Towards Multimodal Sentiment Analysis

Rada Mihalcea

University of North Texas

Joint work with Veronica Perez-Rosas and Louis-Philippe Morency

What is Sentiment Analysis?

- Subjectivity and sentiment analysis focuses on the automatic identification of private states in natural language (Wiebe et al., 2005)

"I love Romania."

Subjectivity Analysis	Subjective			Objective
Sentiment Analysis	Positive	Negative	Neutral	---

Applications

- Businesses and organizations:
  - product and service benchmarking
  - market intelligence
  - Business spends a huge amount of money to find consumer sentiments and opinions.
    - Consultants, surveys and focused groups, etc
- Individuals: interested in other's opinions when
  - purchasing a product or using a service,
  - finding opinions on political topics
- Ads placements: Placing ads in the user-generated content
  - Place an ad when one praises a product.
  - Place an ad from a competitor if one criticizes a product.
- Opinion retrieval/search: providing general search for opinions.

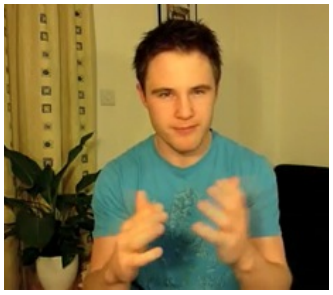
Applications

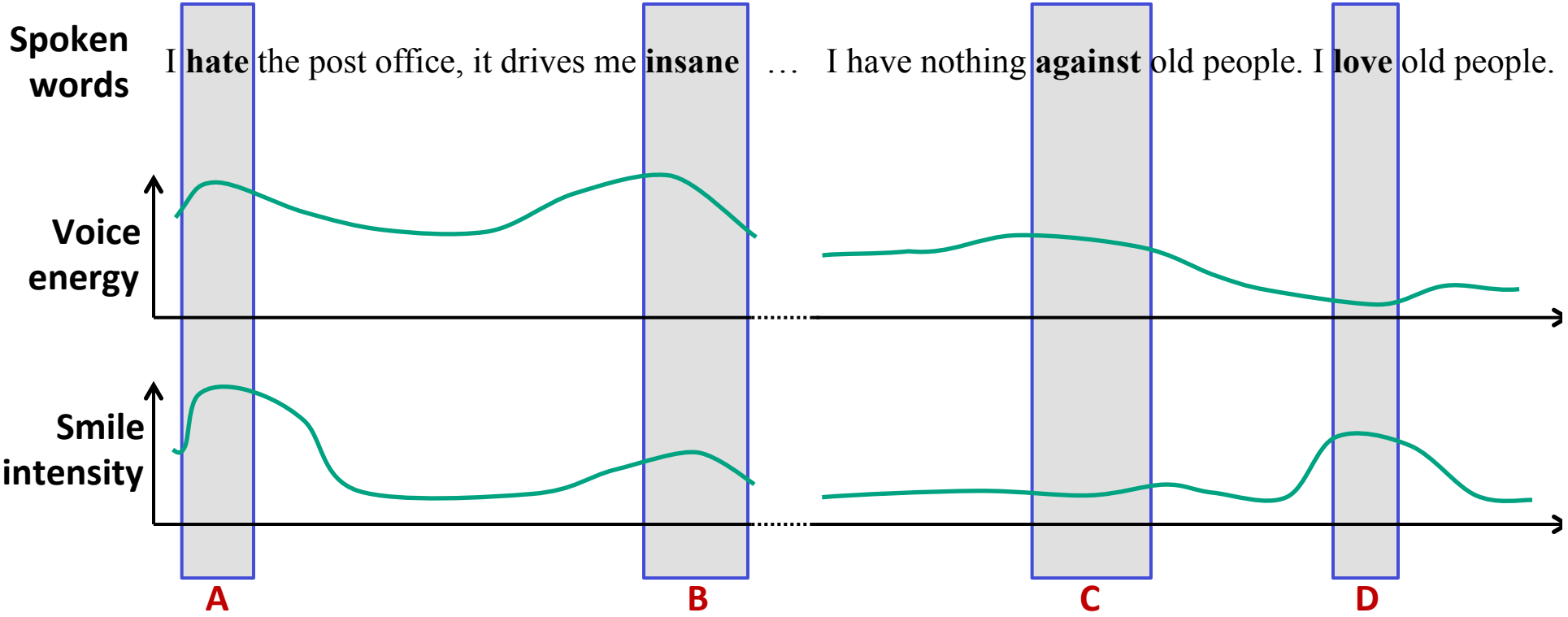
- Tracking sentiment timelines in news (Lloyd et. al, 2005)
- Review classification (Turney, 2002; Pang et. al, 2002)
- Mining opinions from product reviews (Hu and Liu, 2004)
- Expressive text-to-speech synthesis (Alm et. al, 2005)
- Text semantic analysis (Wiebe and Mihalcea, 2006) (Esuli and Sebastiani, 2006)
- Question answering (Yu and Hatzivassiloglou, 2003)
- Text summarization (Carenini et al., 2008)

