Prediction of number of views for YouTube videos using Classification and Linear Regression models on tone analysis data

Anthem Rukiya J. Wingate

Department of Computational Data Science and Engineering, North Carolina Agricultural and Technical State University

ARTICLE INFO

Keywords:

Tone Analysis
View Prediction
Classification Model
Linear Regression Model

ABSTRACT

In order to predict the number of views for a newly posted YouTube video, a python program is utilized to obtain the transcript of the video. The transcript is analyzed for emotional tone using IBM Watson Machine Learning Tone Analysis API. A classification model is applied to the obtained emotional tones of the video in order to predict a possible range of views. Based on that range of views, a more precise prediction can be made using a linear regression model.

1. Introduction

jiNx isn't just software, it's connection. It is a solution predicated on the premise that each individual customer has a unique door to access our world, and we open it for them. The door and everything on the other side is utterly unknown. At jiNx we embrace that because the unknown is the one place engineers are themselves uniquely designed to inhabit.

Why Build jiNx?

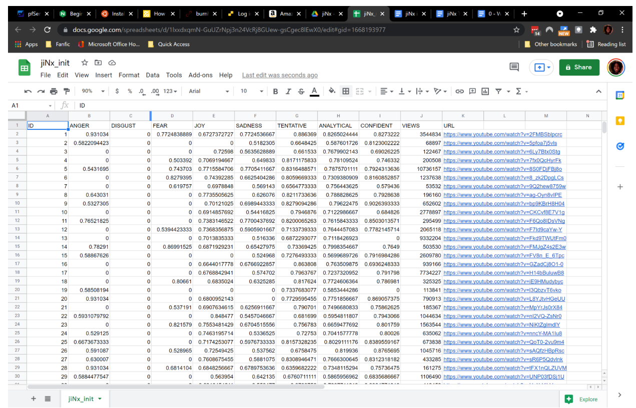
Better to ask, how do we reach each other? How do we connect when we are pulled ever farther apart by distance and ideology? If information is the world's only truly priceless commodity, connection is its most sought after resource. Information may be parsed by qualifiers, like truth and relevance, that will ultimately determine who listens. However, before those qualifiers can be applied, a connection must be leveraged to ensure information is disseminated to the widest possible audience. In the prevailing zeitgeist of global economics, not just businesses, but ideas flourish or flounder given only one parameter - connection.

What if a tool existed that could predict and guide users on how to amplify their connection to the world around them? A consumer seeking to expand their audience and ensure that they can effectively spread their core message is ultimately focused on optimizing connection. jiNx is a platform of interwoven machine learning technologies with one goal in mind - a superlative connection.

Social media prediction is currently based on leveraging metadata. The accuracy of those predictions relies on the assumption that the content itself is an unnecessary source of influence on prediction. jiNx posits that predicting the response to social media content on the basis of the context of the content can be at least as reliable if not more accurate. In the context of the global posture by which we socialize, we are connected most effectively, efficiently, and broadly through social media. No matter what the message, ensuring that it reaches an ever-widening audience is the key concern of those who utilize social media. Businesses, political and social justice movements, even emergency response coordination relies on social media for the purpose of distribution of key messaging. Predicting whether the message will successfully reach its target audience or determining how to shape a message so that it reaches the broadest possible audience is why jiNx is a timely and useful product.

jiNx 1.0 obtains the transcript of a video from a provided Youtube URL. The transcript is obtained utilizing the IBM Speech to Text service. The transcript is in turn sent to the IBM Tone Analyzer service. The Tone Analyzer service outputs ratings of the emotional tone of the transcript based on these eight emotions: Anger, Disgust, Fear, Joy, Sadness, Confident, Analytical, Tentative. The tone analysis ratings and view count are stored as a record in a PostGreSQL database through the IBM Databases as PostGreSQL service.

Once sufficient data has been collected on the PGSQL instance, the database is exported to a local CSV file.



The CSV file is utilized as training data for a Multivariate Regression Machine Learning model.

Once the model has been sufficiently trained, the model is utilized to predict view count of a test video. The test video is analyzed using IBM Watson Natural Language Understanding to provide corrective insight as necessary to the end user.

Architecture

Logical view:

- IBM Watson APIs are used to parse Youtube video into text and analyze the emotional content of the text
- Thin Client displays the outcome of the analysis from IBM Watson.
- Video analysis outcomes are stored in PostgreSQL database.

