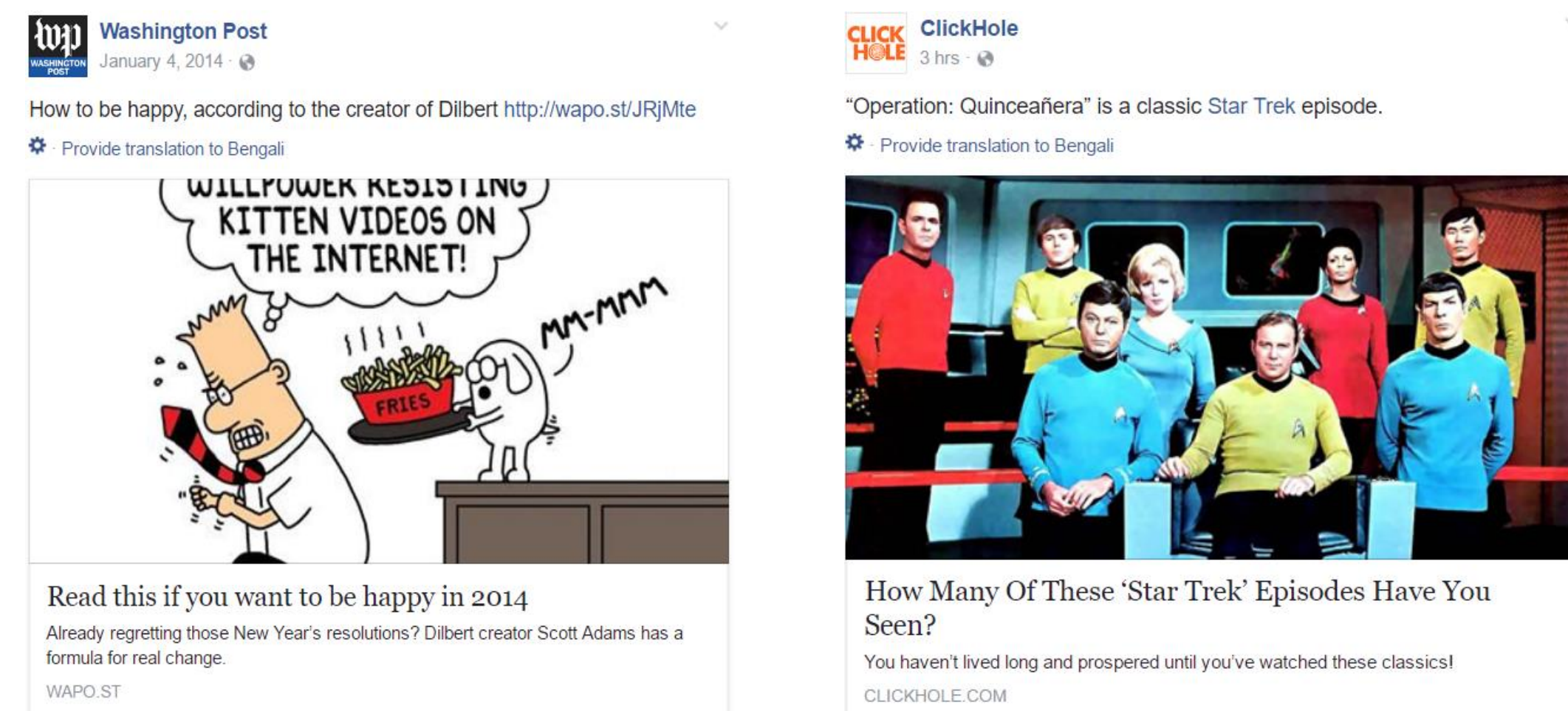# We Dived Deep into Clickbaits: You won't Believe what Happened Next!

Md Main Uddin Rony[1], Naeemul Hassan[1], Mohammad Yousuf[2]
[1]The University of Mississippi, [2]The University of Oklahoma