Multimodal Opinions

- Most work to date applied on text
  - Knowledge-based methods (lexicons)
  - Data-driven methods (large annotated corpora)
- Multimodal information is continuously growing
  - Human communication is by nature multimodal
  - Day-to-day interactions, news, movies, interviews
- Multimodal information online
  - YouTube, Vimeo, VideoLectures, ExpoTV
    - E.g., 10,000 new YouTube videos every day!
  - Flickr, Picasa, Facebook
- Growing number of video opinions




Spoken words
I **hate** the post office, it drives me **insane** … I have nothing **against** old people. I **love** old people.
Voice energy
Smile intensity
A
B
C
D

Initial Study

YouTube Dataset

47 videos from YouTube

- Collection
  - Keyword-based search
    - opinion, review, product review, best perfume, toothpaste, war, job, business, cosmetics review, camera review, baby product review, I hate, I like
  - Truncated to 30-second clips to address introductory titles and multiple topics
- Manual transcriptions
- Sentiment annotations at video level
  - Positive, neutral or negative
  - Three novice annotators
  - 78.7% coder agreement
- 13 positive, 22 neutral, 12 negative

Feature Modeling: Linguistic

- Tri-modal features
- **Polarized words**: positive and negative words
  - Use a large pre-compiled lexicon (MPQA)
    - good, bad, nice, terrible
  - Use valence shifters
    - not, no, without, neither
  - Potential issue: scarcity of the clues in spoken utterances

