**BAX 423 Big Data**

**An Analysis for TechCrunch Blogs**

**Team Super Duper**

**Yuanyuan Ge**

**Yining Hang**

**Yijun Huang**

**Joey Li**

**Yiying Wang**

**Table of Contents**

[**Executive Summary**](#_heading=h.f6fzic8ljprz) **3**

[***Business Objective and Initiative***](#_heading=h.30j0zll) ***4***

[**Measurement**](#_heading=h.olnh6j2lv2wi) **4**

[**Data Characteristics**](#_heading=h.p3oyz6t7azyc) **4**

[**Model Preparation, Selection, Evaluation, and Interpretation**](#_heading=h.k4zqbls2ohnt) **6**

[**Writing Styles Identification**](#_heading=h.j1faqsmasd1f) **6**

[**Topic Modeling**](#_heading=h.k702mmtg9edh) **7**

[**Sentiment Analysis**](#_heading=h.fn5nszoinqic) **9**

[**Regression**](#_heading=h.95xegk17ecyj) **10**

[**Identify barriers to implementation — and strategies to ...**](#_heading=h.7zqgzpeh1tbs) **12**

[**Conclusion**](#_heading=h.j0u3aml809so) **12**

# **Executive Summary**

# **Business Objective and Initiative**

TechCrunch is a leading technology media property, who is dedicated to obsessively profiling startups, reviewing new Internet products, and breaking tech news. It has 13 million followers that generate 26 million monthly page views. More importantly, it owns valuable readership, with 19% considering themselves influential and 44% earn an annual household income of $110K+. However, the problem is how can we transform this valuable readership into value. Like all the media property, TechCrunch’s revenue is mainly driven by the advertisement and the sponsored content in the blogs. Advertisers are concerned with the influence of the blog, and content sponsors are interested in both the influence of the blog and the reader’s reaction. We aim to apply Natural Language Processing (NLP) to identify what factors will affect the influence and reader’s reaction and to provide TechCrunch with insights to attract more advertisers and content providers to generate more value.

# **Measurement**

We use both direct and indirect metrics to measure the success of our project. The direct metric is the revenue generated by the advertisement and the sponsored content. The indirect metrics will be what drives the revenue, which is the price and volume of the advertisement, which are driven by the influence of the blog and the reader’s reaction. The former one could be measured by the page views and the number of comments and shares, and the latter one could be measured by the sentiment analysis of the comment. Additionally, we could measure user acquisition by the number of registration and subscription as well as user engagement by the number of their pageviews, comments, and shares.

# **Data Characteristics**

To comprehensively capture the factors related to the bloggers and blogs, we applied analysis on a series of datasets with multiple metrics. The author dataset contains 107 bloggers of TechCrunch, with computed influence index, MEIBI, which relates to the number of the blog post’s inlinks and its comments, as well as MEIBIX, which relates both the number and age of inlinks and comments. The post dataset contains 19464 posts by those bloggers and the comment dataset contains the 746561 comments which were submitted by the readers of those posts. As we mentioned, we will try to identify writing styles and topics for post dataset, and sentiments for comment dataset.

**The Structure of the Datasets**

A close up of a sign

Description automatically generated

**The Features of the Datasets**

|  |  |
| --- | --- |
| Dataset | Features |
| Author Dataset | Author ID, Author Name, MEIBI, MEIBIX, Average Length of Words, Average Length of Words in posts (without stopwords) |
| Post Dataset | Post ID, Author ID, Author Name, Title, Content, Number of comments, Number of retrieved inlinks, Number of retrieved comments |
| Comment Dataset | Comment ID, Post ID, Content Vote |

# **Model Preparation, Selection, Evaluation, and Interpretation**

# **Writing Styles Identification**

First, we tried to identify the writing styles for the posts. Writing styles could be complex, and we broke it down into three aspects with multiple metrics. The first aspect is lexical features, which are the most basic features that tell us about the structure of the text. The second aspect is vocabulary richness features. Richness refers to the diversity of the vocabulary, for example, posts with low richness have limited vocabulary repeated over and over again. The final aspect is the readability scores. Readability stems from the field of linguistics and researchers have frequently used linguistics’ laws to determine the metrics.