Motivation

Clickbaiting has become widespread
- All types of media practice clickbait.
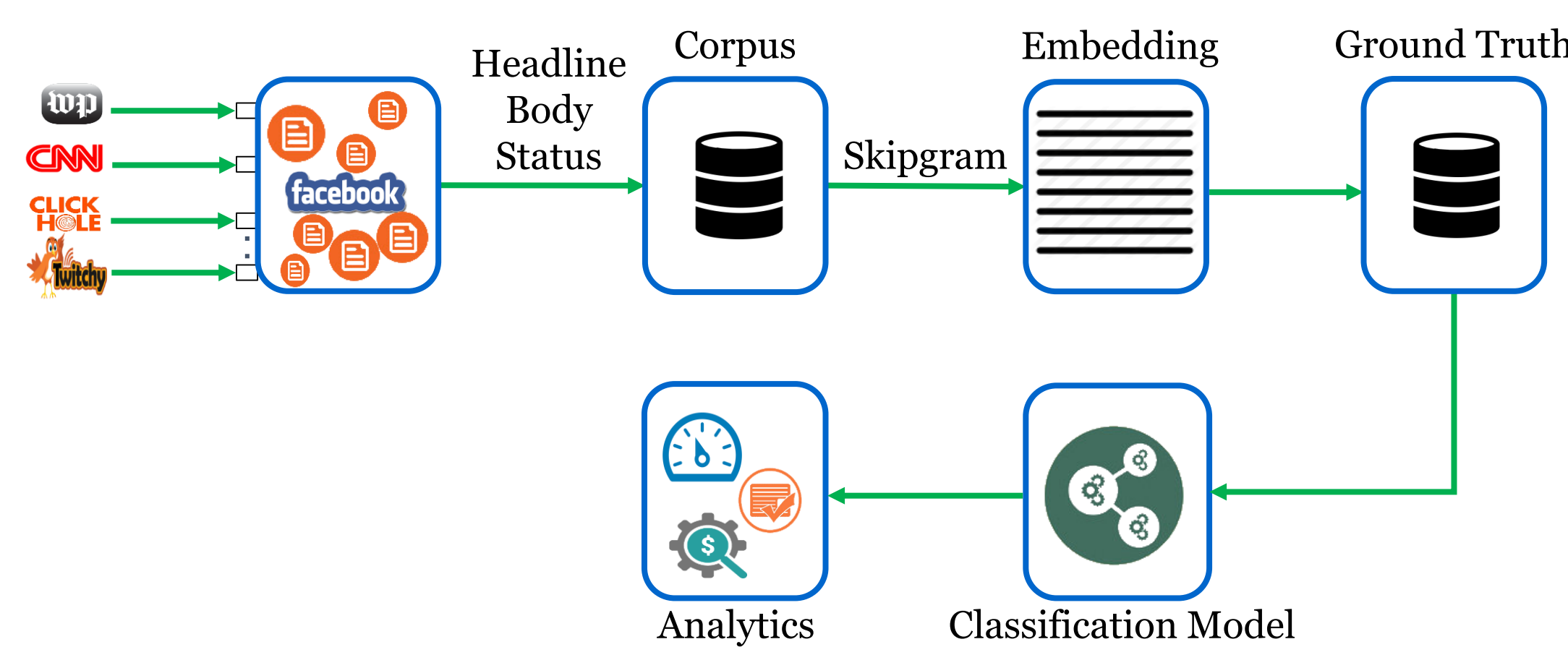- It has become a source of easy revenue.

Negative impact on media ecosystem
- Damages media-user relationship.
- Depletes media brand value.

Lack of research on clickbait practice
- No large scale analysis on practice of clickbait by media organizations.
- No study to show its contribution to public engagement on social media.

Problem Formulation

Two categories of headlines
Clickbait
- *Eek! What's Lurking in the Shadows?! I Have to Know!*
- *I Left My Daughter And THIS Happened!*

Non-Clickbait
- *America's democracy has become illiberal*
- *Twitter Expands Fight Against Abuse*

Goal
- Building a clickbait detection model.
- Performing large-scale study on the clickbait practice by a range of media on social network (Facebook).

Data Collection

Ground Truth
- 32,000 manually labeled news headlines. [1]

Media Corpus
- About 1.7 million Facebook posts.
- From 68 mainstream and 85 unreliable media.
- Considered headlines, bodies, and the statuses.

Media	Category	Link	Video	Total
Mainstream	Broadcast	324028	32924	356952
	Print	516713	14129	530842
Unreliable	Clickbait	371834	4099	375933
	Conspiracy	309122	5841	314963
	Junk Science	51923	649	52572
	Satire	41046	151	41197
Total		1614666	57793	1672459

