An analytics report on using blog with the Blog Authorship Corpus

Chuong Nguyen

Contents

1 Introduction

This study will process and analyze bloggers' information and their blog contents with Natural Language Processing (NLP) techniques. The study will apply descriptive statistics, clustering and analyzing sentiments models to extract insights from the corpus. NLP is a science of combining human languages and computer science. Due to the booming of big data and the rapid growth of data analytics, NLP has had many applications such as speech recognition, natural language generation, processing textual data, etc. In this study, NLP techniques will be used for text mining.

Blogging is one of the earliest social media platforms in the world. The social media industry has increasingly been successful with technology giants such as Facebook, Twitter, Instagram, etc. Understanding the behaviors of social media users could be beneficial for this industry. Although blogging is out of date, trends and patterns in blogging might help social media companies to figure out specific purposes and demands for using social media. From these insights, they can improve their service and even revive the golden age of the online diary with better platforms.

2 Aim

Initially, the study will produce descriptive statistics on the blogger information of the data in terms of the blogger age distribution, the number of bloggers categorized by gender, astrology and industry. Secondly, the number of words and posts and blog contents will be analyzed to find trends in using blog and word frequency respectively. The final purpose is to focus on categorizing and sentiment analyses of blog contents.

3 Methodology

The secondary data will be analyzed in this paper with the quantitative research methods and NLP techniques.

3.1 Data collection

The dataset is open sourced, and it is named 'The Blog Authorship Corpus' and shared in GitHub dataset for NLP. GitHub is a highly reliable online platform with more than 40 million users, where people can share data and their analytics. The dataset is collected from blogger.com in August 2004, and this unstructured data is text-heavy with around 800 thousand words after normalizing. This data has 19,320 files, corresponding to 19,320 bloggers. The files' names include information about identification number, gender, age, industry and astrological sign of each blogger. The files are stored with Extensible Markup Language (XML) format. Each file contains all posts' contents and

publish dates of posts, which were written and posted by bloggers. Totally, there are 658,805 posts in the data.

3.2 Methods of analysis

The study uses Spyder – an Integrated Development Environment in Python language to explore the dataset. The used Python packages include os, pandas, parse, dateutil, re, urllib, nltk, numpy, matplotlib, seaborn, wordcloud and scikit-learn. Firstly, bloggers information from file names and the contents of posts in each file are extracted and transformed to structured data in the data pre-processing step. In this step, file names will be converted to data frame by string methods in io and pandas. Next, bs4, pandas, urllib, re, and nltk packages will be used to extract month, weekday and year from publish dates and normalize text data of posts. The normalizing phase includes removing website links, de-contractions and removing punctuation, number, stop word. Then, pandas is used to count number of words in each posts after cleansing. After the pre-processing step, there are two structured data frames. The first data frame contains bloggers information with more than 19,000 observations. The second data frame has more than 650,000 observations (equivalent to more than 650,000 posts) and 14 variables of bloggers information, publish date and posts contents.

In the data processing step, the descriptive statistics will be used to summarize the main points of trends and patterns in using blogs by distributions, variability, two-way tables. Moreover, the unsupervised machine learning technique with the non-negative matrix factorization (NMF) algorithm in scikit-learn will be applied to find and label topics for each post. Besides, sentiment analyses with Vader from nltk package will be used to measure quantitatively the sentimental opinions in blogs. To show clearly insights about blogging, the study applies data visualization techniques with visual packages of Python including pandas, matplotlib and seaborn.

4 The analysis results

4.1 Analyzing blogger information

The blogger age distribution.

The histogram of blogger age shows that there are 3 age groups in the dataset. The numbers of bloggers from 14 - 17 and 24 - 27 age groups were dominant. The figure for 33-48 age group was far lower than others. The number of 17 years old bloggers was the highest number with around 2500 people. In the teenage group, the number of older bloggers was higher than the youngers. In contrast, there were less older bloggers in cohort from age 23 upward. The blogging seemed to be more attractive with young people than old people. This feature could help social media companies to design effective strategies based on customers' age.

