Fake News, Investor Attention, and Market Reaction

Jonathan Clarke, Hailiang Chen, Ding Du, and Yu Jeffrey Hu
Information Systems Research, forthcoming
2019 09

王念硕

Outline

1. Introduction
   - Background
   - Motivation
   - Research question
   - Related researches
   - Research contents
   - Contribution
   - Framework
2. Research design
   - Variable
   - Data
   - Method
3. Empirical result
   - Investor Attention
   - Commenter and Editor Reaction
   - Linguistic Characteristics
   - Market Reaction
4. Conclusion

1. Introduction

Background

1. Fake news has received considerable press attention in recent years. Facebook, Twitter, and other social media platforms have been criticized for promoting articles that later proved to be false. Based on anecdotal evidence, the reach of fake news stories is staggering.

2. Facebook estimated that fake news stories about the 2016 election reached over 126 million Americans (New York Times 2017) .

1. Introduction

Motivation

1. Although many scholars have studied the influence of fake news, the real impact of fake news in financial markets is less well understood.

1. Introduction

Research question

1. Is fake news more attractive to investors than legitimate news?

   Yes

2. Whether fake news can be seen from the linguistic features of the article?

   Yes

3. Is there any difference between fake news and legitimate news on stock trading and return?

   Yes

1. Introduction

Research Contents

1. First, fake news articles generate 83.4% more page views on average than legitimate news articles.

2. Second, fake news articles do not receive a significantly different number of comments compared with legitimate news articles.

3. Third, we implement several well-known machine learning algorithms based on linguistic characteristics for identifying fake news stories.

4. Finally, there is a significant increase in trading volume on the day a fake news article is released. The volume is less than that observed for legitimate articles.

1. Introduction

Related researches

1. Vosoughi, Roy, and Aral (2018) examine the spread of true and fake rumors on Twitter. Their main finding is that fake news, especially fake political news, diffuse significantly faster and farther than legitimate news articles.

2. Newman et al. (2003) find that liars communicate in qualitatively different ways than sincere story tellers.

3. Hu et al. (2012) show that a textual analysis of sentiment and readability can be used to identify fake product reviews.

4. Kim and Verrecchi (1991) develop a theoretical model of trading and show that precision of news impacts the trading volume response.

1. Introduction

Contribution

1. Our findings are important to social media platforms, regulators, and investors. Our machine learning approach based on linguistic characteristics provides a method for social media platforms and regulators to identify attempts to manipulate the market by releasing fake news stories.

1. Introduction

Framework

Fake news can get more attention.

Neither investors nor editors can tell the truth of the news.

The machine learning method can identify the true and false messages according to the language features.

The market reaction to fake news is lower than that to true news.

2. Research design: Variable

Length is the total number of words in an article.
%NegWord is the fraction of negative words in an article in percentage points..
Premium is an indicator variable denoting whether the author of the article receives monetary compensation from Seeking Alpha, which is determined mainly by the page views of the article.

Investor Attention：
page views (PVs)
the number of unique visitors (Uniques)
the number of times an article was read to the end (ReadToEnd)
the number of page views for which the comments were read (ReadComment)

2. Research design: Variable

Comment is the total number of comments of an article.
%Contradiction is defined as the absolute difference between the fraction of negative words and the average fraction of negative words.

EditorPick denotes whether the article is featured in the Editors’ Picks section.
ConvincingScore: whether the article can share pertinent information.
ActionableScore: whether the article provides new information.
WellPresentedScore: whether the article is easy-to-understand.
AggregateScore is a composite of these three scores.

2. Research design: Variable

We use standard event study methods to estimate the excess return for firm i on day t as:

$$ARet_{i,t} = Ret_{i,t} - Ret_{m,t}$$

We calculate the abnormal trading volume on day t:

$$AVol_{i,t} = \frac{Vol_{i,t}}{\overline{Vol}_{i,t-5,t-65}}$$

We are interested in both short-term and long-term market performances. [0, +1], [0, +2], [+3, +120], [+3, +242].

2. Research design: Data

Data Source： SEC、Seeking Alpha、Compustat、IBES、CRSP.

Period： 2011 August to 2014 March.

Sample: All New York Stock Exchange (NYSE), American Stock Exchange (Amex), and Nasdaq.

