Exploring the Use of Verbs in Viral Hashtags

Project Summary

Hashtag activism has become a popular way for groups and individuals to rally support around a cause. Activists and social scientists are both interested in determining what kinds of hashtags encourage engagement and commitment. Two recent activist hashtags which enjoyed great popularity are #metoo and #takeaknee, but each has a very different voice (passive acknowledgement versus direct command, respectively). This analysis strove to determine if active voice in a hashtag engendered active voice (more engagement) in the overall tweet. The methodology for determining active voice was the mean use of verbs per tweet. However, one of the analyzed hashtags contains an implicit verb, while the other does not. Tweets were collected based on the hashtags, and the analysis was performed both with the hashtag intact ("metoo," "takeaknee") and broken into its component words ("me too", "take a knee").

Data Collection

In October and December, 2017, tweets were collected for each hashtag using Python's Tweepy package and the Twitter REST API. Tweets were filtered before collection to remove replies, media, and retweets and include only English language tweets.

Data Prep

The TextBlob package in Python (an extension of the Natural Language Tool Kit) was used to identify the parts of speech and count each verb (Table 1 and Figure 1), both with and without the hashtag split. In most, but not all, cases, breaking #takeaknee into "take a knee" resulted in an increase in verb counts. Tweets were written to CSV files for further cleaning and analysis in R. In R, the tweets were cleaned by eliminating duplicate authors and tweets, and removing non-native American English speakers. Five thousand of the remaining tweets were randomly selected for analysis.

Analysis

When comparing two independent quantitative variables, we first look to a t-test. However, a t-test generally assumes a normal distribution. A Wilcoxon Rank Sum test is a similar test that can be used when data is not normally distributed. Wild (1997) states that, "In a practical situation in which we are uneasy about the applicability of two-sample t methods, we use both them and the Wilcoxon and feel happiest when both lead to very similar conclusions."

The data in this case was heavily skewed, so both were performed. We have no prior evidence that suggests that either hashtag will have more verbs. Therefore, the null hypothesis for the Wilcoxon test is that the two hashtags have the same distribution, and the alternative is that they do not. The null hypothesis for the t-test is that the 2 tweet groups will have the same median number of verbs, and the alternative is that they will not.

Conclusion

When the hashtags are broken into individual words, there is evidence that both the shape of the distribution ($p\text{-value} < 2.2e\text{-}16$) (Figure 3) and mean number of verbs per tweet ($p\text{-value} < 2.2e\text{-}16$) varies significantly between the two populations. However, when the hashtags are left as is (which results in them primarily be treated as nouns), there is no significant evidence that either the shape of the distribution ($p\text{-value} = .05944$) (Figure 2) nor the mean number of tweets ($p\text{-value} = .04758$) varies significantly between the two populations. These results exemplify the care which must be taken when determining the appropriate method for analysis. Contradictory results can be generated by altering just a single component of the data. Ultimately, while there may be a statistically significant result, there is no practical significance, and no conclusions regarding a relationship between active voice hashtags and active voice tweets should be drawn.

Figures and Tables

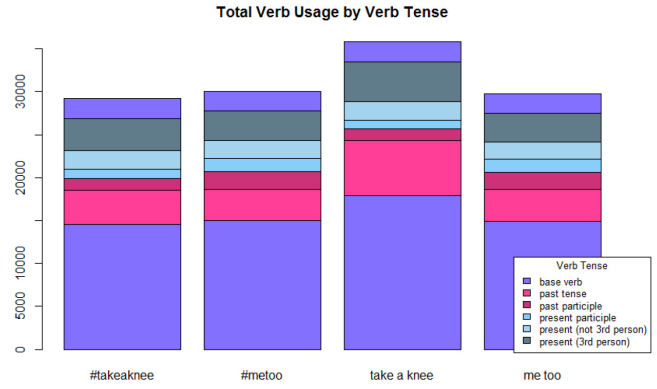

Figure 1 – A stacked bar chart shows that #takeaknee and #metoo tweets have a very similar number of verbs. But, that "take a knee" tweets have significantly more verbs than "me too" tweets. Careful text analysis of tweets would be required to determine when the hashtag #takeaknee was being used as a part of speech, and when it was being used as simply a hashtag. That analysis was not completed for this project.