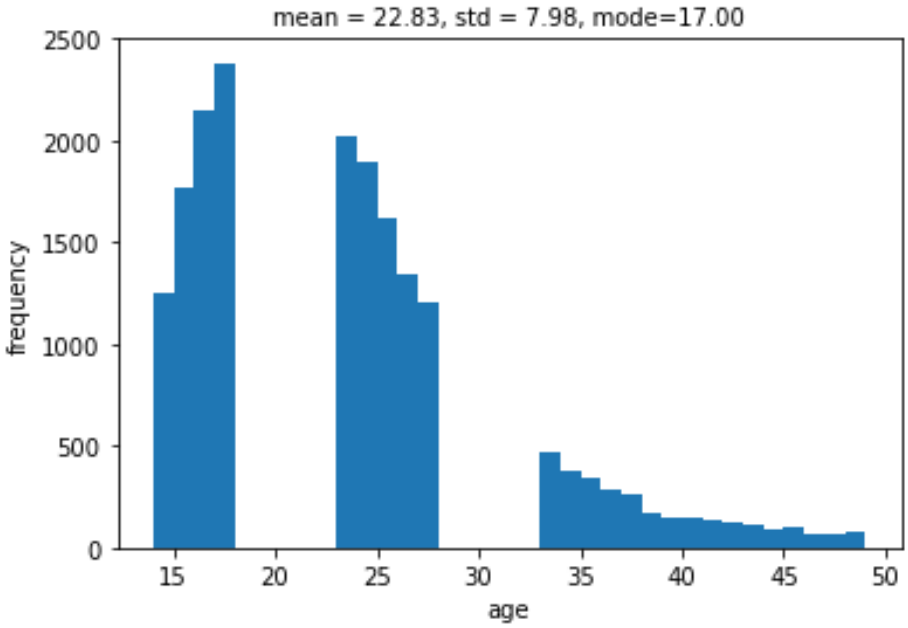

Fig. 1. A histogram showing the blogger age distribution

The number of bloggers with different categorizations

In comparisons between genders and astrological signs, there was no considerable difference between the number of bloggers in 12 astrological groups. Each of groups had from 700 to 900 bloggers for both genders. In terms of gender, the number of male bloggers in Aquarius group was higher than females while the figure for female bloggers in Pisces group was higher than males. In other astrological groups, the numbers of bloggers categorized by gender were slightly different.

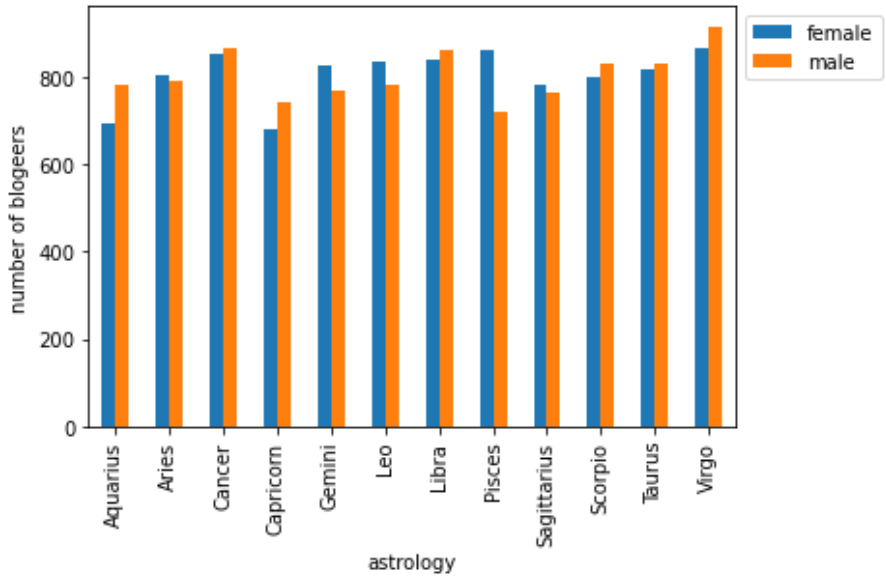

Fig. 2. A bar chart showing the number of bloggers by gender and astrology

With the view on gender and industry, the numbers of bloggers who were marked as unknown industry and students were 7000 and 5000 respectively, which were by far the most in comparison with other industry. It seemed that students had more time for blogging and bloggers did not want to declare their working in their online diaries. The number of male bloggers in Technology was far higher than that of female. This could be reasonable because there were more male employees working in this industry than females and they tended to try this new online platform.

