12-2019

Rating News Claims: Feature Selection and Evaluation

Izzat Alsmadi

Michael J. O'Brien

Mathematical Biosciences and Engineering

MBE, 17(3): 1922–1939.
DOI: 10.3934/mbe.2020101
Received: 17 September 2019
Accepted: 16 December 2019
Published: 19 December 2019

http://www.aimspress.com/journal/MBE

Research article

Rating news claims: Feature selection and evaluation

Izzat Alsmadi [1,] * and Michael J. O'Brien [2]

[1] Department of Computing and Cyber Security, Texas A&M University–San Antonio, San Antonio, Texas 78224, USA

[2] Office of the Provost, Texas A&M University–San Antonio, San Antonio, Texas 78224, USA

* **Correspondence:** Email: ialsmadi@tamusa.edu; Tel: + 1–210-784-2313.

Abstract: News claims that travel the Internet and online social networks (OSNs) originate from different, sometimes unknown sources, which raises issues related to the credibility of those claims and the drivers behind them. Fact-checking websites such as Snopes, FactCheck, and Emergent use human evaluators to investigate and label news claims, but the process is labor- and time-intensive. Driven by the need to use data analytics and algorithms in assessing the credibility of news claims, we focus on what can be generalized about evaluating human-labeled claims. We developed tools to extract claims from Snopes and Emergent and used public datasets collected by and published on those websites. Claims extracted from those datasets were supervised or labeled with different claim ratings. We focus on claims with definite ratings—false, mostly false, true, and mostly true, with the goal of identifying distinctive features that can be used to distinguish true from false claims. Ultimately, those features can be used to predict future unsupervised or unlabeled claims. We evaluate different methods to extract features as well as different sets of features and their ability to predict the correct claim label. By far, we noticed that OSN websites report high rates of false claims in comparison with most of the other website categories. The rate of reported false claims is higher than the rate of true claims in fact-checking websites in most categories. At the content-analysis level, false claims tend to have more negative tones in sentiments and hence can provide supporting features to predict claim classification.

Keywords: feature extraction; information credibility; online social networks; predictive models

1. Introduction

With the evolution of the Internet, online social networks (OSNs) dominate how users exchange

information. Unlike with classical news outlets such as television and newspapers, OSN users can create and exchange news, information, and the like. As a result, problems related to information credibility can begin to grow. False news can be seen as a new form of malware similar to worms in that it spreads fast through the Internet. The payload of false news is not to harm users' machines or to steal their information but rather to harm their knowledge and confuse them or manipulate their opinions and decisions. The problem is more serious with the heavy dependence currently placed on the Internet, OSNs, and smartphones as main sources of information. Current Internet information models do not require content generators to perform any fact-checking. Fact-checking websites such as Snopes, FactCheck, and Emergent represent isolated efforts to deal with information-credibility problems. Importantly, those efforts require significant human resources, and some websites (e.g., Emergent) struggle to stay alive.

Driven by these challenges in information credibility, we explored the use of data analytics in the information-credibility assessment. To that end, we used human efforts in supervised or rated claims, i.e., through websites such as Snopes, FactCheck, and Emergent, to determine what we can learn from claims in terms of predicting their rating based on content and cited websites. We evaluated features related to the contents of the claims and also categorized websites based on the nature of how they report true or false claims.

Our paper makes the following contributions:

- A model to extract hybrid features from claims' content and metadata in order to automatically predict a claim's label (e.g., whether the claim is true or false).
- A new model to rate websites based on the volume of false or true claims they report. Such a rating can be used to weigh the credibility of those websites in reporting claims. The model is dynamic in order to enable such ratings to change continuously or to be updated.

2. Background

Early literature on information-credibility assessment identified five criteria that users should employ in their assessments of the credibility of Internet-based information: accuracy, authority, objectivity, currency, and coverage/scope [1–4]. Whereas for a stable website some of those criteria can be simple to identify, for news articles and OSN posts, identifying criteria such as authority and objectivity is anything but simple.

