Paper Anniversary Surprise for James

Natural Language Processing Analysis on WeChat History Before vs. After Getting Married

Suting Yang[1]

[1] Mechanical and Industrial Engineering, University of Toronto

Abstract

July 10, 2022, marks Zijian (James) Wang and Suting (Bean) Yang's first marriage anniversary, also known as the "paper anniversary". Bean hence prepares this paper as a memorable and unique anniversary gift for James. The paper aims to analyze the Wechat history between the couple, before and after getting married, with the purpose to demonstrate that the marriage has deepened the couple's relationship bond and made them happier than before. 100 WeChat messages are randomly selected pre- and post-marriage as two different sample sets. Natural language processing (NLP) methods are used to tokenize text and analyze the most frequent words. The author also conducts a sentiment analysis with statistical tests to identify any substantial sentiment change before vs. after getting married. The analyses have demonstrated that James is Bean's perfect match, and marrying him is her best decision ever in life (and vice versa). The author will always love James and is more than willing to spend the rest of her life with him.

Keywords: Natural Language Processing; Sentiment Analysis; Data Analytics; Successful Marriage; Paper Anniversary; Best-ever Decision; Key of Marriage; Commitment Strategies

Introduction

Bean (Industrial Engineer) and James (Mechanical Engineer) have been together for 7 years. On June 30, 2015, on that hot steamy summer afternoon, they first met at the Grad School Seminar held by the University of Toronto (UofT). On that day, they got to know each other and exchanged their contact information. Later, they studied for the G1 Driver's Exam every day until they were together on July 10, 2015. They spent much time together in the labs during their third-year Engineering, helping each other with assignments, applying for intern jobs and preparing for interviews and exams together.

For their internship year (2016 to 2017), coincidentally, both of them secured co-op jobs in Mississauga. During that year, they had a chance to step away from all the stress from school. They traveled to different places, which were all wonderful memories.

After their internships, they returned to school to complete their fourth year of study. After a year of hard work, they both successfully graduated with honours from the UofT. James started working for a consulting firm as an acoustical engineer; Bean continued her study as a master student at the UofT, pursuing a career in healthcare engineering and data science. They have been through many ups and downs together, especially during the pandemic, at which time Bean received strong mental and physical support from James.

In 2020, they moved to Aurora and got their first and second puppies, Waffle and Fuwa. On Valentine's Day of 2021, James proposed to Bean, and Bean SAID YES! On July 10, 2021, they officially became husband and wife at the Fantasy Farm in Toronto. That day marked not only their sixth anniversary of being together, but also the beginning of a new chapter of their lives. Since that day, the five individuals have formed a family consisting of two loved ones, two lovely dogs, and a 7-year-old bunny; and since that day, they have enjoyed every single second, minute, hour, day, week, month, season, and year of their lives together.

To celebrate their first wedding anniversary on July 10, 2022, Bean decides to write a paper to James as an anniversary gift. Bean has written so many papers during her undergraduate and graduate studies, but this one is especially for James. This paper aims to analyze their WeChat history before and after getting married to identify differences in the content and sentiment of the WeChat messages. It is a unique gift in the whole world that shows how Bean loves James deeply in her heart.

Method

Data Collection

The study period for this paper ranges from January 2021 to December 2021 and is segmented into 2 time phases, the pre-marriage period (before July 10, 2021) and the post-marriage period (after July 10, 2021). Due to the limitation of the WeChat App, the complete chat history could not be directly extracted. Therefore, the author randomly selects 100 messages during the pre-marriage time period and 100 messages during the post-marriage time period as two different sample sets, under the assumption that the data sets are unbiased. All WeChat messages are manually copied from the App and pasted into the *.csv* files for later analyses.