2. Research design: Method

Propensity Score Matching

Fixed effect regression

Gradient Boosting, Logistic Regression, Naive Bayes, Neural Network, Random Forest, Support Vector Machine

3.1 Empirical result: Propensity Score Matching

Table 2. Propensity Score Matching

VARIABLES	(1) *Fake* Before Matching	(2) *Fake* After Matching
Log(Size)	-0.818***	0.106
	(0.045)	(0.083)
Market-to-Book	-0.001**	0.000
	(0.000)	(0.000)
ROA	0.085**	0.053
	(0.038)	(0.061)
LEV	0.051**	0.015
	(0.026)	(0.072)
$ARet_{-30,-1}$	-0.212**	0.013
	(0.089)	(0.119)
Constant	3.366***	-0.742
	(0.579)	(0.818)
Sector FE	Yes	Yes
Quarter FE	Yes	Yes
Observations	14,518	496
Pseudo R^2	0.616	0.018
Log Likelihood	-592.8	-337.6

Notes: *, **, and *** indicate p<0.1, p<0.05, and p<0.01, respectively. Standard errors are presented in parentheses.

We conduct propensity score matching to find a matched group of legitimate news articles that cover a set of firms similar as those targeted by fake news articles.
the firms covered by the matched legitimate news articles share similar characteristics as those of fake news articles.

3.1 Empirical result: Investor Attention

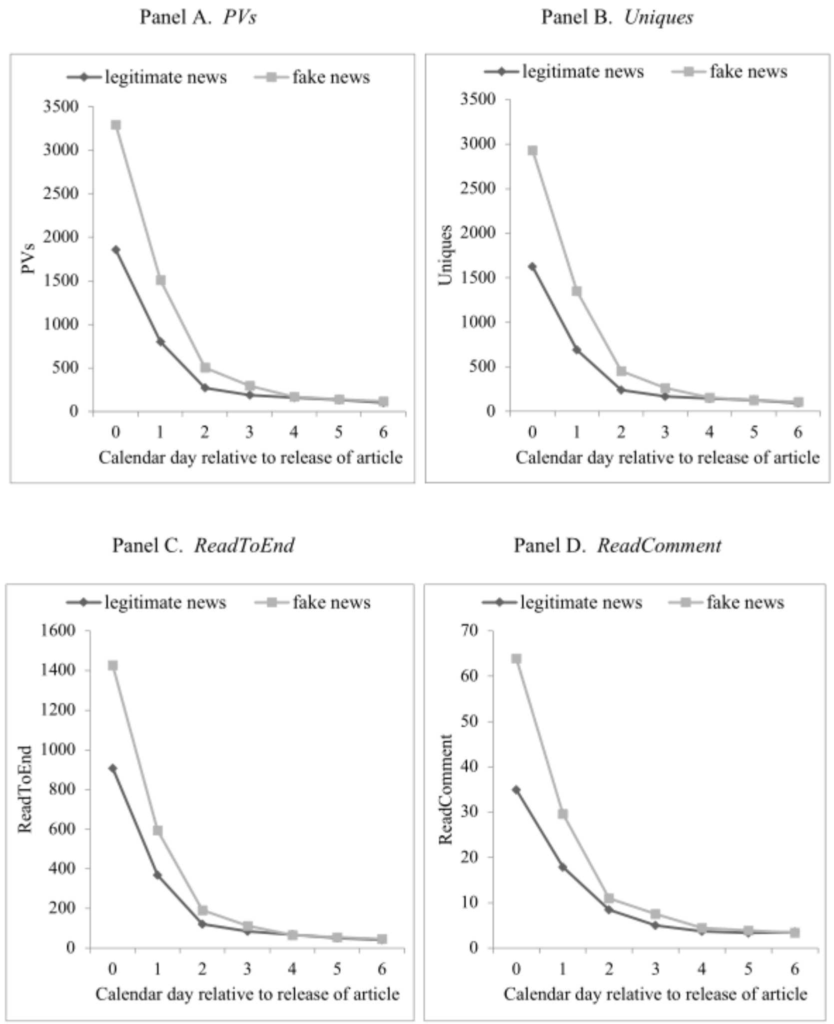

Figure 1. Investor Attention on Fake News

For all these four attention measures, fake news stories generate higher levels of investor attention than legitimate news articles.

