Yash Singh & Madhav Sharan

2269137173, 6096558650

CS535 Fall 2016 - HW1

1. While watching videos of the data set, we observed differences between positively, negatively and neutral annotated video segments. We watched the whole video set multiple times, sometimes muting the audio and few times just listening to audio to gather some understanding on non verbal expressions. While watching we identified certain audio-visual cues which might help us classify these video frames and could be worth analyzing statistically. A brief description about what different thoughts are conveyed by speakers is given below. Details about behavioral cues are given after the description.

   (a) **Positive :**
   Brief overview: We found that most of the video frames which annotators rated positive (1) expressed love, praise, linking, advice, hope or admiration for a thing or person,. E.g.:

   - video 8: The speaker suggests rinsing mouth and brushing teeth after eating/drinking food.
   - Video 19: The speaker admires and congratulates new heavy weight champion.
   - Video 25: The speaker talks about how much he loves movie Avatar
   - Video 10: The speaker expresses liking for cradle.
   - Video 6: In the latter part, speaker expresses liking for a lesbian couple.
   - Video 19: The speaker wishes well being for soldiers.
   - Video 27: The speaker appreciates a face wash

   (b) **Negative :**
   We believe that annotators rated video frames as Negative (-1) which involved speaker expressing dislike, anger, hatred, frustration and complaints. E.g.:

   - Video 3: The speaker expresses his strong dislike for star fox.
   - Video 4: The speaker expresses hatred for the post office because of long queues.
   - Video 6: In the beginning speaker expresses his dislike towards one of his customer who didn't keep her time commitment.
   - Video 7: The speaker expresses disliking for a customer who couldn't differentiate between two different sodas.
   - Video 9: The speaker complains about a product.
   - Video 12: The speaker expresses dislike for exercising.
   - Video 15: The speaker expresses dislike for President Obama.
   - Video 16: The speaker shows his dislike for a company and complains about employees not getting compensation.
   - Video 28: The speaker expresses strong dislike for Mathematics
   - Video 29: The speaker expresses strong dislike for people who ask for advice but do not listen to it.
   - Video 30: The speaker expresses dislike for a song.
   - Video 32: The speaker complains about difficulties faced in his life.
   - Video 33: The speaker complains about the system which favors Jews over Muslims.
   - Video 35: The speaker expresses dislike and opposes the statement given by Sarah Palin.
   - Video 37: The speaker expresses strong hatred for a man.
   - Video 44: The speaker expresses strong hatred and disliking towards people driving rashly.

   (c) **Neutral :**
   The video segments annotated as neutral(0) involved speakers not displaying any affinity or hatred. Rather they speak about themselves, random incidents or generic fact about things happening around. E.g. :

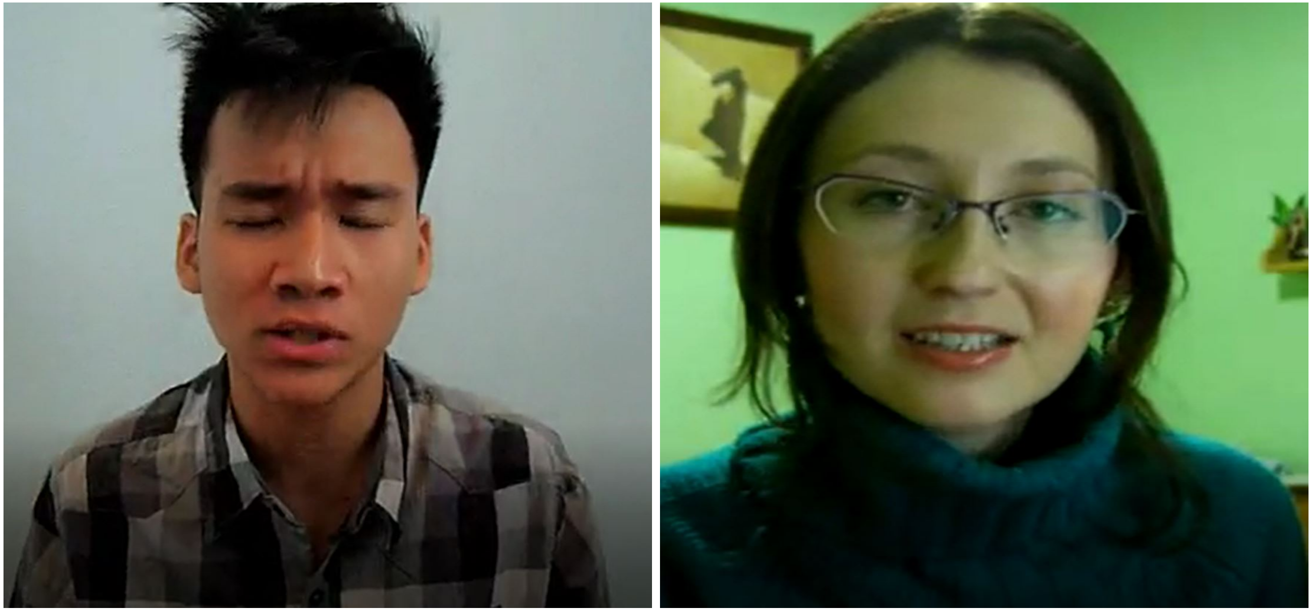
   - Video 1: President Obama talks about himself in the beginning
   - Video 5: Speaker talks about an incident
   - video 13: speaker talks about himself reading a blog

