Sentiment Analysis Based on Frequency of Colour Names on Social Media

Boshuo Guo
University of Leeds

Peihua Lai
University of Leeds

Sentiment Analysis Based on Frequency of Colour Names on Social Media

Boshuo Guo [1,*], Stephen Westland [1,*] and Peihua Lai [1]

[1] The University of Leeds; b.guo1@leeds.ac.uk, s.westland@leeds.ac.uk, ss17pl@leeds.ac.uk
* Corresponding author: b.guo1@leeds.ac.uk, s.westland@leeds.ac.uk

Abstract

This study aims at finding out the sentiment associated with eight colour terms in the context of an overall negative marketplace sentiment during 2020/2021 and how the sentiment varies over time. We focus on the valence aspect of sentiment. We collected two datasets, separated by six months, each containing 18000 mentions of each of the eight colour terms in English from Twitter users around the world. We calculated the weighted average sentiment score of each instance when a colour is mentioned. We find that purple, pink and green have a positive average sentiment score in both observation points. Brown, red and orange are negative in both observation points. We also find that the relative sentiment value associated with the colour terms did not significantly vary over the six months. This finding indicates that there is a strong coherence in the sentiment. Our work contributes to colour perception in marketing communication.

Keywords: *colour, marketing, psychology*

INTRODUCTION

Colour plays an important role in marketing communication. Research shows that colour makes advertisements more attractive (Bohle and Garcia 1986), promotes favourable attitudes (Fernandez and Rosen 2000) and influences customers' information processing (Lee et al. 2014). Understanding the sentiment associated with colours enables marketers to generate positive online word of mouth, improve persuasion (Han, Duhachek and Agrawal 2014) and effectively promote products.

In this study, we aim at finding out the sentiment associated with eight colour terms in the context of an overall negative marketplace sentiment in 2020 and 2021 and how the sentiment varies over time. We focus on the eight chromatic colours that are blue, brown, green, orange, pink, purple, red and yellow; we chose these colour terms because they are the chromatic terms in the 11 basic colour names (the other 11 basic colour names, that we did not include, are white, black and grey). We focus on the valence aspect of sentiment. We collected the first datasets in July 2020 and the second dataset in January 2021. Each dataset contains 18000 mentions of each of the eight colour terms in English from Twitter users around the world. We measure sentiment using the weighted average sentiment score of each instance when a colour is mentioned based on known sentiment scores of other words in the tweet.

We have two main findings. First, our sentiment measurement shows that purple, pink and green are the colours with a positive average sentiment score in both observation points. Purple is the colour with the highest average score, indicating that the users exhibit the most positive sentiment when mentioning purple. Brown, red and orange are negative in both the 2020 and 2021 datasets. Our second finding is that the sentiment value associated with these colour terms did not significantly vary over the six months. When mentioning a colour, twitter users used more than half of the words that are identical between the two observation points. In terms of the words that carry sentiment meanings, the match rate is even higher (>75%.) This finding indicates that there is a strong coherence in the sentiment associated with the mentions of colour within six months.

In the next sections we describe our data, develop our measurement of sentiment and discuss our findings, as well as study limitations and directions for future research.

DATA AND METHOD

We collected two datasets from Twitter, using Twitter API and Rtweet (Kearney 2019) package. Each dataset contains 144000 Twitter posts (18000 posts per colour name x 8 colour names). In order to avoid repeated content in retweets and consequent bias in calculation, we included only original Twitter posts in our datasets and no retweets are included. We group the posts according to mentions of colour. We collected the first dataset in July 2020 and the second dataset in January 2021, in order to assess the coherence and temporal consistency in the sentiment associated with these colour names.

In our datasets, every post is described with 90 variables, including time when the post was created, post text, user's name, hashtags, number of likes and more. However, we simply focus our analysis on post text in this study. We collected the posts from Twitter users around the world and only include in our datasets the posts in English.

