What makes a rumor popular? A case study of Cryptocurrency Rumors on Reddit and Twitter

SBP-BRiMS 2018 Data Challenge

Dian Hu, Advisor: David A. Broniatowski
Department of Engineering Management & Systems Engineering
The George Washington University

Abstract

Introduction:
Fuzzy-Trace Theory (FTT), a leading theory of decision making, predicted that a message with causal explanation of otherwise mysterious adverse events would be more popular on the internet (Reyna, 2008). The landscape model, another related theory, argued that people will more easily comprehend and recall a message with a clear causal structure statement (Van den Broek, 1990). Based on landscape model, we infer that such a message will be also more popular on the internet. These prior works have hinted that the popularity of a message might also be related to the internal causal structures of that message.

Goal:
In this data challenge project, we are interested in how the internal causal structure of rumor would affect its popularity. In particular, we want to understand which internal factors, if any, are predictive of how often the message is upvoted/retweet on Reddit/Twitter.

Methods:
1. Our approach is to use **regression analysis** to explore which independent variables, are predictive of the number of Upvotes/Retweet of each messages. We use Bonferroni Correction (Dunnett, 1955) to control the multiple comparison issue.
2. **Dependent Variables**: In the multiple regression analysis, we ignore negative upvote and transform the number of upvote/retweet using logarithmic functions.
3. We explored four sets of **Independent Variables, 205 in total**:
    I. The proportional weight of each topic in each message, based on Latent Dirichlet Allocation Model.
    II. Linguistic features include number of character, number of lexicon, number of sentences, Flesch Readability, Dale and Chall Readability, Coleman and Liau Index and their normalized or transformed versions. to reduce the dimensions, we decomposed all 19 features into three dimensions using a principal component analysis. Based on the PCA loadings, we conclude that the first component corresponds to the length of the message, the second component corresponds to comprehensibility and the third component corresponds to verbatim features.
    III. The interaction terms.
    IV. Whether the message contains a hyperlink or at least on digit.

Data Explored:
1. In the first dataset, we collected 100000 Reddit messages randomly sampled from June 1st, 2017 to Nov 1st, 2017 with at least one of these 5 keywords "bitcoin," "xrp," "ethereum," "litecoin," "cryptocurrency".
2. In the second dataset, we use one public data collection from Kaggle, we randomly selected 100000 bitcoin-related tweets in the year of 2017 from this tweet source.
3. In the third dataset, we collected another 100000 Reddit messages randomly sampled from June 1st, 2017 to Nov 1st, 2017 with we collected 100000 Reddit messages randomly sampled from June 1st, 2017 to Nov 1st, 2017 featuring several popular conspiracy theories themes. Those conspiracy themes keywords include: "smartmeter and cancer", "Monsanto and GMO", "flat earth", "lizard man", "reptoid" and their lexical variations. We collected the third database for another project, but we use it in this paper as a baseline of Reddit rumors.
4. In summary, we have 3 dataset, each with 100000 messages in this analysis.

Results from 100000 Bitcoin Reddit messages

- In this pilot study, after applying Bonferroni Correction, we found 8 features to be significant based on the multiple regression analysis.
- Consistent with many studies presented in this conference, we found that a message with hyperlink have higher likelihood to be upvoted.
- We found Reddit Post moderated as "archive post" have lower likelihood to be upvoted. This makes sense because this type of post will no longer allow upvote. But this side finding is not very interesting to our goal.
- We found that the interaction term between PCA component 1 (comprehensibility) and several topics are significant. But the other two components are not presented. It's interesting that comprehensibility plays a big role in this dataset. It's also interesting to see that bitcoin topics are better off with high comprehensibility language in terms of upvote.
- We found expression with strong emotion (profane) have higher likelihood to be upvoted.

	coef	std err	z	P>|z|	[0.025	0.975]	Annotation
hyperlink_whether	3.8242	0.504	7.589	0	2.836	4.812	whether the message has at least one hyperlink
pca_1*topic_0	1.3	0.295	4.413	0	0.723	1.877	Interaction between comprehensibility and bitcoin development
pca_1*topic_33	2.8307	0.692	4.091	0	1.474	4.187	Interaction between comprehensibility and bitcoin historical price
pca_1*topic_31	2.2521	0.56	4.021	0	1.154	3.35	Interaction between comprehensibility and profane expression
Topic_39	15.3377	4.076	3.763	0	7.349	23.327	Topic: sharing links
pca_1*topic_25	1.9104	0.526	3.634	0	0.88	2.941	Interaction between comprehensibility and bitcoin transaction
pca_1*topic_6	-2.5003	0.681	-3.674	0	-3.834	-1.166	Interaction between comprehensibility and archive post
Topic_6	-4.326	0.889	-4.868	0	-6.068	-2.584	Topic: archive post