Clickbait Detection

- Distributed subword embedding was used to create sentence vectors.
- Pre-trained embeddings using 477236 words of the Media Corpus.
- Used Softmax function as the classifier.
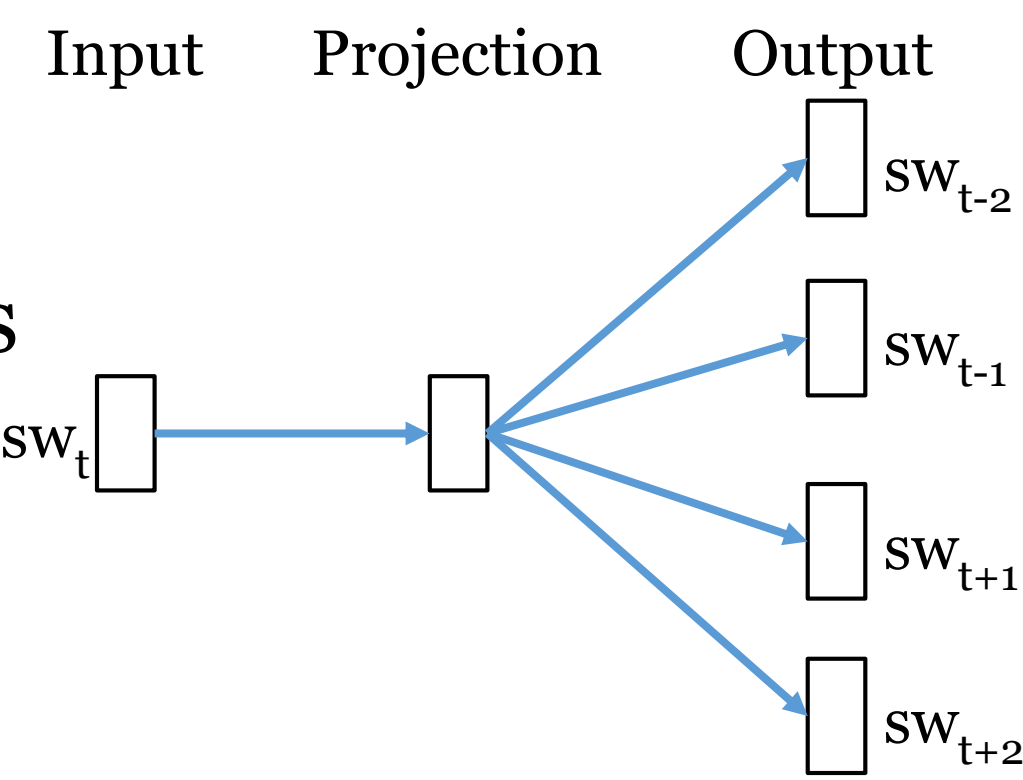
Method		Precision	Recall	F-score	Accuracy
Without Pre-trained Vectors	Chakraborty et al. [1]	0.95	0.90	0.93	0.93
	Ground Truth	0.976	0.975	0.975	0.976
With Pre-trained Vectors	Anand et al. [2]	0.984	0.978	0.982	0.982
	Ground Truth + Google word2vec	0.977	0.977	0.977	0.976
	Ground Truth + Media Corpus	**0.983**	**0.983**	**0.983**	**0.983**

Social Reachability and Impact Analysis

Blue indicates clickbait articles get more attention than non-clickbaits, Green indicates the opposite
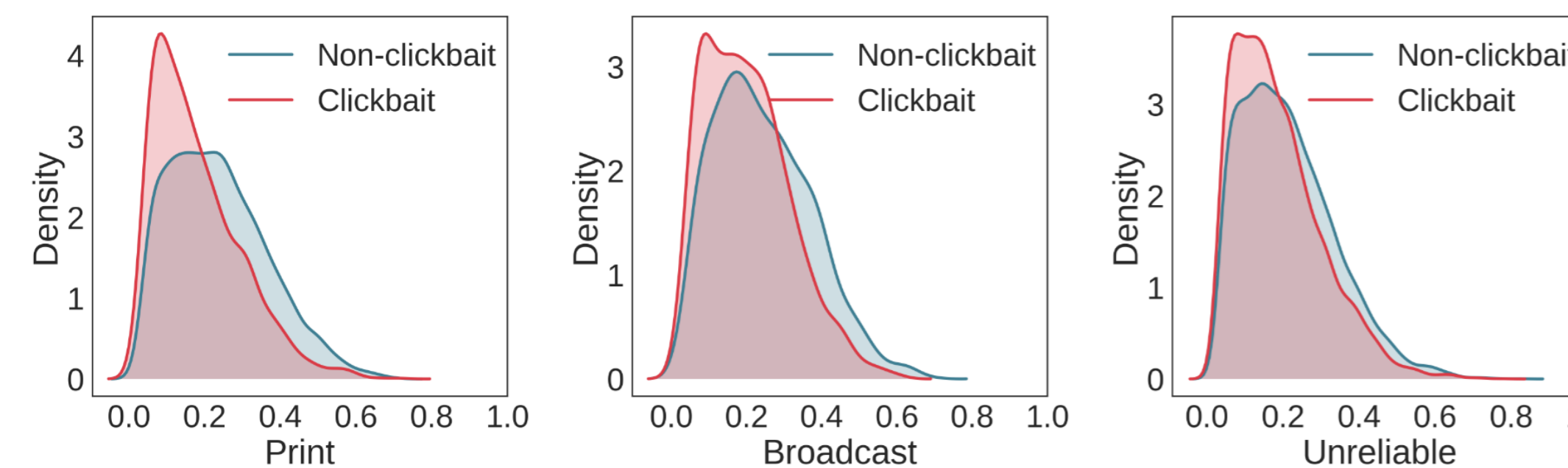
Qualitative Analysis

