# Differences in Health News from Reliable and Unreliable Media

Sameer Dhoju
University of Mississippi
sdhoju@go.olemiss.edu

Md Main Uddin Rony
University of Mississippi
mrony@go.olemiss.edu

Muhammad Ashad Kabir
Charles Sturt University
akabir@csu.edu.au

Naeemul Hassan
University of Mississippi
nhassan@olemiss.edu

## ABSTRACT

The spread of 'fake' health news is a big problem with even bigger consequences. In this study, we examine a collection of health-related news articles published by reliable and unreliable media outlets. Our analysis shows that there are structural, topical, and semantic patterns which are different in contents from reliable and unreliable media outlets. Using machine learning, we leverage these patterns and build classification models to identify the source (reliable or unreliable) of a health-related news article. Our model can predict the source of an article with an F-measure of 96%. We argue that the findings from this study will be useful for combating the health disinformation problem.

## CCS CONCEPTS

• **Applied computing → Health informatics**.

**ACM Reference Format:**
Sameer Dhoju, Md Main Uddin Rony, Muhammad Ashad Kabir, and Naeemul Hassan. 2019. Differences in Health News from Reliable and Unreliable Media. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3316741

## 1 INTRODUCTION

Of the 20 most-shared articles on Facebook in 2016 with the word "cancer" in the headline, more than half the reports were discredited by doctors and health authorities [7]. The spread of health-related hoaxes is not new. However, the advent of Internet, social networking sites (SNS), and click-through-rate (CTR)-based pay policies have made it possible to create hoaxes/"fake news", published in a larger scale and reach to a broader audience with a higher speed than ever [16]. Consequences of misleading or erroneous health news can be very critical. Believing health misinformation may lead to a hazardous health condition. Houston reported a measles outbreak in Europe due to lower immunization rate which experts believed was the result of anti-vaccination campaigns caused by a false news about MMR vaccine [14]. Moreover, misinformation can spoil the credibility of the health-care providers and create a lack of trust in taking medicine, food, and vaccines.

Recently, researchers have started to address the fake news problem in general [21, 32]. However, health disinformation is a relatively unexplored area. According to a report from Pew Research Center [8], 72% of adult internet users search online for information about a range of health issues. So, it is important to ensure that the health information which is available online is accurate and of good quality. There are some authoritative and reliable entities such as National Institutes of Health (NIH) [1] or *Health On the Net* [2] which provide high-quality health information. Also, there are some fact-checking sites such as Snopes.com [3] and Quackwatch.org [4] that regularly debunk health and medical related misinformation. Nonetheless, these sites are incapable of busting the deluge of health disinformation continuously produced by unreliable health information outlets (e.g., RealFarmacy.com, Health Nut News). Moreover, the bots in social networks significantly promote unsubstantiated health-related claims [9]. Researchers have tried developing automated health hoax detection techniques but had limited success due to several reasons such as small training data size and lack of consciousness of users [11, 12, 20, 35].

The objective of this project is **(i)** to develop a health-oriented news dataset that covers both reliable and unreliable media outlets, **(ii)** to identify discriminating features that can potentially separate a reliable health news from an unreliable health news by leveraging a large-scale dataset, and **iii)** use the dataset and the discriminating features for building a classifier. We examine how reliable media and unreliable media outlets conduct health journalism. First, we prepare a large dataset of health-related news articles which were produced and published by a set of reliable media outlets and unreliable media outlets. Then, using a systematic content analysis, we identify the features which separate a reliable outlet sourced health article from an unreliable sourced one. These features incorporate the structural, topical, and semantic differences in health articles from these outlets. For instance, our structural analysis finds that the unreliable media outlets use clickbaity headlines in their health-related news significantly more than what reliable outlets do. Our topical analysis finds that while the reliable outlets discuss "cancer" along with research and studies, in the unreliable outlets "cancer" is associated with autism and vaccination. The semantic analysis shows that on average a health news from reliable media contains more reference quotes than an average unreliable sourced health

*WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA*

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316741

[1] https://www.nih.gov/
[2] https://www.hon.ch/en/
[3] https://www.snopes.com/
[4] http://www.quackwatch.org/

news. We argue that these features can be critical in understanding health misinformation and designing systems to combat such disinformation.

To justify this, we developed a classifier to distinguish unreliable media sourced health news from reliable articles. Along with the word vectors built from the body of the articles we utilized these features from our analysis to build the classification model. Experiment results demonstrated promising accuracy of our approach. Further, we identified and analyzed the most-important features behind the model. Our analysis shows that 4 of our engineered features were among the 20 most-important features.