In short, the first aspect differentiates the complexity of the text, the second one differentiates the diversity of the vocabulary, and the final one differentiates the understandability. We processed the text and calculated more than twenty metrics for further modelling.

**The Features of the Writing Styles**

|  |  |
| --- | --- |
| Aspects | Features |
| Lexical Features | Average Word Length, Average Sentence Length By Word, Average Sentence Length By Character, Average Syllable per Word, Functional Words Count, Punctuation Count, Special Character Count |
| Vocabulary Richness Features | Hapax Legomenon, Hapax DisLegemena, Honores Measure, Brunets Measure, Yules Characteristic, Shannon Entropy\*, Simpson’s Index\* |
| Readability Scores | Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning Fog Index, Dale Chall Readability Formula, Shannon Entropy\*, Simpson's Index\* |

\* Shannon Entropy and Simpson's Index are both Vocabulary Richness Features and Readability Scores

# **Topic Modeling**

Secondly, we tried to extract the topics for the posts with topic modeling. It is challenging to extract high-quality topics that are clear, segregated, and meaningful, for we need to consider the quality of text processing and the algorithm of modeling. In this report, we applied a popular topic modeling algorithm called Latent Dirichlet Allocation(LDA). LDA considers each post as a collection of topics and each topic as a collection of keywords, and it rearranges the topics and keywords distribution when the number of topics is given.

**The mechanism of Latent Dirichlet Allocation**

**A picture containing star

Description automatically generated**

Firstly, we cleaned up and tokenized the posts. Secondly, we created bigrams and trigrams models and lemmatized. Bigrams refer to two words frequently occurring together in the post, and trigrams refer to three words. Lemmatization refers to grouping inflected or variant forms of the same word. Then, we created two main inputs to the LDA topic model, the dictionary and the corpus. Finally, we developed the LDA model and optimized the number of topics by the coherence score.