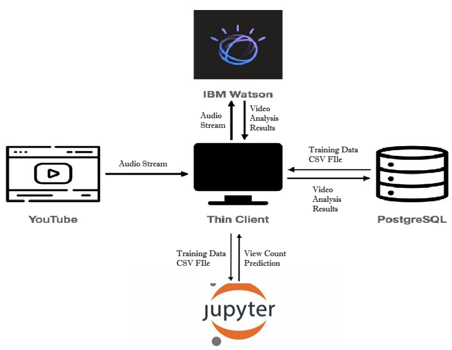
Setup Process view:

- Video analysis of training data automatically sent to PostgreSQL database
- PostgreSQL stores the analysis statistics in TranscriptStats table
- PostgreSQL database automatically exports TranscriptStats table to CSV File
- CSV File used to train Machine Learning Models in Google Collaboratory Notebook
- Machine Learning model exported to core code

End User Process view:

- Thin Client gets transcript from YouTube.
- The Thin Client sends transcript to IBM Watson.
- Thin Client receives video analysis outcomes
- Thin Client gets views prediction
- Thin Client displays full video analysis

Physical view:



Development view:

- Programming language: Python.
- IBM Watson Speech to Text API
- IBM Watson Tone Analysis API
- IBM Watson Natural Language Understanding
- Database: PostgreSQL.
- User interface: Flask App thin Client.
- Machine Learning Model: Jupyter Notebook with GPT2 Algorithm and Multivariate Regression

Scenarios:

- User launches the Thin Client on localhost and posts the url to a youtube video, and gets the Video Analysis Statistics displayed on the web client.
- User launches the Thin Client on localhost and posts the path to a locally stored video and get the Video Analysis Statistics displayed on the web client including a prediction for view count.

jiNx is a tongue in cheek reference to "the aforementioned", but with better utilization of inherent characteristics for the purpose of improved functionality. We take what has been said, and make it more informative by leveraging the context of the content for additional insight.

2. Literature Review

Core Implementation Phase I, jiNx was DiTTo. DiTTo in its first incarnation was intended to be a real-time streaming in-browser video captioning service. The first demo showed that the speech to text functionality was resource-inefficient and there was a near 33% observed inaccuracy of the transcript for word recognition. In addition, jiNx at the time of the demo was only compatible with video files and not streaming video. Given the announcement of the Google Captions Extension on March 18, 2021, the decision was made to shift focus in another direction.

After consideration, it was determined that it was not necessary to build a product that scales efficiently for high throughput. A generic or all-purpose approach was not what jiNx would provide in a landscape where it has long been discovered that one size does not fit all. The kind of analysis jiNx would be redesigned for would be infinitely more productive if it could offer user-specific insight after a single use. For those who argue that this is a poor business model, the response is simple, and eloquently devised by Cambridge Analytica:

"Communication has changed. Blanket advertizing no longer provides viable ROI for every campaign. Big data revolutionized the way organizations identify and locate their best prospects. But data alone isn't enough. Cambridge Analytica is building a future where every individual can have a truly personal relationship with their favorite brands and causes by showing organizations not just where people are, but what they really care about and what drives their behavior." - *https://cambridgeanalytica.org/*

A customer that could finetune jiNx's machine learning algorithm with their own data, is looking for a single outcome that will have a substantial impact across their entire business. jiNx could pivot from being a generic tool that spits out subtitles for any user who opts to download a Google extension to something far more useful.

IBM Cloud Services

IBM Cloud provides APIs free of charge which made it possible for jiNx to be developed as a completely thin client. This means that jiNx can be run by any user, from anywhere without concern about meeting minimum hardware requirements on their local environment. cloud.ibm.com

Python/Flask Vs Java/SpringBoot/JavaFx/Gradle/Camel:

During development it was elected to make a shift from a Java software development kit configured to utilize SpringBoot, JavaFx, Gradle and Camel to a Python software development kit utilizing Flask for quicker frontend development. Simply put, at the time jiNx pivoted from a captioning service to a video analysis prediction service, there was concern about technical debt and on-time completion. Developing in Python meant that development activities could proceed at a much faster pace comparatively speaking.

At the completion of phase one of jiNx development, it was determined that it was necessary to analyze the trending behavior that created jiNx in its current incarnation. This is critical as a means of identifying prototype viability throughout the software development life cycle of jiNx. Prototype viability was defined as a function of validation and verification of the performance of each iterative prototype. This was accomplished via tracking those metrics which corresponded to the time to implement each solution and the time to compile the completed solution for a quantifiable outcome. Time to complete and time to compile were framed as metrics using the GQM process. Data collection and statistical validation of the results were conducted using Intellij, Travis CI, Codecov, Radon, and other commonly implemented open source development tools.

The first prototype required roughly 12 programming days, the second 17 programming days, and the final prototype required 11 programming days. Time to compile for each prototype respectively was 179s, 132s, and 92 s.

Overall, jiNx was more expedient in comparison to the preceding implementations as might be expected. With each iterative cycle of development, time applied to research decreases as domain knowledge is expanded. This outcome extends to time to compile as the build time for jiNx was significantly improved over the prior two prototypes. It follows that the benefit of this iterative prototyping development approach can be attributed to progressively improved efficiency and efficacy of the code base.

3. Materials and Methods

Proposed Methodology

How is Parasocial Interaction Relevant to jiNx?

