Developing an Information Source Lexicon

Aseel Addawood*, Rezvaneh Rezapour†, Shubhanshu Mishra†, Jodi Schneider*†, Jana Diesner*†
*Illinois Informatics Institute, †School of Information Sciences, University of Illinois at Urbana-Champaign

Introduction

Many users share links to content when posting on social media platforms like Twitter. This is motivated by either their need to circumvent the content limit imposed by the platform (for Twitter this was 140 characters at the time of our study), or to back up their opinion [2]. These links when used to support one's claim can provide useful evidence about the information source on which the user grounds their claim. Furthermore, information sources are varied in the way they support a claim e.g. scientific article provide a logical stance on an argument, while blogs and social media posts provide an opinionated stance. Recent events of the proliferation of fake news in social media and its impact on social systems like presidential elections, provide an incentive to examine these information sources used by social media users. Furthermore, quantifying the usage of different information source types can be utilized to measure the proliferation of different types of content about a given topic. Herein, we present a lexicon based approach for identifying and categorizing different types of information sources. Using a corpus of tweets about Measles, Mumps and Rubella (MMR) vaccine debate, we demonstrate the application of our lexicon by contrasting the distribution of various information sources. In previous work, manual annotation of information sources was used to identify information source types [6, 7]. However, this approach is limited because (a) it requires extensive labor and (b) biased towards the small sample of data. To improve these aspects of the research, we developed a large-scale information source lexicon by combining data from various open resources. The focus of the lexicon is on identifying different content types present identified by domain of their URL e.g., video, social media, blog, news, fake news, and scientific communication. Our lexicon allows for a simple and high recall identification of information source types present in social media content. Moreover, it can be used to develop new tools to help social media users in vetting, verifying, fact-checking, filtering, and flagging debatable information, which may result in raising public awareness regarding online misinformation. The following section discusses the development of the lexicon and its use in an experimental study.

Development of the Lexicon

To construct the lexicon, we first identified the main categories of information sources used online. Based on prior literature [6] and close readings of user-generated texts on social media, we found that the initial information source types are news outlets, blogs, fake news, social media, commercial, videos, and scientific references. We then retrieved the domain names for each category from existing lexicons or all the items indexed under a similar category in Wikidata [9]. Moreover, since the main objective of this analysis was to identify the types of information sources used on Twitter, we utilized a corpus of tweets about the topic of MMR vaccine debate (described in prior work [10]), to enhance the lexicon and improved the initial set of categories. For each tweet, we extracted the destination of all

Table 1 Number of instances of each information source type in our lexicon

Type	Counts	Description	Example
Blog	194	All blogging platforms indexed in Wikidata	wordpress.com
Commercial	55	All commercial websites indexed in Wikidata	amazon.com
Fake news	518	1) A list developed by Melissa Zimdars and her research team at Merrimack College [1] 2) A list of fake news websites from Wikipedia [3] 3) FakeNewsChecker [4]	naturalnews.com
News	1,988	1) News sources indexed by Wikidata 2) List of trusted news domains created by Facebook [5]	nytimes.com
Scientific	2,962	All scientific publishers indexed by Wikidata	springer.com
Social media	87	All social media domains indexed by Wikidata	facebook.com
Twitter	1	Links to other tweets, twitter hosted images, videos	twitter.com
Videos	13	All video sharing services from Wikipedia [7, 8]	vimeo.com

mentioned URLs, by following all redirects (we refer to this process as URL expansion). For each URL, its domain was extracted e.g. in the URL *https://en.wikipedia.org/wiki/Emily_Dickinson, en.wikipedia.org* is the domain. We decided to use domain information instead of the full URL because URLs from same domain are very likely to exhibit similar types of contents, although the distribution of contents type in different domains is likely to vary, e.g. blog can contain scientific content, humor, sarcasm, or opinionated content but scientific articles are very likely to only contain peer-reviewed scientific findings. Every domain absent from our lexicon was indexed under all applicable information source types based on the manual inspection of its content. The assignment of information source types to domains is not exclusive, i.e., a domain can be categorized under multiple information source types. For example, *youtube.com* indexed under video and social media. Similarly, *tumblr.com*, is indexed under social media and blog. However, *wordpress.com* is only indexed under blog. A description of each information source type along its count in our lexicon and one example instance is presented in Table 1. In total, we have 5,818 domain names in our lexicon.