## 2 RELATED WORK

Fake news is an emerging topic which draws continuous attractions of the researchers. There are already some approaches for detecting fake news computationally. Potthast et al. use lexical and syntactic features, capture specific writing styles to build their fake news detection model [27]. Rubin et al. also capture manipulation in writing styles to identify fake news [2]. Apart from writing styles, Gupta et al. use visual features to identify fake images that are created intentionally [13]. [23, 39] employ external knowledge bases to verify the claims in the news contents. All the mentioned approaches explored fake news identification problem from a general perspective. None of them examined domain specific features for example features related to health misinformation which could be instrumental to identify fake news more effectively.

There has been extensive work on how scientific medical research outcomes should be disseminated to general people by following health journalism protocols [4, 5, 18, 31, 33]. For instance, Lopes et al. suggest that it is necessary to integrate journalism studies, strategic communication concepts, and health professional knowledge to successfully disseminate professional findings. Some researchers particularly focused on the spread of health misinformation in social media [22]. For example, Chou et al. discuss the gap in understanding the effect of health misinformation, challenges in developing a framework for research and practice on social media [3]. [11] analyzes Zika [5] related misinformation in Twitter. In particular, it shows that tracking health misinformation in social media is not trivial, and requires some expert supervision. It exploited crowdsource to annotate a collection of Tweets and used the annotated data to build a rumor classification model. One limitation of this work is that the used dataset is too small (6 rumors) to make a general conclusion. Moreover, it didn't consider the features in the actual news articles unlike us. [12] examines the individuals on social media that are posting questionable health-related information, and in particular promoting cancer treatments which have been shown to be ineffective. It develops a feature based supervised classification model to automatically identify users who are comparatively more susceptible to health misinformation. There are other works which focus on automatically identifying health misinformation. For example, [19] developed a classifier to detect misinformative posts in health forums. One of the limitations of this work is that the training data is only labeled by two individuals. Researchers have also worked on building tools that can help a user to easily consume health information. [20] developed the "VAC Medi+board", an interactive visualization platform integrating Twitter data and news coverage from a reliable source called MediSys[6]. It covers public debate related to vaccines and helps users to easily browse health information on a certain vaccine-related topic.

Our study significantly differs from these already existing researches. Instead of depending on a small sample of health hoaxes like some of the existing works, we take a different approach and focus on the source outlets. This gives us the benefit of investigating with a larger dataset. We investigate the journalistic practice of reliable and unreliable health outlets, an area which has not been studied to the best of our knowledge.

## 3 DATA PREPARATION

For investigating how reliable media outlets and unreliable outlets portray health information, we need a reasonably sized collection of health-related news articles from these two sides. Unfortunately, there is not an available dataset which is of adequate size. For this reason, we prepare a dataset of about 30, 000 health-related news articles disseminated by reliable or unreliable outlets within the years 2015 – 2018. Below, we describe the preparation process in detail.

### 3.1 Media Outlet Selection

The first challenge is to identify reliable and unreliable outlets. The matter of reliability is subjective. We decided to consider the outlets which have been cross-checked as reliable or unreliable by credible sources.

*3.1.1 Reliable Media.* We identified 29 reliable media outlets from three sources– **i)** 11 of them are certified by the Health On the Net [25], a non-profit organization that promotes transparent and reliable health information online. It is officially related with the World Health Organization (WHO) [36]. **ii)** 8 from U.S. government's health-related centers and institutions (e.g., CDC, NIH, NCBI), and **iii)** 10 from the most circulated broadcast [30] mainstream media outlets (e.g., CNN, NBC). Note, the mainstream outlets generally have a separate section for health information (e.g., https://www.cnn.com/health). As our goal is to collect health-related news, we restricted ourselves to their health portals only.

*3.1.2 Unreliable Media.* Dr. Melissa Zimdars, a communication and media expert, prepared a list of false, misleading, clickbaity, and satirical media outlets [38, 40]. Similar lists are also maintained by Wikipedia [37] and informationisbeautiful.net [15]. We identified 6 media outlets which primarily spread health-related misinformation and are present in these lists. Another source for identifying unreliable outlets is *Snopes.com*, a popular hoax-debunking website that fact-checks news of different domains including health. We followed the health or medical hoaxes debunked by *Snopes.com* and identified 14 media outlets which sourced those hoaxes. In total, we identified 20 unreliable outlets. Table 1 lists the Facebook page ids of all the reliable and unreliable outlets that have been used in this study.

