyzing sentiment in text.
 - Lexicons, datasets, etc.
- Much of the research done on sentiment in speech is conducted in ideal, scientific environments.

Creating a Trimodal dataset

- 47 2-5 minute youtube review video clips were collected and annotated for polarity.
 - 20 female/27 male, aged 14-60, multiple ethnicities
 - English
- Majority voting between the annotations of 3 annotators:
 - 13 positive, 22 neutral, 12 negative
- Percentile rankings were performed on annotated utterances for the following audio/video features:
 - Smile
 - Lookaway
 - Pause
 - Pitch



Figure 1: Selected snapshots from our new video dataset.



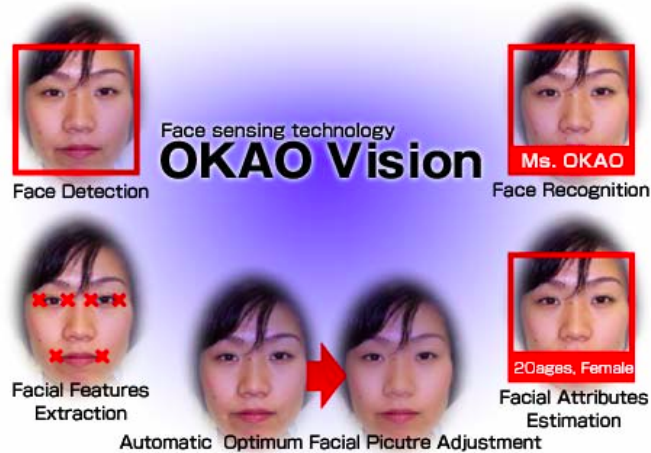
Features and Analysis: Polarized Words

- Effective for differentiating sentiment polarity
- However, most utterances don't have any polarized words.
 - For this reason we see that the median values of all three categories (+/-/~) is 0.
- Word polarity scores are calculated through use of two lexicons
 - MPQA, used to give each word a predefined polarity score
 - Valence Shifter Lexicon, polarity score modifiers
- Polarity score of a text is the sum of all polarity values of all lexicon words, checking for valence shifters within close proximity (no more than 2 words)

OKAO Vision

"OKAO Vision" Face Sensing Technology

Visual information plays a significant role in face-to-face communication. Clearly, communication between people and machines would be more comfortable if a machine could understand people visually in much the same way as people do. "OKAO Vision," which stands for "face vision" in Japanese, is the collection of OMRON's cutting edge technologies in this area. By visually sensing and extracting useful information from face images, OMRON aims to provide various kinds of services optimized for each individual. These services will match their interfaces and contents to user's capabilities, preferences, condition, attributes, and applicability.



Facial tracking performed by OKAO Vision

Features and Analysis: Smile feature

- a common intuition that a smile is correlated with happiness
- smiling found to be a good way to differentiate positive utterances from negative/neutral utterances

- Each frame of the video is given a smile intensity score of 0-100
- Smile Duration
 - Given the start and end time of an utterance, how many frames are ID'd as “smile”
 - Normalized by the number of frames in the utterance

Features and Analysis:

Lookaway feature

- people tend to look away from the camera when expressing neutrality or negativity
- in contrast, positivity is often accompanied with mutual gaze (looking at the camera)

- Each frame of the video is analyzed for gaze direction
- Lookaway Duration
 - Given the start and end time of an utterance, how many frames is the speaker looking at the camera
 - Normalized by the number of frames in the utterance

Features and Analysis:

Audio Features

- OpenEAR software used to compute voice intensity and pitch
- Intensity threshold used to identify silence
- Features extracted in 50ms sliding window
 - **Pause duration**
 - Percentage of time where speaker is silent
 - Given start and end time of utterance, count audio samples identified as silence
 - Normalize by number of audio samples in utterance
 - **Pitch**
 - Compute standard deviation of pitch level
 - Speaker normalization using z-standardization
- Audio features useful for differentiating neutral from polarized utterances
 - Neutral speakers more monotone with more pauses

Results

- Leave-one-out testing

HMM	F1	Precision	Recall
Text only	0.430	0.431	0.430
Visual only	0.439	0.449	0.430
Audio only	0.419	0.408	0.429
Tri-modal	0.553	0.543	0.564