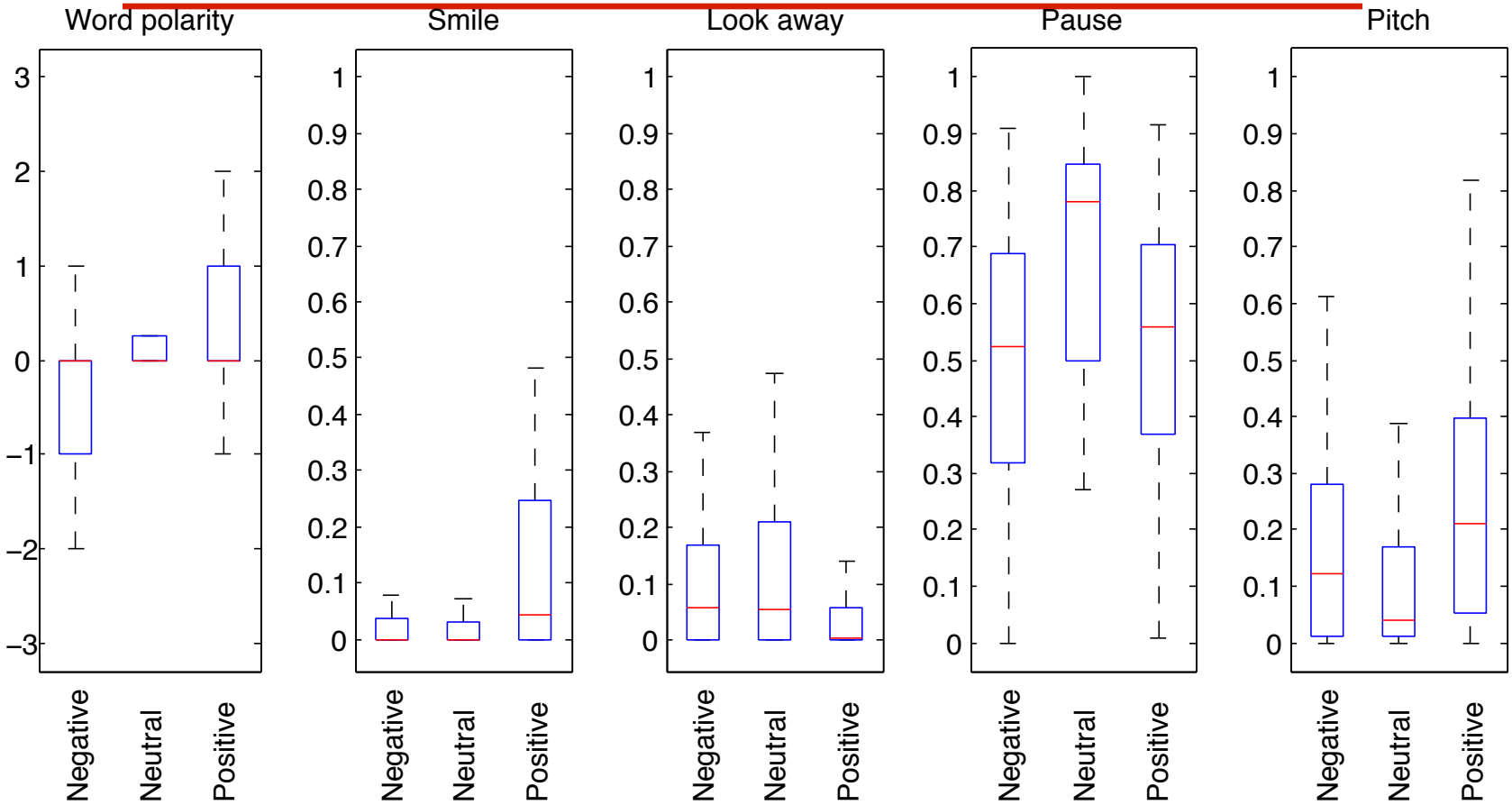
Feature Modeling: Audio

- **Pause duration**: Given the start and end time of the utterance, how many audio samples are identified as silence
  - Normalize by the number of audio samples in the utterance.
  - Feature represents percentage of the time where the speaker was silent
- **Pitch**: Compute the standard deviation of the pitch level for the spoken utterance
  - Measure represents the variation of voice intonation during the same utterance

Feature Modeling: Video

- **Smile duration**: Given the start and end time of an utterance, how many frames are identified as smile
  - Two thresholds for intensity: 50 and 75
- **Look-away duration**: Given the start time and end time of the utterance, in how many frames is the speaker looking at the camera
  - Horizontal and vertical angular thresholds were experimentally set to 10 degrees