Most data-analytic, programmatic-based models for credibility assessment focus on enumerating features related to the content, author, source, and followers and then evaluate the impact of those features on the final credibility assessment of the content [e.g., 5]. A 2007 summary of surveys on information-credibility assessment [6] found that regular users rarely checked for any of the criteria outlined above, even if the effort needed was minimal. Another study [7] focused on methods to detect rumors in OSNs. As one of the major categories of inaccurate or unreliable information, rumors have their own characteristics and goals. With the Internet and OSNs, rumors can spread fast, similar to malicious worms. Users are driven by different motives to spread such rumors, especially when they match/support their own minds, desires, and beliefs. The authors in [7] proposed an approach to detect or predict rumors based on data extracted from Twitter. The approach starts from an original rumor-set baseline (labeled manually by users) and then tries to search for identical or similar rumors. Their models, however, showed low precision and, in

many cases, false detections.

Freggeri et al. [8] used a dataset collected from Snopes to track the process of rumor propagation, share, and evolution. They showed that around 25% of collected Snopes rumors were actually accurate or true and that almost half of the collected dataset represented false or inaccurate information. The latter included gray-area information that was partially accurate/inaccurate. Even when misconceptions or inaccurate information are cleared by different sources, there is no guarantee that the same audience that received and spread the original misconceptions will receive and confirm the corrected versions.

Shao et al. [8,9] proposed Hoaxy as a framework for evaluating misinformation sources collected from news websites and OSNs. The system was developed to act as a search engine where users could submit queries, and relevant or related results would be retrieved from "low credibility" websites and "fact-checking" websites. In general, fact-checking websites index a host of low-credibility articles and rate them, which means that just by existing in a fact-checking website, an article will not be credible.

Jin et al. [10] focused on studying rumor Tweets around the 2016 U.S. election. In addition to collecting a large dataset of Tweet rumors, the researchers injected manually verified rumors as a training dataset. They subsequently distinguished between temporal or transient fake news, which lasts for a short time, and fake news, which lasts for a long time. They also identified top-5 fake-news websites based on the volumes of referrals from social-media websites, particularly Facebook and Twitter, to articles on those fake-news websites. As political orientations are typically stable and may not change frequently, users can be driven by a bias to their political orientations to buy into fake news.

Kim et al. [11] discussed the idea of leveraging public users to help in detecting and reducing fake news. The approach involves an algorithm (Curb [learning.mpi-sws.org/curb]) that decides whether an article should be sent for public credibility assessment. This can prevent credibility-assessment systems from being part of the problem and spreading any piece of information created by anyone. Clearly, this will overwhelm users volunteering in credibility assessment. Dealing with such a problem as an optimization issue requires defining a clear goal. In one online-credibility-assessment model that we are evaluating (community-based information-credibility assessment), friends of online users can be the main credibility-assessment team for their friends. Specialized fact-checking websites can be used as a second-stage credibility assessment when certain claims reach a "dispute level." Such an approach can solve another problem discussed by Kim et al. [11], which is that once fake-news stories become large, they are difficult to control, and there is no guarantee that all audiences who were misinformed will be correctly informed by fact-checking the information. If stories can be handled by the people closest to their roots, this will be efficient in terms of using the most relevant public assessors for each possible fake story.

Androniki and Psannis [12] described different types of challenges in OSN analysis. In particular, they focused on challenges related to the automatic analysis of unstructured human text and how to train machines to learn and predict human behavior. This approach can be broadly subsumed under content-based analysis and metadata-based analysis. Data preprocessing and cleaning stages play important role in deciding the overall quality of such analysis.

3. Online-information-credibility-assessment models

Different information-credibility-assessment models exist, and we present just one as an example of those relevant to building automated data-oriented models. Compared to approaches such as ours, which evaluate a claim's input, CRAAP (see below) is an output assessment model that defines high-level goals (e.g., currency and relevance) and evaluates claims based on those goals.

3.1. Goal-oriented models: The CRAAP assessment model

The CRAAP assessment model focuses on five metrics: *C*urrency, *R*elevance, *A*ccuracy, *A*uthority, and *P*urpose.