Nonverbal cues: On the visual inspection of the videos, we found a number of features which could be interesting in classifying videos as positive, negative or neutral. Explanation of these cues is provided below:

i. **H1:Eyebrow frowns**

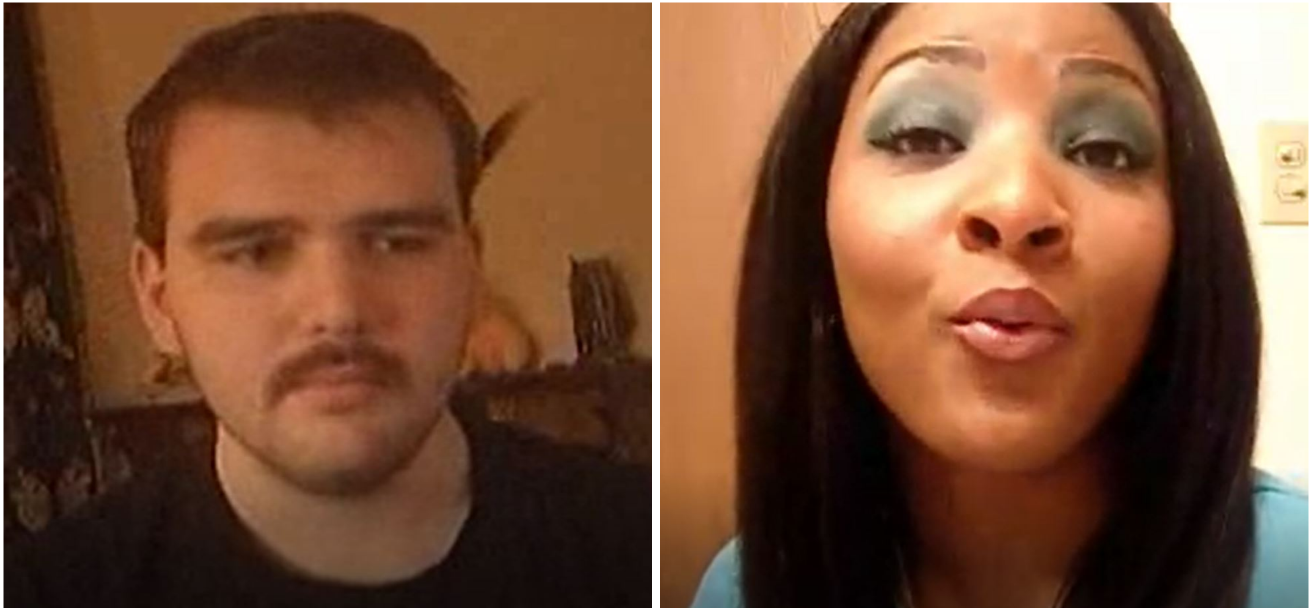
A. We found that speakers in negative video frames tend to have middle of the eyebrows pulled down which slope inwards. It often showed that the person was angry or frustrated. It also indicated intense concentration. On the other hand, positively annotated speakers seemed to be relaxed and did not display pulled eyebrow expression that often. This can be seen in the given below image:

Figure 1: Eye brow frown variation: frown for negative(left), no frown in positive/neutral(right)