"A parasocial relationship, first defined by Horton and Wohl, consists of one-sided and one-way mediated symbolic interactions between a viewer and a media persona (presenter, actor, singer, or other celebrity), whereby the media persona is perceived to be a "friend" by the viewer. A parasocial relationship is, thus, built from sequential parasocial one-way interactions from the viewer to the media persona. Many scholars have applied these theoretical constructs of parasocial interactions and relationships to their study of the audience-media relationship." (Horton &Wohl, 1956)

Why is this connected?

"Viewers interactions with media characters foster positive reactions towards the media product that have been reported, in turn, to be strongly linked to viewing motives, attitudes, and activity levels. Hence, stimulating viewers' symbolic interactions with a TV programme and their parasocial relationships with the programme characters may lead to high levels of involvement that may be connected to media producers successfully promoting their product." (Dibble et al., 2016; Liebers & Schramm, 2019)

There are two key types of parasocial interaction that are of relevance to jiNx. These are Referential and Critical Involvement. Referential Involvement is the symbolic connection between the media character's experience and the viewer's personally lived experience (Liebes & Katz, 1986). Critical Involvement is the viewer's aesthetic engagement with the media product as a whole (Liebes & Katz, 1986)

In a Korean television study published by Kim and Sintas in 2021, the researchers established the prevalence of parasocial interaction in social media platforms. In this study, Twitter was scraped to find evidence of parasocial interactions with television shows that fall within one of three classes of popularity, i.e. low, moderate, or high. Directly thereafter, the study identified definitive outcomes for confirmed parasocial interactions. During this process, topic clouds were created based on the scraped social media sources which reflect the specific type of parasocial outcome. The distribution of parasocial outcomes based on the proportion of data that specifically reflects each topic cloud of relevance as pertains to jiNx are as follows:
Topic 2: Empathy 31%
Topic 3: Criticism 19%
Topic 10: Behavior 17%
Topic 12: Expression of emotion 10%
Topic 8: Referential Involvement 4%
Topic 6: Cognitive 4%

The final step of the Kim and Sintas study was to establish a correlation between the defined involvement outcome and level of popularity of the television show, i.e. less, moderately, or more popular. In this final step, a linear model was created correlating the popularity of an episode of television with the prevalence of data scraped from social media referencing that episode. The regression model for the six key parasocial interaction outcomes of relevance for jiNx all show positive correlation with popularity with the exception of the topics of behavior and empathy.

Kim and Sintas concluded that the popularity of a media product is critically influenced by the Cognitive, Referential, Criticism, and Expression of Emotion topics directly reflecting parasocial interaction outcomes of the audience. This conclusion became the basis for modeling the data obtained by jiNx. The topic modelling percentages determined in the Kim and Sintas study were described as the equivalent for the explained variance of the Principal Component Analysis. Converting topic percentages to PCA yields an accounting of 82% of the variance. This was used as a predictor of overall success of jiNx. A similar level of cumulative accuracy in modeling would be interpreted as confirmation that jiNx is an effective means of predicting views.

Thus the following percentages were determined for each tone as a parallel attribute to the topics investigated in the Kim and Sintas study. An acceptable measure of accuracy for the multivariate regression model should approximate these values within a statistically significant margin of error.

- *Criticism*: 19/36 = 0.5378 => Attribute 54% of variance to Analytical, Confident, Tentative
- *Expression of Emotion*: 10/36 = 0.2778 => Attribute 28% of variance to Joy, Sadness, Anger, Fear
- *Cognitive*: 4/36 = 0.1111 => will remain unattributed
- *Referential Involvement*: 4/36 = 0.1111 => will remain unattributed

Datasets and Pre-Processing

The complexity of the problem presented by jiNx is not determined to embody the entirety of its core definition so much as the underlying considerations. It is certainly possible to identify and predict behavior of a specific demographic. This is accomplished quite successfully when contemplating the efforts expended by corporations in pursuit of consumers. A clear observation of this is within politics. Marketing analysts are able to quantify human behavior down to how voters will respond to a hair cut, a chosen catch phrase or the color of a suit jacket worn by a candidate. Thus, the complexity of the question of predicting the popularity of a YouTube video based on the tone analysis of the transcript of said video, is not held in determining whether these are adequate parameters. In fact, the complexity is in cataloguing and modeling human response to specific stimuli. The collected sample of data is likely sufficient to make fairly accurate predictions. However, a dimension of the analysis is missing. Without having a parallel model to determine typical human response to each individual emotional tone, the data loses a great deal of impact as a means to predict a video's potential for attracting views.

With this in mind, it is possible to assess the relevance of this single dimension of the data in regards to the proposed question. Meaning, any level of accuracy at or above fifty percent is highly favorable, assuming that the other dimension is equally weighted. Given this understanding, there are a few initial assumptions that can be declared based on the illuminating results of the pre-processing step.