**The Optimization of the Number of Topics**

A close up of a map

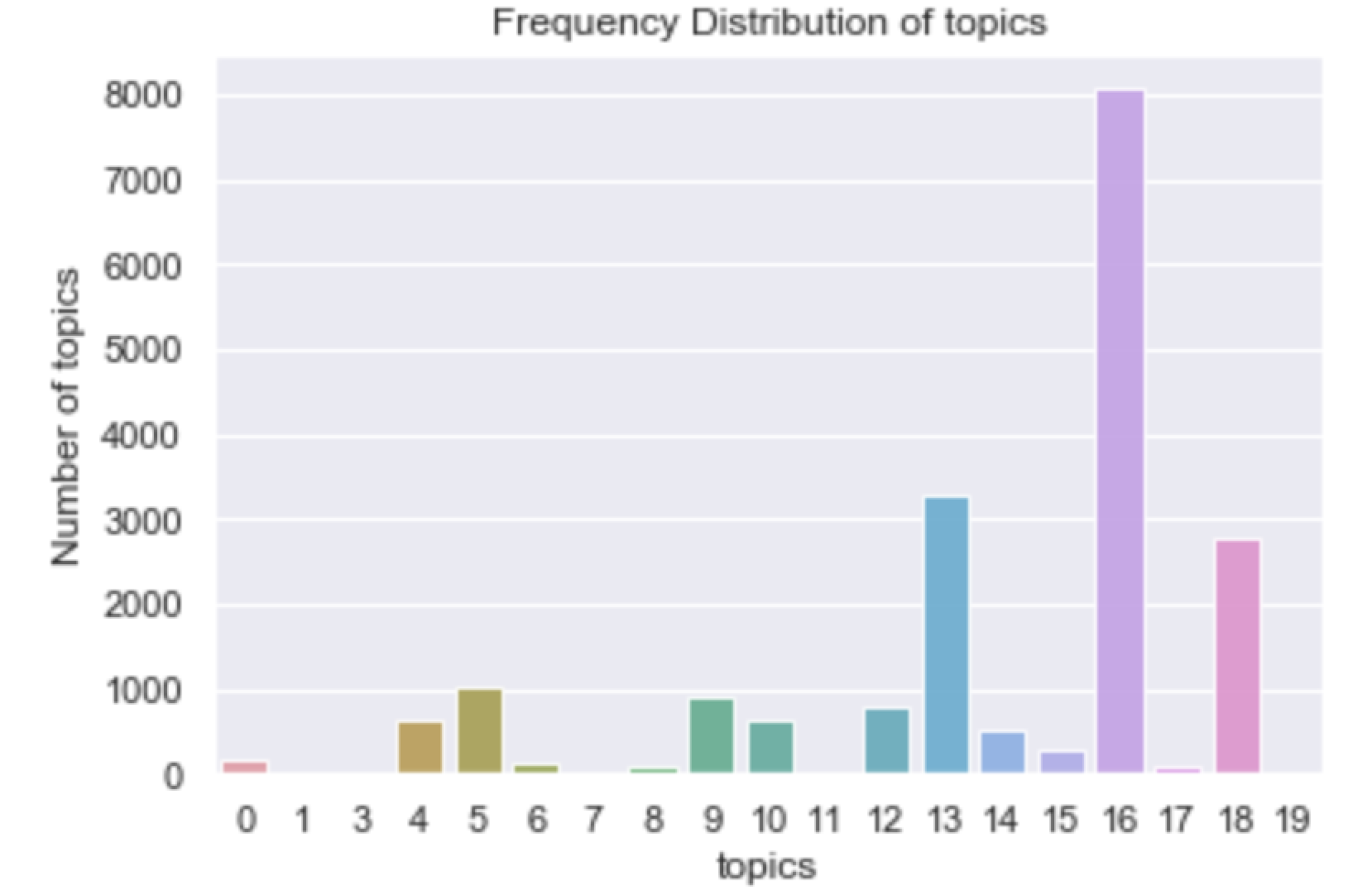
Description automatically generated

As the graph is shown, we chose 20 different topics. With the LDA model, each topic has a combination of keywords and each keyword contributes a certain weightage to the topic.