[5] https://en.wikipedia.org/wiki/Zika_virus

[6] http://medisys.newsbrief.eu

**Table 1: List of Facebook page ids of the reliable and unreliable outlets. Some of them are unavailable now.**

| | |
|---|---|
| Reliable | everydayhealth, WebMD, statnews, AmericanHeart, BBCLifestyleHealth, CBSHealth, FoxNewsHealth, WellNYT, latimesscience, tampabaytimeshealth, philly.comhealth, AmericanHeart, AmericanCancerSociety, HHS, CNNHealth, cancer.gov, FDA, mplus.gov, NHLBI, kidshealthparents, ahrq.gov, healthadvocateinc, HealthCentral, eMedicineHealth, C4YWH, BabyCenter, MayoClinic, MedicineNet, healthline |
| Unreliable | liveahealth, healthexpertgroup, healthysolo, organichealthcorner, justhealthylifestyle1, REALfarmacyCOM, thetruthaboutcancer, BookforHealthyLife, viralstories.bm, justhealthyway, thereadersfile, pinoyhomeremedies, onlygenuinehealth, greatremediesgreathealth, HealthRanger, thefoodbabe, AgeofAutism, HealthNutNews, consciouslifenews, HealthImpactNews |

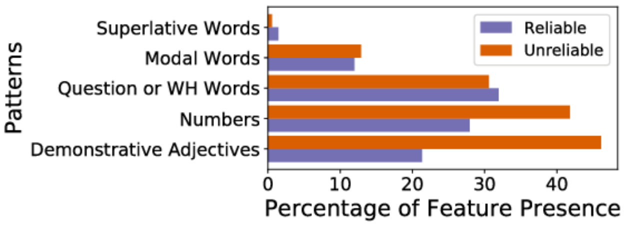


**Figure 1: Distribution of clickbait patterns**

## 3.2 Data Collection

The next challenge is to gather news articles published by the selected outlets. We identified the official Facebook pages of each of the 49 media outlets and collected all the link-posts [7] shared by the outlets within January 1, 2015 and April 2, 2018 [8] using Facebook Graph API. For each post, we gathered the corresponding news article link, the status message, and the posting date.

*3.2.1 News Article Scraping.* We used a Python package named Newspaper3k [9] to gather the news article related data. Given a news article link, this package provides the headline, body, author name (if present), and publish date of the article. It also provides the visual elements (image, video) used in an article. In total, we collected data for 29, 047 articles from reliable outlets and 15, 017 from unreliable outlets.

*3.2.2 Filtering non-Health News Articles.* Even though we restricted ourselves to health-related outlets, we observed that the outlets also published or shared non-health (e.g., sports, entertainment, weather) news. We removed these non-health articles from our dataset and only kept *health*, *food & drink*, or *fitness & beauty* related articles. Specifically, for each news article, we used the document categorization service provided by Google Cloud Natural Language API [10] to determine its topic. If an article doesn't belong to one of the three above mentioned topics, it is filtered out. This step reduced the dataset size to 27, 589; 18, 436 from reliable outlets and 9, 153 from unreliable outlets. We used this health-related dataset only in all the experiments of this paper.

# 4 CONTENT ANALYSIS

Using this dataset, we conduct content analysis to examine structural, topical, and semantic differences in health news from reliable and unreliable outlets.

## 4.1 Structural Difference

The headline is a key element of a news article. According to a study done by American Press Institute and the Associated Press [17], only 4 out of 10 Americans read beyond the headline. So, it is important to understand how reliable and unreliable outlets construct the headlines of their health-related news. According to to [1], a long headline results in significantly higher click-through-rate (CTR) than a short headline does. We observe that the average headline length of an article from reliable outlets and an article from unreliable outlets is 8.56 words and 12.13 words, respectively. So, on average, an unreliable outlet's headline has a higher chance of receiving more clicks or attention than a reliable outlet's headline. To further investigate this, we examine the *clickbaityness* of the headlines. The term clickbait refers to a form of web content (headline, image, thumbnail, etc.) that employs writing formulas, linguistic techniques, and suspense creating visual elements to trick readers into clicking links, but does not deliver on its promises [10]. Chen et al. [2] reported that clickbait usage is a common pattern in false news articles. We investigate to what extent the reliable and unreliable outlets use clickbait headlines in their health articles. For each article headline, we test whether it is a clickbait or not using two supervised clickbait detection models– a sub-word embedding based deep learning model [28] and a feature engineering based Multinomial Naive Bayes model [24]. Agreement between these models was measured as 0.44 using Cohen's $\kappa$. We mark a headline as a clickbait if both models labeled it as clickbait. We observe, 27.29% (5,031 out of 18,436) of the headlines from reliable outlets are click bait. In unreliable outlets, the percentage is significantly higher, 40.03% (3,664 out of 9,153). So, it is evident that the unreliable outlets use more click baits than reliable outlets in their health journalism.