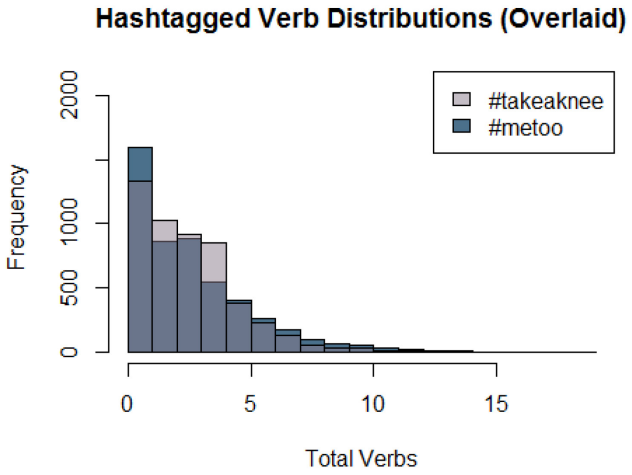

Figure 2 - Overlaying the distribution of #takeaknee tweets on #metoo tweets allows the viewer to visualize how the shape of the distribution varies. While both datasets had the same minimum (0), there was a difference in the maximum number of verbs. Visually, there appears to be more tweets in the 2-3 range for #takeaknee, but test results indicate the difference is not significant.

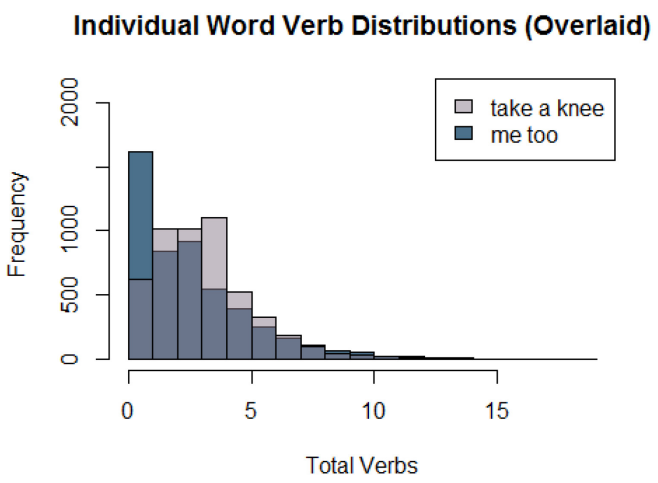

Figure 3 – Comparing Figure 2 to Figure 3, you can see that including the component parts of the hashtag drastically reduced the zero verb tweets for "take a knee," causing a shift in the distribution. Note that "take" can also be treated as a noun by a speech analyzer, depending on the preceding word. For example, "the take a knee campaign" would result in "take" being considered a noun. In rare instances, splitting #metoo into "me too" also resulted in an increase in verbs. The example in Table 1 illustrates an instance where this occurred. The misspelled word "occurances" was treated as a noun when #metoo was a hashtag, and treated as a verb when it was not.

	Sample Tweets, before and after splitting the hashtag			
Parts of Speech	"i allowed metoo occurances because of the teachings of the lord"	"i allowed me too occurances because of the teachings of the lord"	"takeaknee just stand up"	"take a knee just stand up"
Total Verbs	1	2	1	2
Base Verb	0	0	1	2
Past Tense	1	1	0	0
Past Participle	0	0	0	0
Present Participle	0	0	0	0
Present (not 3rd person)	0	1	0	0
Present (3rd person)	0	0	0	0

Table 1: Sample tweets from the dataset, with verb counts.

References

Docs - Twitter Developers. (n.d.). Retrieved October 15, 2017, from https://developer.twitter.com/en/docs

Ford, C. (2017). The Wilcoxon Rank Sum Test. Retrieved October 25, 2017, from http://data.library.virginia.edu/the-wilcoxon-rank-sum-test/

Natural Language Toolkit. (n.d.). Retrieved October 20, 2017, from http://www.nltk.org/

TextBlob: Simplified Text Processing. (n.d.). Retrieved October 20, 2017, from https://textblob.readthedocs.io/en/dev/index.html

Tweepy Documentation. (n.d.). Retrieved October 15, 2017, from http://docs.tweepy.org/en/v3.5.0/index.html

Wild, C. (1997). The Wilcoxon Rank-Sum Test. Retrieved October 25, 2017, from https://www.stat.auckland.ac.nz/~wild/ChanceEnc/Ch10.wilcoxon.pdf