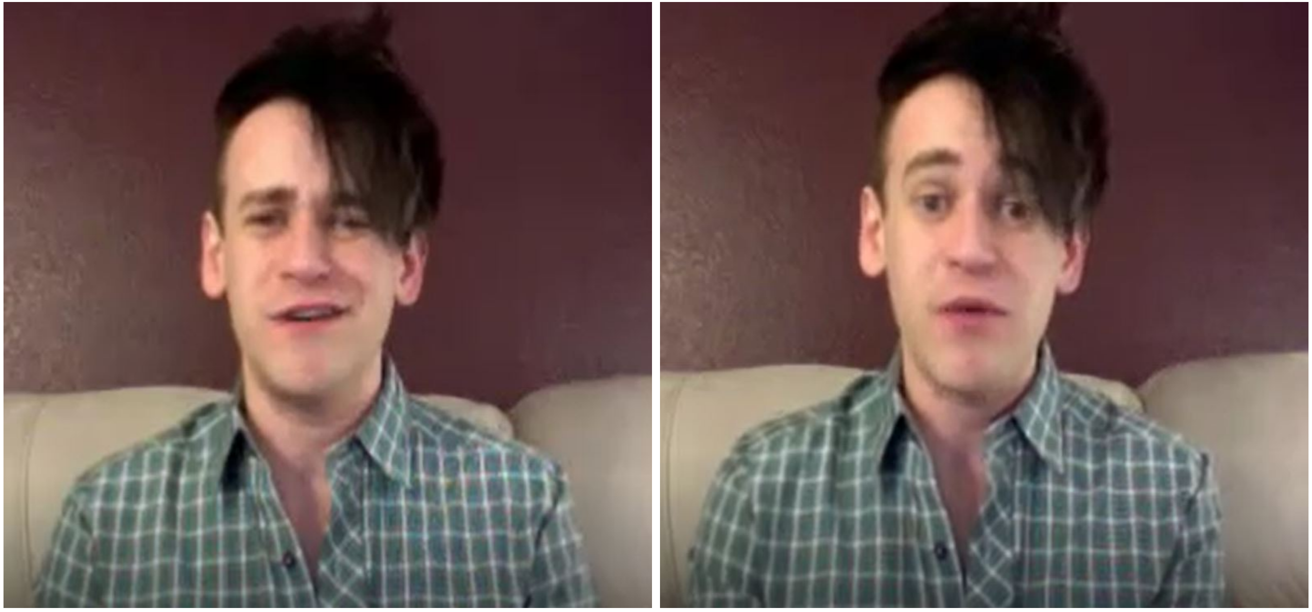
ii. **H2: Eyebrow raise** We also observed that in positively annotated videos, speakers tend to raise their eyebrows quite often. This can be attributed to surprise, joy or happiness. On the other hand negatively annotated speakers do not raise eyebrows that often. This can be attributed to tension / frustration/ deep thought. This can be seen in the images below:

Figure 2: Eye brow raise variation: lowered eyebrow for negative(left), raised brow in positive(right)



iii. **H3: Eyes openness** We also observed that positively annotated people tend to have more eye openness. On the other hand negatively annotated people tend to have lesser eye openness. Below is the snapshot of frames for same person which have been negatively and positively annotated:

Figure 3: Eye Openness: Negatively annotated on left, Positively annotated on right



iv. **H4: Head nods** We observed that positively and negatively annotated people tend to nod their head more. On the other hand neutral video frames did not display much up down motion. This can be explained as : positive and negative class people are considerably more expressive as compared to neutral people. The expressiveness can be observed as head movement specially as head nod. Hence we observed more head nods in positively and negatively annotated segments.

v. **H5: Smile** We feel that since positively annotated users tend to express admiration or likeliness they are more likely to smile as compared to people expressing hatred, anger and dislike. Although we didn't find a strong correlation between smile and the classification, we thought that it would be interesting to analyze this feature statistically.

vi. **H6: Voice tenseness** Audio in positively annotated video frames were found to be more tense as compared to negative video frames. This can be explained because while expressing love, likeliness or admiration towards something, speaker seems to take time to express it. On the other hand negative class speakers, while expressing dislike, hatred or frustration, tend to let their feelings go away. In this process their voice tend to possess less tenseness.

E.g. - Segments in videos 18, 19, 25, 27 and 10 are positively annotated and speaker in these have tense voice. On the other hand, 28, 29, 44, 30 are negatively annotated and speakers in these expressed anger and had lesser tense voice.

vii. **H7: Word elongation** While listening to speakers in video we observed difference in way they lengthen on some words. We saw people in positive class elongate more than people in negative class. For example in video 4 as soon as speaker shifted to positive class frequency of lengthening increases. From 30-33 seconds speaker stretches on words as "u:p me: do they lo:ve a post office they lo:ve i:t"

Verbal cues: We also observed that speakers who express frustration, use obscene words very frequently. Hence segments containing more curse words are more likely to be negative. An example from video 28 where speaker talk about mathematics - ”*in the ass i hate it its pointless it makes no sense everything about math PISSES me off. mainly because it's USELESS outside of school i mean if i'm going to be going to school to learn something at least let it be useful like how to suck a dick*”

Word cloud given below summarizes the word distribution:

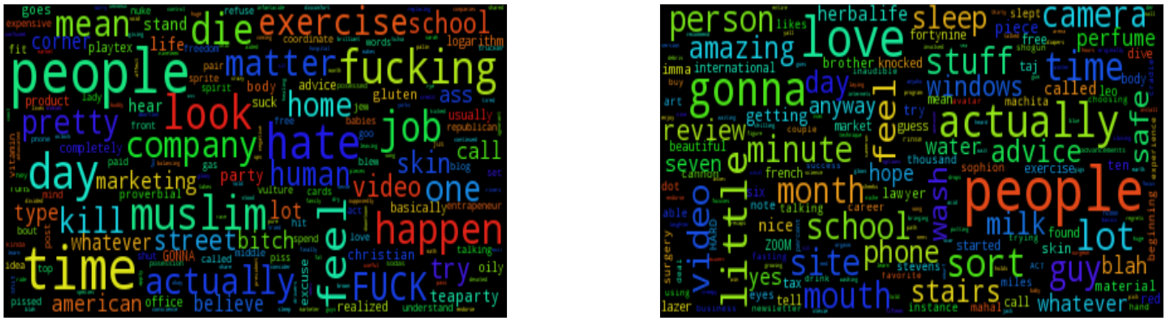


Figure 4: Word cloud for negative(left) and positive (right) segments

2. Mean Variance per item: 0.210714286 Mean: -0.066666667 Variance: 0.710687326 As taught in class

$$\alpha = 1 - \frac{ErrorVariance}{TotalVariance}$$

$$\alpha = 1 - \frac{0.210714286}{0.710687326}$$

$$\alpha = 0.703506341$$

As stated by Krippendorff : $\alpha > .8$ is a good reliability and with $.67 < \alpha < .8$ we can draw tentative conclusions. Since in our case k = 0.7035, we can draw tentative conclusions from our data set.

3. After selecting features which looked interesting, we conducted ANOVA test and analyzed if they are statistically significant. We didn't choose T-Test because we found that these features are not following normal distribution. Also, we chose standard deviation for aggregating frames for annotated segments. This is because standard deviation can tell us about the variation of the feature in that segment. And in the given cases, we are interested in analyzing that variation.

Below are the details regarding the statistical analysis of features which were interesting:

(a) **Anger:** This feature was extracted from the Shore output file. It had p-value = 0.0409475 and F-value = 3.232612453. Below is its box plot:

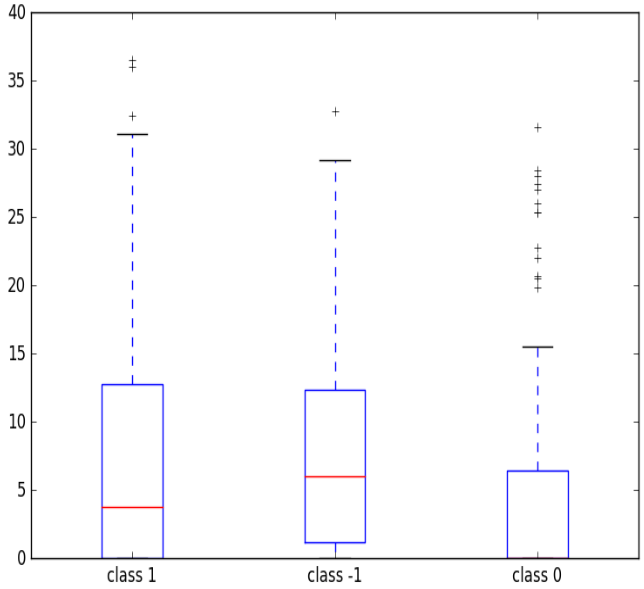


Figure 5: Anger: As can be seen negatively annotated speakers are found to be more angry an compared to positive and neutral

(b) **Eyebrow Frown:** This feature was extracted from CLM output file. Below is the image of format that CLM follows. We can consider the eye frown as variation of distance between point 22,23.