We further investigate the linguistic patterns used in the clickbait headlines. In particular, we analyze the presence of some common patterns which are generally employed in clickbait according to [1, 26]. The patterns are-

- Presence of demonstrative adjectives (e.g., this, these, that)
- Presence of numbers (e.g., 10, ten)
- Presence of modal words (e.g., must, should, could, can)
- Presence of question or WH words (e.g., what, who, how)
- Presence of superlative words (e.g., best, worst, never)

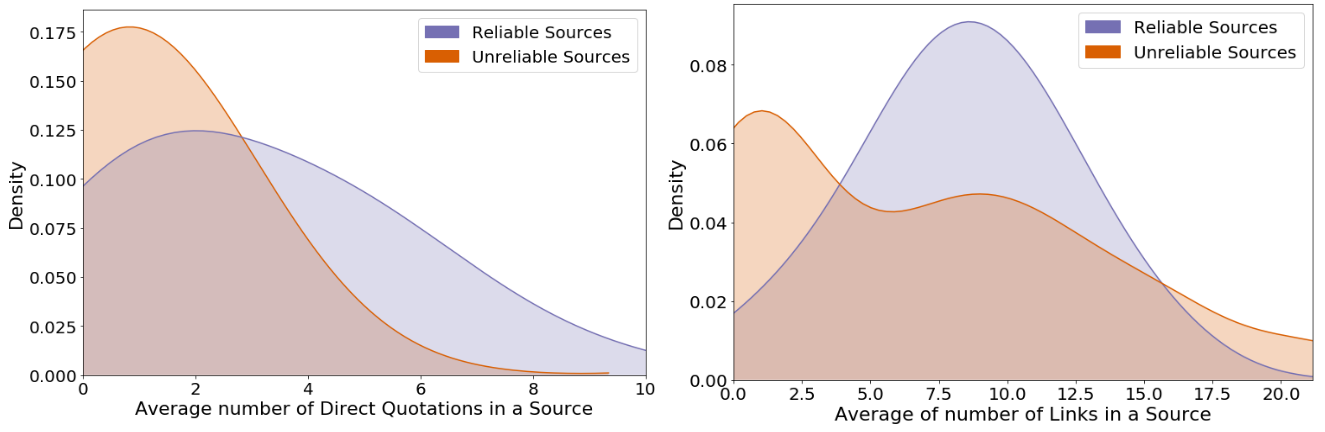
Figure 1 shows the distribution of these patterns among the clickbait headlines of reliable and unreliable outlets. Note, one headline may contain more than one pattern. For example, this headline *"Are these the worst 9 diseases in the world?"* contains four of the above patterns. This is the reason why summation of the percentages isn't equal to one. We see that unreliable outlets use demonstrative adjective and numbers significantly more compared to the reliable outlets.

[7] Facebook allows posting status, pictures, videos, events, links, etc. We collected the link type posts only.

[8] After that, Facebook limited access to pages as a result of the Cambridge Analytica incident.

[9] https://newspaper.readthedocs.io/en/latest/

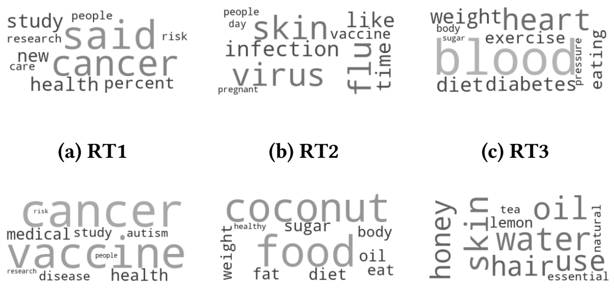
[10] https://cloud.google.com/natural-language/