Though there is a preponderance of zeroes in the DISGUST and FEAR parameters, these are not null values. The expectation could be, when considered in reference to viewer reaction, that limited expressions of disgust and fear in the transcript will increase viewership. Thus the values must be retained as they are relevant data points.

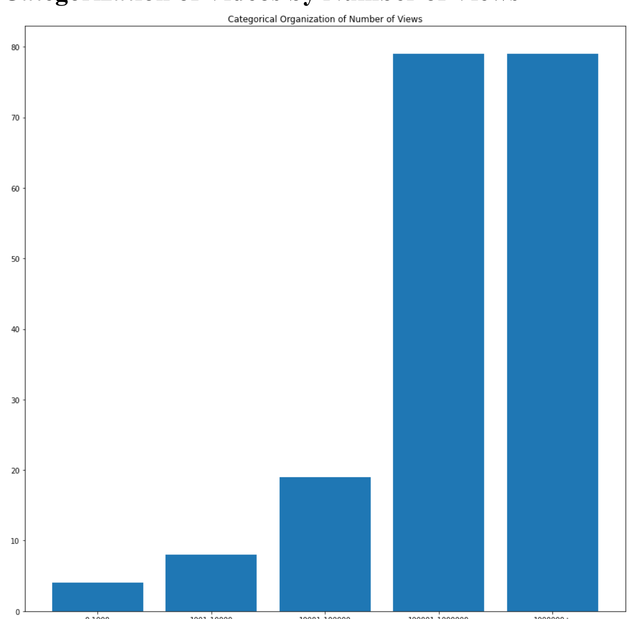
In addition, correlations roughly between 0.5 and 0.8 can be observed for the SADNESS and ANALYTICAL emotional tones. The expectation could be, when considered in reference to viewer reaction, that viewers tend to identify greatly with videos that convey both sadness and rationale. Instructional videos and emotional testimonies both tend to be highly sought after by viewers.

```
Training data shape: (226, 10)
RangeIndex: 226 entries, 0 to 225
Data columns (total 10 columns):
 #  Column      Non-Null Count  Dtype
--- ------      --------------  -----
 0  ID          226 non-null    int64
 1  ANGER       226 non-null    float64
 2  DISGUST     226 non-null    int64
 3  FEAR        226 non-null    float64
 4  JOY         226 non-null    float64
 5  SADNESS     226 non-null    float64
 6  TENTATIVE   226 non-null    float64
 7  ANALYTICAL  226 non-null    float64
 8  CONFIDENT   226 non-null    float64
 9  VIEWS       226 non-null    int64
dtypes: float64(7), int64(3)
memory usage: 17.8 KB
```

Feature Engineering and Training Data

A standardization followed by a PCA analysis limited to two components will likely assist significantly in finding an adequate fit for the data. In addition, Though initial intentions are to make predictions using linear regression, a polynomial regression model may yield better results. Based on the outcome of the preprocessing, it will be imperative to utilize classification as a means of determining a range of values for accurate predictions prior to applying the trained regression model to new samples of data.

Categorization of Videos by Number of Views



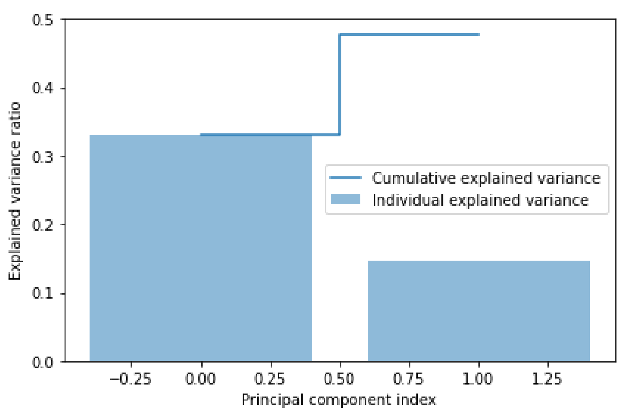
Explained Variance
[0.32993115 0.14740921]

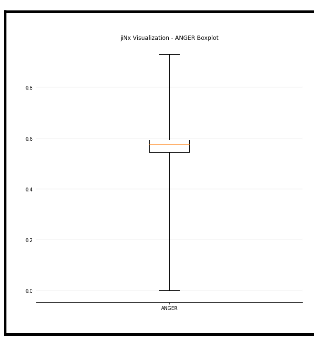
PCA Components
[[-0.39516389 0.3397766 0.39713613 0.40135344 0.30132884 0.50582859 0.22150684 -0.11776573]
[-0.21213783 0.12162536 0.33400427 0.13364595 -0.48856766 -0.01174173 -0.36134923 0.66436076]]