3.1 Empirical result: Investor Attention

VARIABLES	(1) *Log(PVs)*	(2) *Log(Uniques)*	(3) *Log(ReadToEnd)*	(4) *Log(ReadComment)*
Fake	0.834***	0.855***	0.853***	-0.287
	(0.135)	(0.132)	(0.137)	(0.387)
Log(Length)	0.198*	0.196*	-0.082	0.333
	(0.118)	(0.119)	(0.125)	(0.315)
%NegWord	0.162*	0.152	0.154	-0.051
	(0.095)	(0.099)	(0.102)	(0.288)
Premium	-0.029	-0.021	0.003	-0.130
	(0.102)	(0.100)	(0.101)	(0.304)
EditorPick	0.021	-0.040	-0.090	-1.130***
	(0.172)	(0.178)	(0.193)	(0.383)
Log(Comment)	0.312***	0.301***	0.341***	0.434***
	(0.050)	(0.050)	(0.052)	(0.133)

For the first three measures of investor attention, we find that fake news articles generate more investor attention than legitimate news stories.

We find no statistically significant difference in the number of page views that also read article comments between fake news articles and legitimate news articles.

3.2 Empirical result: Commenter Reaction

Table 4. Commenter Reaction to Fake News

VARIABLES	(1) *Comment*	(2) *%Contradiction*
Fake	-3.621	-0.004
	(2.377)	(0.059)
Log(Length)	9.103**	-0.150**
	(4.248)	(0.064)
%NegWord	2.367*	0.113*
	(1.328)	(0.061)
Premium	5.249***	0.225***
	(1.579)	(0.058)
EditorPick	3.262	0.017
	(4.605)	(0.110)
Log(Comment)		-0.158***
		(0.035)
Log(CommentWord)		-0.098
		(0.062)
%NegFactiva	1.207	-0.004
	(1.697)	(0.082)
Factiva	2.650	0.096
	(1.688)	(0.079)

Fake news articles do not generate more comments or more disagreement than legitimate news articles.

3.2 Empirical result: Editor Reaction

Table 5. Editor Reaction to Fake News

VARIABLES	(1) *EditorPick*	(2) *Convincing Score*	(3) *Actionable Score*	(4) *WellPresented Score*	(5) *Aggregate Score*
Fake	-0.081***	-0.172***	-0.165***	-0.102*	-0.180***
	(0.030)	(0.044)	(0.052)	(0.055)	(0.050)
Log(Length)	0.168***	0.296***	0.219***	0.205***	0.311***
	(0.033)	(0.046)	(0.049)	(0.055)	(0.049)
%NegWord	0.032	0.059**	-0.031	-0.014	0.037
	(0.021)	(0.029)	(0.035)	(0.036)	(0.032)
Premium	-0.018	-0.035	-0.008	0.019	-0.033
	(0.026)	(0.041)	(0.046)	(0.051)	(0.047)
Log(Comment)	-0.002	-0.007	0.025	-0.013	-0.014
	(0.014)	(0.022)	(0.022)	(0.026)	(0.023)
%NegFactiva	-0.031	0.024	-0.094**	-0.056	0.012
	(0.021)	(0.047)	(0.041)	(0.043)	(0.052)
Factiva	0.004	0.001	0.082	0.009	0.028
	(0.027)	(0.046)	(0.056)	(0.058)	(0.054)
PosFA	-0.155***	-0.165*	-0.198*	-0.210*	-0.243**

Neither editors nor commenters can reliably detect fake news.
While the coefficient on the Fake indicator is also negative and statistically significant at the 1% level, the magnitude of the result is again not economically meaningful.

3.3 Empirical result: Linguistic Characteristics

Table 6. Linguistic Characteristics of Fake News

	Fake news (N=248)		Legitimate news (N=248)		t-test	
LIWC variables	mean	stdev	mean	stdev	t-stat	p-value
Word count	1644.1	581.8	1333.7	1067.9	4.02	0.000
Analytical thinking	92.4	5.14	93.3	5.36	-1.85	0.065
Clout	50.5	6.13	54.5	8.50	-6.01	0.000
Authentic	29.1	11.9	28.5	14.2	0.47	0.639
Emotional tone	63.5	14.3	56.8	18.2	4.59	0.000
Words per sentence	25.2	3.73	22.7	3.93	7.31	0.000
Words>6 letters	27.9	3.40	27.7	3.69	0.36	0.718
Dictionary words	77.4	4.14	75.5	4.68	4.65	0.000
Function Words	43.7	3.28	42.4	3.95	4.14	0.000
Total pronouns	6.41	1.78	6.39	1.97	0.15	0.884
Personal pronouns	1.40	0.84	2.05	1.22	-6.93	0.000
First-person singular	0.74	0.53	0.91	0.75	-2.93	0.004
First-person plural	0.23	0.30	0.48	0.67	-5.30	0.000
Second person	0.14	0.26	0.16	0.27	-0.69	0.493
Third-person singular	0.054	0.11	0.11	0.27	-2.91	0.004
Third-person plural	0.24	0.26	0.40	0.41	-5.28	0.000
Impersonal pronouns	5.01	1.31	4.34	1.31	5.74	0.000
Articles	8.55	1.12	8.09	1.24	4.31	0.000
Prepositions	14.7	1.20	14.9	1.33	-2.07	0.039
Auxiliary verbs	6.88	1.24	6.53	1.47	2.88	0.004
Common adverbs	3.10	0.88	2.76	1.07	3.87	0.000
Conjunctions	5.35	0.84	4.66	0.90	8.83	0.000

