This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News

Benjamin D. Horne and Sibel Adalı

Why this work matters:

- Misinformation may cause political and societal decisions that run counter to a society's best interest
- While truth is hard to detect computational, early warning systems can assist both fact-checkers and news readers in their decisions
- Additionally, understanding what features distinguishes types of news can help us understand how users are persuaded

Q: Is there any **systematic stylistic** and **other content differences** between **fake** and **real** news?

We use three types of news:

✓ **Real** ✗ **Satire** ✗ **Fake**

- Use **2** different **independent** data sets of political news:

  Data set 1: Buzzfeed 2016 Election, highly engaged on Facebook

  Data set 2: Random political news

- Use **ANOVA** and **Wilcoxon** tests to find significant shifts in feature distributions
- Use **linear kernel SVMs** on top 4 most statistically significant features to show predictive power of features

	Real	Fake	Satire
Data set 1	36	35	0
Data set 2	75	75	75

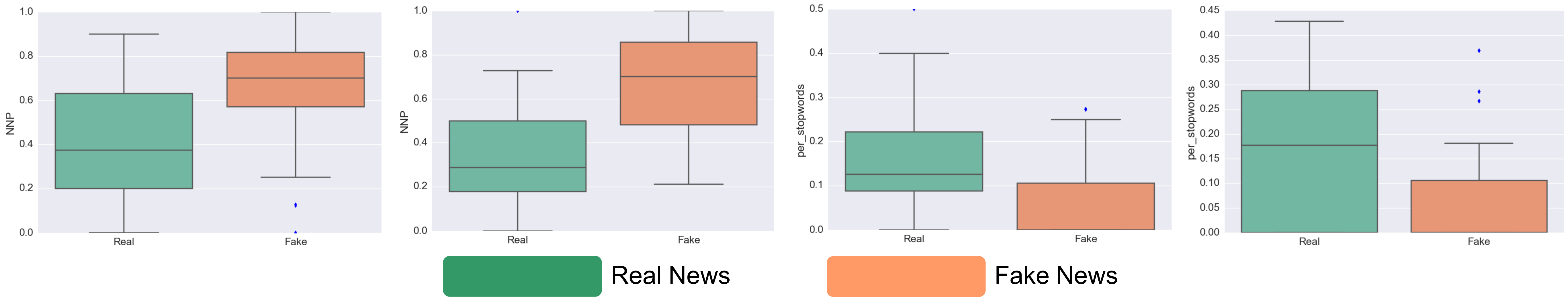
Published at The International Workshop on News and Public Opinion at ICWSM 2017

Types of features:

- **Complexity Features** : Natural language features that capture the intricacy of article and title (*Ex: readability, depth of syntax tree, fluency, lexical redundancy*)
- **Psychology Features** : Based on well studied psychological meaning of words using Linguistic Inquiry and Word Count (*Ex: analytic, insight, emotional tone, certainty, personal concerns*)
- **Stylistic Features** : Simple natural language features that capture the sentence structure and grammatical elements (*Ex: word count, quotes, verb phrases, personal pronouns, informal words*)

Number of Proper Nouns in Title — Data set 1, Data set 2

Percent Stop-words in Title — Data set 1, Data set 2



Body Content

Feature	Data set 1	Data set 2
Word Count	Real > Fake	Real > Fake > Satire
Fluency	Fake > Real	Satire = Fake > Real
Avg Word Len	Real > Fake	Real > Fake = Satire
Quotes	Real > Fake	Real > Fake = Satire
Readability		Real = Satire > Fake
Person pronoun		Satire > Fake > Real
Adverb		Satire = Fake > Real
Punctuation	Real > Fake	Real > Fake = Satire
You		Satire > Fake > Real
We		Fake > Real = Satire
Redundancy	Fake > Real	Satire > Fake > Real
Neg Emotion	Fake > Real	
Analytic words		Real > Fake = Satire
Syntax depth	Fake > Real	

Title Content

Feature	Data set 1	Data set 2
Word Count	Fake > Real	Fake > Real = Satire
Fluency	Fake > Real	Satire > Fake = Real
Avg Word Len		Real > Fake = Satire
All Caps	Fake > Real	Satire > Fake > Real
Readability	Real > Fake	Real > Satire = Fake
Person pronoun	Real > Fake	
Noun	Real > Fake	Real > Satire > Fake
Proper Noun	Fake > Real	Fake = Satire > Real
Posses Pronoun	Real > Fake	
Determiner	Real > Fake	
% Stop-words	Real > Fake	Real > Satire = Fake
Exclaim		Fake > Real = Satire
Analytic words	Fake > Real	
# Verb Phrases		Fake > Real = Satire

Features that significantly differ between news types. Significance determined by Wilcoxon Rank-Sum or ANOVA tests depending on the normality of the feature distribution. Significance is considered for p-values less than 0.05 and large F-values.

▭ **Strongest significance**

What we found out:

- Fake content differs in **word count**, **word length, # quotes,** and **redundancy**
- Fake titles use more **proper nouns**, **verb phrases**, & fewer **stop-words, nouns**

Fake Title: BREAKING BOMBSHELL: NYPD Blows Whistle on New Hillary Emails: Money Laundering, Sex Crimes with Children, Child Exploitation, Pay to Play, Perjury

Real Title: Preexisting Conditions and Republican Plans to Replace Obamacare

- Fake content is more similar to satire than to real, findings consistent on 3rd data set
- Real news persuades through arguments, while fake news persuades through heuristics

Linear SVM Classification Results:

	Baseline	Fake vs Real	Satire vs Real	Satire vs Fake
Body	50%	71%	91%	67%
Title	50%	78%	75%	55%

What we plan to do in the future:

- More data
- Create unsupervised ground truth
- Include news from across the spectrum

Acknowledgments

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053 (the ARL Network Science CTA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

Research Supported by the Army Research Lab

Twitter: @benjamindhorne Medium: @benjamindhorne314