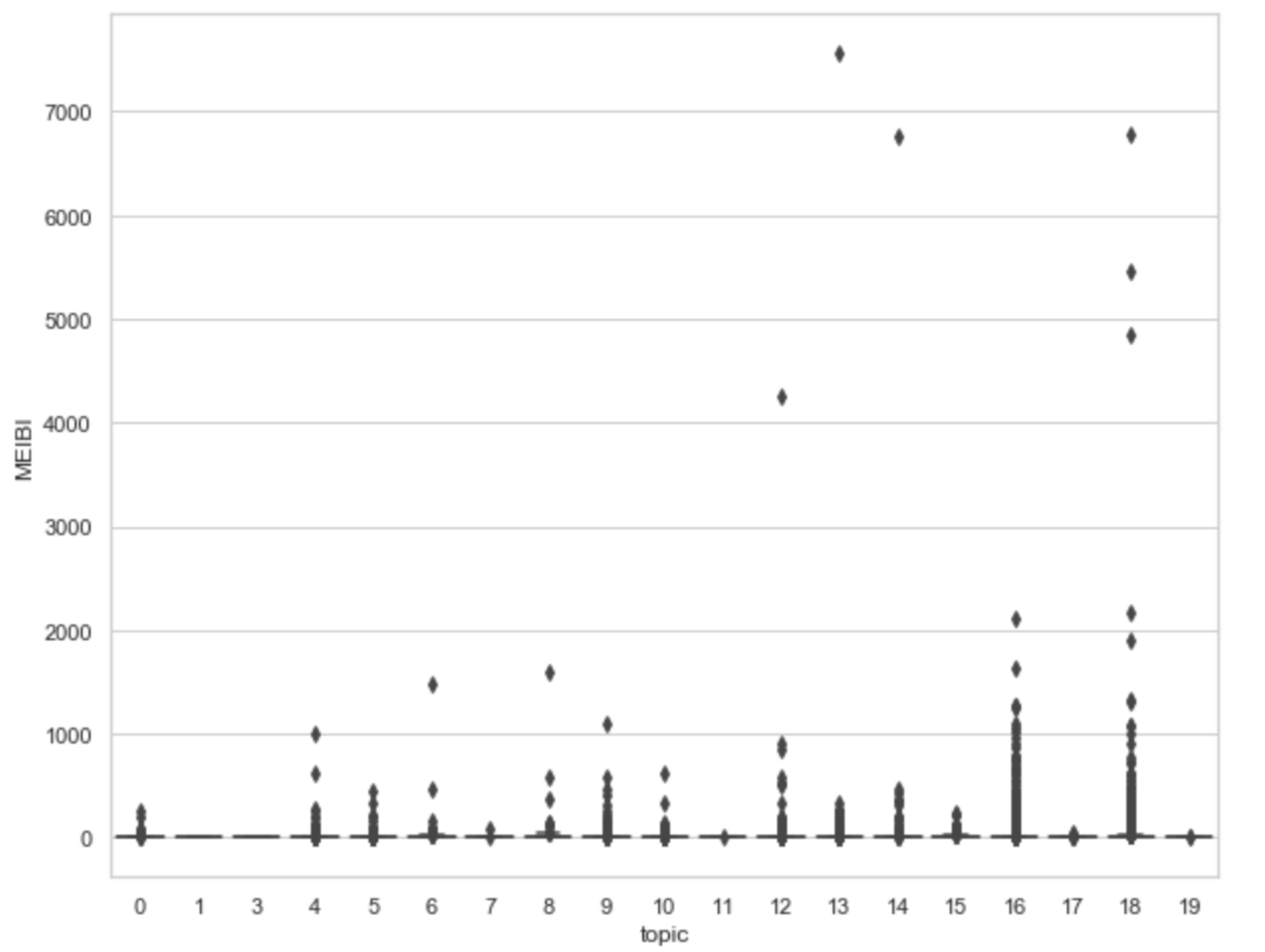
For example, topic 16 has the summarized keywords of “event”, “startup”, “launch”, “conference”, so The posts labeled as this topic are related to the new feature or new platform announcement.

**The Frequency Distribution of Topics**



As the frequency distribution of topics shown above, topic 16 (new product announcement) are the most popular topics, topic 13 and topic 18 are also very popular in the website.

**The Relationship between Topics and MEIBI**



As the relationship between Topics and influential index MEIBI is shown, some topics such as topic 16 and 18 have higher influence than other topics on the website.

# **Sentiment Analysis**

Then, we tried to determine readers attitude towards the posts with sentiment analysis. Sentiment analysis studies the subjective information in an expression, that is, the opinions, emotions, or attitudes towards a topic. We assigned a sentiment score to each comment, with 1 indicating positive, -1 indicating negative, and 0 indicating neutral. Then, we calculated the average sentiment score for each post.