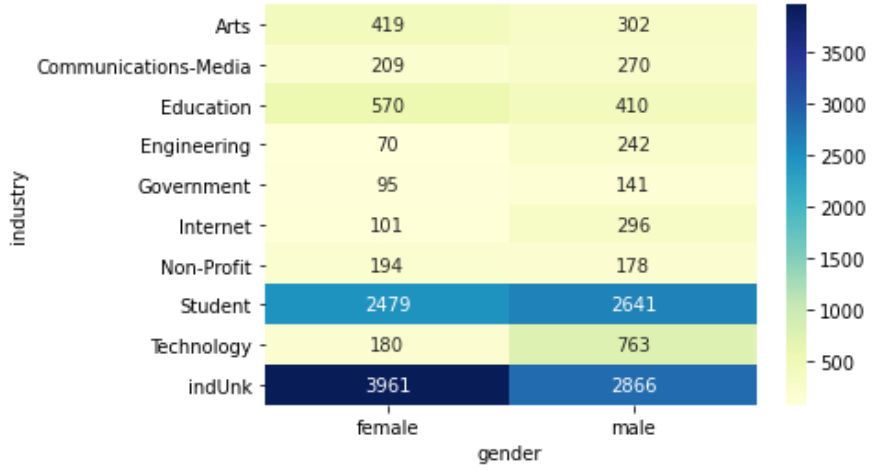

Fig. 3. A heatmap showing the number of bloggers by gender and industry

4.2 Analyzing number of post and post length

Word number distribution.

The histogram shows that most of the posts in this corpus had from 1 to 300 words after normalizing. The distribution indicates that bloggers tended to write short blogs, which contained below 50 words per post. Posts with 4 words had the highest frequency (equivalent to around 12000). The standard deviation implies that the word number in blogs spread out from the mean with a wide range. These numbers might be meaningful for imposing the limitation of contents in social media platforms.

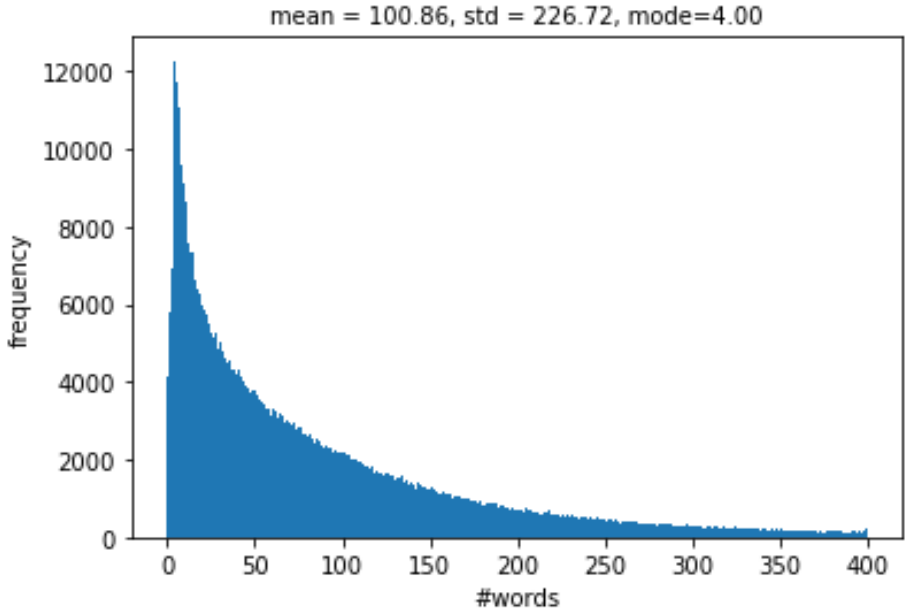

Fig. 4. Histogram showing the distribution of words number in posts

Monthly number of posts and post length.

The monthly number of posts increased gradually to more than 20000 from January 1999 to April 2004. After only 3 months, this number jumped suddenly to around 140000 posts in July 2004. The reason of this booming could be the re-design of blogging functions. Since 2004, bloggers were able to post by email, comment and archive posts, etc. (Wikipedia, 2020). This booming could demonstrate significant influences on the usage of online platform when social media companies improve their products for users' convenience.

