Engineering the Loudest #Tweet

Andrew Zhen

Weilun Yao

Rongxin Zhu

Hypothesis

Does a tweet become viral because of the number of hashtags it contains? Are people more likely to retweet something that has more characters or less? For this research report, we predict that tweets that have hashtags will attract more attention than those without them. But we also anticipate that once the number of hashtags exceeds a certain amount — for example, three — the popularity will not be as high. We hypothesize that even though hashtags allow a tweet to be more likely to be seen by users who are interested in the hashtag topic, tweets with too many hashtags may be overseen and become uninteresting. We also predict that tweets with lengths ranging from 100 to 200 characters will get retweeted more often because tweets with too few characters may not contain enough information to be engaging and too many characters may be too long for users to engage with.

Introduction

Twitter is a social network where users post and interact with messages, called a "tweet." Today, Twitter is one of the most popular social networks with 328 million active monthly users and proven to be the largest source of breaking news. Its popularity influences many fields such as in fashion, sports and even in politics. A phrase known as "trending topic" is often hashtagged after a tweet by the users who express their opinions about a trending event. These topics help other users from various locations to understand what is happening in the world and what people's views are. Sometimes, tweets from certain celebrities can influence or even alter the trending public opinion. As a result, the power of a 200-character tweet is hard to be ignored.

Our team is interested in finding out the patterns that popular tweets have. The popularity of a tweet, in our case, is mainly determined by its number of retweets as a proportion of the user's total followers, where a retweet is defined as a tweet shared by another user. The goal of our project is to figure out, despite its content, factors that make other users want to retweet more. We will focus on two major elements of the tweet: the number of hashtags and the number of characters. A hashtag is inserted before a particular word to categorize their tweet so it can show up on the newsfeed in a related topic. Using hashtags is one way people use to gain publicity for their tweets, so the number of hashtags should have a strong influence towards its popularity. Another thing that sets Twitter apart from other social media is its word limit. Before September of 2017, Twitter had a 140 character limit for tweets. But since the new update, that limit has been doubled, and we figured that one reason behind this change relates to Twitter's popular platform for opinion sharing.

We think that our project will have significant contributions because with the great influence that Twitter has on public opinions, many individuals, groups, and companies will want to take advantages of it to gain attention and fame for their benefits. And by following the patterns that our research will discover about popularity gain for tweets, these groups of people can increase their chances of sharing "successful" tweets.

Proposed Methods

We will be using a few fundamental data methods to conduct our research. These methods include data mining, statistical analysis and finally, data visualization. Because we will base a particular tweet's success off of its rate of being retweeted — which is determined by dividing the number of retweets by the total number of followers, we need to apply data mining to extract this information from a tweet along with its length. Furthermore, some additional information that we can extract from a tweet that may be useful is the time of day and day of the year and whether or not the user holds a verified account—meaning that an account of public interest is authentic, such as in the case of celebrities. Time and day can be insightful factors because statistically speaking; there should be an ideal time that a tweet receives the most exposure within a user's newsfeed, like say, Thursday afternoon in July. The date variable can also be used to determine if that time of year is when events are particularly interesting, such as during holidays, elections and for breaking news. Understanding this along with who is tweeting will help us narrow down tweets that will contribute to our research in determining what important factors are at play.

Our first step is to accumulate the necessary sample of tweets that will help determine our research conclusion. We can approach this process by creating a script that scrapes Twitter for tweets that have been retweeted many times and gather the necessary information. One area of Twitter we can target is the "trending now" section where tweets are more likely gain traction. Another approach is to identify a few popular hashtags and select a handful from all categories to use as our datasets. Moreover, we can also accomplish the task of collecting our data using existing Twitter datasets available online. Next, after we have properly acquired the data in the correct formats and placed it in an organized structure, we may proceed to visualize the data. Since we hypothesized that tweets with around 100 to 200 characters in length would have the most success in being retweeted, we expect to see a certain shape to our visualization graphs. By plotting the length of a tweet against the number of retweets on a scatter plot chart, we hope to notice a normal distribution, or otherwise known as a bell-shaped curve. With the curve peaking slightly past the 100 character mark and with either side sloping downwards, we will have identified a correlation between the two variables. Keeping these results in mind, we will proceed to thoroughly plot the rest of the variables against each other to spot any anomalies and to further examine the relationship between these attributes. For example, it is possible that a tweet may have received plenty of attention in the form of comments and likes but simply was not retweeted enough for us to categorize it as “popular.” Therefore, we will make sure to account for this in our process of statistical analysis and trying to generalize our findings.

Discussion

We chose to do retweets as a proportion of the user’s number of followers instead of merely the number of retweets because the twitter account has more than 10,000 followers is more likely to have a tweet got retweeted 100 times than the one with only 100 followers. But in the 100-follower account case, almost 100% of the followers retweet while the 10,000-follower account just has 1%. Therefore we think the proportion of the user’s number of followers is more appropriate in determining the popularity of the tweet.

Since nearly 80% of Twitter accounts are outside of the United States, many of the tweets are in languages other than English. The limitation here is that we only study the tweets in English, so the result can only be applied to the English tweets. But we still expect to see the similar result in other languages because we think people across the world should have a similar attitude toward the length of the tweet and the number of hashtags in it.

Also, one could argue that the result of a tweet’s popularity is due to its content instead of its length.We would not be able to collect contents of tweets as data because there are a lot of elements involved, such as tone, which has different keywords for different situations and events.