PCA Fit Training Data

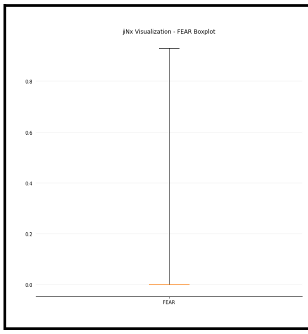
	PC1	PC2	Class
0	-1.347729	-0.732881	5
1	2.262030	-0.143108	4
2	-0.630956	-0.348620	5
3	0.193846	-0.131892	4
4	-1.287144	-1.471555	4
5	-1.820851	-0.281026	4
6	-1.376238	-1.206764	3
7	1.867579	-0.013575	4
8	-0.278123	1.438477	5
9	-1.286136	-1.288535	4
10	2.220024	0.119416	5
11	-1.204638	-1.476113	4
12	-3.559449	2.843018	5
13	2.368239	-0.219133	5
14	1.048383	0.960514	5
15	-1.236725	-1.137886	4
16	-1.437353	-0.922738	5
17	-0.398152	-1.230729	4
18	-2.970373	1.548011	5
19	2.373505	0.150892	4

Cumulative Explained Variance

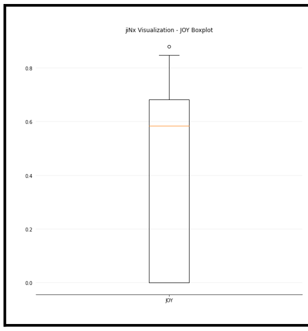




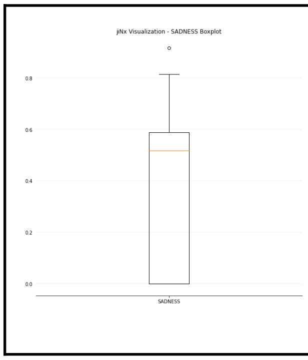
ANGER summary statistics
Min: 0.0
Mean: 0.46
Max: 0.93
25th percentile: 0.54
Median: 0.58
75th percentile: 0.59
Interquartile range (IQR): 0.05



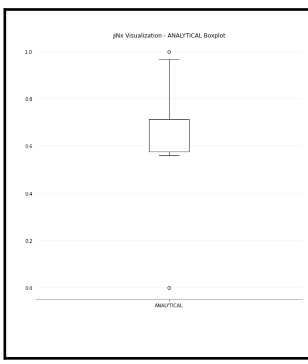
FEAR summary statistics
Min: 0.0
Mean: 0.14
Max: 0.93
25th percentile: 0.0
Median: 0.0
75th percentile: 0.0
Interquartile range (IQR): 0.0



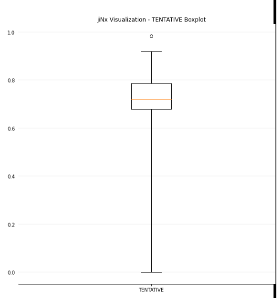
JOY summary statistics
Min: 0.0
Mean: 0.38
Max: 0.88
25th percentile: 0.0
Median: 0.59
75th percentile: 0.68
Interquartile range (IQR): 0.68

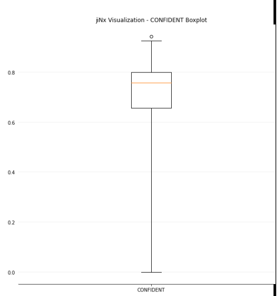


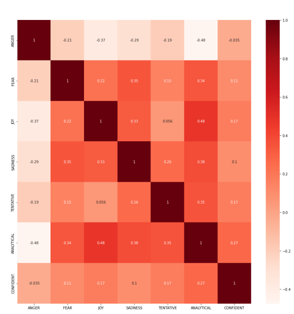
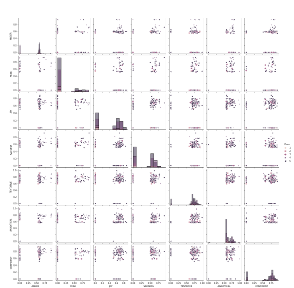
SADNESS summary statistics
Min: 0.0
Mean: 0.31
Max: 0.92
25th percentile: 0.0
Median: 0.52
75th percentile: 0.59
Interquartile range (IQR): 0.59



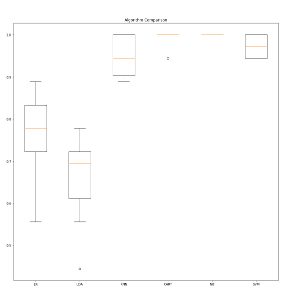
ANALYTICAL summary statistics
Min: 0.0
Mean: 0.64
Max: 1.0
25th percentile: 0.58
Median: 0.59
75th percentile: 0.71
Interquartile range (IQR): 0.13

	**TENTATIVE summary statistics** **Min: 0.0** **Mean: 0.67** **Max: 0.98** **25th percentile: 0.68** **Median: 0.72** **75th percentile: 0.79** **Interquartile range (IQR): 0.11**