**(a) Quotation**

**(b) Link**

**Figure 2: Distribution of average number of quotation/link per article from reliable and unreliable outlets.**



**(a) RT1** **(b) RT2** **(c) RT3**

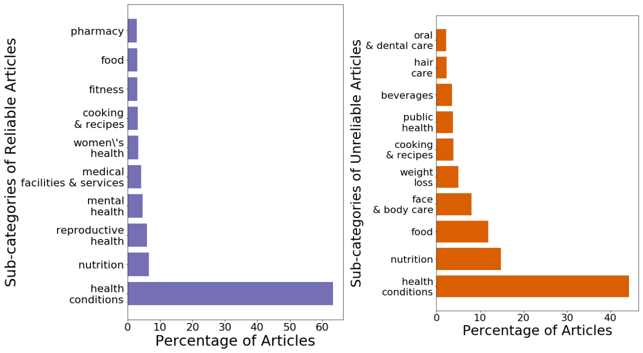
**(d) UT1** **(e) UT2** **(f) UT3**

**Figure 3: Topic modeling ($k = 3$) of articles from reliable outlets (top, denoted as RT) and from unreliable outlets (bottom, denoted as UT).**

## 4.2 Topical Difference

All the articles which we examined are health-related. However, the health domain is considerably broad and it covers many topics. We hypothesize that there are differences between the health topics which are discussed in reliable outlets and in unreliable outlets. To test that, we conduct an unsupervised and a supervised analysis.

*4.2.1 Topic Modeling.* We use *Latent Dirichlet Allocation(LDA)* algorithm to model the topics in the news articles. The number of topics, $k$, was set as 3. Figure 3 shows three topics for each of the outlet categories. Each topic is modeled by the top-10 important words in that topic. The font size of words is proportional to the importance. Figure 3a and 3d indicate that "cancer" is a common topic in reliable and unreliable outlets. Although, the words *study*, *said*, *percent*, *research*, and their font sizes in Figure 3a indicate that the topic "cancer" is associated with research studies, facts, and references in reliable outlets. On the contrary, unreliable outlets have the words *vaccine*, *autism*, and *risk* in Figure 3d which suggests the discussion regarding how vaccines put people under autism and cancer risk, an unsubstantiated claim, generally propagated by unreliable media [11, 12]. Figure 3e and 3f suggest the discussions about weight loss, skin, and hair care products (e.g., essential oil, lemon). Topics in Figure 3b and 3c discuss mostly flu, virus, skin infection, exercise, diabetes and so on.



**(a) Reliable**

**(b) Unreliable**

**Figure 4: Top-10 topics in reliable and unreliable outlets.**

*4.2.2 Topic Categorization.* In addition to topic modeling, we categorically analyze the articles' topics using Google Cloud Natural Language API [13]. Figure 4 shows the top-10 topics in the reliable and unreliable outlets. In the case of reliable, the distribution is significantly dominated by *health condition*. On the other hand, in the case of unreliable outlets, percentages of *nutrition* and *food* are noticeable. Only 4 of the 10 categories are common in two outlet groups. Unreliable topics have *weight loss*, *hair care*, *face & body care*. This finding supports our claim from topic modeling analysis.

[11] https://www.webmd.com/brain/autism/do-vaccines-cause-autism
[12] https://www.skepticalraptor.com/skepticalraptorblog.php/polio-vaccine-causes-cancer-myth/
[13] https://cloud.google.com/natural-language/

## 4.3 Semantic Difference

We analyze what efforts the outlets make for a logical and meaningful health news. Specifically, we consider to what extent the outlets use quotations and hyperlinks. Use of quotation and hyperlinks in a news article is associated with credibility [6, 34]. Presence of quotation and hyperlinks indicates that an article is logically constructed and supported with credible factual information.

*4.3.1 Quotation.* We use the Stanford QuoteAnnotator [14] to identify the quotations from a news article. Figure 2a shows density plots of the number of quotations per article for reliable and unreliable outlets. We observe that unreliable outlets use less number of quotations compared to reliable outlets. We find that the average number of quotations per article is 1 and 3 in unreliable and reliable outlets, respectively. This suggests that the reliable outlet sources articles are more credible and unreliable outlets are less credible.

*4.3.2 Hyperlink.* We examine the use of the hyperlink in the articles. On average, a reliable outlet sourced article contains 8.4 hyperlinks and an unreliable outlet sourced article contains 6.8 hyperlinks. Figure 2b shows density plots of the number of links per article for reliable and unreliable outlets. The peaks indicate that most of the articles from reliable outlets have close to 8 (median) hyperlinks. On the other hand, most of the unreliable outlet articles have less than 2 hyperlinks. This analysis again suggests that the reliable sourced articles are more credible than unreliable outlet articles.

## 5 SOURCE CLASSIFICATION

In this section, we describe our classification model design process. The model leverages the above mentioned features and uses other generic text features.