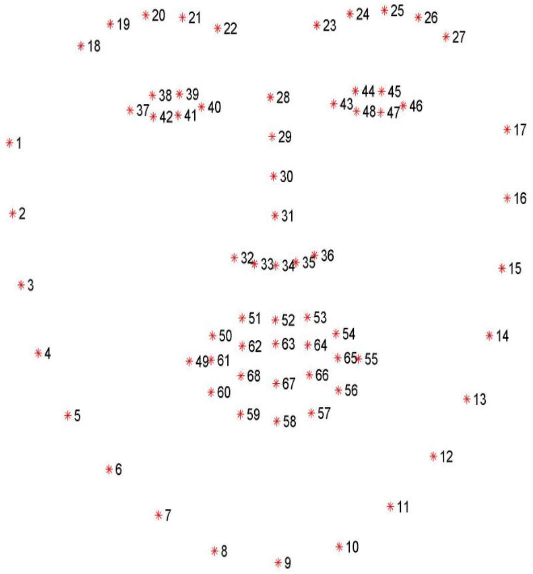


Figure 6: points captured in CLM output

On conducting ANOVA over brow frown we found p = 0.009970944 and F = 4.685595783. Below is its box plot:

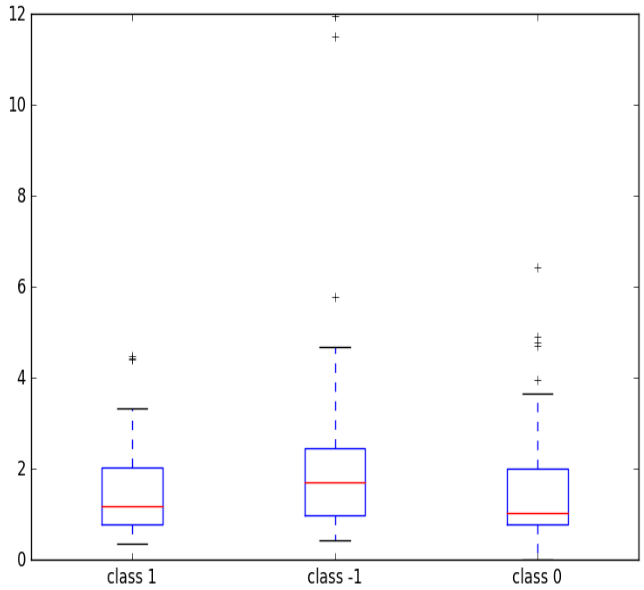


Figure 7: Eyebrow frown: as can be seen negatively annotated speakers tend to frown more often

(c) **Eyebrow raise** This feature was also calculated using the features from CLM file. i.e. distance between point 20,38 and 25,45. On conducting ANOVA, we found p = 0.097296628 and F = 2.349700082. Below is its box plot:

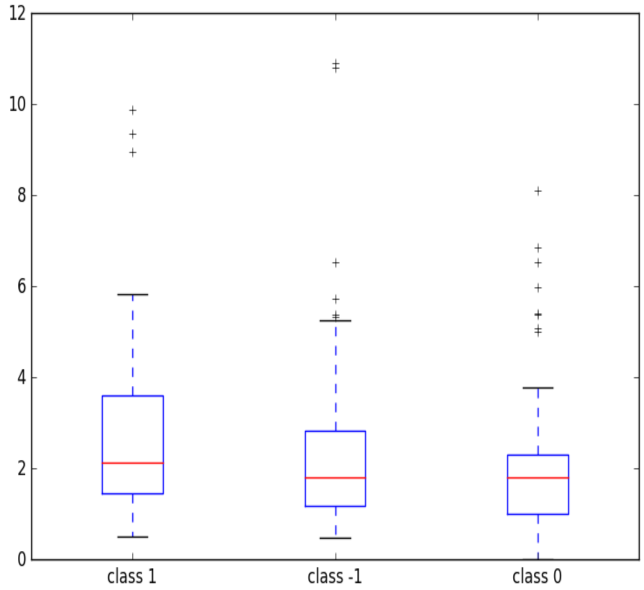


Figure 8: Eyebrow raise: as can be seen there isn't any statistically significant difference in the positively and negatively annotated distributions

(d) **Right Eye openness** This feature is available in OKAO file. On conducting ANOVA, we found p = 0.011640962 and F = 4.525591583. Below is its box plot:

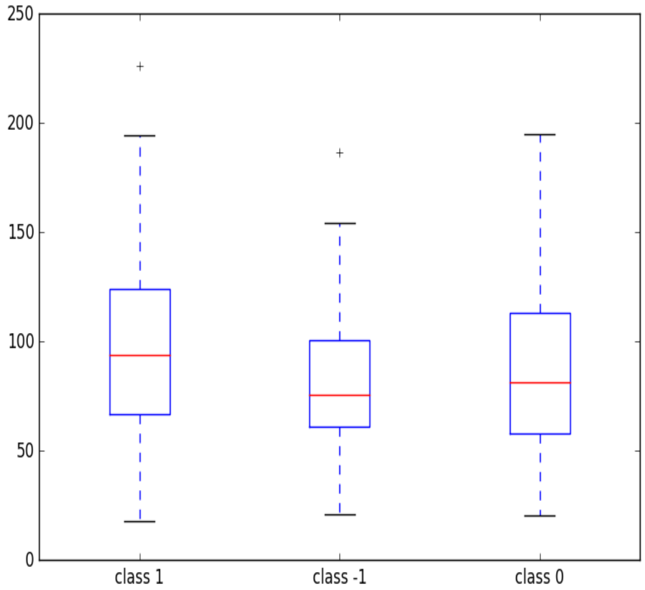


Figure 9: Right eye openness: as can be seen positively annotated speakers used to open up their eyes more often