	**CONFIDENT summary statistics** **Min: 0.0** **Mean: 0.64** **Max: 0.94** **25th percentile: 0.66** **Median: 0.76** **75th percentile: 0.8** **Interquartile range (IQR): 0.14**

Heat Map of Tones	**Pair-wise Scatterplot of Tones**
	

3. Machine Learning Model Development

Spot Check Algorithms and Compare Models	
	**LR: 0.755556 (0.100000)** **LDA: 0.661111 (0.100769)** **KNN: 0.950000 (0.046148)** **CART: 0.994444 (0.016667)** **NB: 1.000000 (0.000000)** **SVM: 0.972222 (0.027778)**

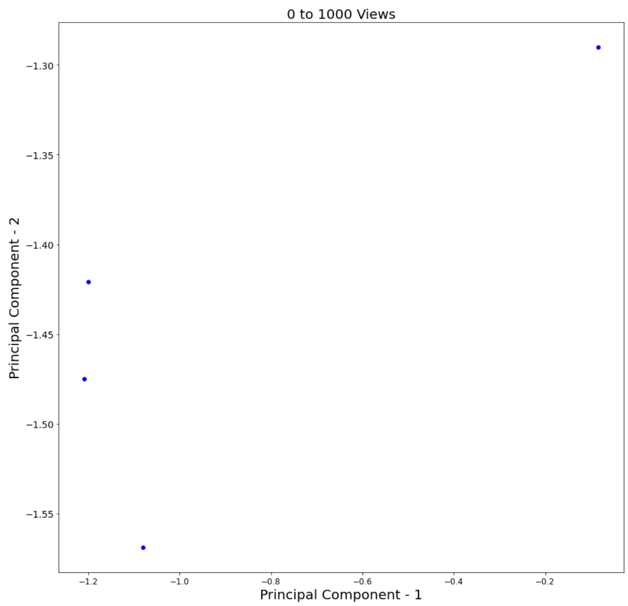
```
Logistic Regression Predictions

Accuracy:  0.8913043478260869


Confusion Matrix:
 [[ 0  1  0  0  0]
 [ 0  0  1  0  0]
 [ 0  0  2  3  0]
 [ 0  0  0 13  0]
 [ 0  0  0  0 26]]


Classification Report:
              precision    recall
f1-score   support

           1       0.00      0.00
0.00         1
           2       0.00      0.00
0.00         1
           3       0.67      0.40
0.50         5
           4       0.81      1.00
0.90        13
           5       1.00      1.00
1.00        26

    accuracy
0.89        46
   macro avg       0.50      0.48
0.48        46
weighted avg       0.87      0.89
0.87        46
```

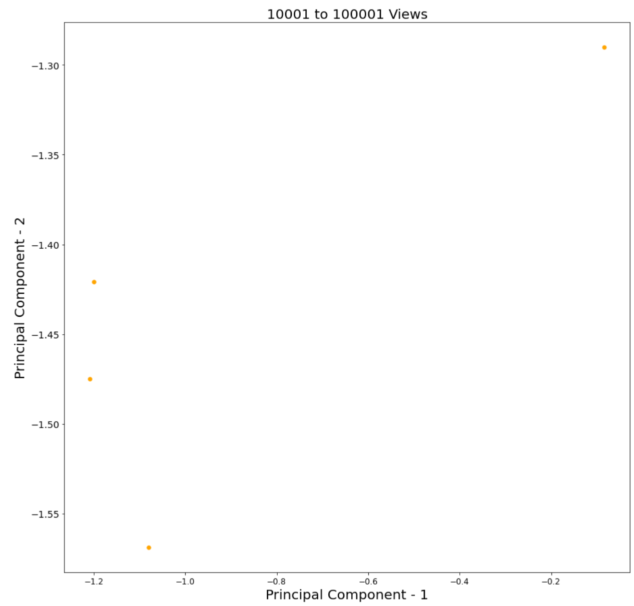
```
Linear Discriminant Predictions

Accuracy:  0.6304347826086957


Confusion Matrix:
 [[ 0  0  0  1  0]
 [ 0  0  0  1  0]
 [ 0  0  0  5  0]
 [ 0  0  0 10  3]
 [ 0  0  0  7 19]]
```