Statistical Analysis

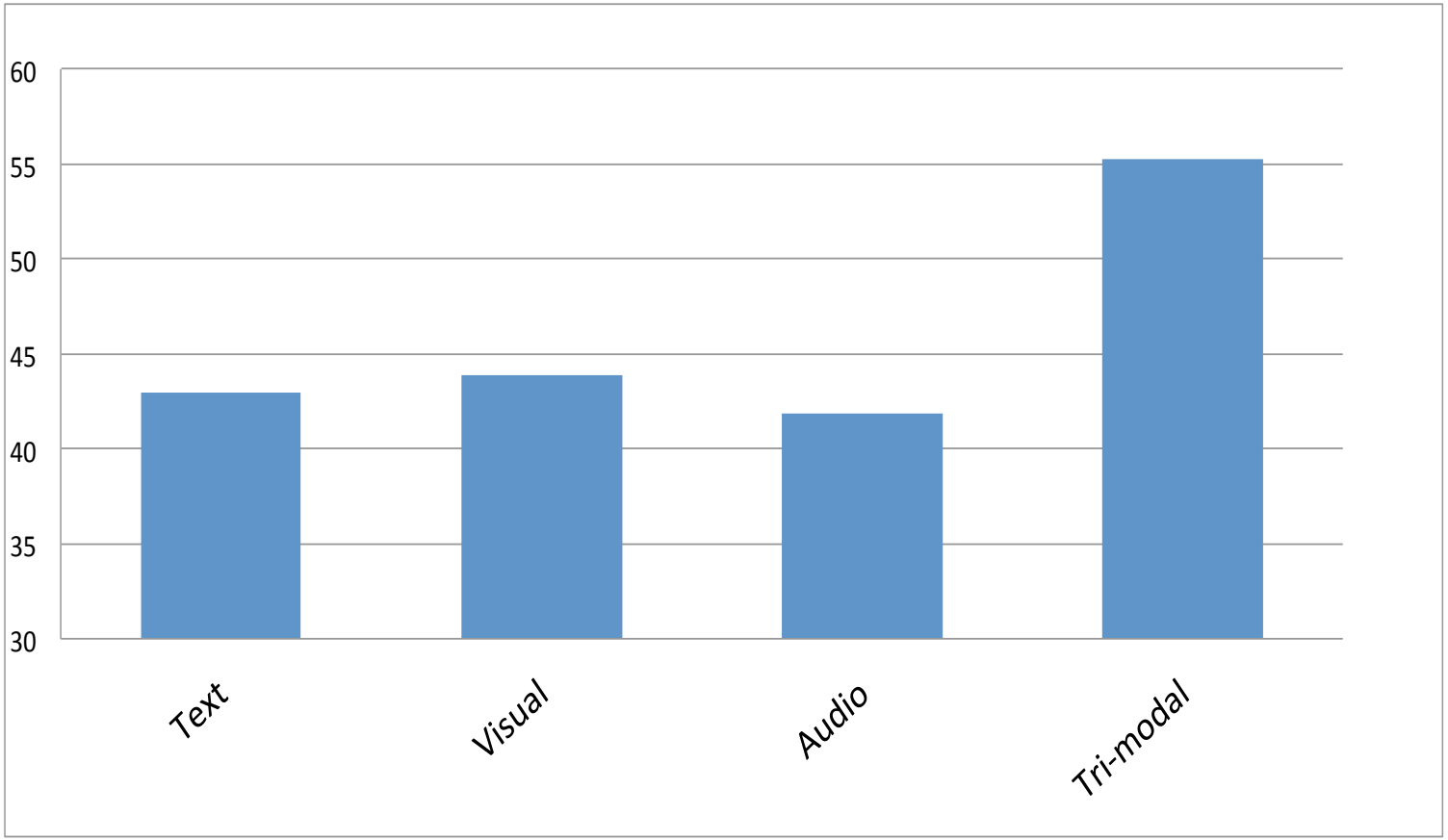


Word polarity is a great way to differentiate sentiment but many utterances do not contain polarized words as shown by all three medians equal to zero. The visual features (smile and look away) are good ways to differentiate positive utterances from neutral or negative utterances. Audio features (pauses and pitch) are great ways to differentiate neutral utterances from positive or negative utterances.

Classifier

- HMMs that take as input tri-modal features summarizing each utterance
  - Each YouTube video clip is modeled using a Markov chain where each element of the chain represents one spoken utterance
- Leave-one-out testing where one video clip is kept for testing and all remaining 46 clips are used for training and validation

