**Natural Language Processing**

**Q.1. What is so unique about Ali Wong’s comedy that makes her one of the best standup comedians?**

**Q.2. Can machine learning creates comedy transcript?**

A. Data Gathering: web scraping using **requests** and **BeautifulSoup**

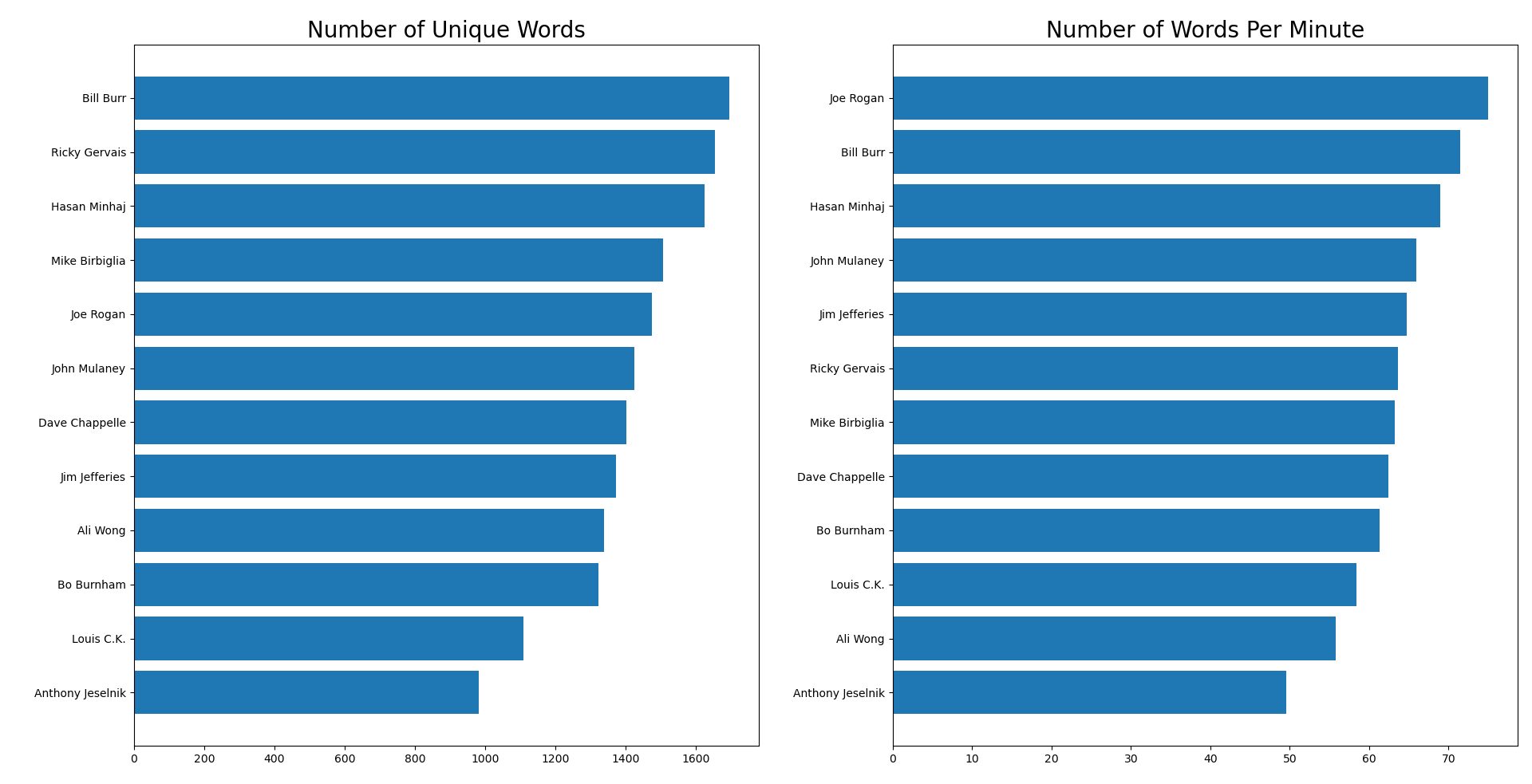
B. Data Cleaning: **Corpus** (ordered text) and **Document-Term Matrix** (bag-of-words, order of words irrelevant)