Data Processing

Once the data is collected, the author applies the NLP techniques to convert each text into analyzable words. The NLP process consists of the following steps:

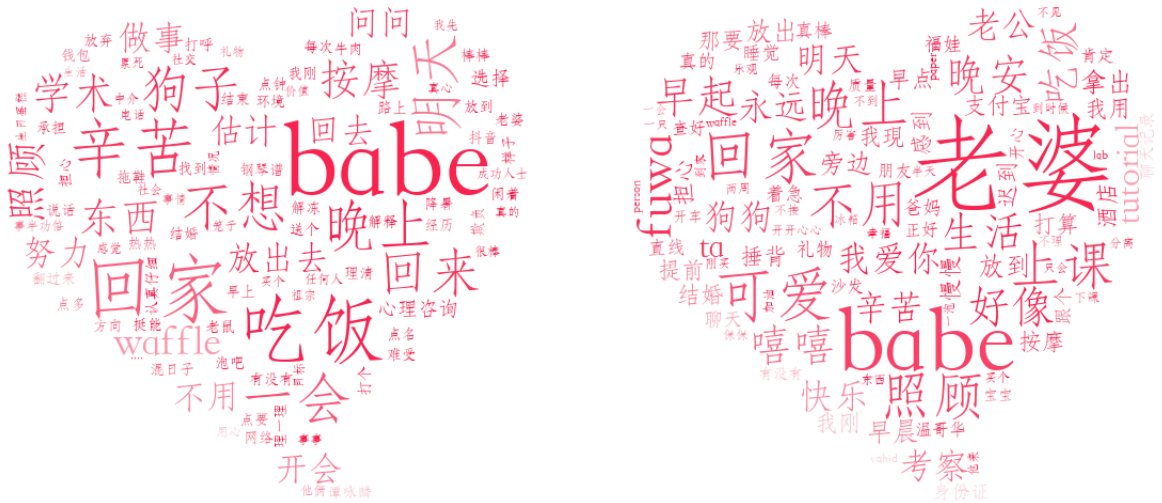
1. **Word tokenization:** word tokenization is the process of splitting a text into words. The author uses a built-in `Jieba` (**Neutrino**, 2020) package in *Python* that segments each Chinese message into individual words.
2. **Removal of stopwords:** the stopwords in NLP are the most common words in data that we do not want to use to describe the topic of the content. The author uses a stopword corpus (available online, with both Chinese and English stop words) as the reference to remove stop words.
3. **Removal of single-character words**: the author removes the words with only a single Chinese character.

Statistical Analysis

The statistical analysis consists of two parts. First, the author analyzes the most frequent words. At this stage, WordCloud images are created to visually represent the top words. Cloud creators are used to highlight popular words and phrases based on frequency and relevance. Second, the author performs sentiment analysis on the text content before and after getting married. At this stage, the author uses `SnowNLP` (**Wang**, 2020), a built-in *Python* module for the sentiment analysis of Simplified Chinese words. The built-in algorithm in this package uses a pre-trained sentiment prediction model such that given an input text, it outputs a sentiment score ranging from 0 to 1, with 0 representing the most negative sentiment and 1 representing the most positive sentiment. The author then analyzes the distributions of sentiment scores of WeChat messages for pre- and post-marriage time periods and applies a *t*-test to identify the significance of the difference in sentiment value distributions.

