The Effect of Hashtag Usage on Retweet Probability in Dutch Tweets

Niek Biesterbos
s4744497

Abstract

In a world where a big chunk of all communication is digital, it is important to understand how certain factors influence the reach of a message or post. This study examines the impact of hashtag usage on the likelihood of receiving retweets in Dutch tweets. Using the Dutch Tweet Corpus from Rijksuniversiteit Groningen, we explore whether a higher number of hashtags in a tweet increases its probability of the tweet being retweeted. We found that 1, 4 or 6 hashtags is the optimal amount of hashtag usage in a tweet to achieve the highest proportion of receiving at least one retweet.

1 Introduction

In the rapidly evolving world of social media, retweets serve as a key indicator of the reach and impact of a tweet. With how commonly hashtags are used on platforms like Twitter, it is essential to explore whether hashtags affect the likelihood of a tweet being retweeted and in what manner. How many retweets a tweet receives indicates both its popularity and its reach. This study hypothesizes that a higher number of hashtags in a tweet increases its chances of being retweeted.

2 Related Work

The work by Suh et al. explores the influence of various features, such as URLs and hashtags, on the retweetability of tweets. Their research, similar to my project, aims to understand the mechanisms behind information diffusion on Twitter, but it goes beyond by establishing a predictive model for proportion of retweets. Their conclusions, particularly regarding the strong relationship between hashtags and retweetability, strengthen the hypothesis of my project, namely, that hashtags significantly influence the frequency of retweets.

What's really interesting about their paper is that they did not find any connection between how many tweets someone has sent before and whether their new tweet will get retweeted (Suh et al., 2010). This means that it doesn't matter how much a user has tweeted in the past. The most important factor is the content of the current tweet.

In conclusion, the study is a comprehensive examination of the factors affecting retweetability (Suh et al., 2010). While their work is broader in scope, the findings related to hashtag usage are directly relevant and provide a good backbone for my project. The inconsistency also present opportunities for further exploration, especially concerning the role of a user's tweeting history in predicting retweet frequency.

3 Data

This research utilizes the Dutch Tweet Corpus provided by the Rijksuniversiteit Groningen, covering a span from 2012 to 2016. The dataset was selectively trimmed to include tweets from every 1st, 7th, 14th, and 21st day of each month, captured at 12:00 each day. We tried to collect as much data as possible in a random way, while maintaining reproducible. The key features examined in each tweet were hashtags and retweet identifiers. These fields are selected by the tweet2tab tool. In the python code itself we extracted the features using regular expressions and counted them.

The independent variable is the number of hashtags used in a tweet, and the dependent variable is the likelihood (measured as proportion) of the tweet being retweeted.

The tweet2tab tool is a simple way to get information from a lot of Twitter data. It's hosted on the Karora RUG server and it offers a lot of usecases. A user can use it by typing a command in its folder. For example, it can find user names or the text of tweets from a certain date and time. This is done by connecting the scraped Tweets to the tweet2tab tool and then asking it to execute some sort of manipulation on it. For instance, you can get all user names from a specific date and time. This tool helps in studying and understanding data the tweets the University of Groningen fetched over the years.

Pre-processing Since the research is primarily concerned with numerical and Boolean data, the data pre-processing was straightforward: each line (representing a tweet) was split into separate words, facilitating the counting of hashtags and the determination of the presence of a retweet-id. Table 1 shows how the data was structured after the pre-processing.

4 Analysis and Results

The study aimed at determining the correlation between hashtag usage and retweet frequency. It was found that tweets without hashtags were retweeted about 15.06% of the time, while tweets with hashtags were retweeted about 33.73% of the time. While the proportion of retweets is not the same as the probability of a tweet receiving a retweet, the proportion does provide an indication of how the probability could look like.

On examining the proportion of retweets across different numbers of hashtags used, it was found that there seemed to be multiple optimal numbers of hashtags that maximized the chance of a tweet being retweeted. However, due to the low number of tweets with more than 12 hashtags, results beyond this point were not seen as representative.

Results As shown in table 2, the difference in proportion of retweets between 0 retweets and 1 is quite high. Adding one hashtag to a tweet in comparison with no hashtags increased the percentage of retweets by 134.5%. Although this is not 1:1 the same as probability, it does show that adding hashtags to a tweet increased the rate of retweets.
Up until the representative amount of hashtags (12) each amount of hashtags has a higher proportion of retweets compared to no hashtags. We deemed results with 13 hashtags or higher not representative as the amount of tweets is not significant enough. This statement is supported by the total proportion of retweets with hashtags which is 0.3373401061752279 compared to the total proportion of retweets without hashtags which equals 0.150623704801594.

Although some differences in proportion are small, the most optimal amounts of hashtags are 1, 4 and 6 with respective proportions of 0.353204, 0.347614 and 0.351081.

5 Conclusion

This study aimed at understanding the correlation between hashtag usage and the likelihood of receiving at least one retweet in Dutch tweets. The findings moderately connect with the hypothesis that a higher amount of hashtags improves the retweet probability as they suggest an existence multiple optimal numbers of hashtags to acquire the highest chance on receiving at least one retweet. Although the study does not account for the actual content of tweets or the number of retweets, the insights provide a stepping stone for further exploration in this area.

In further work we could look into the actual content of a tweet, the author and the amount of retweets to further understand how much influence the usage of hashtags has on the proportion of retweets.

One thing to also keep in mind is that hashtag content and relevancy is not included in this research. Adding a layer of research to the content and relevancy of hashtags could further improve the correct rate of predicting if a tweet receives a retweet.

References

Suh, B., L. Hong, P. Pirolli, and E. H. Chi (2010). Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *2010 IEEE second international conference on social computing*, pp. 177–184. IEEE.

Tweet ID	Hastags	Retweet ID
16939...19488	#toekomst	16938...62240
16939...75617	#HEMA #fijnewinkel	
16939...31360	#valentijnvriendinnetjes	
16939...90785		16938...70560
16939...35584	#goedbezig	
16939...92352		16938...64096

Table 1: The Initial Data Obtained from the Tweet2tab Tool

Amount of Hashtags	Amount of Tweets	Proportion of Retweets
0	1.68444e+07	0.150624
1	2.51065e+06	0.353204
2	940210	0.308414
3	355241	0.297373
4	111086	0.347614
5	44988	0.343558
6	17805	0.351081
7	8466	0.338531
8	4084	0.346719
9	2301	0.326380
10	1239	0.317191
11	778	0.223650
12	383	0.240209
13	177	0.265537
14	159	0.446541
15	54	0.444444
16	57	0.578947
17	24	0.416667
18	7	0.428571
19	7	0.428571
20	2	0.000000
21	1	0.000000
22	2	0.000000
24	4	1.000000
28	1	0.000000
35	1	0.000000

Table 2: Proportion of Retweets for Different Number of Hashtags