## 5.1 Feature Extraction

We extract two types of features from the news articles. They are described below.

**Word (W):** We used n-grams ($n = 1, 2$) that are present in the body of the articles to build *tf-idf* word features. We discarded rare n-grams that appeared in less than five articles. We also set the maximum number of n-grams to $5,000$ to avoid over-fitting. These $5,000$ n-grams are the most-frequent in the whole corpus. We didn't apply stemming. However, we removed the English stop words. We carefully removed any mention of the source of the article from the features. We also removed the author names. These steps ensure that our model is not learning news outlet or author categorization.

**Extracted Features (EF):** This set of features were extracted from the news articles following the section 4. The features are- *headline length*, *5 linguistic patterns from headline*, *clickbaitiness*, *number of direct quotes*, and *number of hyperlinks*. In addition, we considered the presence of social media mention as a source in the news article as a feature. To identify whether an article contains any social media mention, we applied the approach described in [29].

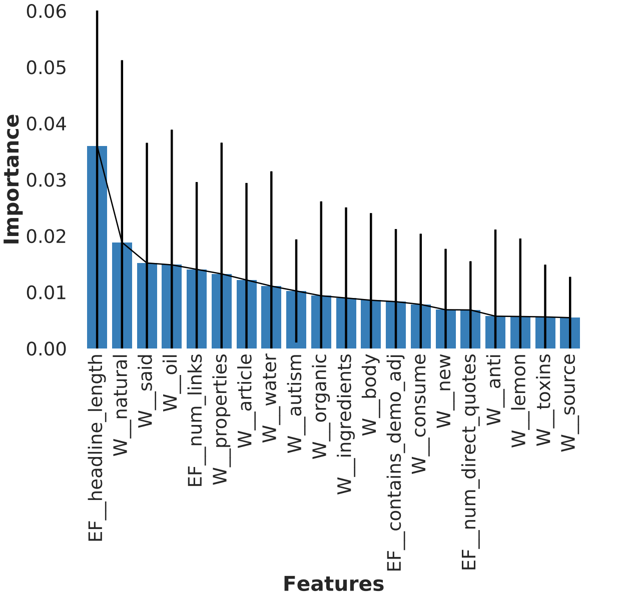


**Figure 5: Feature Importance**

## 5.2 Feature Importance

To compare the effectiveness of the features, we measured the importance of each feature. We trained a random forest classifier for which we used 50 estimators to measure the importance of features in constructing each decision tree. The overall importance of a feature is its average importance over all the trees. Figure 5 shows the importance of the 20 most-important features in the forest. The black solid lines indicate the standard deviations of the importance values. To distinguish between ngram features (W) and extracted features (EF), we use their short form as prefixes. We find that 4 out of 10 EF's make to top 20 including the top spot. Headline length, number of direct quotes, and hyperlinks proves to be very important features. It is unsurprising that words such as autism, toxic, lemon, mixture are present in the most important features.

## 5.3 Classification

Our dataset contains 27,589 health related articles. We preformed 5-fold cross-validation using several classical machine learning methods, including Multinomial Naive Bayes, Linear Support Vector Classifier (*SVC*), Random Forest, and Logistic Regression. Among them, Multinomial Naive Bayes and *SVC* were set with default parameters. We tried Logistic Regression and Random Forest with random state 0. For Random Forest, the number of estimators and the maximum depth were set to 200 and 3 respectively. We conducted experiments with three combinations of features - Words or n-grams (*W*), Extarcted Features (*EF*), and Word + Extracted Features (*W + EF*). For all the combinations, *SVC* outperformed other models. So, we reported here the performance of *SVC* only. Table 2 shows the performance in terms of precision, recall, and f-measure. Feature set, *W* outperforms *EF* by a large margin. The combination of both feature sets (*W+EF*) improves the overall performance, 96% average (macro) F-measure.