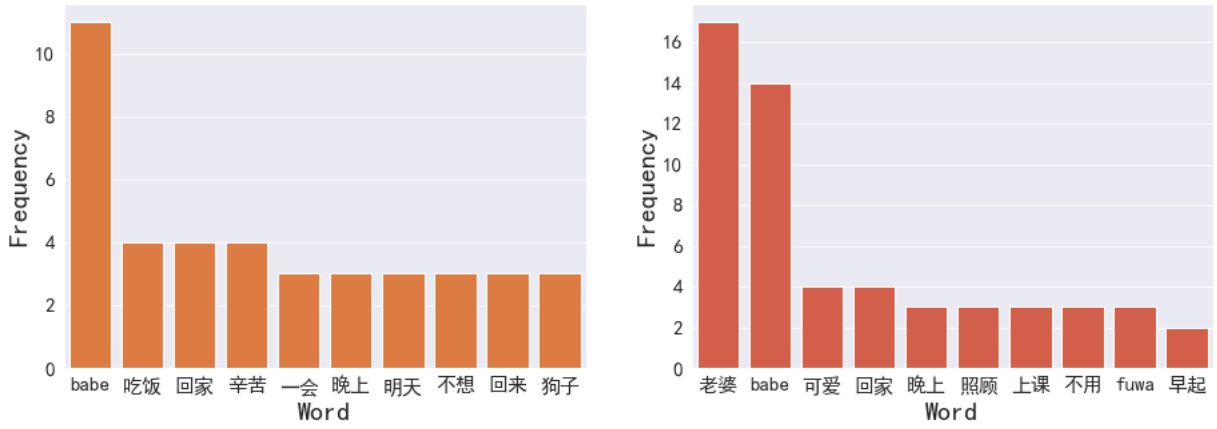
Result

Figure 1 shows WordClouds representing the frequency of text words within pre- and post-marriage time periods. Figure 2 shows bar charts of the top 10 frequent words. We can infer that the top words presented in both time periods are (1) “babe” - an affectionate term used between the couple; (2) “return back home” - a word frequently used to show the couple’s eagerness to stay with each other after a busy day of work; (3) “dogs” - a frequent topic talked between the couple because they love Waffle and Fuwa so much; (4) “take care of” (mostly from James) - a word that shows James’ determination of taking care of Bean and the dogs. It is noteworthy that the top word which uniquely appears after getting married is “wifey”, a term often used when James calls Bean, as she has officially become his wife. Another top word that only appears after getting married is “cutie” (mostly from James to describe Bean), maybe because James has found the truth that Bean has become even cuter than before. It is also noted that some negative words such as “extremely tired” no longer appear as top words after getting married. Most likely, it is because that their marriage have brought them more happiness, making these negative thoughts disappear as if the vapor in a desert. Moreover, the main finding is that after getting married, they say “I love you” to each other more frequently than before. According to the Greek philosopher Aristotle, “I love you” is an important word that shows their trust and commitment to marriage.



(a) Pre-marriage **(b)** Post-marriage

Figure 1. WordClouds



(a) Pre-marriage **(b)** Post-marriage

Figure 2. Bar charts of Frequent Words

Table 1. Descriptive statistics of sentiment scores

	Pre-marriage	Post-marriage
Mean (STD)	0.51 (0.24)	0.61 (0.27)
Min	0.00023	0.274
Median	0.48	0.64
IQR	0.33-0.69	0.39-0.87
Max	0.995	0.996

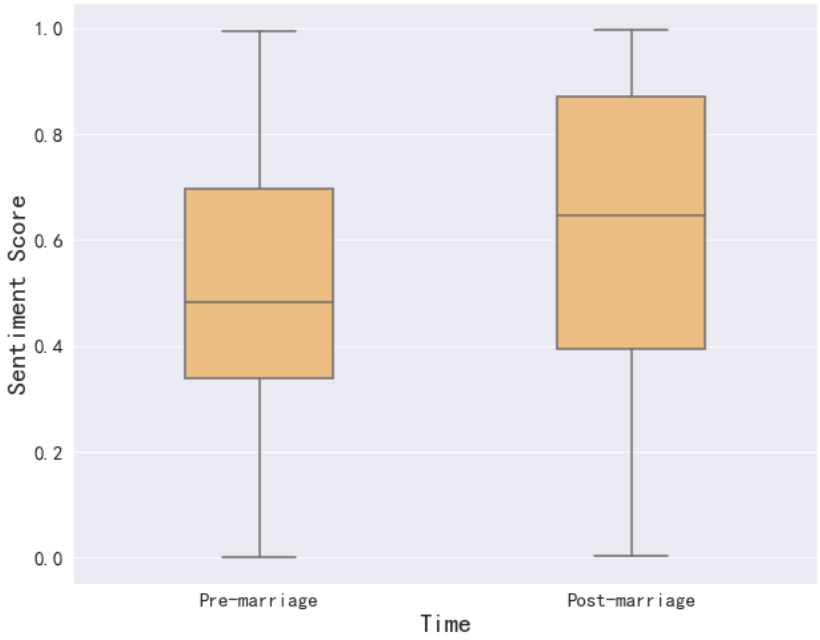


Figure 3. Boxplot of sentiment distributions

Table 1 summarises the distributions of sentiment scores for pre- and post-marriage time periods in terms of descriptive statistics, and Figure 3 visualizes the distribution in terms of the box plot. There is a significant increase in the average sentiment score (0.61 post-marriage vs. 0.51 pre-marriage, 20% increase), indicating that the marriage has made the couple happier than before. The t-test has proved that such increase is significant (t-statistic $= -2.9$, p-value < 0.01).