We cleaned the datasets and made the post text into "tidy text" in the following steps. First, we removed URLs, mentioned user's names and hashtags from the text. This is to avoid the bias in calculating the sentiment value, since, although some URLs and users' names contain words that carry sentiment meanings, quoting these elements does not uncover the sentiment of the post author. Stop words are common words that do not carry meaning, for example, "is" and "the". We follow the routine of natural language processing and removed these words from text too. We split the remaining text into single words and calculated the frequency of every word in the 18000 posts of each colour.

We derive sentiment value of each word using AFINN (Nielsen, 2011) lexicon. AFINN lexicon rates the valence of English word using a score from -5 to +5. A positive score indicates positive emotion, while a negative score indicates negative sentiment. A score with a bigger absolute value indicates a stronger emotion.

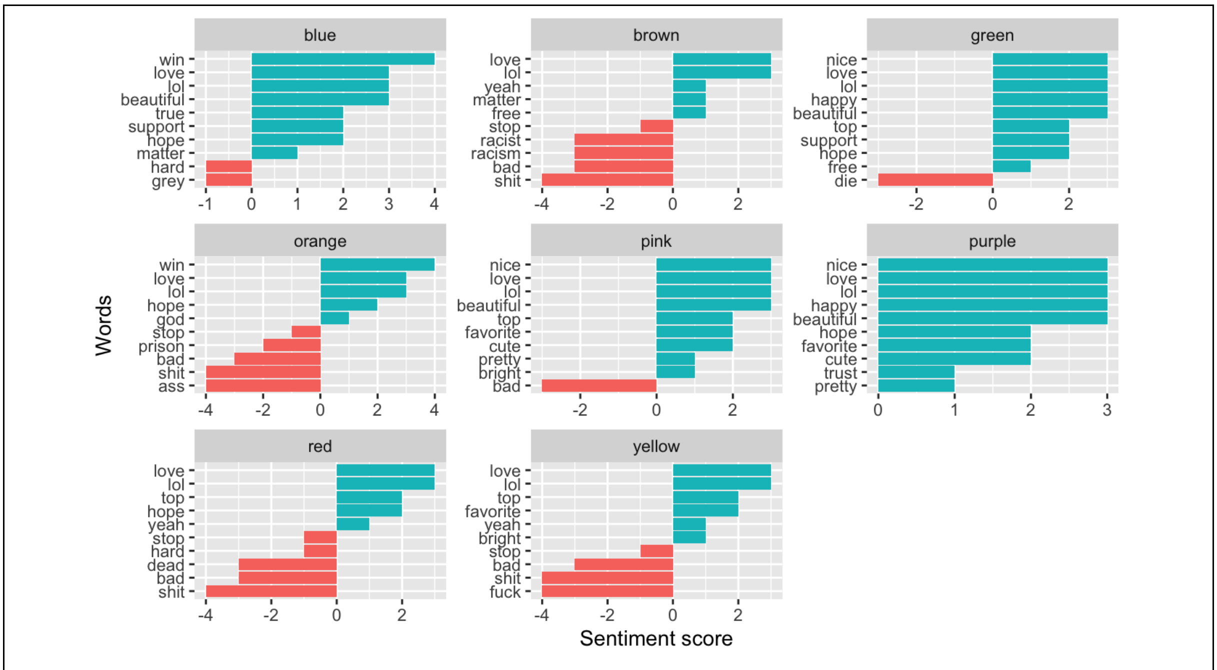

Figure 1: The ten most frequently used words (and their sentiment scores) for each of the colour words in the 2020 July dataset.