**Sentiment Score for Each Comment**

A screenshot of a cell phone

Description automatically generated

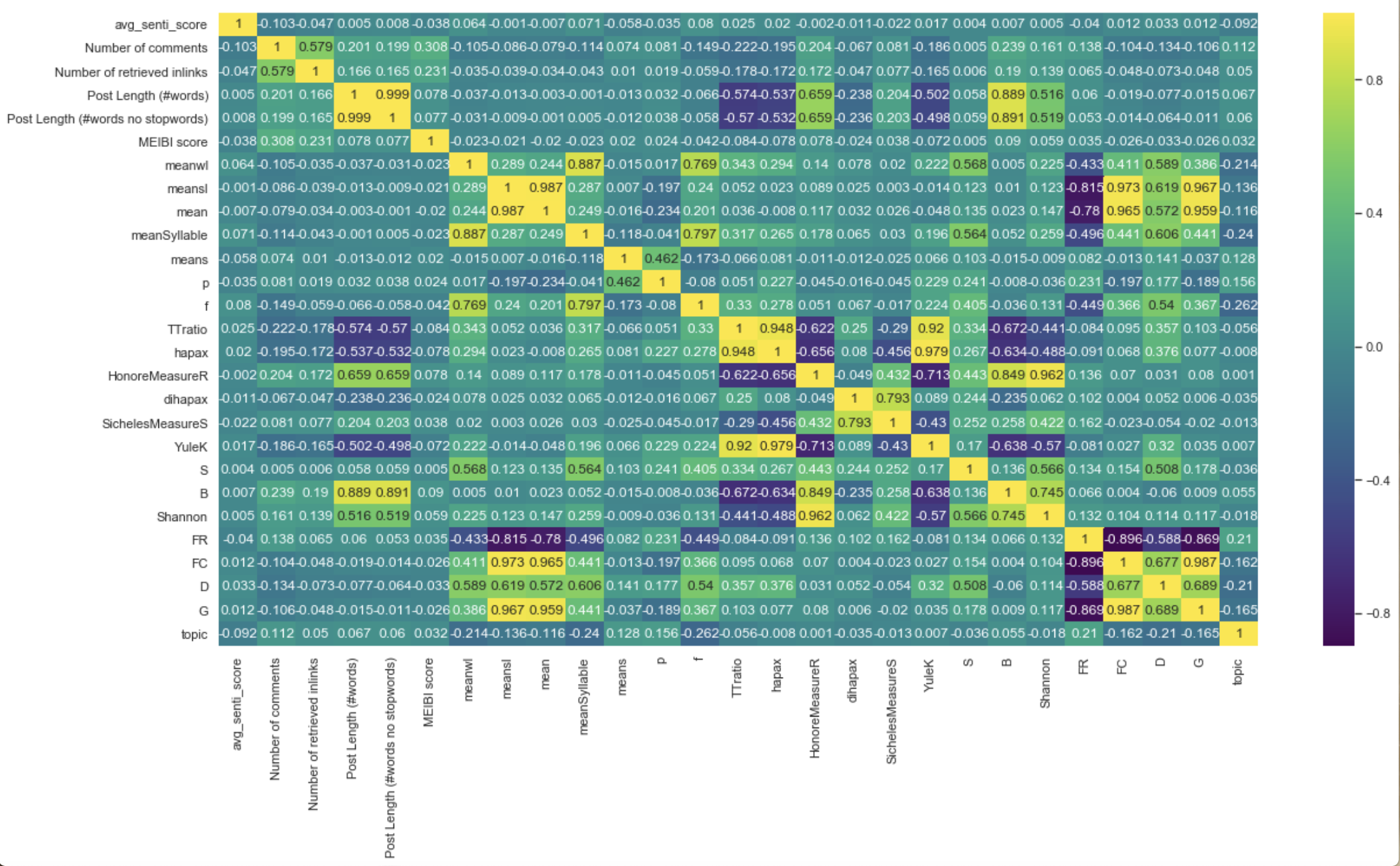
**Sentiment Score for Each Post**

A screenshot of a cell phone

Description automatically generated

# **Regression**

Finally, we developed multiple regression models with the influence index average sentiment score as the dependent variable and the features we created with writing styles identification, topic modeling, and sentiment analysis as independent variables. We first explored the correlation between the independent variables and dependent variables and excluded some variables to improve the performance of the model.



From the heatmap above, we can notice that the 'Post Length (#words)', and 'Post Length (#words no stopwords)' are very similar, so we only keep 'Post Length (#words)'.

#### For the Lexical Features: 'Average Word Length', 'Average Syllable per Word', and 'Functional Words Count' are highly correlated; 'Average Sentence Length By Word' and 'Average Sentence Length By Character' is highly correlated; 'Shannon' stands for 'Shannon Entropy' and 'S' is the 'Simpson's Index', and they both reflects the diversity or disorder of the system, here we keep one of them. So for Lexical Features, we only keep 'Average Word Length', 'Average Sentence Length By Word', 'Special Character Count', 'Punctuation Count', and 'S'(Simpson’s Index).