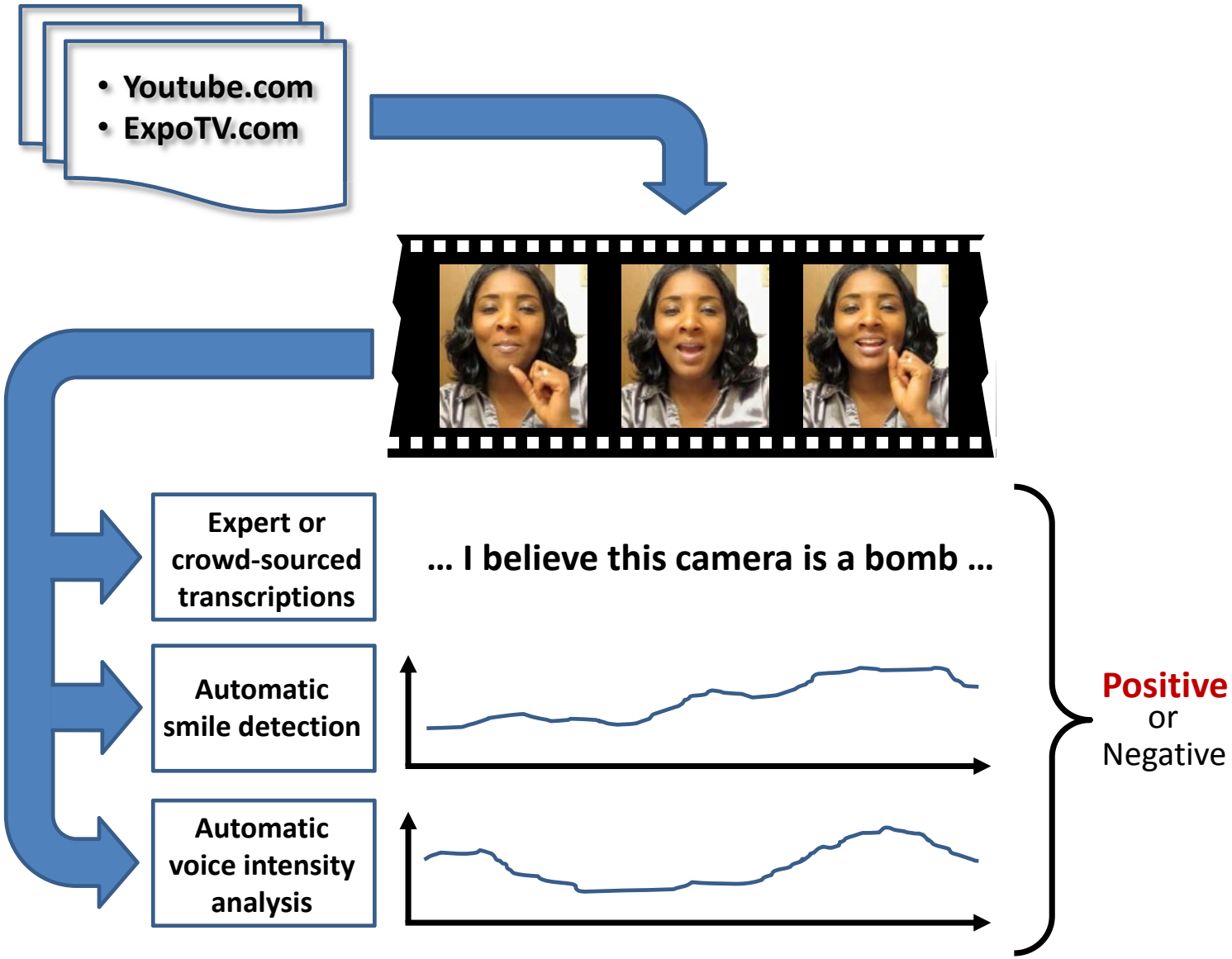
Results



Automatic sentiment analysis for one modality at a time and for a multimodal model

Second Study

* Focus on sentiment only
* Larger dataset
* Language portability


• Youtube.com
• ExpoTV.com
Expert or crowd-sourced transcriptions
… I believe this camera is a bomb …
Automatic smile detection
Automatic voice intensity analysis
Positive
or
Negative

A Spanish Multimodal Dataset

- 105 videos from YouTube
- Only videos with opinions (positive, negative)
- Collection
  - Keyword-based search
    - mi opinion (my opinion), mis products favoritos (my favorite products), me gusta (I like), no me gusta (I dislike), products para bebe (baby products), mis perfumes favoritos (my favorite perfumes), peliculas recomendadas (recommended movies), opinion politica (politic opinion), video juegos (video games) and abuso animal (animal abuse)
  - Truncated to 30 sec. clips to address introductory titles and multiple topics

A Spanish Multimodal Dataset

- Speakers
  - 21 males, 84 females
  - 15-60 years old
  - Various Spanish-speaking countries

Sentiment Annotations

- 30 sec. clips manually transcribed
  - 550 utterances, 10,000 words
- Sentiment annotations:
  - Two expert annotators
  - Inter-annotator agreement of 92%, kappa = 0.84
- Final annotation: 47 positive, 54 negative, 4 neutral
  - Neutral annotations are not used in the experiments
- Baseline of 51%

Feature Modeling: Linguistic

- (Almost) the same tri-modal features as before
- **Polarized words**:
  - No large polarity lexicon for Spanish
  - Learned from the data
  - Unigrams occurring in training data