- **Currency**: An article can be inaccurate because the information in it is outdated. Users should always determine when an article was published/created and correlate that date with the article's content. Between manual assessment and automated assessment, currency can be easily extracted by both methods from evaluated articles.
- **Relevance**: Similar to currency, evaluating relevance in articles is possible when computed by data-analytic tools and even by humans. For example, many data-analytic algorithms exist in document classification, where they can automatically classify a document based on its content to one of the possible different classifiers (e.g., politics, sports, and entertainment). Why can relevance be important to credibility? Because it can reflect knowledge/experience in the article subject. In the information context, relevance can be defined as the relation among the information resource (e.g., the article), users' queries, and perceived information needs [13].
- **Accuracy**: Accuracy can be defined as the degree to which information is correct and trustworthy [14]. Further, is information factual or verifiable? This attribute is the one most relevant to our credibility goal, especially as the definitions of both accuracy and credibility are closely related.
- **Authority:** Authority refers mainly to the content author and content publisher and their own credibility. Credible authors and websites are expected to create credible content and vice versa.
- **Purpose:** With respect to purpose, we are investigating any possible bias/motivation in the content. We evaluate content sentiment as a possible purpose indicator, the goal being to determine whether some significant associations exist among content category, rating, and polarity.

3.2. Credibility assessment in online social networks

In the context of microblogging websites—online social networks (OSNs)—several models take into consideration the nature of the OSN. For example, some Twitter models divide credibility assessment in Twitter among three main entities: event, post/content, and user levels [e.g., 15,16]. A fourth category (propagation-based features) is also often included. In general, credibility assessment for OSNs can be more structured in comparison to (free-form) webpages. The built-in constructs in those OSNs (e.g., Tweet size, hashtags, number of retweets, and followers) can be used as credibility-assessment features.

Unlike high-level reference models that describe general constructs, microblogging, or OSN, models use features specific to each website. For example, whereas all OSN websites include general entities such as content, user, followers, and source, features that belong to each one of those entities vary. For example, Twitter has followers, followings, retweets, and hashtags; Facebook has likes, comments, and friends.

3.3. A credibility-reference model

Evaluation of the different credibility-assessment models described in the literature showed there were six commonalities, each of which can have a unique name in the different OSNs. Additionally, each one of those entities can have different features to be included in the model.

1. **Hosting website**: No doubt, there are many websites *known* to post credible information and many other websites known to post false/inaccurate information. In other words, we should be able to collect features about the hosting website that can be used to assess information credibility.
2. **Content author**: Also, many Internet content authors are known to create/post credible information, and many others are known to create/post the opposite. In credibility claims, authors may not always be known.
3. **Content**: Whereas most researchers focus on the content itself as the main credibility distinguisher, it is not the only criterion in credibility assessment.
4. **Cited references** (e.g., in-links): Cited references can have their own elements to be assessed.
5. **Outlinks**: These can include followers, re-posters, and re-tweeters. They also have their own elements to be assessed.
6. **Article-credibility evaluator**: On credibility websites, users/experts evaluate the credibility of articles and claims' credibility. Knowledge about evaluators can also be used to help understand the human credibility-assessment process. Our ultimate goal is to develop an automated, unsupervised model to predict credibility as an output.

4. A reference model for malicious contents: learn from other models for detecting malicious content

As an alternative to the CRAAP information-credibility, goal-oriented model, we propose the TICO false-information-assessment model: false-information *T*olerance mechanisms, false-information *I*mpact, false-information *C*ontent, and false-information *O*rigin).

In the current context of the information world, false information takes many forms for similar goals of tricking users into believing in some information and taking actions accordingly. Here our goal is to examine how information models of different domains can learn from each other. We use the term "malicious content" to refer to this broad scope. Others use the term "fake news," which may indicate a narrower focus that may not include malicious content, such as spam emails and messages, phishing hyperlinks, and malware. There exist for these domains mature algorithms/models for automatic detection.