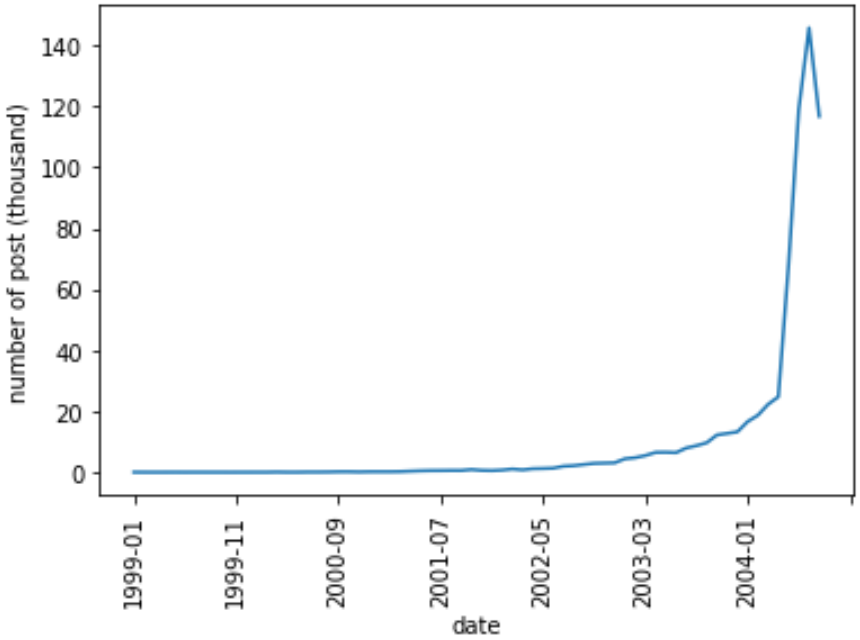

Fig. 5. Line chart showing the number of posts per month

Following the same pattern, there were below 3 million posted words per month from January 1999 to April 2004. Because of the heavy usage of blogs from May 2004 to August 2004, this number soared dramatically to peaks of 7 to 15 million words per month.

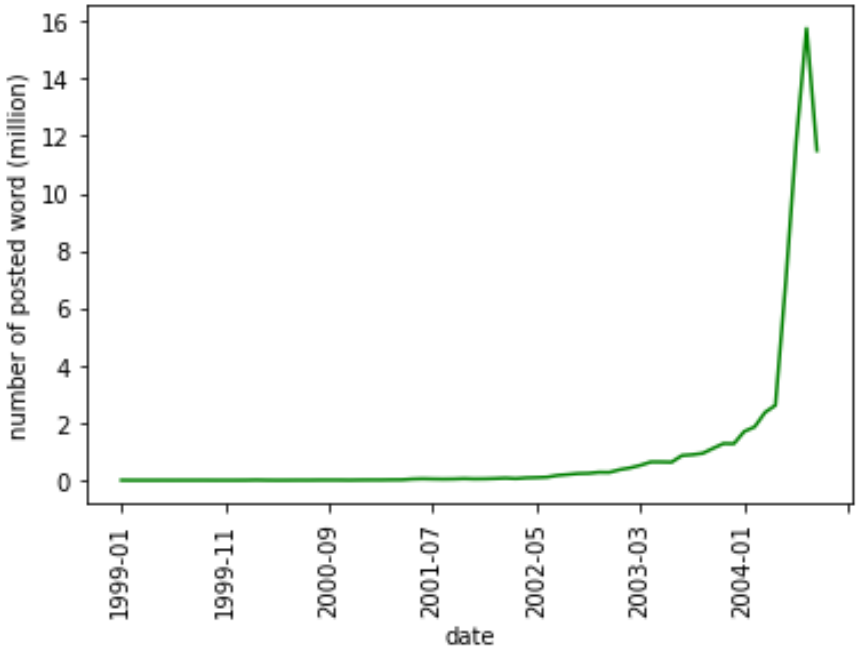

Fig. 6. A line chart showing the number of posted words per month

Average words number per post between weekday and month.

The bubble chart is used to express the relationship among three variables: average word per post, weekdays and months. From the bubble chart, the patterns in average posted words on weekdays are pointed out. Over the weekends, bloggers posted on average 100 to 110 words per day. From Tuesday to Thursday, this number decreased to the range of 90 to 100 words per day. These patterns show that blogging seemed to be a hobby in bloggers' leisure time and they likely wrote extensively on blogs in the weekends.

In terms of month, bloggers used on average 105 to 110 words per post in July and this number was slightly higher than others. Conversely, the average number of words per post in October was lower than others with under 90 words.