Topic Modeling
- Used Biterm Topic Modeling (BTM).
- Clickbait headlines in print and broadcast media represent more personalized, sensationalized and entertaining topics.
- Non-clickbait headlines highlight topics of collective problems such as public policies.


Headline-Body Relevance
- Hypothesis: Clickbait headlines are less relevant to the body content.
- Cosine similarity was used to measure the relevance between a headline and the body.

Quantitative Analysis

Media	Category	Clickbait	Non-Clickbait	Clcikbait(%)
Mainstream	Broadcast	169752	187200	47.56
	Print	128022	402820	24.12
Unreliable	Clickbait	172271	203662	45.82
	Conspiracy	90389	224574	28.7
	Junk Science	23637	28935	44.96
	Satire	21798	19399	52.91

**% of clickbaits in various media**
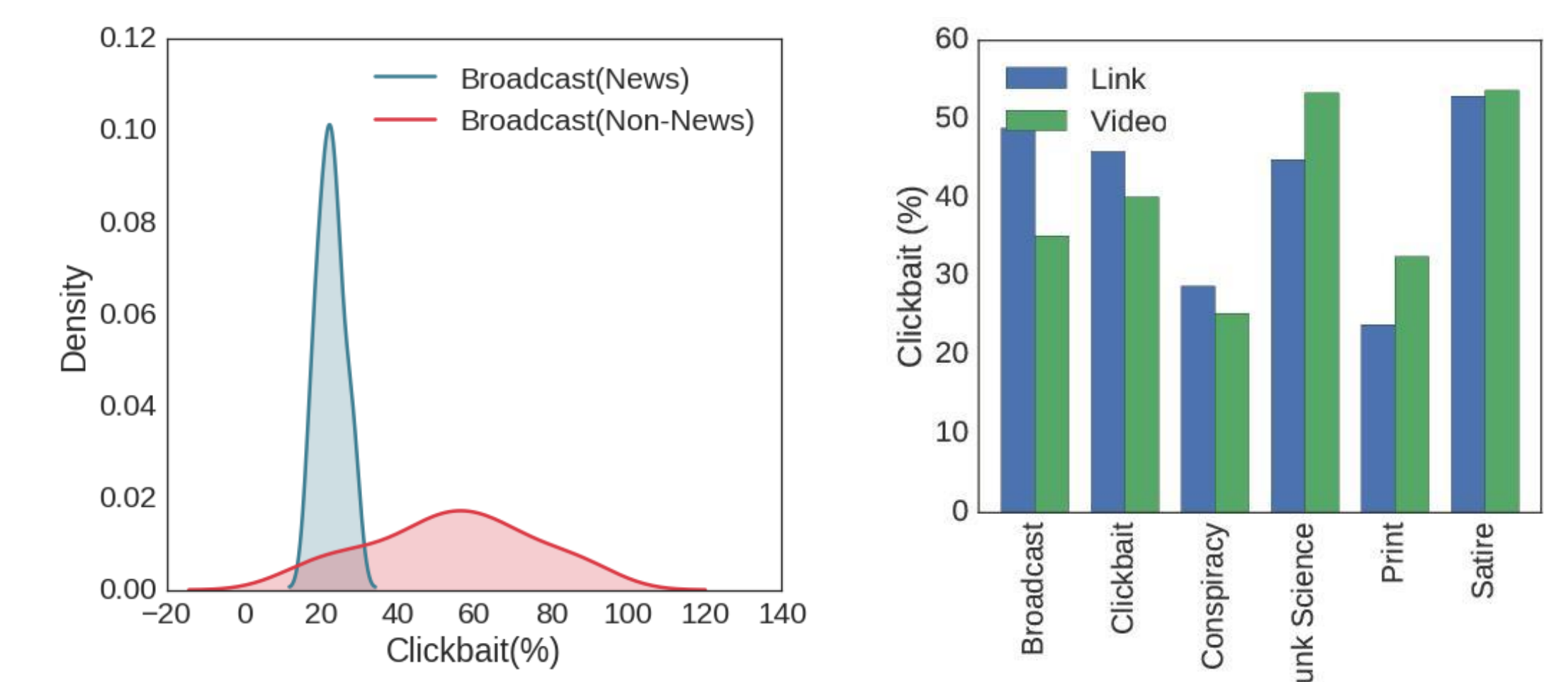
% of clickbait in news & non-news   **% of clickbait in link & video**