Experimental Study

In order to demonstrate the utility of out lexicon, we considered an existing corpus of 40,713 tweets about MMR vaccine debate [10]. 57.2% of the tweets in this dataset contain a URL, with a total of 24,143 URLs. We expanded each URL and extracted its domain named as described above.

provides the top three domains of the different information source types and the probability of finding it among those tweets with a URL. We observe that news domains are referred to the most, while scientific domains have the least reference. These results indicate that online users mainly rely on news sources to support their statement. This may have happened since news articles are easier to access and comprehend in comparison to scientific articles, most of which are behind a paywall and written for a specific audience. Furthermore, the probability of *fakenews* domain is quite high (compared to scientific articles). In fact, "fakenews" and blog domain probabilities are quite comparable indicating that Twitter users in our dataset are very likely to share opinionated or *fakenews* content when discussing a controversial issue like vaccine. This may have happened since the users have limited knowledge about the sources that they obtain and share.

Table 2 Top three domains of each type and their probability in a dataset of tweets about vaccines

Information source type (x)	Top three domains	P (source type=x)
Blogs	truthinmedia.com, wordpress.com, paraven.net	0.062
Commercial	vaxxedthemovie.com, amazon.com, video214.com	0.035
Fake news	naturalnews.com, truthkings.com, infowars.com	0.069
News	nytimes.com, washingtonpost.com, forbes.com	0.191
Scientific	ncbi.nlm.nih.gov, cdc.gov, healio.com	0.007
Social media	youtube.com, facebook.com, periscope.tv	0.134
Twitter	twitter.com	0.227
Videos	youtube.com, instagram.com, vimeo.com	0.071

These results highlight the importance of utilizing an information source lexicon by demonstrating ease of use and high coverage results. Online users can use the developed lexicon as a confirmation step before sharing unverified sources via social media. This step will limit the circulation of misinformation and assist users in gaining better and healthier digital literacy practices when looking for information online. As mentioned earlier, the developed lexicon also can help researchers to better understand the behavior of social media users by analyzing the content they share on various social media platforms such as Facebook and Twitter. Moreover, the lexicon can be used to develop tools that our lexicon can help online users with accurately identifying the types of information sources shared online and that having a credible social media ecosystem is the responsibility of everyone, requiring a tremendous collaborative work from all online users. Being able to think critically and vet information before sharing them in social media is an important skill these days, and our developed lexicon can be considered as one of the first steps toward having that environment.

References

[1] Opensources. Retrieved Oct 23 2017 from http://www.opensources.co/.

[2] Kinsella, S., Wang, M., Breslin, J. and Hayes, C.(2011).Improving categorisation in social media using hyperlinks to structured data sources. *The Semanic Web: Research and Applications*390-404

[3] List of fake news websites, Wikipedia. Retrieved Oct 23 2017 from https://en.wikipedia.org/wiki/List_of_fake_news_websites.

[4] FakeNewsChecker. Retrieved Oct 23 2017 from www.FakeNewsChecker.com.

[5] Information about trending topics, facebook newsroom. Retrieved Oct 23 2017.from http://newsroom.fb.com/news/2016/05/information-about-trending-topics/.

[6] Addawood, A.A. and Bashir, M.N.(2016). "What is your evidence?" a study of controversial topics on social media. Proceedings of the 3rd Workshop on Argument Mining. at *54th Annual Meeting of the Association for Computational Linguistics (ACL)*. Munich, Germeny (ACL).

[7] Rinott, R., Dankin, L., Perez, C.A., Khapra, M.M., Aharoni, E. and Slonim, N.(2015). Show me your evidence-an automatic method for context dependent evidence detection. In *Proceedings of the EMNLP*.

[8] List of video hosting services, Wikipedia. Retrieved Oct 23 2017 from https://en.wikipedia.org/wiki/List_of_video_hosting_services.

[9] Wikidata. Retrieved Oct 23 2017 from https://www.wikidata.org/.

[10] Addawood, A., Rezapour, R., Abdar, O., & Diesner, J. (2017). Telling apart tweets associated with controversial versus non-controversial topics. Proceedings of 2nd Workshop on NLP and Computational Social Science (NLP+CSS) at *55th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 32-41). Vancouver, Canada (ACL).