4.1. False-information tolerance: detection/protection mechanisms

In classical malware, several methods of detection exist: dictionary-based detection, signature or rule-based detection, and behavior-based detection. Can we apply signature-based detection to uncover false claims? Do different claim originators or websites create identical or similar claims? Similar to rumors, how often do false claims spread, and can they return in different language or wording? Can we measure the similarity between a claim with a known rating and another claim with an unknown rating, and based on the extent to which they match, predict the rating of the

unknown claim? How can we measure content and context similarity, when in some cases, two more letters added to another claim (e.g., "did" and "didn't") can negate a statement and hence change the meaning?

4.2. False-credibility impact/payload

Unlike other malicious contents, the problem with false claims is that their impact, or payload, is indirect and thus difficult to measure. For example, the question of the extent to which Russian OSN interference in the 2016 U.S. elections played a role in the outcome will probably always remain open.

4.3. False-information-content originators/distributors

Malicious-content originators have different goals or drivers, and understanding them might help us detect the content. Marketing and financial gains, for example, are popular drivers in spam/phishing cases. Malicious hyperlinks and some malware such as ransomware are also popular in spam/phishing. In false-claim cases, claims are generated by drivers such as political orientations and/or the desire to attract attention. Similar to dictionary-based detection methods for classical malicious content, can we create dictionary-based or white–blacklists for false claims? Can we learn patterns of behavior about malicious-content originators? Can those patterns be used for future false-claims detection?

To answer those questions, we evaluated a model (below) to track volumes of websites for posting false and true claims. We wanted to determine (a) whether there were patterns related to websites that had larger volumes of one type (e.g., either posting largely true or largely false claims) and (b) whether we could use this to predict future claims originating in those websites.

4.4. False-information content assessment

Can we predict if a file or link, for example, is malicious by investigating its content? In spam and hyperlinked malicious models, researchers evaluate flags, hints, or certain keywords in the content that can be used to predict the nature of that content. Text analysis uses methods such as frequent words, bag-of-words models (BoWs), *n*-grams, named-entity recognition, and word clouds to extract content-based features. This is one of the main focuses below, where we evaluate different methods.

5. Goals and approaches

Again, one of our major motivations is to build predictive analytic models that alert us to the possibility of false claims or fake news based on the content of the claim and/or on where the claim originated or was cited. Similar systems in other domains exist with different levels of accuracy. For example, many email systems or applications employ automatic spam-detection engines that use different techniques related to email contents, source, or origin. Security controls or systems to automatically detect malicious hyperlinks or websites also exist and continuously improve their engines of detection. Similarly, malware-detection engines can automatically detect and block

suspicious malware using dictionary-, signature-, rule-, or behavior-based techniques.

Inspired by those systems, we believe that the current websites that provide human-based claims evaluation, such as Snopes, FactCheck, and Emergent, are impractical. Our goal is to show how feature-based, rule-based, or signature-based techniques can help in the design of automatic detection engines for false claims. In assessing our models, we use human-rated claims based on datasets of rated claims extracted from fact-checking websites.

5.1. Predicting claim rating based on top-reporting websites

Obviously, credible websites should report only credible information/claims, and the opposite is true. Although we understand that some credible websites report false claims for different reasons, we want to evaluate whether there is a clear distinction between websites that report either false or true claims. We developed the following steps to create features from those websites.

- We used Google and Bing search engines to report top websites that reported a claim (excluding fact-checking websites).
- We focused on claims that were false (false or mostly false) or true (true or mostly true). We ignored other categories of claims rating.
- Each website that was listed in a false claim received –1 point, and each website that was listed in a true claim received +1 point. In the end, each website received a "claims-based credibility rank" based on how many true/false claims it had in the dataset.
- Websites were listed as features, with their overall cumulative points as their feature values (e.g., credibility rank). Tables 1 and 2 show a sample of the collected dataset with top-referenced websites crawled from Google and Bing search engines. We note the following observations when collecting top websites that cite claims from Bing and Google:

(1) Search engines enforce different types of rate limitations when crawling webpages. By using two of the most popular search engines, we hoped to reduce some of the bias in using only one engine. The first dataset was built using 5834 claim queries in Bing and 1584 claim queries in Google. For false/true/other claims, the Bing dataset had 3266/793/1775 rated claims and the Google dataset 811/223/549. For each claim, the top 10 websites were retrieved from Bing and the top 20 from Google.