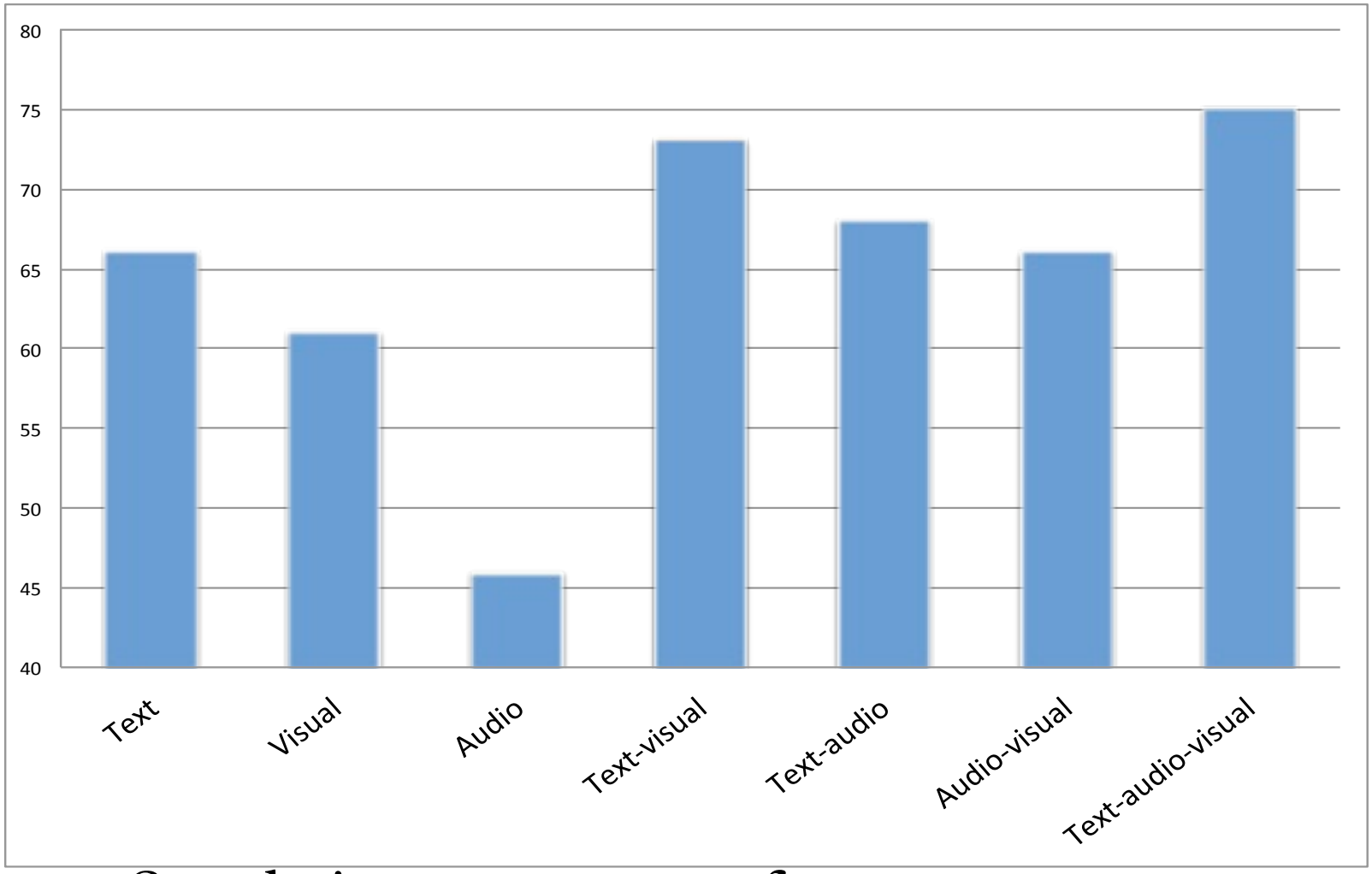
Feature Modeling: Audio

- **Pause duration**: Given the start and end time of the utterance, how many audio samples are identified as silence
- **Pitch**: Compute the standard deviation of the pitch level for the spoken utterance
- **(+)Intensity**: Measure the sound power of the spoken utterances in the video. We compute the average voice intensity over the whole video
- **(+)Loudness**: Determine the perceived strength of the voice factored by the ear's sensitivity. We compute the average loudness measure over the entire video