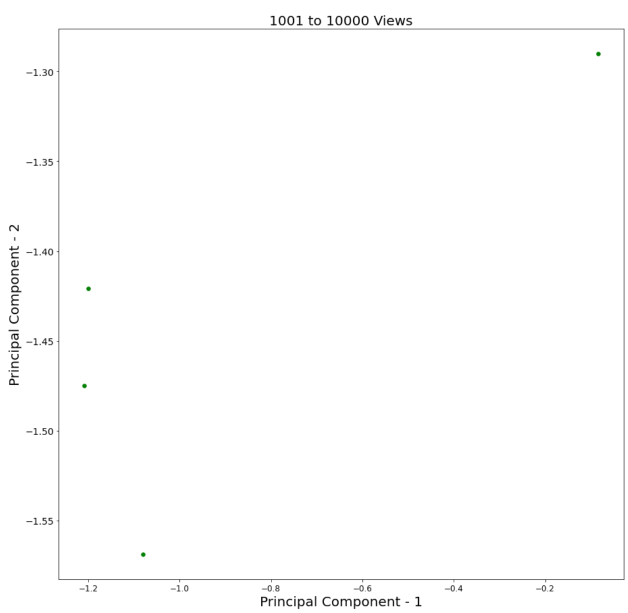
```
Classification Report:
               precision    recall
f1-score   support

           1       0.00      0.00
0.00         1
           2       0.00      0.00
0.00         1
           3       0.00      0.00
0.00         5
           4       0.42      0.77
0.54        13
           5       0.86      0.73
0.79        26

    accuracy
0.63        46
   macro avg       0.26      0.30
0.27        46
weighted avg       0.61      0.63
0.60        46
```

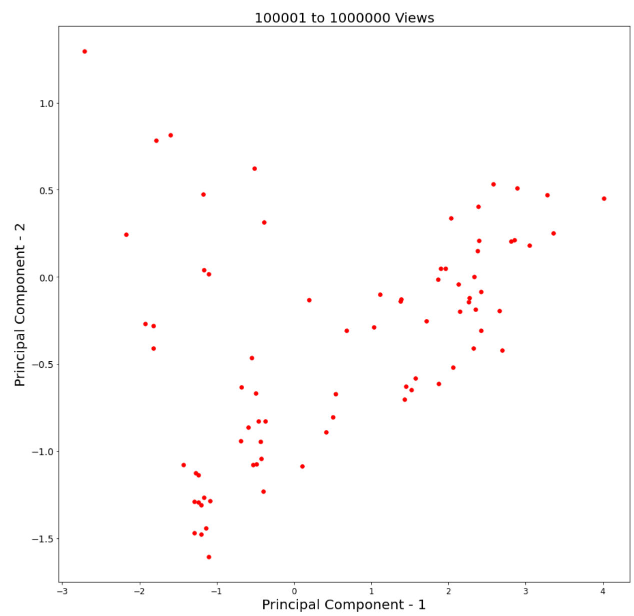
```
k-Nearest Neighbors Predictions

Accuracy:  0.9347826086956522


Confusion Matrix:
 [[ 1  0  0  0  0]
 [ 0  1  0  0  0]
 [ 0  0  2  3  0]
 [ 0  0  0 13  0]
 [ 0  0  0  0 26]]


Classification Report:
               precision    recall
f1-score   support

           1       1.00      1.00
1.00         1
           2       1.00      1.00
1.00         1
           3       1.00      0.40
0.57         5
           4       0.81      1.00
0.90        13
           5       1.00      1.00
1.00        26

    accuracy
0.93        46
   macro avg       0.96      0.88
0.89        46
weighted avg       0.95      0.93
0.92        46
```

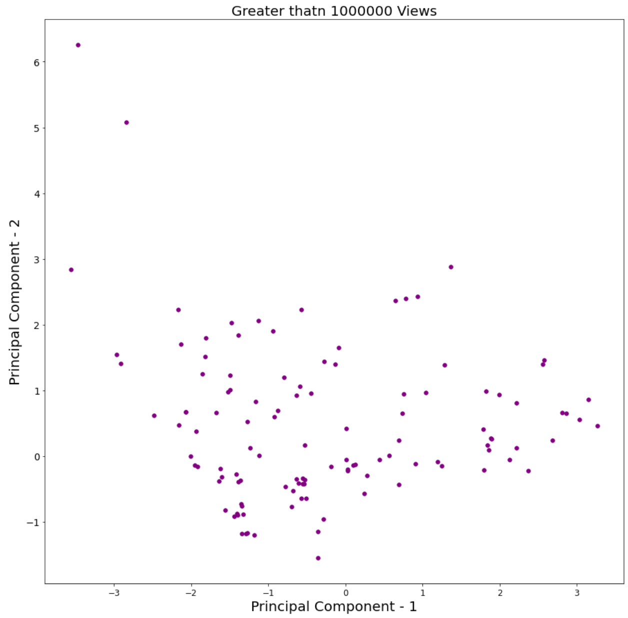
```
Support Vector Machines Predictions

Accuracy:  0.9565217391304348


Confusion Matrix:
 [[ 0  0  1  0  0]
 [ 0  0  1  0  0]
 [ 0  0  5  0  0]
 [ 0  0  0 13  0]
 [ 0  0  0  0 26]]


Classification Report:
               precision    recall
f1-score   support

           1       0.00      0.00
0.00         1
           2       0.00      0.00
0.00         1
           3       0.71      1.00
0.83         5
           4       1.00      1.00
1.00        13
           5       1.00      1.00
1.00        26

    accuracy
0.96        46
   macro avg       0.54      0.60
0.57        46
weighted avg       0.93      0.96
0.94        46
```