#### For Vocabulary Richness Features: 'TTratio' measures word richness, 'hapax' reflects the number of words that show up only once, 'YuleK' depends on words with all frequencies, and 'HonoreMeasureR' dependent on 'hapax'; they are intuitively highly correlated; 'SichelesMeasureS' depends on 'dihapax', which is the number of words show up twice; so for Vocabulary Richness Features, we only keep 'TTratio', 'dihapax', and 'B'.

#### For Readability Scores: from the heatmap, it is obvious that 'FR', 'FC', 'D', 'G' are highly correlated. These are all features that measure the readability of the text using different equations. So, we only keep 'D', which has the lowest correlations with other independent variables.

Then we built linear regression, tree regression as well as gradient boosted tree regression model. By looking into RMSE to measure the goodness of fit, it is shown that gradient boosted tree regression slightly outperformed the other two regressions. Thus, we chose the gradient boosted tree regression as the regression model and test the relationship between sentiment score and other independent variables.

**The RMSE of Regression Models**

|  |  |
| --- | --- |
| Model | RMSE |
| Linear Regression | 0.0806914 |
| Tree Regression | 0.0802835 |
| Gradient Boost Tree Regression | 0.0801473 |

As a result, we developed the importance of different features to the posts and those significant features will be used to select future influential blogs. We set 0.05 as our threshold and therefore we derived 12 important variables. For example, the number of comments and retrieved links and MEIBI are the most important variables to identify the influence as they represent the popularity of the posts. Besides, topics are also influential in blog matching. Some topics can arouse more empathy and positive evaluations from the audience and hence are more suitable for advertisements. Lexical features (Average Word Length, Average Sentence Length By Character, Punctuation count), vocabulary richness features (Hapax Legomenon, Hapax DisLegemena) and readability features (Dale Chall Readability, Simpson's Inde) features in writing styles also play a vital role in securing audience for the website.

**Feature Importance of the Model**

|  |  |
| --- | --- |
| Feature | Importance |
| Number of comments | 0.11065 |
| Number of retrieved inlinks | 0.08365 |
| MEIBI score | 0.08333 |
| topic\_0 | 0.07735 |
| Post Length (#words) | 0.07583 |
| dihapax ( Hapax DisLegemena(Sichel’s Measure)) | 0.06864 |
| meanwl (Average Word Length) | 0.06770 |
| meansl (Average Sentence Length By Character) | 0.06214 |
| D (Dale Chall Readability) | 0.06178 |
| S (Simpson's Index) | 0.05941 |
| p (Punctuation Count) | 0.05757 |
| TTratio (Hapax Legomenon) | 0.05397 |

# **Possible Impediment areas**

We tried to identify potential barriers to the implementation of our model — and think about some strategies to cope with these scenarios to make sure a better realization of our goals.

First, we used several perspectives in the model development, which involves stylometry features and different NLP techniques. Thus, a lack of understanding of the logic and the method we employed in our model might make it hard for the company to adopt our model and solution to the business problem.

Our method to deal with this potential obstacle includes adding enough and clear comments in the code, writing this report which goes through our methods used in developing the model, and making presentations to illustrate the impact our project can bring about.

Second, since the application of the model is directly associated with the revenue of the company, they might tend to be cautious in taking radical changes in how they match and put advertising contents on the website. Thus, our model needs to be examined first, the adoption will take some time and there can be some necessary further improvement of the model.

A strategy we can think about is suggesting the company to try out our approach by A/B experiment. By assigning the new advertising method to a group of users and the old way to another group, we can clearly tell how our approach works, and try to identify any improvements we need to do with our model.

# **Conclusion**

In order to identify influential blogs to improve the matching for advertising and sponsored content, we first used natural language processing to identify the writing style of posts, topics of posts, and sentiment analysis of comments. After that, we built a gradient boosted tree regression to test the relationship between the sentiment score and other features of the posts such as writing styles, topics, number of links and other influential indexes. From the analysis, we identified new features from natural language processing to measure the influence and likeness of the blogs, which can be used in future advertisement matching and increase the click through rate of the advertisement.