(e) **Left Eye openness** This feature is available in OKAO file. On conducting ANOVA, we found p = 0.25231465 and F = 1.383947116. Below is its box plot:

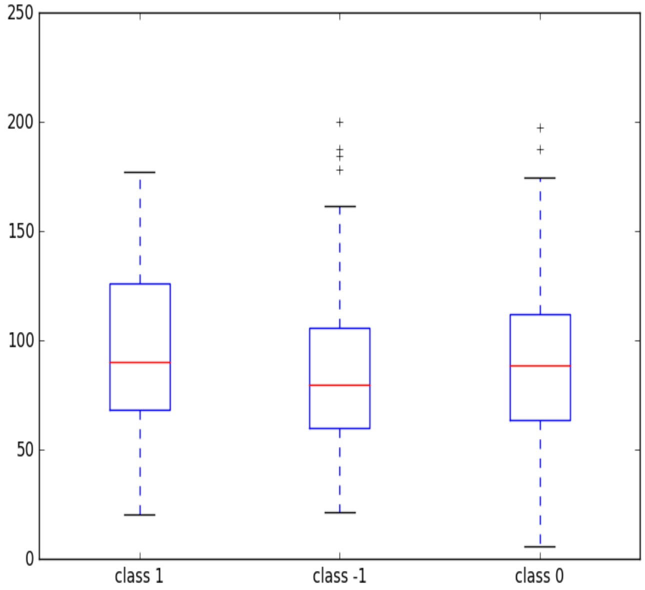


Figure 10: Left eye openness: figure shows that there is difference in mean between positive and negative class but on analyzing ANOVA p-value, we observed that the difference is not statistically significant.

(f) **Smile** This feature is available in OKAO file. Corresponding p-value = 0.808139757 and F-value = 0.213184171. Below is the corresponding box plot:

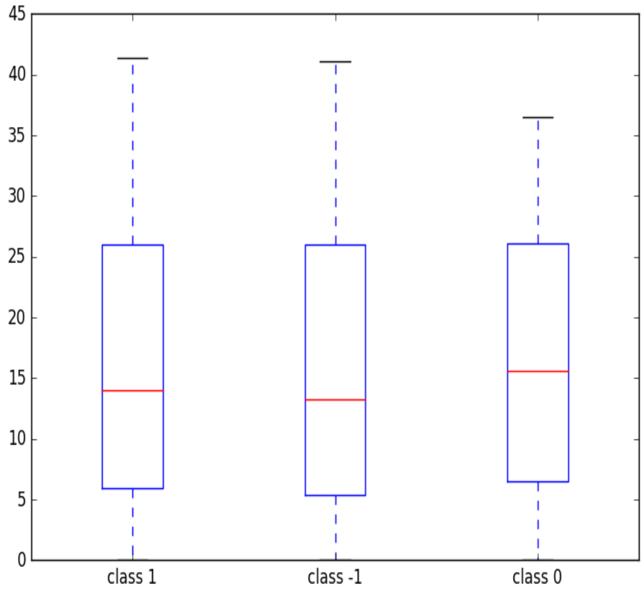


Figure 11: Smile: as can be seen there isn't statistically significant difference between the distribution of positive and negative speakers smile variation

(g) **Head nod** This feature is available as face up down feature in OKAO file. Corresponding p-value = 0.034638331 and F-value = 3.403951288. Below is the corresponding box plot:

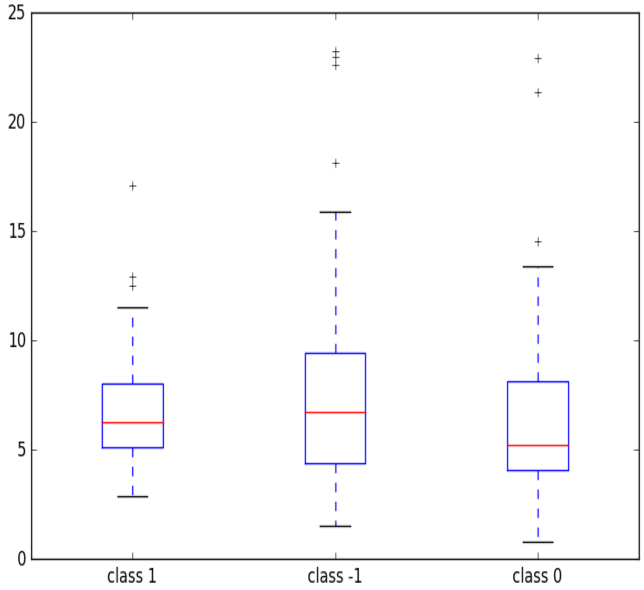


Figure 12: Head nod: as can be seen positively and negatively annotated speakers tend to nod their head more frequently as compared to neutral speakers