C. Data Exploration:

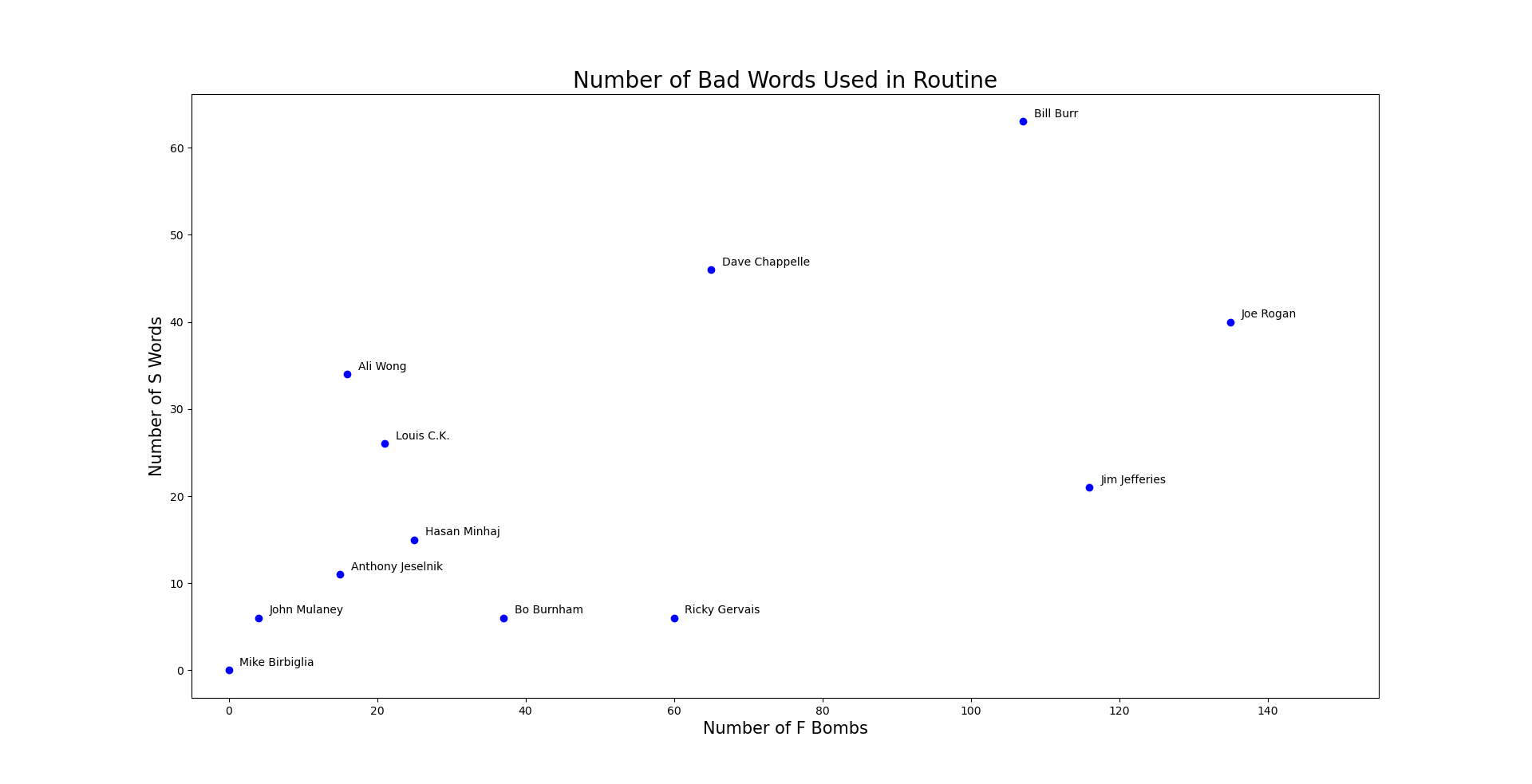
1. Top words (wordcloud)