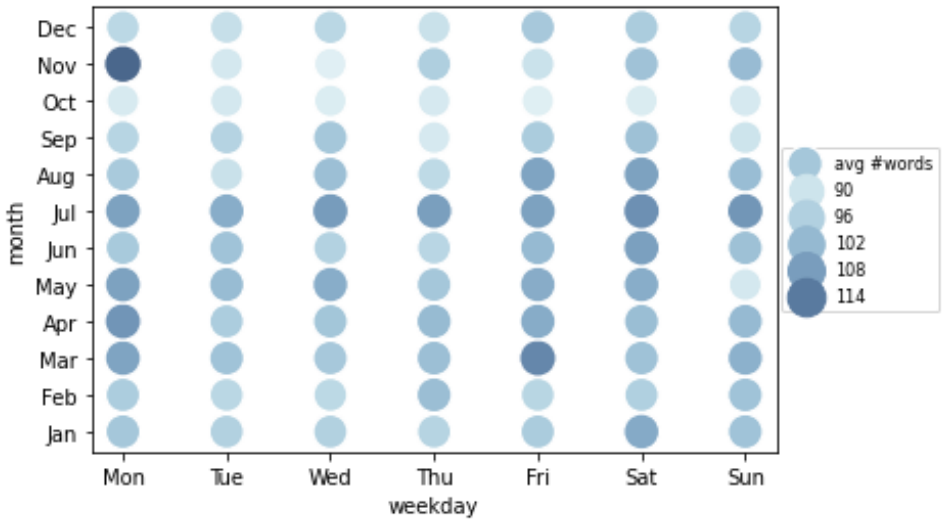

Fig. 7. A bubble chart showing the average number of words per post between month and days

Frequently used words.

The most used words were last, night, said, course, even, said and thought, which indicate that bloggers seemed to write about their past experiences in life. Besides, the high usage of word ‘urllink’ shows that bloggers were interested in sharing online resources with each other.

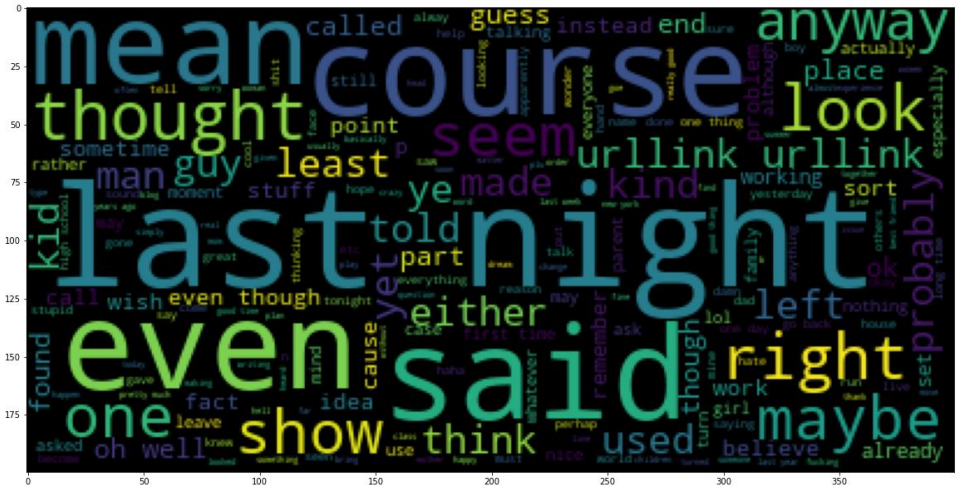

Fig. 8. A word cloud of the blog contents

4.3 Clustering and analyzing blog contents

Model for clustering.

In this section, Term Frequency – Inverse Document Frequency (TF-IDF) Vectorizer and NMF algorithm in scikit-learn are applied to classify words in the corpus. Firstly, the corpus is vectorized and embedded by TF-IDF Vectorizer to a matrix. Next, the NMF algorithm will factorize the corpus matrix into two matrices with optional number of components. From the factorized matrix, the top ten frequently posted words in each component will be showed for clustering.

In this study, the chosen number of components is three. Basing on the top ten words in each component from the figure, the purposes of using blog could be predicted with the meanings and relevance of these words. In each purpose, these words not only were posted regularly but also appeared together in posts. In the first purpose, blogger likely shared information with links, pictures and quizilla – a popular online platform for bloggers to create quizzes by themselves in the 2000s. The second purpose has verbs in past tense with words such as home, time, night and work, which indicate that bloggers shared diaries in their blogs. In the last purpose, the words like feel, think, love, like and people show that blog was a place, where bloggers wanted to share their emotions. The data will have a new variable for purpose labeled by the NMF model.

```
Top 10 words for purpose #0:
['new', 'posted', 'check', 'site', 'link', 'picture', 'com', 'brought', 'quizilla',
'urllink']


Top 10 words for purpose #1:
['home', 'time', 'night', 'work', 'good', 'going', 'went', 'day', 'today', 'got']


Top 10 words for purpose #2:
['really', 'things', 'feel', 'life', 'think', 'love', 'want', 'people', 'like',
'know']
```

Fig. 9. A figure showing top ten popular words in each purpose

Trends in purposes of blogging.

The bar chart show that bloggers mostly used blog for sharing their diaries and emotions. Both female and male bloggers posted from 120000 to 130000 blogs for these purposes. The number of posts for sharing information was far lower than other purposes with below 60000 posts written by each gender. In terms of gender, female bloggers wrote more posts for sharing diary than males did. Contrarily, the number of posts for sharing information from male bloggers was far higher than that of female. There was no different between the numbers of posts for sharing emotion from males and females. Basing on these trends, social media company could give out more emotional icons or the autocorrect for past tense verbs, which meet the demand for sharing emotions and past experiences.