(h) **Voice tenseness** We considered NAQ feature available in feat files for voice tenseness. We got the corresponding p-value as 0.010443366 and F-value = 4.637745991. Below is the corresponding box plot:

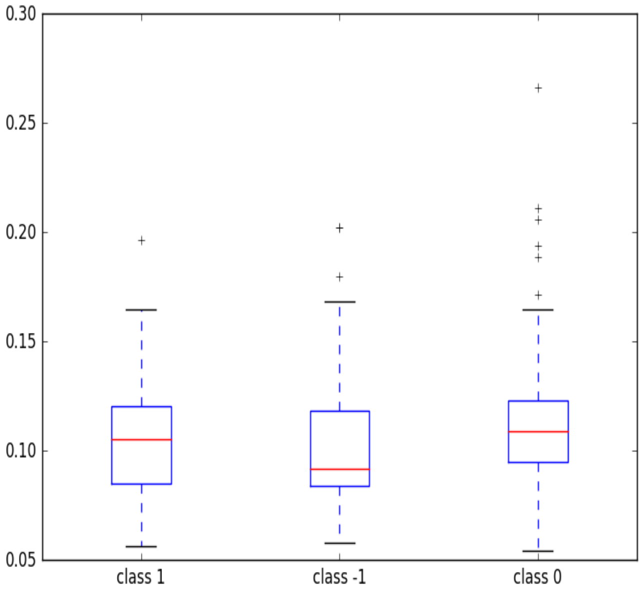


Figure 13: Voice tenseness: as can be seen positively and neutral annotated speaker showed more variations in voice tenseness compared to negatively annotated speakers

(i) **Word elongation-** We could observe statistically that speakers while talking positively elongate more than while talking negatively. Below is a excerpt from transcribed file from video 4 notice colons -

```
<Sync time="30.559"/>
u:p me: do they lo:ve a post office they
<Sync time="32.995"/>
/
<Sync time="33.802"/>
lo:ve i:t
<Sync time="34.322"/>
/
<Sync time="34.894"/>
they're dra:wn to it
<Sync time="35.986"/>
```

Some statistical numbers and box plot-

```
p-value = 0.006274897 and f-value = 5.165183207
Mean of standard deviation of elongation per utterance per segment -
 Positive = 0.777
 Negative = 0.667
Maximum frequency of elongation in a segment -
 Positive = 2.5 per second
 Negative = 1.9 per second
```

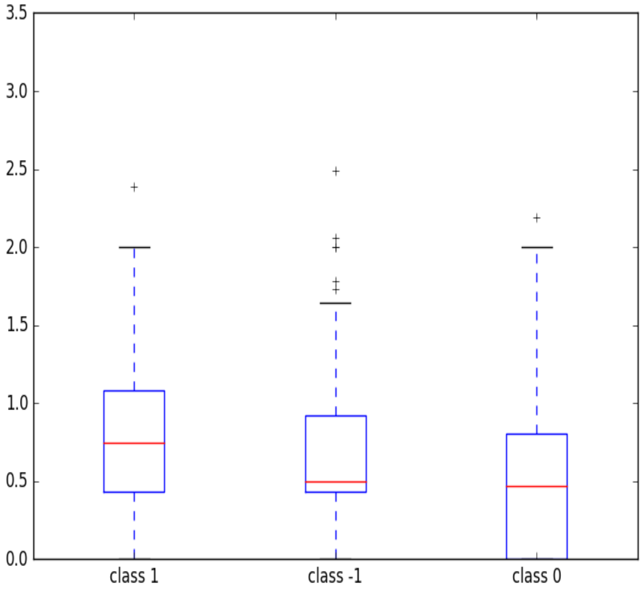


Figure 14: Word elongation: as can be seen positively annotated speaker tend to stretch more as compared to negatively annotated speakers

(j) **Obscene words** We could prove statistically that obscene/curse words impacts highly of negativity of conversation. There were instances when speaker used such word in positive annotated segments but negative segments dominated this feature. Below are some numbers supporting the argument and then a histogram depicting frequency per segment.

```
p-value = 0.000104929 and f-value = 9.472074218
 Total count of obscene words. Taken from [1]
 Positive = 12
 Negative = 60
Mean of occurrences of obscene words per segment-
 Positive - 0.136
 Negative 0.576
 Standard deviation of occurrences -
 Positive - 0.4312
 Negative - 1.245
```

One thing to note here is that obscene words (at least captured by used dictionary) are not uniformly distributed. Hence every negative segments don't have a curse word.

