**All-Star, Influencers, or Others?**

Yuwei (Johnny) Meng & Yiteng (Ethan) Zhang

2023 Winter – JSC270S

Professor Carolina Nobre

7 Apr 2023

**Introduction & Literature Review**

After the first programmable digital computer, named ENIAC, was built in 1945, humans entered the era of information. In the next merely 80 years, digital computers and cell phones started pervading everyone’s life, while social media overthrew face-to-face interaction and became the main platform of socialization. Among these, Twitter was found in 2006 and soon evolved into one of the most popular social media across the globe. On Twitter, a very important feature is that users can “follow” another. Nowadays, people ask, why would users with many followers, or “social influencers,” have followers? Do they say different things than us? The present study aims to address these curiosities by proposing the research question: *Can we predict whether the author of a tweet is a social influencer based on the tweet they send?*

Previous studies have attempted to address the popularity of social influencers by various quantities. For example, Anger & Kittl, researchers at a technology company, measured social influence on Twitter by retweets, followers, and mentions. They claimed that Twitter is “rather content-oriented with networks evolving around topics and subjects” (Anger & Kittl, 2011). This confirmed that there is some correlation between the number of followers and the content of tweets, which supported the importance of our research question. Furthermore, a group of researchers from the University of Michigan and Yahoo analyzed several measurements of URLs in Twitter posts. They built a predictive model to quantify social influence based on these measurements (Bakshy et al., 2011). This paper corroborated the previous one that the content of tweets is integral to social influence. To further verify this result, the current study takes another approach and aims to shed light onto this research topic by finding some correlation between Twitter posts and social influence, measured by the number of followers, using NLP.

**Methodology**

The dataset used in this study was extracted from Twitter using a Twitter API by Cait. It consists of 10,000 observations and 8 columns, including the two variables of interest: The original content of the tweets and the number of followers of the users. To fit the NLP model, the data will be explored and preprocessed first. Underlying features in the tweets will be extracted. Multiple models will be constructed for comparison. The analysis will be carried out using Python and Jupyter notebook, with use of Python libraries including pandas, numpy, and spacy.

**Exploratory Data Analysis & Preprocessing**

The data were taken to plot histograms and compute summary statistics. For the number of followers in particular, the distribution is strongly right-skewed, which means that most users in the dataset have very few followers; only a few users are social influencers. The summary statistics confirm that the median number of followers is 1,348 and the maximum number of followers is 755,203.

The data were subsequently preprocessed. Since it is impossible to predict the exact number of followers of a user because it is constantly changing, the users were classified by the number of followers into “Few”, “Medium”, and “Many” groups. The classification is based on the standards proposed by Smith (2022) that users with fewer than 1,000 followers are regular users; users with more than 1,000 and fewer than 10,000 followers are nano influencers who are highly engaged in social platforms but still have a regular lifestyle; users with more than 10,000 followers are famous influencers who specialize in at least one field (Smith, 2022). After the classification, a bar chart showed that there are as many users with “Few” followers as users with “Medium” followers in the dataset, but users with “Many” followers are significantly fewer. This might be a limitation of our model and will be addressed later in this paper. Moreover, probably due to API issues, most tweets in the dataset were truncated somewhere. This might be another limitation since we actually had incomplete data to work with.

Next, the tweets were preprocessed first by removing the link to the original post at the end of each tweet. Then, the tweets were lemmatized and tagged with spacy. The following analysis will be based on these lemmas and tags of the words.

**Feature Extraction & Model Building**

21 features were selected for extraction. Most of them were counts, such as the number of hashtags, the number of personal pronouns, and the number of proper nouns. We believe that these features would measure the speaking habits of each user. In addition, we included the Bristol norms and the Warriner norms as measurements for reference. These norms are based on research and measure the underlying characteristics of every lemma, such as the age of acquisition, familiarity, arousal, and dominance levels (Stadthagen-Gonzalez & Davis, 2006; Warriner et al., 2013). For each tweet, we computed the mean and standard deviation for each of these measurements.

The model was subsequently fit on the count features and the norms to predict the follower class that a user belongs to, which is a supervised model. Several models, including a random forest, a support vector machine (SVM), and an AdaBoost classifier, have been built and evaluated. Among these, the random forest had the highest training accuracy of 0.997, so this was chosen as the final model of this study, since this model nearly got every prediction correct on the training set.

We would like to provide some basic information about random forests and why it was a suitable choice in this case. A random forest fits multiple decision trees on subsamples of the data and takes the average parameters across all decision trees. As a result, a random forest would have a higher accuracy than a decision tree and work better on imbalanced data. The drawback is that a random forest takes longer to fit because it is an iterative process. In our case, a random forest fit well because our data were imbalanced because there were fewer users with “Many” followers. Our random forest could also go deeply into splitting nodes according to various features and measurements that we had.

**Evaluation & Conclusion**

The random forest was evaluated on the testing set and resulted in an accuracy of 0.5025. This accuracy was not nearly as good as the training accuracy, but it was acceptable because a random guess would only have an accuracy of 0.33. Additionally, the dramatic decrease in accuracy showed that overfitting was a problem. This was expected since decision trees were notorious in overfitting. Even though random forests take averages across parameters, they are essentially still composed of decision trees.

However, this model still provided insights into the research question: We can predict the follower class that a user belongs to based on the tweets they post. In other words, there is some correlation between the tweets and the number of followers of Twitter users. This conclusion was in alignment with previous literatures. More specifically, this result confirmed with Anger & Kittl (2011) that the number of followers is content-oriented. Namely, followers usually decide whether to follow a social influencer based on the tweets that the social influencer posts. This NLP model also built upon the research conducted by Bakshy et al. (2011) because this model evaluated the entire post rather than just analyzing URLs.

There are some limitations to this NLP model though. As mentioned previously, the tweets in the given dataset were mostly truncated and the data were incomplete. This manifested that the model actually learned less from the dataset than expected. This limitation was particularly seen in the feature of the number of hashtags. Since users usually put hashtags at the end of their posts, truncated tweets lose all these information that were otherwise available. This might result in inaccurate prediction. Furthermore, the number of users in each labeled class was not nearly equal. The consequence was that according to the confusion matrix, the model rarely predicts a user in the testing set to have “Many” followers because of the fewer users in the “Many” group in the training set, which makes the predictions error-prone.

It is recommended for future researchers to gather a complete dataset for analysis to eliminate, or at least mitigate, the aforementioned limitations. In addition, future researchers should consider NLP more as their means of data analysis because, as mentioned previously, Twitter is overall content-oriented.

In conclusion, despite the limitations, there is some correlation between the number of followers of users and the content of their tweets.

**References**

Anger, I., & Kittl, C. (2011). Measuring influence on Twitter. *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, 1–4. https://doi.org/10.1145/2024288.2024326

Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone’s an influencer: Quantifying influence on twitter. *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, 65–74. https://doi.org/10.1145/1935826.1935845

Smith, R. (2022, March 1). 19 types of influencers based on follower count and niches they serve. *Affable.Ai*. https://www.affable.ai/blog/types-of-influencers#:~:text=There%20are%20five%20different%20types%20of%20influencers%20based,follower%20count%3A%20all-star%2C%20mega%2C%20macro%2C%20micro%2C%20and%20nano.

Stadthagen-Gonzalez, H., & Davis, C. J. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, *38*(4), 598–605. https://doi.org/10.3758/BF03193891

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x