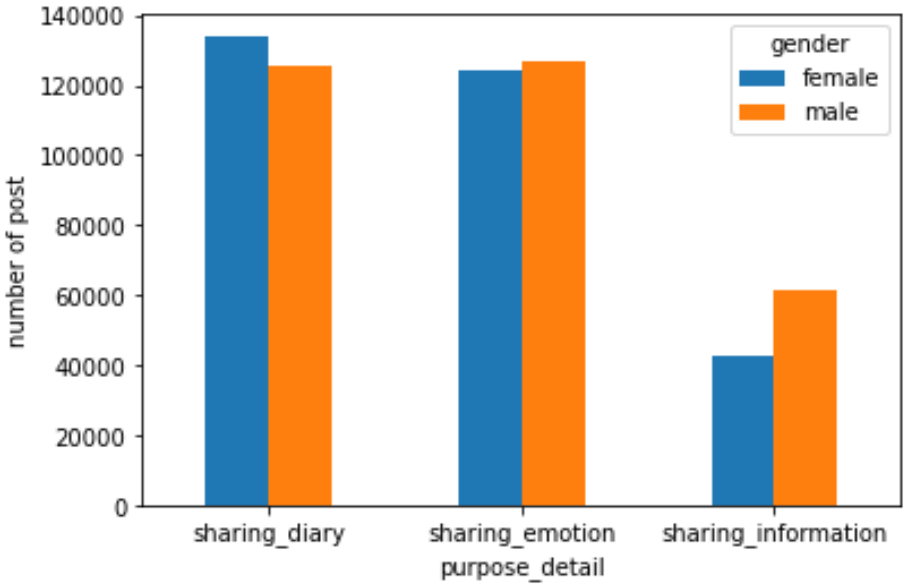

Fig. 10. A bar chart show number of posts by gender and purpose

Similarly, the number of words being posted for sharing emotion and diary was dominant in 2004. From January to April, the total number of words in posts for sharing diary went up slowly from 500 to around 850 thousand words per month. This number increased rapidly to 2 million words in May and reached a peak at 5 million words in July. The figure for sharing emotion and sharing diary have the same patterns and volumes. The number of words in posts for sharing information was extremely minor and the highest number was only around 600 thousand in the peak of July.

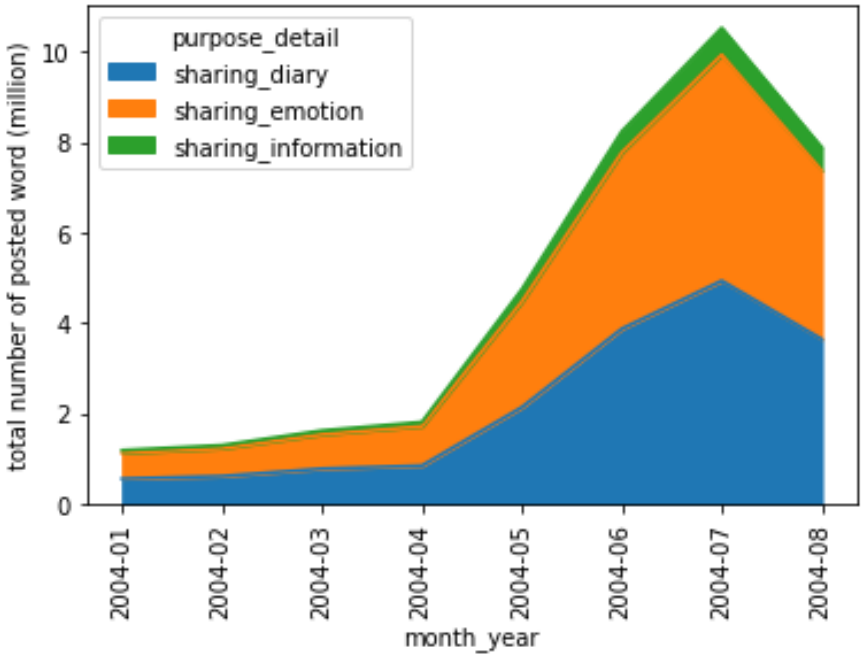

Fig. 11. An area chart showing total number of words in posts by purpose in 2004

4.4 Sentiment analyses

Scoring and categorizing sentiment in posts.