Discussion

Many studies have shown that family, particularly marriage, would appear to provide rewards in terms of life satisfaction, mental well-being, mental health, and physical health (**Gove, Hughes, & Style**, 1983). This study, through NLP and statistical analysis, is consistent with those results. The top words in the couple's WeChat messages have shown that (1) there is a strong bond in their relationship; (2) they love their dogs so much that they treat them as their family members; (3) they trust each other and are committed to their marriage. The sentiment analysis for pre- and post-marriage periods has provided convincing evidence that the couple's marriage positively affects their psychological well-being and life fulfillment.

This paper is innovative in a way that it utilizes NLP and statistical analysis to scientifically and logically prove that marrying James is Bean's best decision ever (and vice versa). To our knowledge, this is the first study that uses data science methods to analyze the couple's text messages. Despite its strength and uniqueness, this paper still has the following limitations. First, due to the limitation of the WeChat App, the author only manually extracts a small sample (100 messages pre-marriage and 100 messages post-marriage). As such, the sample data sets might not represent the whole chat history. Second, when comparing the outcome (sentiment scores), the author applies a univariate analysis method with only one exposure variable (i.e., whether the time is post-marriage). Since the data collected belongs to time-series data, timestamps might have impacted the dependent variable. As such, a mixed-effect model with the timestamp as a random effect variable would have been more appropriate. Third, the sentiment prediction model in the `SnowNLP` package was trained on comments made when purchasing a product. Since the author uses the trained model to predict the sentiment score on other domains, the scores might not be precise enough (**Wang**, 2020). A more accurate approach is to prepare the customized data set by collecting the positive and negative sample sentences and re-train the model based on the new data set. Despite these limitations, the author still believes that the insights from this study are valid and convincing.

The future directions of this study include the following. First, the author could extract the frequencies of a typical word on a regular basis and perform an interrupted time series (ITS) analysis (**McDowall, McCleary, & Bartos**, 2019) to evaluate the influence of their marriage on such frequencies. For example, Bean could extract the frequencies of them saying "I love you" each day/week, and then use ITS to detect a trend difference (i.e., whether after getting married, they have statistically higher frequency of saying "I love you"). Second, as previously mentioned, the author could apply a mixed-effect model with temporal features as random effects so that the insights gained are risk-adjusted. Third, on their decade-long anniversary, Bean could re-perform this retrospective study to discover the changes in their text messages over a decade.

Conclusion

The author performs a natural language analysis study on Bean and James' WeChat history before and after getting married, which has provided convincing evidence that they are fully committed to each other and that their marriage has made them happier than before. Despite the time limit and other few limitations of this paper, the author believes that the insights gained are valid and convincing, and it is a unique and memorable gift to James. Finally, the author wants to conclude the paper using a quote from one of her favourite books "*Flipped*" (**Draanen**, 2001):

"*Some of us get dipped in flat, some in satin, some in gloss,* ***but every once in a while you find someone who's iridescent, and once you do, nothing will ever compare.***"

Acknowledgements

I would like to express my sincere gratitude to everyone who helps me with this special paper. This work would not have been possible without all your support. I want to thank my friends - Sherry, Cynthia, Hachio, Vivian, Austin, Cass, Dexter, Katee, Terry, Elena, Yubo, Han - for spending time to peer review this paper and providing comments. Special thanks to Sherry for providing your insights and expertise on NLP techniques, and to Cynthia for your valuable suggestions on the future directions. I would also like to thank my manager, Saba, for your strong interest and willingness to review this paper. Last but not least, thank you James for always loving me and giving me support whenever needed. I will keep loving you as always.

- Bean

References

Draanen, W. V. (2001). *Flipped*. Ember.

Gove, W. R., Hughes, M., & Style, C. B. (1983). Does marriage have positive effects on the psychological well-being of the individual? *Journal of health and social behavior*,(), 122–131.

McDowall, D., McCleary, R., & Bartos, B. J. (2019). *Interrupted time series analysis*. Oxford University Press.

Neutrino. (2020). *jieba.* `https://github.com/fxsjy/jieba`. GitHub.

Wang, R. (2020). *Snownlp.* `https://github.com/isnowfy/snownlp`. GitHub.