Results from 100000 Reddit Conspiracy messages

- In this data, after applying Bonferroni Correction, we found 21 features to be significant based on the multiple regression analysis.
- Once again, we found that a message with hyperlink have higher likelihood to be upvoted. We also found expression with strong emotion (profane) have higher likelihood to be upvoted.
- We found that Illuminati, Flat-earth, "Trump and Conspiracy", are among the most popular topics in conspiracy messages.
- We found that the interaction term PCA components and topic model has very small coefficient, which indicates that in conspiracy data, people pay more attention to the topic itself and care less of the semantic features.
- We found that having digit, having internal Reddit link, and sub-forum rules have less likelihood to be upvoted.

	coef	std err	z	P>|z|	[0.025	0.975]	Annotation
digit_whether	-0.1164	0.0099	-11.7353	0	-0.1358	-0.0969	whether the message has at least one digit
46	-0.2674	0.0278	-9.6286	0	-0.3218	-0.213	Topic: Reddit Internal Link
27	-0.7131	0.096	-7.4261	0	-0.9013	-0.5249	Topic: Stopwords
9	-0.3927	0.056	-7.0092	0	-0.5025	-0.2829	Topic: Reddit Subforum Rules
pca_2*topic_46	-0.0021	0.0003	-6.1684	0	-0.0027	-0.0014	Interaction between message length and Reddit Internal Link
pca_0*topic_46	-0.0003	1.00E-04	-5.8155	0	-0.0004	-0.0002	Interaction between message verbatim and Reddit Internal Link
10	-0.3545	0.0664	-5.3422	0	-0.4845	-0.2244	Topic: Evidence and Proof
12	-0.2508	0.0549	-4.5719	0	-0.3583	-0.1433	Topic: Astrology
pca_1*topic_19	0.0042	0.001	4.2288	0	0.0022	0.0061	Interaction between comprehensibility and Flat-earth
pca_0*topic_21	0.0009	0.0002	4.3961	0	0.0005	0.0013	Interaction between length and Flat-earth
38	0.3116	0.0696	4.4749	0	0.1751	0.4482	Topic: Profanity message
30	0.347	0.0712	4.8742	0	0.2074	0.4865	Topic: "a friend", "a story", a typical narrative component
pca_1*topic_21	0.0061	0.0012	5.2545	0	0.0038	0.0083	Interaction between comprehensibility and Flat-earth
pca_1*topic_46	0.0005	1.00E-04	5.2632	0	0.0003	0.0007	Interaction between comprehensibility and Reddit Internal Link
42	0.3478	0.0643	5.4112	0	0.2218	0.4737	Trump related message
35	0.3769	0.0583	6.4588	0	0.2625	0.4912	Conspiracy in general
pca_2	0.0024	0.0004	6.5527	0	0.0017	0.0031	PCA component 2: verbatim
pca_0*topic_17	0.004	0.0006	6.963	0	0.0029	0.0052	Interaction between message length and Illuminati related message
17	1.7503	0.2444	7.1617	0	1.2713	2.2293	Illuminati related message
21	0.7381	0.0837	8.8173	0	0.574	0.9022	Flat-earth related message
hyperlink_whether	1.0343	0.03	34.4562	0	0.9755	1.0932	whether the message has at least one hyperlink

Results from Tweet Data

• During our analysis, we used a tweet collection from Kaggle. After applying the same methods used in the other two dataset, we didn't find particular interesting results. And I couldn't interpret the LDA results with enough confidence.

• This is probably a result of the pool quality of the tweet data I used in the first place. I suspect much data is generated by bots in this collection.

Limitation and Future Work

• Based one the finding of comprehensibility in dataset 1, I have more confidence that comprehensibility will affect the popularity of a message, which is consistent with prior theory from earlier works in reading comprehension (Van den Broek, 1990). But this preliminary finding need more rigorous testing in quasi-experiment settings.

• I should be more careful in the future when analyzing Twitter data by removing bot generated messages.

• During this data challenge, I didn't have enough evidence showing that having causal structure will positively affect popularity, which is hypothesized by related theories.

• Even within the same social media platform, each topic or keyword will attract many different communities and have different way of expression. I should be more careful of the limitation from keyword based approach.

References

1. Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. Journal of Applied Psychology, 60(2), 283.
2. Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. Educational research bulletin, 37–54.
3. Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association, 50(272), 1096–1121.
4. Flesch, R. (1948). A new readability yardstick. Journal of applied psychology, 32(3), 221.
5. Reyna, V. F. (2008). A theory of medical decision making and health: fuzzy trace theory. Medical Decision Making, 28(6), 850–865.
6. van den Broek, P. (1990). Causal inferences and the comprehension of narrative texts. Psychology of learning and motivation, 25, 175–196.