We find that 65 out of these 93 linguistic characteristics are significantly different at the 5% level between fake news articles and legitimate news articles. This implies that linguistic styles can be potentially helpful in detecting fake news.

3.3 Empirical result: Linguistic Characteristics

Table 7. Performance of Six Machine Learning Algorithms for Detecting Fake News

Measure	Median	Mean	Std. Dev.	Min	Max
		Panel A. Gradient Boosting			
Precision	0.871	0.871	0.031	0.769	0.939
Recall	0.907	0.905	0.027	0.833	0.981
F1	0.889	0.887	0.018	0.834	0.936
		Panel B. Logistic Regression			
Precision	0.857	0.852	0.032	0.765	0.928
Recall	0.900	0.897	0.033	0.804	0.962
F1	0.873	0.873	0.024	0.814	0.932
		Panel C. Naïve Bayes			
Precision	0.756	0.756	0.037	0.672	0.844
Recall	0.716	0.715	0.049	0.606	0.825
F1	0.731	0.734	0.033	0.652	0.818
		Panel D. Neural Network			
Precision	0.848	0.839	0.070	0.654	0.974
Recall	0.918	0.864	0.123	0.436	0.991
F1	0.855	0.842	0.056	0.596	0.935
		Panel E. Random Forest			
Precision	0.883	0.883	0.030	0.804	0.955
Recall	0.880	0.880	0.030	0.804	0.957
F1	0.881	0.881	0.019	0.839	0.927
		Panel F. Support Vector Machine			
Precision	0.835	0.834	0.033	0.750	0.933
Recall	0.873	0.868	0.036	0.750	0.933
F1	0.848	0.850	0.024	0.785	0.905

Notes: Results for each machine learning algorithm are based on 100 experiments.

The gradient boosting classifier based on linguistic styles achieves the best performance among the six classifiers and is effective in detecting fake news stories. On average, this classifier has achieved a precision of 87.1%, a recall of 90.5%, and an F1 score of 88.7%.

3.3 Empirical result: Linguistic Characteristics

Table 7. Performance of Six Machine Learning Algorithms for Detecting Fake News

Measure	Median	Mean	Std. Dev.	Min	Max
Panel A. Gradient Boosting					
Precision	0.871	0.871	0.031	0.769	0.939
Recall	0.907	0.905	0.027	0.833	0.981
F1	0.889	0.887	0.018	0.834	0.936
Panel B. Logistic Regression					
Precision	0.857	0.852	0.032	0.765	0.928
Recall	0.900	0.897	0.033	0.804	0.962
F1	0.873	0.873	0.024	0.814	0.932
Panel C. Naïve Bayes					
Precision	0.756	0.756	0.037	0.672	0.844
Recall	0.716	0.715	0.049	0.606	0.825
F1	0.731	0.734	0.033	0.652	0.818
Panel D. Neural Network					
Precision	0.848	0.839	0.070	0.654	0.974
Recall	0.918	0.864	0.123	0.436	0.991
F1	0.855	0.842	0.056	0.596	0.935
Panel E. Random Forest					
Precision	0.883	0.883	0.030	0.804	0.955
Recall	0.880	0.880	0.030	0.804	0.957
F1	0.881	0.881	0.019	0.839	0.927
Panel F. Support Vector Machine					
Precision	0.835	0.834	0.033	0.750	0.933
Recall	0.873	0.868	0.036	0.750	0.933
F1	0.848	0.850	0.024	0.785	0.905

Notes: Results for each machine learning algorithm are based on 100 experiments.

While neither commenters nor editors are able to reliably identify fake news, our results from machine learning and textual analysis imply that linguistic styles can be particularly informative in detecting fake news.

