A Large-Scale Study of Social Media Sources in News Articles

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

Motivation

Widespread use of social media content as source

- Journalists quote and paraphrase contents from social media pages regularly.
- "Convenient, cheap and effective" news source.
  [Marcel Broersma and Todd Graham. 2013]

Lack of research on social media sourcing practice

- No large scale analysis on the extent to which news media use social media as news source.
- Experiment on sourcing practice focusing US media is absent.

Systematic analysis will be beneficial

- To explain how social media becomes a primary source.
- To understand the role of social network used by news media in spreading misinformation.

Research Questions

- **RQ1:** How often do mainstream and unreliable news websites use Facebook and Twitter content in articles?
- **RQ2:** To what extent do mainstream and unreliable news media process Facebook and Twitter content used in articles?
- **RQ3:** Does the use of social media source vary for different news topics?

Data Collection

- Curated Facebook posts dated from January 1, 2013 to December 31, 2017

| Media Type	# of Organizations	# of Posts	# of Samples
Mainstream	68	29,656	29,656
Unreliable	85	76,997	29,700
Total	**153**	**106,653**	**59,356**

Extraction of Social Media Source

- Identified 212 (Facebook: 134, Twitter: 78) citation patterns.
- Developed a regex based classifier using the patterns.
- Categorized the source usages into - *Quotation, Paraphrase and Embedding*

Most Frequent Twitter and Facebook Patterns
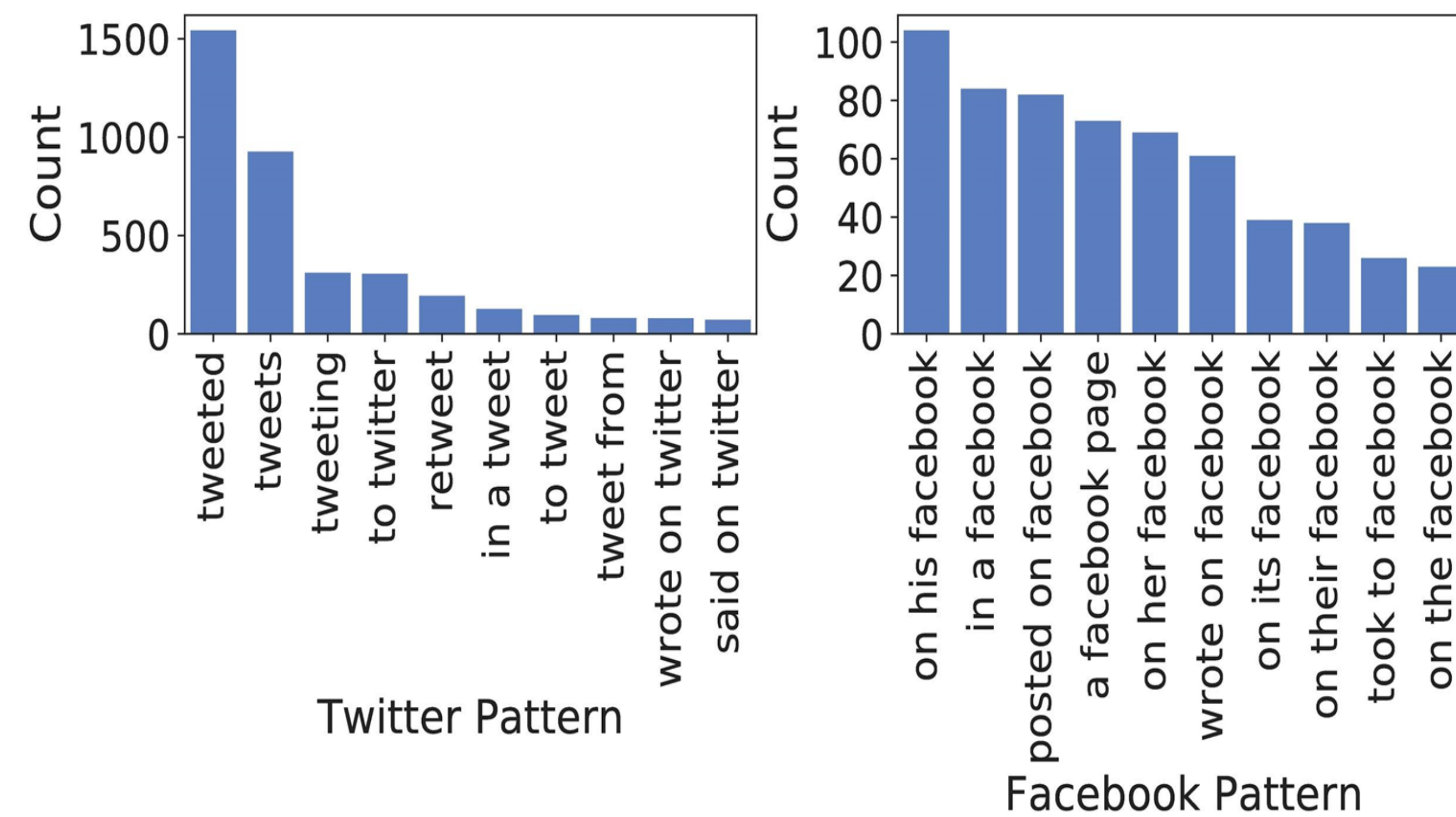
Category	Precision	Recall	F1
Quotation	89.80%	73.33%	86.21%
Paraphrase	94.34%	79.37%	80.73%
Embedding	100%	100%	100%
Macro-average	**94.71%**	**84.23%**	**88.98%**
Micro-average	**97.85%**	**92.62%**	**95.16%**