We tried few other features that looked interesting but not everything worked like-
- Count of words with higher volume - There were occurrences when speaker said a certain word in a higher volume and we could get it from transcript. In total positive class have 15 occurrence and negative segments have 45. On an average positive class have 0.170 high volume word per segment and negative have 0.432. But as feature is not uniformly distributed and it's very sparse it didn't give very good results it can be observed that standard deviation of such words is very high positive 1.07 negative - 2.36.
-Number of words spoken per second - only on +1 and -1 Avg 3.12414722233 2.74853314794 Std 2.09088559933 2.21125581755

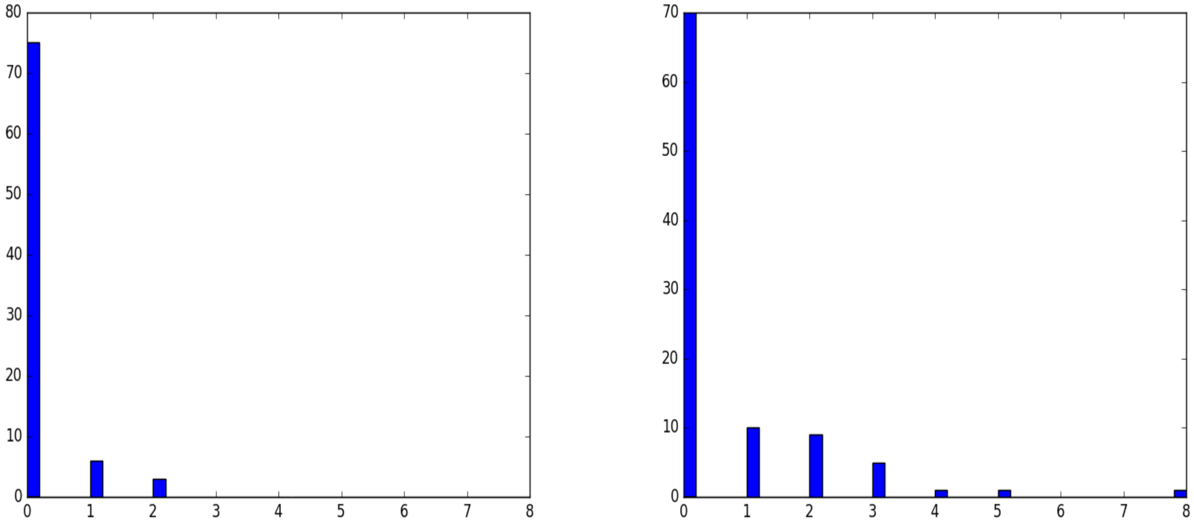


Figure 15: Histogram of obscene word usage per segment: as we can see negative(RIGHT) annotated speakers have higher occurrences of curse words than positive(LEFT)

Table 1: Results of Statistical analysis

Feature Type	**Feature Format**	**Name of Feature**	**Anova - p value**	**Anova - F value**
Visual	SHORE	Anger	**0.0409475**	**3.232612453**
Visual	CLM	Eyebrow frown	**0.009970944**	**4.685595783**
Visual	CLM	Eyebrow raise	0.097296628	2.349700082
Visual	OKAO	Right Eye openness	**0.011640962**	**4.525591583**
Visual	OKAO	Left Eye openness	0.25231465	1.383947116
Visual	OKAO	Smile	0.808139757	0.213184171
Visual	OKAO	Head nod	**0.034638331**	**3.403951288**
Acoustic	FEAT	Voice tenseness	**0.010443366**	**4.637745991**
Acoustic	TRS	word elongation	**0.006274897**	**5.165183207**
Verbal	TRS	obscene words	**0.000104929**	**9.472074218**

Conclusion: As can be seen in the table above, we analyzed many features which looked interesting. On conducting ANOVA test, some of them turn out to be promising. These features are - 'Anger', 'Eye brow frown', 'Right eye openness', 'Head nod', 'voice tenseness', 'word elongation' while speaking, obscene word utterance. The other feature were not that good. Hence we propose that these features can be used for classifying video segments effectively.

References - 1. https://www.cs.cmu.edu/ biglou/resources/bad-words.txt