Feature Modeling: Video

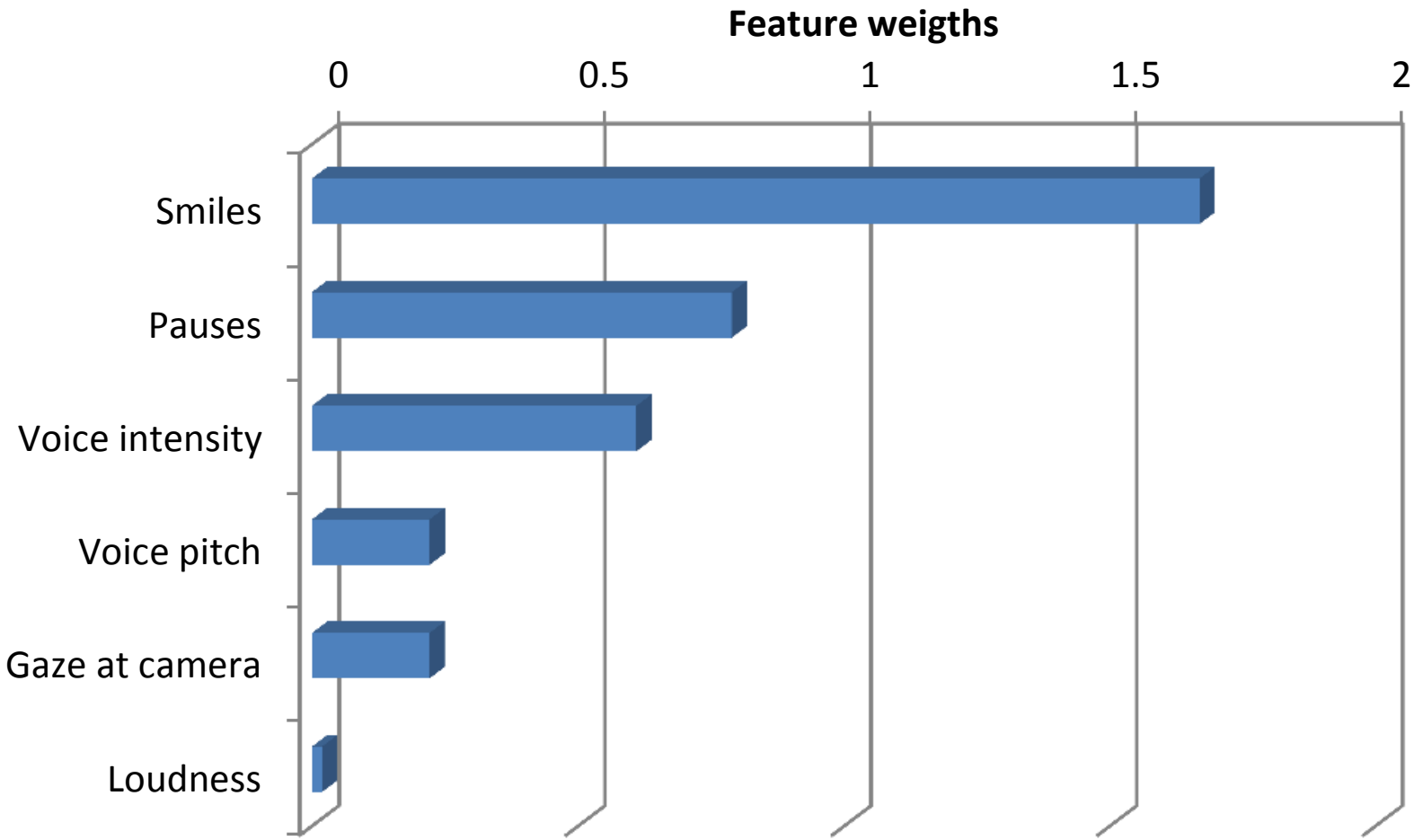
- **Smile duration**: Given the start and end time of an utterance, how many frames are identified as smile
- **Look-away duration**: Given the start time and end time of the utterance, in how many frames is the speaker looking at the camera

Results



- Steady improvements from one to two to three modalities at a time

Feature Analysis



Utterance-level Sentiment Analysis

- Finer-grained sentiment analysis by analyzing reviews at utterance level
- Can we reliably distinguish between positive and negative opinions when the video segments are small?
- Do we still obtain similar improvements through the use of multiple modalities?
- Are there any differences between the analysis performed at video and utterance level?

Dataset

- Each video (30 sec. length) was segmented at utterance level
  - Manual segmentation, based on pauses and complete sentences
  - Each video has eight utterances on average
  - Each utterance has an average duration of 3 sec.
- Final dataset: 498 utterances
    - 202 positives
    - 248 negatives
    - 48 neutral

Segmentation Sample

(start time, end time, utterance, sentiment annotation)