Conclusion

- Showed that integration of multiple modalities significantly increases performance
- First task to explore these three modalities
- Relatively small data size (47 videos)
 - Sentiment judgments only made at video level
- No error analysis
- Future work
 - Expand size of corpus (crowdsource transcriptions)
 - Explore more features (see next paper)
 - Adapt to different domains
 - Attempt to make process less supervised/more automatic

Questions

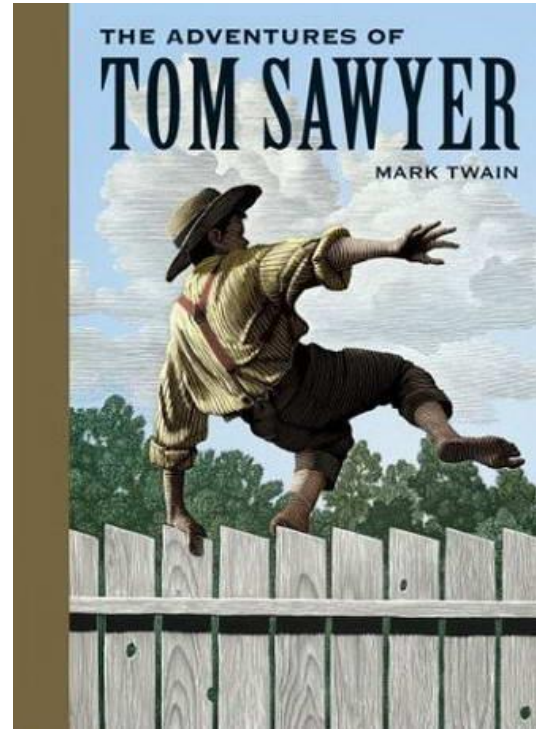
- How hard would it really be to filter/annotate emotional content on the web? There was a lot of hand selection here.
 - Probably very difficult, not very adaptable/automatic
- What about other cultures? It seems like there'd be a lot of differences in features, especially video ones.
 - Again, hand feature selection probably limits adaptability to other languages/domains
- What do you think about feature selection? combination? the HMM model?
 - Good first pass, but a lot of room for expansion/improvement

More Questions

- What does the similarity in unimodal classification say about feature choice? Do you think the advantage of multimodal fusion would be maintained if stronger unimodal (e.g. text-based) models were used?
 - I suspect multimodal fusion advantage would be reduced with stronger unimodal models
 - Error analysis comparing unimodal results would be enlightening on this issue
- Is the diversity of the dataset a good thing?
 - Yes and no, would be better if the dataset was larger

Correlation analysis of sentiment analysis scores and acoustic features in audiobook narratives

Using an audiobook and other spoken media to find sentiment analysis scores.



Why audiobooks?

Turns out audiobooks are pretty good solutions for a number of speech tasks:

- easy to find transcriptions for the speech
- great source of expressive speech
- more listed in Section I

Data

- Study was conducted on Mark Twain's *The Adventures of Tom Sawyer*
 - 5119 sentences / 17 chapters / 6.6 hours of audio
- Audiobook split into “prosodic phrase level chunks”, corresponding to sentences.
 - Text alignment was performed using software called LightlySupervised (Braunschweiler et al., 2011b)

Sentiment Scores (i.e. the book stuff)

- Sentiment scores were calculated using 5 different methods:
 - IMDB
 - OpinionLexicon
 - SentiWordnet
 - Experience Project
 - a categorization of short emotional stories
 - Polar:
 - probability derived from a model trained on the above sentiment scores
 - used to predict the polarization score of a word

Acoustic Features (i.e. the audiobook stuff)

Again, a number of acoustic features were used, fundamental frequency (F0), intonation features (F0 contours) and voicing strengths/patterns

- F0 statistics (mean, max, min, range)
- sentence duration
- Average energy (Σs^2) / duration
- Number of voicing frames, unvoiced frames, and voicing rate
- F0 contours
- Voicing strengths

Feature Correlation Analysis

The authors then ran a correlation analysis between all of the text and acoustic features.

Strongest correlations found were between *average energy / mean F0* and *IMDB reviews / reaction scores*.