To analyze sentiment in each post, the study will use Vader to score and label sentimental level from these scores. Vader is the abbreviation of Valence Aware Dictionary for Sentiment Reasoning. Vader is a model based on a specific dictionary for sentiment. This dictionary has around 7500 sentimental words and elements, which are rated on a scale from -4 (extremely negative) to +4 (extremely positive) by independent raters (Ying, 2020). Vader model can simply compute the sentiment score without training models and machine learning algorithms. Due to the ability of scoring sentiment on punctuation, capitalization, adverb and contrastive conjunctions, the model will score the original posts before normalizing.

Basing on the dictionary for sentimental scores, the sentiment will be labeled from the compound scores, which are calculated as the below formular:

$$compound\ score = \frac{x}{\sqrt{x^2+\propto}}$$

x is the total score of words and other elements in posts (each word or element is scaled from -4 to +4)

α is equal 15

compound score is from -1 (most negative) to +1 (most positive)

The sentimental categorization will be labeled as the scale below:

- positive sentiment if compound score >= 0.05
- neutral sentiment if compound score < 0.05 and > -0.05
- negative sentiment if compound score <= -0.05

After labeling, the dataset will have two new variables. A numeric variable for sentiment score from the compound score and a categorical variable for sentiment.

Sentimental patterns in using blog.

Following the bar chart, the use of language in blogging was likely positive with around 200 thousand positive sentiment posts from each gender. These numbers were nearly three times higher than that of negative posts. Male bloggers posted more negative blogs than females did. There were below 40 thousand neutral blogs from each gender, which were far lower than others.

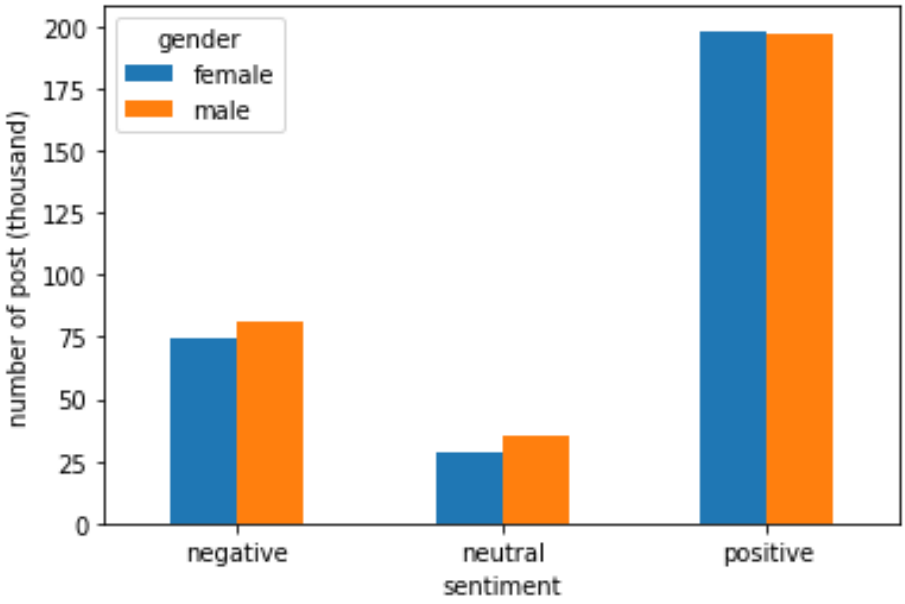

Fig. 12. A bar chart showing the number of posts by gender and sentiment

In comparison of the average sentiment score, the heatmap shows that there were noticeable differences in average sentiment score between gender and among astrological signs. In general, the written language of female bloggers was likely more positive than that of males. The average sentiment score for blogs posted by female bloggers was often from 0.35 to 0.4, which was higher than that for blogs of male bloggers. These average scores were contrastive between gender in each astrological sign. Exceptionally, there were no difference in the average sentiment scores of posts between gender from Virgo and Cancer bloggers.