3.3 Empirical result: Linguistic Characteristics

Table 8. Top 30 Most Important Linguistic Features for Fake News Detection

Feature	Examples	Importance
Word count	--	6.96%
Colons	--	3.81%
Money	audit, cash, owe	3.50%
Words per sentence	--	2.81%
First-person plural	we, us, our	2.56%
Biological Processes	eat, blood, pain	2.52%
Periods	--	2.47%
Netspeak	btw, lol, thx	2.20%
Personal pronouns	I, them, her	2.11%
All Punctuation	--	1.97%
Numbers	--	1.93%
Reward focus	take, prize, benefit	1.87%
Sexuality	horny, love, incest	1.86%
Conjunctions	and, but, whearas	1.69%
Achievement	win, success, better	1.64%
Emotional tone	score (0 to 100) for positive vs. negative emotion	1.57%
Function Words	it, to, no, very	1.57%
Parentheses (pairs)	--	1.56%
Affiliation	ally, friend, social	1.53%
Future focus	may, will, soon	1.49%
Health/illness	clinic, flu, pill	1.48%
Impersonal pronouns	it, it's, those	1.45%

To identify the most important linguistic characteristics, we employ the XGBoost package to generate the feature importance scores for the 93 LIWC linguistic features.
We find that Word Count and Words per Sentence are two of the most important feature components.

3.4 Empirical result: Market Reaction

Table 9 Market Reaction to Fake News – Trading Volume

VARIABLES	(1) $Log(AVol_{0,1})$	(2) $Log(AVol_{0,2})$	(3) $Log(AVol_{3,120})$	(4) $Log(AVol_{3,242})$
Fake	-0.408***	-0.421***	-0.055*	-0.079***
	(0.080)	(0.081)	(0.029)	(0.021)
$Log(AVol_{0,2})$			0.030	0.026*
			(0.018)	(0.015)
Log(Length)	0.008	0.029	0.052*	-0.009
	(0.094)	(0.096)	(0.031)	(0.024)
%NegWord	-0.055	-0.071	-0.033	0.019
	(0.054)	(0.053)	(0.020)	(0.019)
Premium	-0.003	0.017	-0.055**	-0.027
	(0.076)	(0.078)	(0.028)	(0.022)
EditorPick	0.100	0.055	-0.001	-0.023
	(0.154)	(0.161)	(0.047)	(0.033)
Log(Comment)	0.056	0.067*	-0.015	-0.016
	(0.037)	(0.038)	(0.012)	(0.010)

We find a negative and significant coefficient estimate on fake, suggesting that fake news stories induce a lower trading volume reaction than legitimate news stories.

3.4 Empirical result: Market Reaction

Table 10. Market Reaction to Fake News – |Abnormal Return|

VARIABLES	(1) $\lvert ARet_{0,1} \rvert$	(2) $\lvert ARet_{0,2} \rvert$	(3) $\lvert ARet_{3,120} \rvert$	(4) $\lvert ARet_{3,242} \rvert$
Fake	-0.018*	-0.025**	-0.104**	-0.204***
	(0.010)	(0.010)	(0.042)	(0.050)
$\lvert ARet_{0,2} \rvert$			0.442***	0.161
			(0.157)	(0.214)
Log(Length)	0.004	0.017	0.016	0.006
	(0.013)	(0.012)	(0.037)	(0.054)
%NegWord	0.010	0.009	-0.012	-0.047
	(0.007)	(0.007)	(0.027)	(0.034)
Premium	0.020*	0.015	-0.040	0.040
	(0.010)	(0.010)	(0.041)	(0.056)
EditorPick	-0.016	-0.013	0.049	-0.033
	(0.017)	(0.017)	(0.065)	(0.073)

Thus, the magnitude of the stock price reaction to fake news articles is less than that of legitimate news articles.
In sum, the results in Tables 9 and 10 suggest that the market discounts the release of fake news.

4. Conclusion

1. Investor attention is significantly higher for fake news articles than a matched control sample of legitimate news articles.
2. Article commenters or Seeking Alpha editors can’t successfully identify fake news articles.
3. Fake news articles differ from legitimate news articles in word choices.
4. Both the abnormal volume reaction and the magnitude of the abnormal return response to fake news articles are significantly less than that of legitimate news articles.

4. Comment & Inspiration

1. The author thinks that both investors and editors can't identify the fake news, but the market can identify the true and fake news. The reason is not explained.