Other acoustic features were found to have little to no correlation with sentiment features

- no correlation between F0 contour features and sentiment scores
- no relation between **any** acoustic features and sentiment scores from lexicons

	Acoustic features	
Sentiment scores	Energy	mean_F0
ImdbEmphasis	0.51	0.38
ImdbPolarity	-0.33	-0.31
Teehee	0.29	0.13
Wow	-0.17	-0.30
Polar	-0.13	-0.14

Table 2: Pairwise correlation between sentiment scores and acoustic features.

Bonus Experiment!

Predicting Expressivity

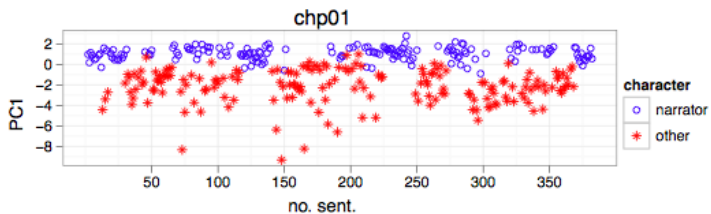
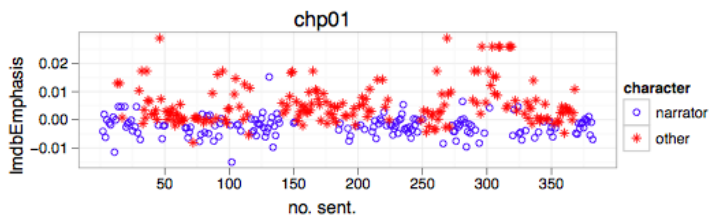
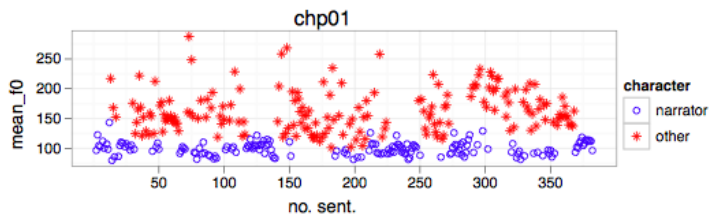
Using sentiment scores to predict the “expressivity” of the audiobook reader.

- meaning the difference between the reader’s default narration voice, and when s/he is doing impressions of characters.

Expressivity quantified by the first principal component (PC1), the result of using Principal Component Analysis on the acoustic features of the utterance.

- according to Wikipedia, “a statistical procedure that uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.”

PC1 scores vs other Sentiment Scores



Empirical findings:

- PC1 scores ≥ 0 corresponded to utterances made in the narrators default voice
- PC1 scores < 0 corresponded to expressive character utterances.