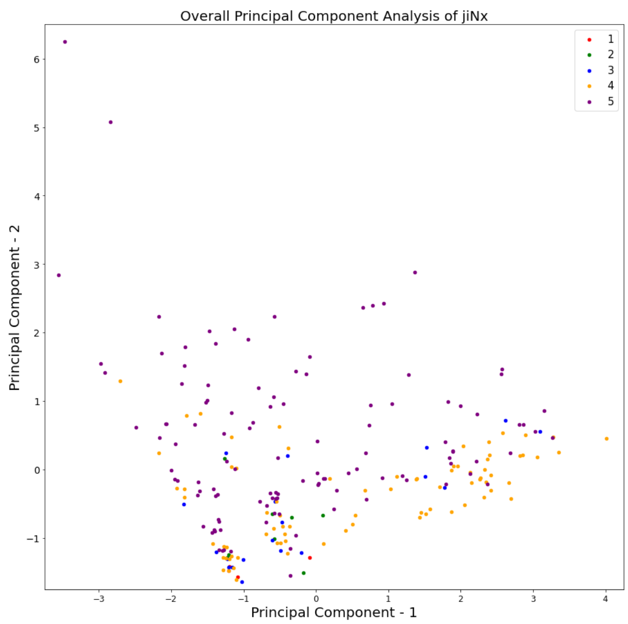

0 to 1000 Views
Principal Component - 1
Principal Component - 2


10001 to 100001 Views
Principal Component - 1
Principal Component - 2


1001 to 10000 Views
Principal Component - 1
Principal Component - 2


100001 to 1000000 Views
Principal Component - 1
Principal Component - 2





4. Discussion

Observation of Algorithm Comparison Indicates Further Modeling Steps

The results observed in the box plots indicate that SVM and KNN are both very well-fit classification models.

Assuming that the outcomes of the classification step are accurate, these models can be used to pre-predict the range of expected views prior to applying a regression model. A Polynomial Regression is implemented under the assumption that a third order polynomial could be derived from the model fit to the data in keeping with the results discussed in the literature by Kim & Sintas.

5. Conclusion

Considering the model as a whole

Evaluating a Support Vector Machines Classification model yields accuracy results of 0.9565217391304348.

Evaluating KNN on the dataset with polynomial features transform of degree=3 yields the best regression fit with mean accuracy of 0.599 and standard deviation of 0.047.

These models would ideally be employed in tandem. Given any tone analysis sample, the source video can be classified into a range of expected views with 95 percent accuracy. Directly thereafter, the exact number of expected views can be predicted for the source video with an expected precision of 60 percent within a standard deviation of 5 percent.

Remaining Steps

In order to improve the fit, it will be necessary to obtain the model from the Kim & Sintas dataset to give a more well-rounded view of the impact of parasocial interaction on media product popularity. Including the insight of the Kim and Sintas model could improve the fit for the regression model enough to lend additional accuracy as well as precision to predictions.

A further step would entail employing a text completion model. OpenAI provides a pretrained Machine Learning model with easily integratable libraries and API access. The true potential of jiNx as a prediction service is in its ability to leverage the power of a massive language prediction model. With an opportunity to utilize GPT3, jiNx could evolve from simply predicting views to something far more impactful for social media influencers. The addition of GPT3 would allow jiNx to employ the added functionality of predicting what specific content results in increased views. A key scenario would be to provide suggested changes to the analyzed transcript of the source video with an accompanying predicted percent increase in view count.

When to employ the model

Notably, with fewer views, a polynomial fit seems the most likely solution. However, as views increase by each order of ten, the distribution becomes more and more disordered. This is what we would expect as outside factors not related to emotional tone can play a significant part on the number of views the source video can and will receive.

The best time to employ the model as a fairly accurate predictor of views should be before the current number of views for the source video has reached 100000. After 100000 views, the influence of non-parasocial promoters, namely YouTube's inherent algorithms, begin to skew outcomes.

6. References

Horton, D., & Wohl, R. R. (1956). Mass Communication and para-social interaction. Psychiatry Journal of Study on Interpersonal Processes, 19, 215-229. https://doi.org/10.1080/00332747.1956.11023049

Kim, J., & Sintas, L. (2021). Social TV viewers' symbolic parasocial interactions with media characters: A topic modelling analysis of viewers' comments. Social Sciences & Humanities Open 3, 1000129. https://doi.org/10.1016/j.ssaho.2021.100129

Liebers, N., & Schramm, H. (2019). Parasocial interactions and relationships with media characters-An inventory of 60 years of research. Communication Research Trends, 32, 4-31.

Liebes, T., & Katz, E. (1986). Patterns of involvement in television fiction: A comparative analysis. European Journal of Communication, 1, 151-171. https://doi.org/10.1177/0267323186001002004