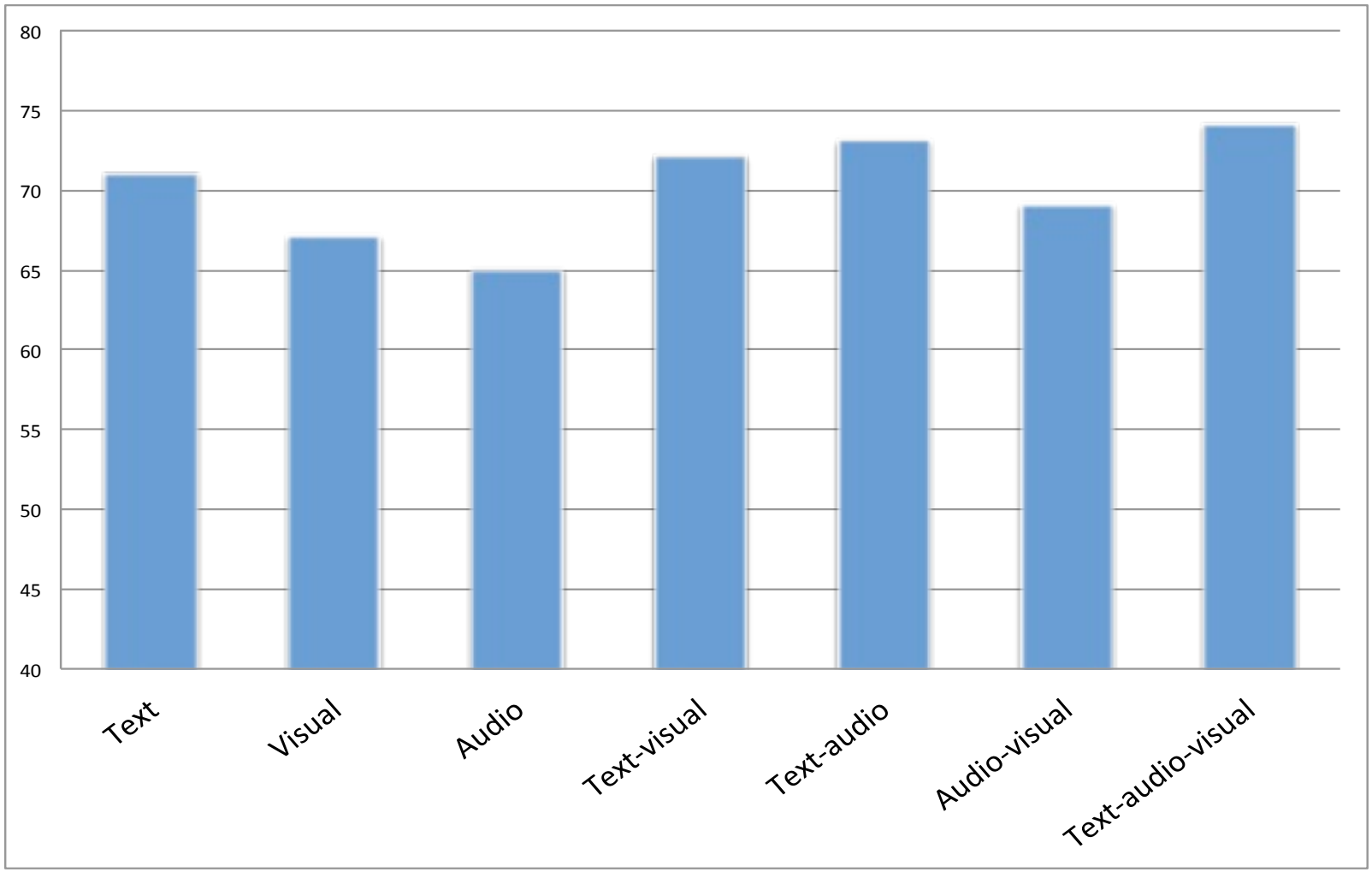
0.000;3.540; "mm: y- hara su efecto aunque: no lo notes";"1"

3.540;13.890; "pero lo que no me gusta de este: de este producto, es que cuando les das, a ver si lo puedo hacer, sale muchi:simo, disparado en linea recta.";"-1"

13.890;21.122; "En vez de ser un spray que se difumine y te caiga mas producto en el pelo, te cae solamente en un sitio del pelo muchisima cantidad de producto.";"-1"

21.122;26.010;"Asi que: sera buen producto pero yo no: eh: no me gusta ";"-1”

Results



- Steady improvements from one to two to three modalities at a time

Conclusions

- Multimodal analysis of sentiment in online videos
  - Large number of videos uploaded everyday on social media and elsewhere
- Textual, visual, and audio features complement each other
- Similar effect observed on English and Spanish videos
- Similar effect observed on longer (full video) and shorter (utterance) levels

Questions?