Building a PC1 predictor

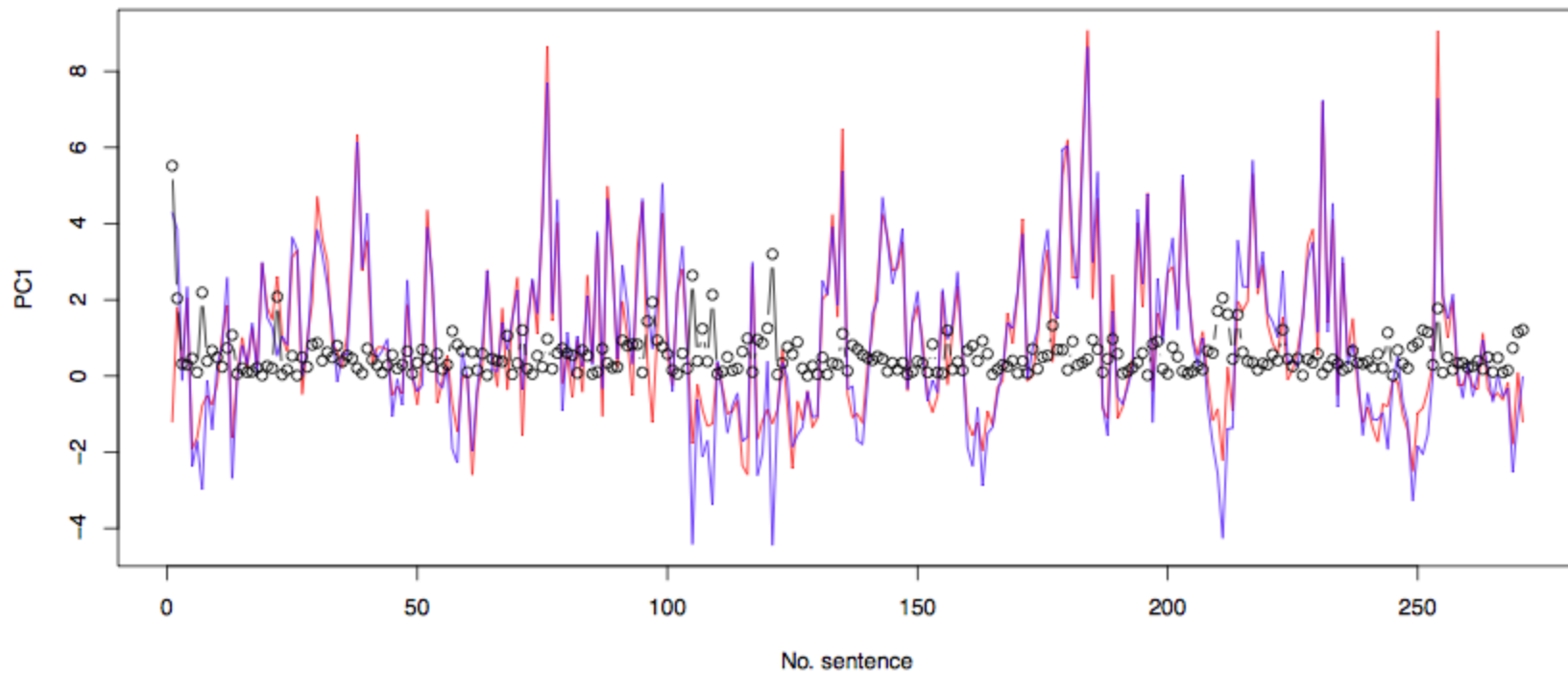
R was used to perform Multiple Linear Regression and Sequential Floating Forward Selection on all of the sentiment score features used in the previous experiment, producing the following parameter set:

$$\begin{aligned} PC1 = & -1.64 + 0.12 \times num_words_sentence \\ & - 48.0 \times ImdbEmphasis + 11.3 \times ImdbPolarity \\ & + 2.24 \times SentiWordNetNeg - 1.78 \times Teehee \\ & - 3.66 \times Understand - 1.17 \times OpinionLexicon \\ & + 0.6 \times Hugs + 0.44 \times SentiWordNetPos \end{aligned} \quad (5)$$

Model was tested on Chapters 1 and 2, which were annotated, and trained on the rest of the book.

Adding sentence length as a predictive feature helped to improve prediction error (1.21 --> 0.62)

chp02 PC1 prediction – blue: real red: predicted black: error



Results

The PC1 model does okay modeling speaker “expressivity”

Variations in performance between chapters

- Argued as owing to two observations:
 - higher excursion in Chapter 1 than in Chapter 2
 - Average sentence length was shorter in Chapter 1 than in Chapter 2
- These observations apparently confirm that shorter sentences tend to be more expressive

	Chapter 01		Chapter 02	
Character	Narrator	Other	Narrator	Other
Narrator	79.8	30.1	92.0	34.0
Other	20.2	69.9	8.0	66.0
Diagonal	73.3%		81.5%	

Table 3: Character prediction for chapters 01 and 02 using number of word, sentiment scores and the learnt model in equation 5.