Performance of Social Media Source Identification

RQ1: Use of Social Media as Source
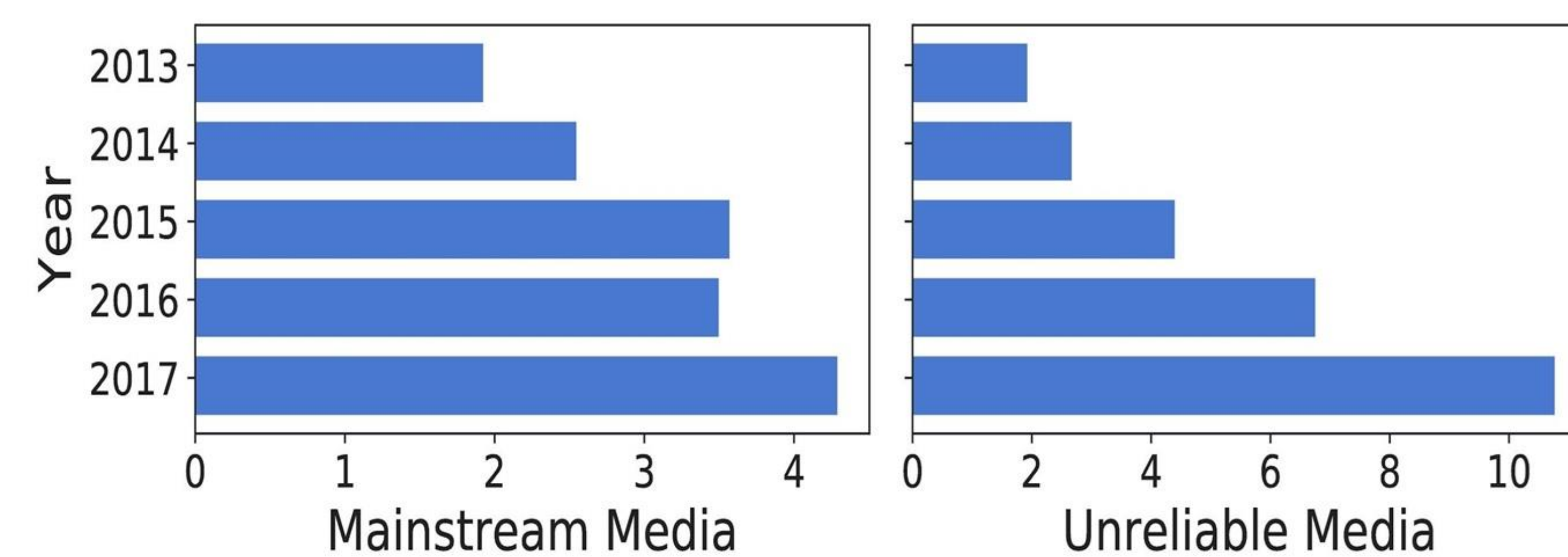
Media Type	# Direct Quote (Avg. Per Article)	# Social Media Source	Ratio
Mainstream	201924 (6.81)	4207	1:48
Unreliable	185182 (6.23)	12436	1:14.89

Social Media Source vs. all direct quotations

RQ2: Social Media Content Processing

- Mainstream media uses more *Quotation* than *Paraphrases*.

Source	Media	# of Articles	Quotation	Paraphrase	Embedding	Total
Twitter	Mainstream	1654	1065 (28.22%)	866 (22.95%)	1843 (48.83%)	3774 (89.71%)
	Unreliable	3170	1137 (9.41%)	1130 (9.35%)	9814 (81.24%)	12081 (97.15%)
	Total	**4824**	**2202**	**1996**	**11657**	**15855**
Facebook	Mainstream	377	228 (52.66%)	205 (47.34%)	--	433 (10.29%)
	Unreliable	324	178 (50.14%)	177 (49.86%)	--	355 (2.85%)
	Total	**701**	**406**	**382**		**788**

Social media content usage as a source by mainstream and unreliable media

RQ3: Relation with News Topic

Topic	Mainstream			Unreliable		
	Quotation	Paraphrase	Embedding	Quotation	Paraphrase	Embedding
Politics	377 (47.7%)	238 (30.1%)	175 (22.2%)	665 (9.8%)	648 (9.5%)	5495 (80.7%)
Entertainment	303 (26.8%)	241 (21.4%)	585 (51.8%)	197 (14.7%)	116 (8.7%)	1026 (76.6%)
Sensitive Subject	187 (32.8%)	165 (29%)	217 (38.1%)	217 (13.8%)	172 (10.9%)	1185 (75.3%)
Law & Govt	69 (34.8%)	73 (36.9%)	56 (28.3%)	75 (11.3%)	74 (11.1%)	516 (77.6%)
Sports	97 (17.1%)	380 (66.8%)	92 (16.2%)	30 (17.1%)	28 (15.9%)	118 (67%)
People & Society	22 (18.2%)	38 (31.4%)	61 (50.4%)	64 (11.8%)	66 (12.2%)	413 (76.1%)

Processing of social media sources in different news topics

Media Type	Topic	# Articles	# Articles with Social Media Source
Mainstream	Arts & Entertainment	5943	491 (8.26%)
	Sensitive Subjects	3391	300 (8.85%)
	Law & Government	2793	112 (4.01%)
	Sports	2592	213 (8.22%)
	Politics	2389	369 (15.45%)
Unreliable	Politics	7104	1711 (24.09%)
	Sensitive Subjects	3790	484 (12.77%)
	People & Society	2889	171 (5.92%)
	Law & Government	2835	228 (8.04%)
	Health	2546	51 (2%)

Social Media source usage in the top-5 news topics for each media