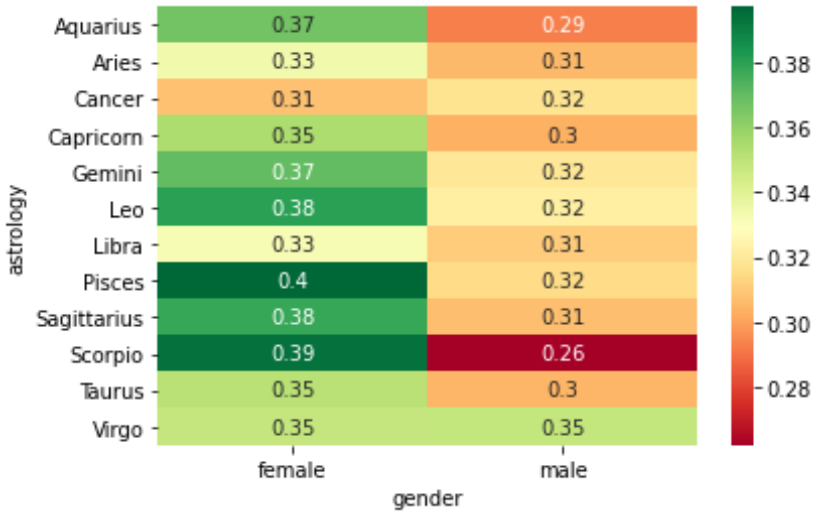

Fig. 13. A heatmap showing average sentiment score by gender and astrology

The bubble chart show that bloggers seemed to use language more positively in weekend and Monday than other days. From Tuesday to Thursday, the average sentiment score was from 0.30 to 0.35 while this number in weekends was higher with the range from 0.34 to 0.42. These higher scores indicate that weekends could help bloggers to have better moods, which were expressed in their posts. In terms of month, the blogs posted in December were likely more positive than other months with the average sentiment score from 0.34 to 0.41. The reason of these higher scores in December could be the long Christmas holiday with its positive emotions. From these behaviors, social media companies could organize online events or campaigns in the weekend or in the end of the year to take advantage of positive emotions.

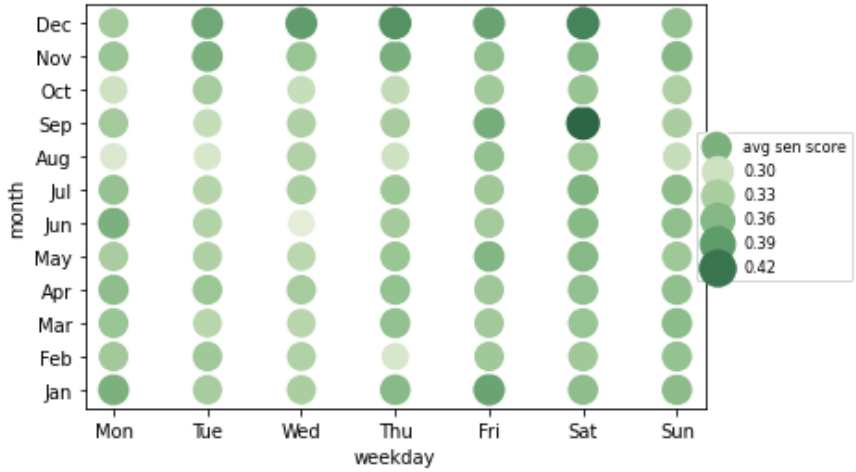

Fig. 14. A bubble chart showing the average sentiment score between weekday and month

5 Conclusion

In conclusion, the study showed trends and patterns in bloggers' information and their blogging. The number of bloggers was analyzed in different aspects including gender age, astrological signs and industry. The number and length of posts were also visualized and explored to find the insights. Moreover, the model for clustering blogs' contents and sentiment analyses gave more imaginations about the purposes of blogging and expressed sentiments in posts.

However, the study lacks the accuracy evaluation of the model for clustering. To evaluate the accuracy of the model, there should be a more detailed study on coherence scores, which benchmark the fitness of the clustering model. Furthermore, the model could be not accurate with long posts and complex contents. Besides, there could be potential biases in scoring sentiment by Vader, which might derive from raters' biases when they compiled the sentimental dictionary. Moreover, the dataset might not represent to the population due to the data sampling method. This may lead to the biases toward the study if the data collection has biases.

6 References

1. ‘Blogger (service)’ (2021), *Wikipedia*. Available at: https://en.wikipedia.org/wiki/Blogger_(service) (Accessed: 06 April 2021).
2. ‘Natural language processing’ (2021) *Wikipedia*. Available at: https://en.wikipedia.org/wiki/Natural_language_processing (Accessed: 06 April 2021).
3. Git Hub (no date) ‘Blog Authorship Corpus’. Available at https://u.cs.biu.ac.il/~koppel/BlogCorpus.htm (Accessed: 06 April 2021).
4. GitHub Guides (2020) ‘Hello World’, 24 July. Available at https://guides.github.com/activities/hello-world/ (Accessed: 06 April 2021).
5. Ying, M. (2020) ‘NLP: How does NLTK.Vader Calculate Sentiment?’, *Medium*, 05 Feb. Available at: https://medium.com/ro-data-team-blog/nlp-how-does-nltk-vader-calculate-sentiment-6c32d0f5046b (Accessed: 06 April 2021).