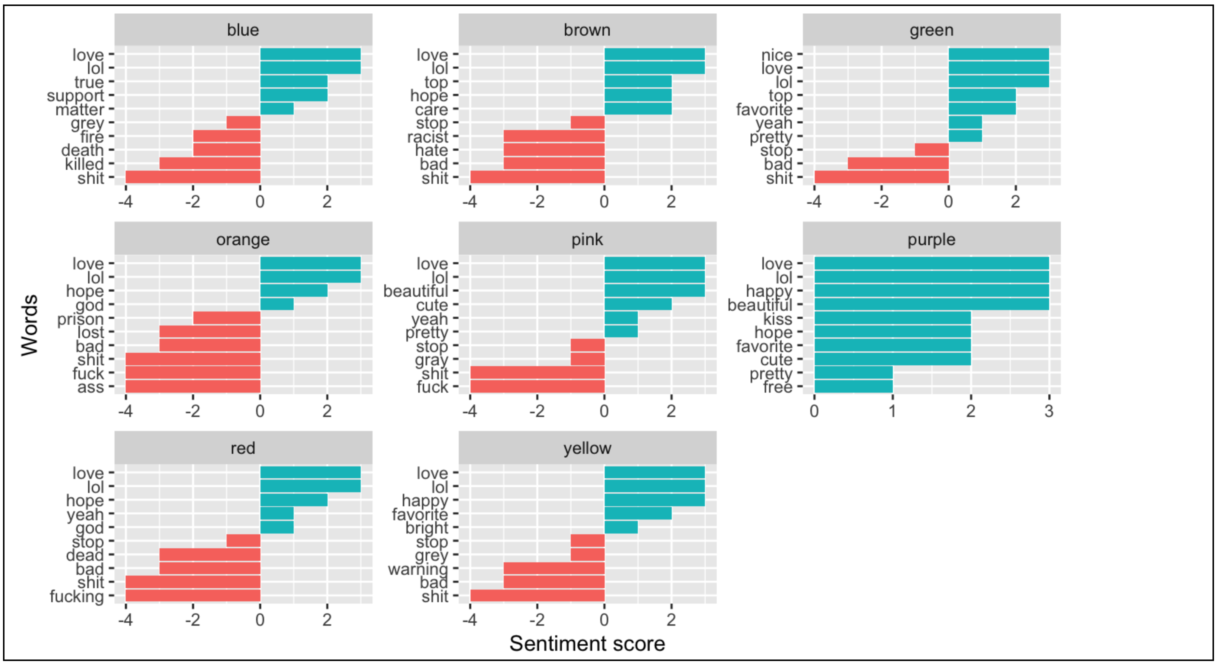

Figure 2: The ten most frequently used words (and their sentiment scores) for each of the colour words in the 2021 January dataset.

Figure 1 and Figure 2 show the ten most frequently used words and their sentiment scores in the group of posts of every colour in each of the datasets. Figure 1 and Figure 2 display words that contribute the most sentiment score in the posts of every colour. In both figures, we use green bars to indicate positive sentiment scores, whilst red bars are used for negative sentiment score.

We develop the measurement of colour associated sentiment as a weighted average sentiment score, as documented in Formula 1.

$$colour\ associated\ sentiment = \frac{\sum(Score * frequency)}{\sum frequency} \quad (1)$$

We use the word frequency as the weight, since a word contributes more to the overall sentiment value if it is used more often in the posts. We multiply every word's sentiment score by the number of times this word was used in all the posts mentioning a certain colour. We sum up this value of all the words in the posts mentioning a certain colour and divided it by the total word frequency.

RESULTS

We present the colour associated sentiment values in Figure 3.

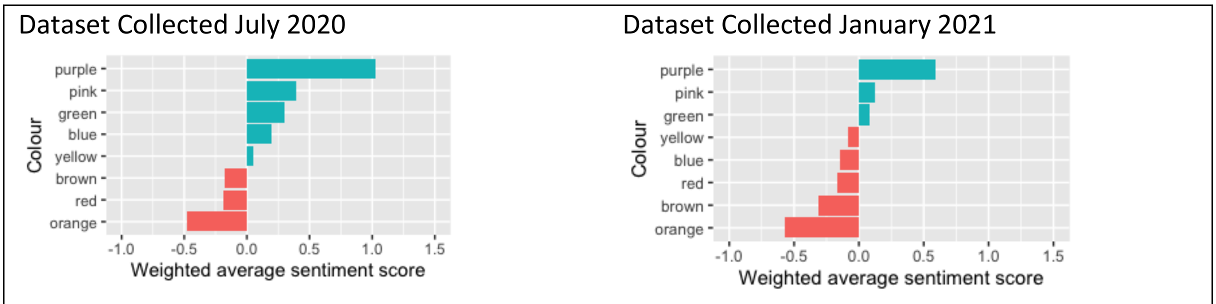

Figure 3: Mean sentiment scores for each of the colour names in 2021 and 2021.