Character	Predicted_PC1	Text
narrator	3.00	Soon the free boys would come tripping along on all sorts of delicious expeditions, and they would make a world of fun of him for having to work -- the very thought of it burnt him like fire.
narrator	2.24	He then held a position at the gate for some time, daring the enemy to come outside, but the enemy only made faces at him through the window and declined.
narrator	1.69	If one moved, the other moved -- but only sidewise, in a circle; they kept face to face and eye to eye all the time.
narrator	0.58	So she lifted up her voice at an angle calculated for distance and shouted:
	...	
narrator	0.05	Spare the rod and spile the child, as the Good Book says.
narrator	0.04	I reckon you're a kind of a singed cat, as the saying is -- better'n you look.
narrator	0.01	If you was to tackle this fence and anything was to happen to it -- "
other	-0.00	Another pause, and more eying and sidling around each other.
other	-0.00	Ben ranged up alongside of him.
other	-0.02	He opened his jacket.
other	-0.04	"Tom, it was middling warm in school, warn't it?"
other	-0.05	At this dark and hopeless moment an inspiration burst upon him!
	...	
other	-1.87	"Nothing."
other	-1.96	"Aw -- take a walk!"
other	-1.97	I'll learn him!"
other	-1.98	"By jingo!
other	-2.11	"You can't."
other	-2.13	Course you would!"
other	-2.16	"Y-o-u-u TOM!"
other	-2.17	Oh, what a hat!"
other	-2.17	"Well why don't you?"
other	-2.18	Why don't you DO it?"
other	-2.99	"Nothing!"
other	-3.20	Ting-a-ling-ling!
other	-3.20	Chow-ow-ow!
other	-3.20	Ting-a-ling-ling!
other	-3.20	SH'T!

Table 4: Predicted PC1 value and corresponding text for some sentences of chapter 01.

Conclusions

Findings:

- correlations exist between Acoustic Energy/F0 and movie reviews/emotional categorizations
- sentiment scores can be used to predict a speaker's expressivity

Applications:




- automatic speech synthesis

Future Work

- Train a PC1 predictor to be able to predict more than two styles

Sentiment Analysis of Online Spoken Reviews

Sentiment classification using manual vs automatic transcription

		
<p>My Favorite Winter Nail Polish!</p> <p>Reviewer: Danielle T. Brand: Essie Category: Makeup</p>	<p>Gillette Fusison Proglide Power</p> <p>Reviewer: Pete M. Brand: Gillette Category: Shavers</p>	<p>Not a fan of this tool set.</p> <p>Reviewer: Casey C. Brand: Craftsman Evolv Category: Tools & Hardware</p>

Goals of the paper

- Build sentiment classifier for video reviews using transcriptions only
- Compare accuracy of manual vs automatic transcriptions
- Compare spoken reviews to written reviews

Dataset

- English ExpoTv video reviews
 - 250 fiction book reviews
 - 150 cell phone reviews
- Each video includes star rating
- Average length 2 minutes
- Amazon reviews



Two Transcription Methods

- Manual transcriptions through MTurk
- Automatic transcriptions through Google's YouTube API
 - Unable to automatically transcribe 22 videos

Sentiment Analysis

- Unigrams (no improvement found with ngrams)
- Group words into sentiment classes using OpinionFinder, LIWC, WordNet Affect

Class	Words
Opinion Finder	
POSITIVE	abundant, eager, fortunate, modest, nicely
NEGATIVE	abandon, capricious, foul, ravage, scorn
NEUTRAL	absolute, certain, dominant, infectious
LIWC	
OPTIM(ISM)	accept, best, bold, certain, confidence
TENTAT(IVE)	any, anyhow, anytime, bet, betting
SOCIAL	adult, advice, affair, anyone, army, babies
WordNet Affect	
ANGER	wrath, umbrage, offense, temper, irritation
JOY	worship, adoration, sympathy, tenderness
SURPRISE	wonder, awe, amazement, astounding

Results

Manual vs automatic - Loss of 8-10%

Features	Cellphones		Fiction Books	
	Manual	Automatic	Manual	Automatic
Uni	73.23	62.58	75.42	67.76
Uni+LIWC	74.64	63.94	74.15	67.79
Uni+OpF	72.53	61.90	74.15	66.94
Uni+WA	72.53	62.58	75.00	67.79
Uni+LIWC+OpF+WA	75.35	65.98	72.88	67.37

Spoken vs Written

Features	Cellphones		Fiction Books	
	Spoken	Written	Spoken	Written
Uni	73.23	71.12	75.42	84.32
Uni+LIWC	74.64	76.05	74.15	86.01
Uni+OpF	72.53	71.83	74.15	83.89
Uni+WA	72.53	71.83	75.00	84.32
Uni+LIWC+OpF+WA	75.35	75.35	72.88	86.01

Conclusion

- Sentiment classification of video reviews can be done using only transcriptions
- 8-10% accuracy is lost using automatic transcriptions instead of manual
- Spoken reviews lead to equal or lower performance compared to written
 - Likely due to reliance on untranscribed cues
- Future work: compare video reviews to spoken (non video) reviews