[14] https://stanfordnlp.github.io/CoreNLP/quote.html

**Table 2: Classification Report for different combination of features**

| Features | Labels | Precision | Recall | F-1 |
|---|---|---|---|---|
| Word (W) | Unreliable | 0.94 | 0.92 | 0.93 |
| | Reliable | 0.96 | **0.97** | **0.97** |
| | **Macro-Avg** | 0.95 | **0.95** | 0.95 |
| Extracted Features (EF) | Unreliable | 0.76 | 0.47 | 0.58 |
| | Reliable | 0.78 | 0.93 | 0.85 |
| | **Macro-Avg** | 0.77 | 0.70 | 0.72 |
| W + EF | Unreliable | **0.95** | **0.93** | **0.94** |
| | Reliable | **0.97** | **0.97** | **0.97** |
| | **Macro-Avg** | **0.96** | **0.95** | **0.96** |

## 6 CONCLUSION AND FUTURE WORK

In this paper, we closely looked at structural, topical, and semantic differences between articles from reliable and unreliable outlets. Our findings reconfirm some of the existing claims such as unreliable outlets use clickbaity headlines to catch the attention of users. In addition, this study finds new patterns that can potentially help separate health disinformation. For example, we find that less quotation and hyperlinks are more associated with unreliable outlets. To show the effectiveness of the features we discovered, we built supervised classification models using the features which showed satisfactory performance. We also conducted an experiment for measuring the importance of the features which revealed the potentiality of the extracted features in distinguishing reliable sourced health news from unreliable sourced health news. However, there are some limitations to this study. For instance, we didn't consider the videos, cited experts, comments of the users, and other information. In the future, we want to overcome these limitations and leverage the findings of this study to combat health disinformation.

## REFERENCES

[1] Chris Breaux. (accessed September 28, 2018). *"You'll Never Guess How Chartbeat's Data Scientists Came Up With the Single Greatest Headline"*. http://blog.chartbeat.com/2015/11/20/youll-never-guess-how-chartbeats-data-scientists-came-up-with-the-single-greatest-headline/

[2] Yimin Chen, Niall J Conroy, and Victoria L Rubin. 2015. Misleading online content: Recognizing clickbait as false news. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*. ACM, 15–19.

[3] Wen-Ying Sylvia Chou, April Oh, and William MP Klein. 2018. Addressing health-related misinformation on social media. *Jama* 320, 23 (2018), 2417–2418.

[4] Nicole K Dalmer. 2017. Questioning reliability assessments of health information on social media. *Journal of the Medical Library Association: JMLA* 105, 1 (2017), 61.

[5] Irja Marije de Jong, Frank Kupper, Marlous Arentshorst, and Jacqueline Broerse. 2016. Responsible reporting: neuroimaging news in the age of responsible research and innovation. *Science and engineering ethics* 22, 4 (2016), 1107–1130.

[6] Juliette De Maeyer. 2012. The journalistic hyperlink: Prescriptive discourses about linking in online news. *Journalism Practice* 6, 5-6 (2012), 692–701.

[7] Katie Forster. (accessed October 30, 2018). *Revealed: How dangerous fake health news conquered Facebook.* https://www.independent.co.uk/life-style/health-and-families/health-news/fake-news-health-facebook-cruel-damaging-social-media-mike-adams-natural-health-ranger-conspiracy-a7498201.html

[8] Susannah Fox. (accessed October 30, 2018). *The social life of health information.* http://www.pewresearch.org/fact-tank/2014/01/15/the-social-life-of-health-information/

[9] Gaby Galvin. (accessed October 30, 2018). *How Bots Could Hack Your Health.* https://www.usnews.com/news/healthiest-communities/articles/2018-07-24/how-social-media-bots-could-compromise-public-health

[10] Bryan Gardiner. (accessed September 28, 2018). *"You'll Be Outraged at How Easy It Was to Get You to Click on This Headline"*. https://www.wired.com/2015/12/psychology-of-clickbait/

[11] Amira Ghenai and Yelena Mejova. 2017. Catching Zika Fever: Application of Crowdsourcing and Machine Learning for Tracking Health Misinformation on Twitter. In *Healthcare Informatics (ICHI), 2017 IEEE International Conference on.* IEEE, 518–518.

[12] Amira Ghenai and Yelena Mejova. 2018. Fake Cures: User-centric Modeling of Health Misinformation in Social Media. In *2018 ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW).* ACM.

[13] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. 2013. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web.* ACM, 729–736.

[14] Muiris Houston. (accessed October 31, 2018). *Measles back with a vengeance due to fake health news.* https://www.irishtimes.com/opinion/measles-back-with-a-vengeance-due-to-fake-health-news-1.3401960

[15] informationisbeautiful.net. 2016. Unreliable/Fake News Sites & Sources. https://docs.google.com/spreadsheets/d/1xDDmbr54qzzG8wUrRdxQl$_C$1dixJSIYqQUaXVZBqsJs.