We find that Purple is the most positively mentioned colour in both January 2021 and July 2020. Purple, pink and green are the colours that are positive in weighted average sentiment scores in both datasets. Blue and yellow are positive in the July dataset while negative in the January dataset. Red, brown and orange are negative in both datasets.

We find that, in both datasets, purple is the only colour where the top ten frequently used words are all positive, as illustrated in Figure 1 and Figure 2. Pink and green have more positive words than negative in both datasets, with significantly more positive words in July dataset. Besides, blue and yellow have more positive words in July dataset, while in January dataset both colours have half of the words positive.

Figure 1 and Figure 2 show that the most frequently used words in the mentions of every specific colour are likely to be highly coherent between July 2020 and January 2021. In order to understand how coherent the words used in the posts are between the two datasets, we derive the word stem of every word used in the posts. Within the group of posts mentioning one colour, we count the number of matches between the word stems in January dataset and the July dataset. We find that more than half of the word stems in January 2021 dataset find a match in July 2020 dataset. This is true for all colours in our datasets. It indicates that, when mentioning a colour, the twitter users used more than half of the words that are identical between the two observation points. In terms of the words that carry sentiment meanings, the match rate is even higher than 75%. This finding indicates that there is a strong coherence in the sentiment associated with the mentions of colour between these different observation time points.

DISCUSSION

Our study contributes to the colour research area by uncovering the coherence of sentiment associated with eight chromatic colour names in Twitter over a six moth scope. To the best of our knowledge, this study is one of the pioneering effort to explore the sentiment coherence of colour terms in social media using natural language processing methods. The coherence that we observed suggests that it is possible to develop tools to automatically monitor and predict the sentiment associated with colour terms in social media.

Our study contributes to marketing communication by suggesting that marketers, before using a colour term to feature grammatical marketing communications on social media, can analysis the sentiment associated with this colour term on this social media platform in order to make more cautious managerial decisions on the marketing communication.

There are several limitations in our study. First, this study is only based on the data collected from Twitter. According to the past studies in consumer sentiment analysis, brand sentiment depends on the social media platforms. We assume that colour sentiment metric can also depend on the platform. Further research can be developed on a more comprehensive analysis of data collected from multiple social media platforms.

Second, we focus only on the grammatical mention of colour names in text. Since processing text is considered as a higher construal level, compared with processing pictures, in consumers' information process, we assume the colour sentiment metric can be different in pictures. Further studies can explore the colour sentiment associated with the colour featured photos on social media.

Third, our study is not limited to the colour sentiment in any specific country or specific industry. Past studies suggest that the meaning of colours depends on country, culture, industry and more. Future studies could develop a more comprehensive analysis on colour sentiment in different markets and different industries.

REFERENCES

Bohle, R. H., and M. R. Garcia. 1986. *Reader Reactions to Color in Newspapers*.

Fernandez, K. V., and D. L. Rosen. 2000. The effectiveness of information and color in yellow pages advertising. *Journal of Advertising* 29(2): 61-73.

Garcia, M. R., and R. Bohle. 1986. Poynter research findings. *Color in American newspapers*: 32-42.

Han, D., A. Duhachek, and N. Agrawal. 2014. Emotions shape decisions through construal level: The case of guilt and shame. *Journal of Consumer Research* 41(4): 1047-1064.

Kearney, M. W. 2019. rtweet: Collecting and analyzing Twitter data. *Journal of open source software* 4(42): 1829.

Lee, H., X. Deng, H. R. Unnava, and K. Fujita. 2014. Monochrome forests and colorful trees: The effect of black-and-white versus color imagery on construal level. *Journal of Consumer Research*, *41*(4), 1015-1032.