2. Size of Vocabulary (Bar Plot)

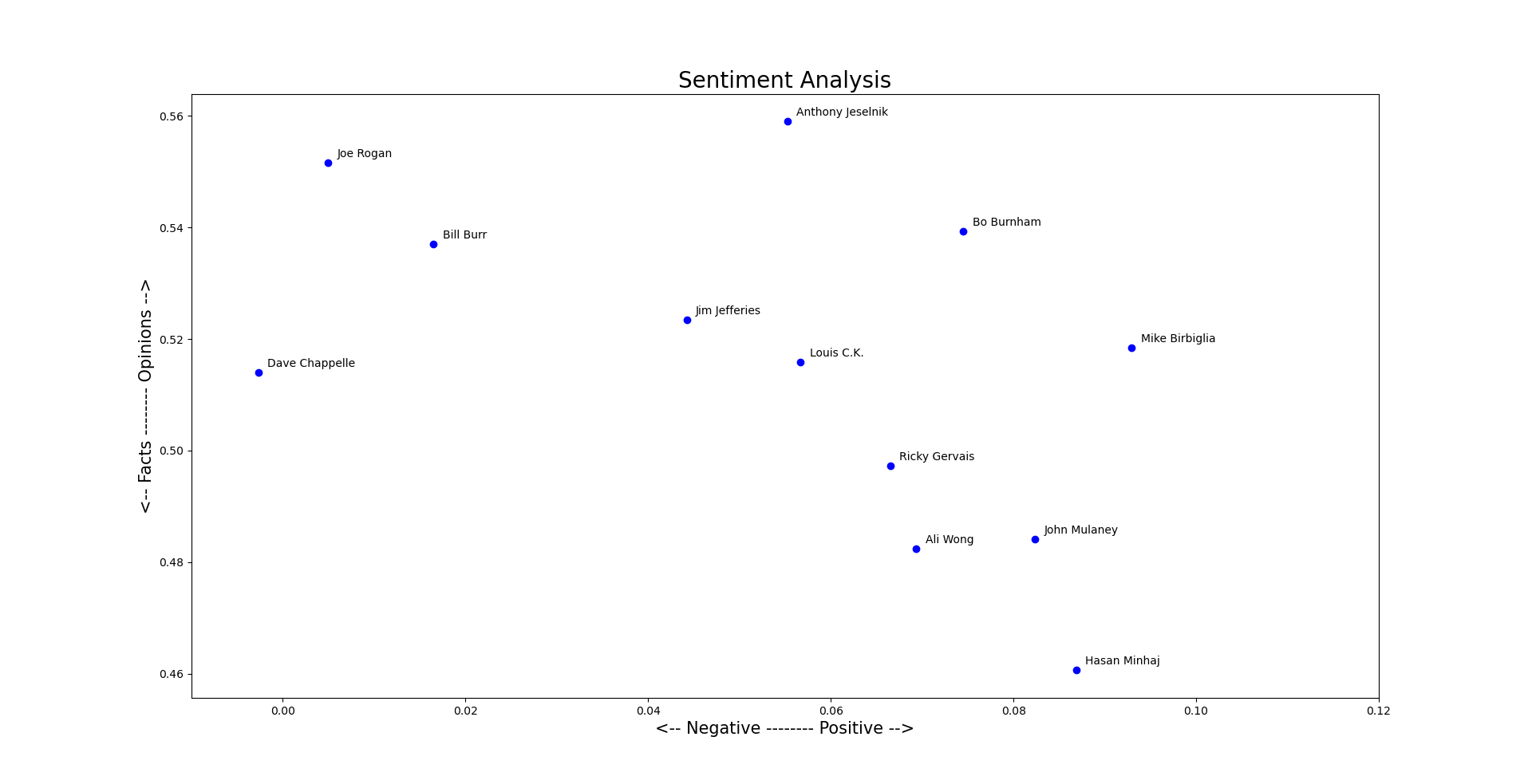


3. Amount of Profanity (Scatter Plot)

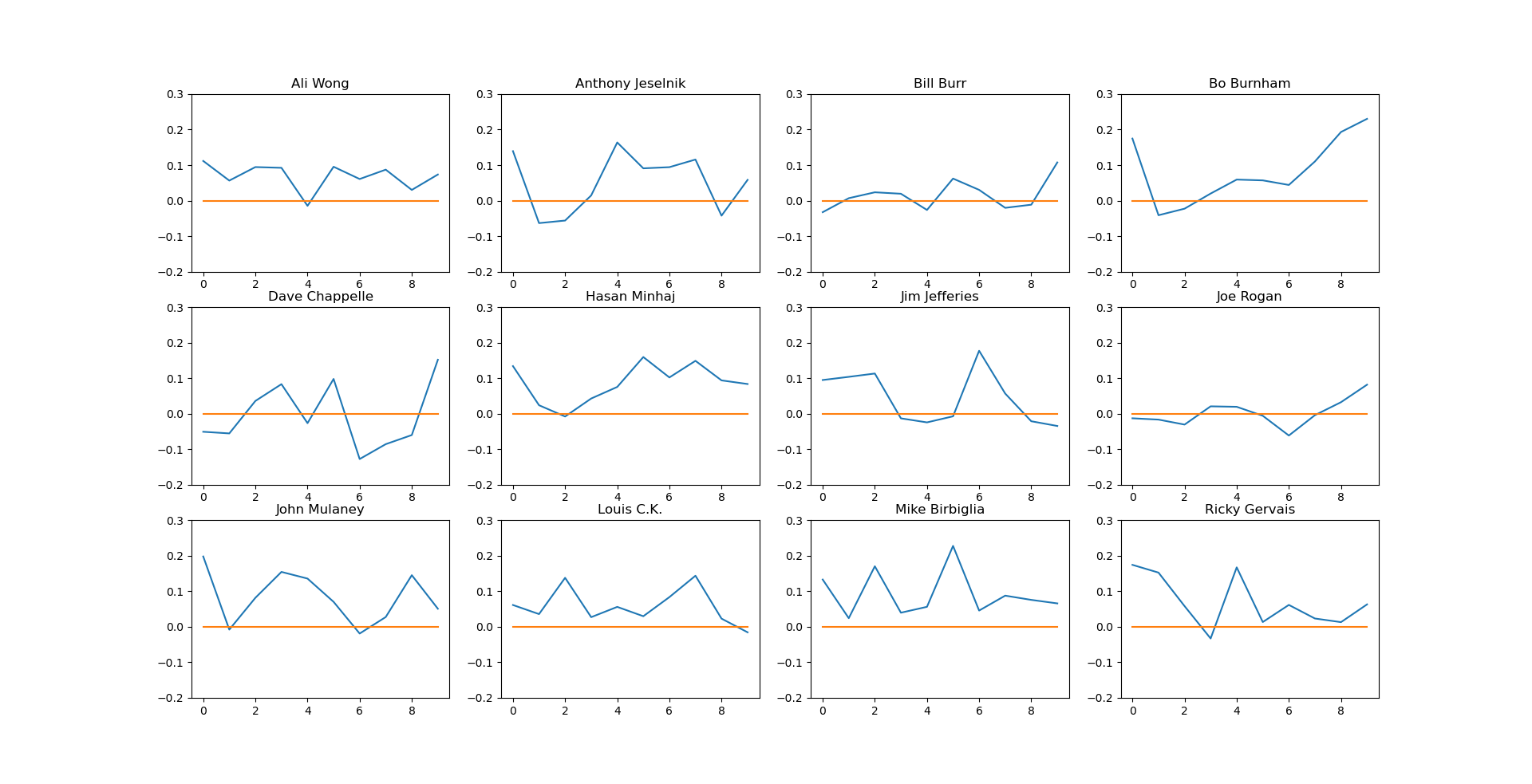


D. NPL Techniques

1. Sentiment Analysis: using **TextBlob** (built on **nltk**) to evaluate polarity and subjectivity, can be improved with classification techniques like **Logistic Regression**, **Naïve Bayes**, etc.



Mean Sentiment



Sentiment Evolution

2. Topic Modeling: using **genism** library (**Latent Dirichlet Allocation**), can be improved using matrix factorization techniques like **Latent Semantic Indexing (LSI)** and **Non-Negative Matrix Factorization (NMF)**

3. Text Generation: using **Markov Chain**, can be improved with **Deep Learning (Long Short-Term Memory, LSTM)**

In general, **SpaCy** library is better in performance compared to **nltk**, and might be replaced in near future.