Media	Category	Clickbait Status	Non-clickbait Link	Clickbait Status (%)
Mainstream	Broadcast	84192	176177	32.34
	Print	164669	379504	30.26
Unreliable	Clickbait	91747	157886	36.75
	Conspiracy	46851	190477	19.74
	Junk Science	12764	28349	31.05
	Satire	7425	14453	33.94

% of clickbait in Facebook Status
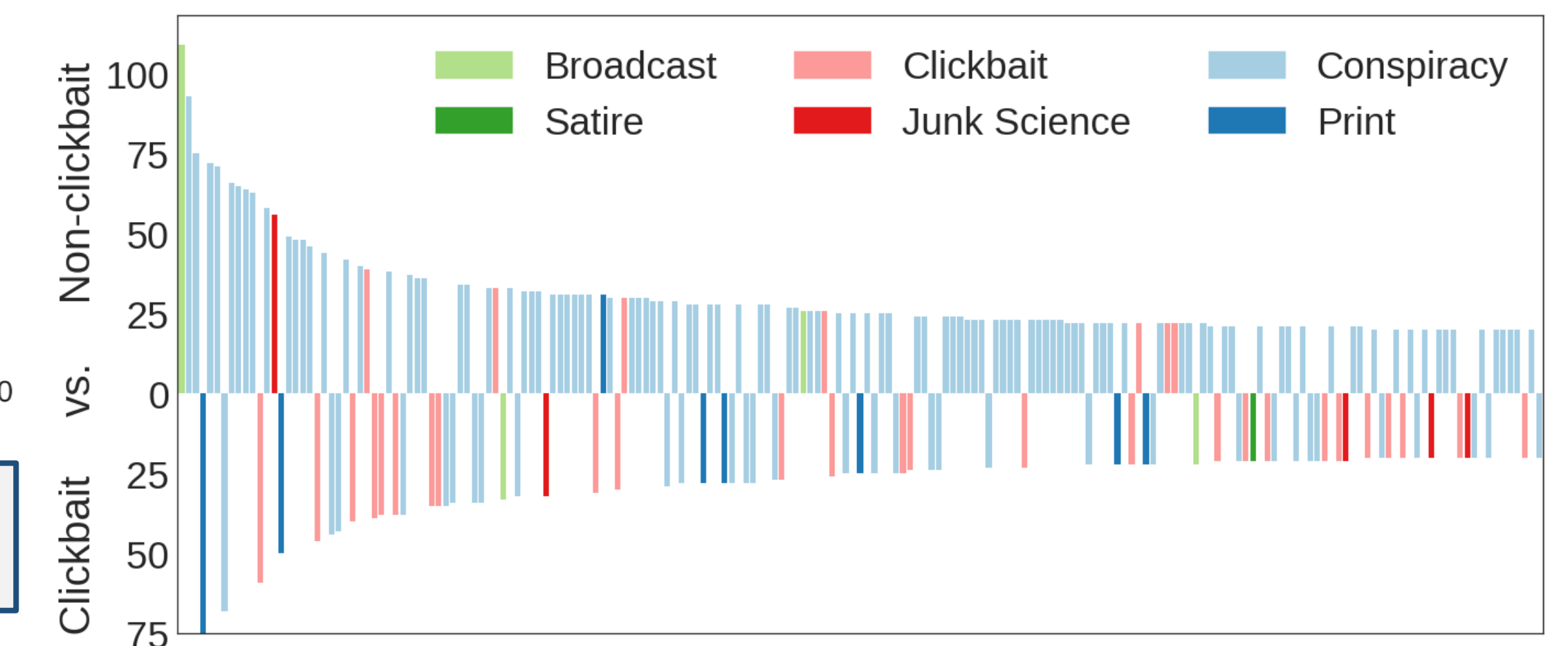
Frequency of re-post by different media

References

1. A. Chakraborty, B. Paranjape, S. Kakarla, and N. Ganguly. Stop clickbait: Detecting and preventing clickbaits in online news media. In Advances in Social Networks Analysis and Mining (ASONAM), 2016.

2 A. Anand, T. Chakraborty, and N. Park. We used neural networks to detect clickbaits: You won't believe what happened next! ArXiv preprint arXiv:1612.01340, 2016.