[16] Mathew Ingram. (accessed October 30, 2018). *The internet didn't invent viral content or clickbait journalism âĂŤ there's just more of it now, and it happens faster.* https://gigaom.com/2014/04/01/the-internet-didnt-invent-viral-content-or-clickbait-journalism-theres-just-more-of-it-now-and-it-happens-faster/

[17] American Press Institute and the Associated Press-NORC Center for Public Affairs Research. (accessed September 28, 2018). *The Personal News Cycle: How Americans choose to get their news.* https://www.americanpressinstitute.org/publications/reports/survey-research/how-americans-get-news/

[18] Marjorie Kagawa-Singer and Shaheen Kassim-Lakha. 2003. A strategy to reduce cross-cultural miscommunication and increase the likelihood of improving health outcomes. *Academic Medicine* 78, 6 (2003), 577–587.

[19] Alexander Kinsora, Kate Barron, Qiaozhu Mei, and VG Vinod Vydiswaran. 2017. Creating a Labeled Dataset for Medical Misinformation in Health Forums. In *Healthcare Informatics (ICHI), 2017 IEEE International Conference on.* IEEE, 456–461.

[20] Patty Kostkova, Vino Mano, Heidi J Larson, and William S Schulz. 2016. Vac medi+ board: Analysing vaccine rumours in news and social media. In *Proceedings of the 6th International Conference on Digital Health Conference.* ACM, 163–164.

[21] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[22] Felisbela Lopes, Teresa Ruão, Zara Pinto Coelho, and Sandra Marinho. 2009. Journalists and health care professionals: what can we do about it?. In *2009 Annual Conference of the International Association for Media and Communication Research (IAMCR),"Human Rights and Communication".* 1–15.

[23] Amr Magdy and Nayer Wanas. 2010. Web-based statistical fact checking of textual documents. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents.* ACM, 103–110.

[24] Saurabh Mathur. (accessed September 24, 2018). *Clickbait Detector.* https://github.com/saurabhmathur96/clickbait-detector

[25] HEALTH ON THE NET. (accessed September 24, 2018). . https://www.hon.ch/en/

[26] Matthew Opatrny. (accessed September 28, 2018). *"9 Headline Tips to Help You Connect with Your Target Audience".* https://www.outbrain.com/blog/9-headline-tips-to-help-marketers-and-publishers-connect-with-their-target-audiences/

[27] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638* (2017).

[28] Md Main Uddin Rony, Naeemul Hassan, and Mohammad Yousuf. 2017. Diving Deep into Clickbaits: Who Use Them to What Extents in Which Topics with What Effects?. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017.* ACM, 232–239.

[29] Md Main Uddin Rony, Mohammad Yousuf, and Naeemul Hassan. 2018. A Large-scale Study of Social Media Sources in News Articles. *arXiv preprint arXiv:1810.13078* (2018).

[30] Michael Schneider. (accessed September 24, 2018). *Most-Watched Television Networks: Ranking 2016's Winners and Losers.* https://www.indiewire.com/2016/12/cnn-fox-news-msnbc-nbc-ratings-2016-winners-losers-1201762864/

[31] Gary Schwitzer. 2008. How do US journalists cover treatments, tests, products, and procedures? An evaluation of 500 stories. *PLoS medicine* 5, 5 (2008), e95.

[32] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.

[33] Miriam Shuchman and Michael S Wilkes. 1997. Medical scientists and health news reporting: a case of miscommunication. *Annals of Internal Medicine* 126, 12 (1997), 976–982.

[34] S Shyam Sundar. 1998. Effect of source attribution on perception of online news stories. *Journalism & Mass Communication Quarterly* 75, 1 (1998), 55–68.

[35] Emily K Vraga and Leticia Bode. 2017. Using Expert Sources to Correct Health Misinformation in Social Media. *Science Communication* 39, 5 (2017), 621–645.

[36] World Health Organization (WHO). (accessed September 24, 2018). . http://www.who.int/

[37] Wikipedia. (accessed September 24, 2018). *List of fake news websites*. https://bit.ly/2moBDvA

[38] Wikipedia. (accessed September 24, 2018). *Wikipedia:Zimdars' fake news list*. https://bit.ly/2ziHafj

[39] You Wu, Pankaj K Agarwal, Chengkai Li, Jun Yang, and Cong Yu. 2014. Toward computational fact-checking. *Proceedings of the VLDB Endowment* 7, 7 (2014), 589–600.

[40] Melissa Zimdars. 2016. My 'fake news list' went viral. But made-up stories are only part of the problem. https://www.washingtonpost.com/posteverything/wp/2016/11/18/my-fake-news-list-went-viral-but-made-up-stories-are-only-part-of-the-problem.