(2) Search engines have their own complex algorithms to rank retrieved search results for a particular query. The overarching element of those algorithms is website popularity, which may have more importance than the relevance of the retrieved results to the query. For our dataset, we did not collect all matched/retrieved weblinks/websites but, again, only the top 10 for Bing and the top 20 for Google. This means that many of those websites that were highly ranked in reporting false or true claims were ranked highly not only because they reported those claims but, more importantly, because they are popular websites.

(3) Using Bing as an example (Table 2), from 3266 false claims, 1422, or about 44%, were reported by Wikipedia in its top 10. Additionally, 357 of 793 true claims, or about 44%, were reported by Wikipedia in its top 10. In other words, given almost the same percentage of reporting between false and true claims, Wikipedia is a reference that gives almost equal weight to false and true claims. The bottom line is, we should look for websites that report false/true claims with a significant percentage difference.

(4) In Bing, the websites Blogspot, Snopes, WordPress, Wikipedia, YouTube, Yahoo, Archive,

and Pinterest report high rates of false claims. What is common among all those websites? They are based on user-driven content, which means that users can create and disseminate almost any content.

Table 1. Website credibility ranking using Google claims search queries.

Website	False	True	Total	Rate
Wikipedia	741	232	−509	3.19
Youtube	205	44	−161	4.65
Quizlet	25	2	−23	12.5
Reddit	60	17	−43	3.52
Cbsnews	80	26	−54	3.07
Washingtonpost	123	49	−74	2.51
Nytimes	252	88	−164	2.86
Theguardian	131	55	−76	2.38
Quora	49	4	−45	12.25
Truth or fiction	76	7	−69	10.85
Dailymail.co.uk	102	30	−72	3.4
Facebook	110	15	−95	7.33
Pinterest	174	38	−136	4.57
USAToday	90	27	−63	3.33
Answers. Yahoo	0	0	0	0
Thesun.co.uk	14	4	−10	3.5
CNN	121	53	−68	2.28
NBCnews	79	28	−51	2.82
Inquisitr	17	3	−14	5.66
NPR.org	66	23	−43	2.86

Table 2. Website credibility ranking using Bing claims search queries.

Website	False	True	Total	Rate %
Blogspot	16872	7854	−9018	2.15
Snopes	8729	3792	−4937	2.30
WordPress	5944	2552	−3392	2.33
Wikipedia	5269	2778	−2491	1.90
YouTube	3081	1320	−1761	2.33
Yahoo	3504	1826	−1678	1.92
Archive	2582	1352	−1230	1.91
Pinterest	3400	2212	−1188	1.54
Daily mail	2307	1154	−1153	2.00
Issuu	2124	1090	−1034	1.95
Reddit	2326	1322	−1004	1.76
Quizlet	2411	1462	−949	1.65
Washington post	1620	774	−846	2.09
Scribd	1472	666	−806	2.21
Quora	1426	720	−706	1.98

5.1.1. Lessons learned

Our preliminary analysis indicates that we can use websites that cite claims as a source for extracting significant features. One of the main goals of many false-news categories, such as rumors, is to impact public views or opinions by reaching more users and websites to cite or mention false news. The process is very dynamic, where false claims will impact the reputation of cited websites and the reputation of cited websites can impact the automatic decision to categorize a claim as false or true. Our experiments showed that from claims-rating assessment, websites should be categorized under different categories, where each category should be handled differently:

- Fact-checking websites: As this is their main role, no wonder we find many claims, false and true. So, from a data-analytics perspective, fact-checking websites should not be rated based on what claims they are citing. However, we can use them for other goals, such as looking for the human-based claim rating on those websites.
- News websites: These are expected to rate more true claims. However, we noticed that when the claim category is politics, for example, some news websites may have political orientations that can impact the nature of the claims they report. In other words, those websites can be used for all types of claims categories except politics. In the politics category, their cited claims can be used to evaluate their orientations.
- OSNs: Websites such as Facebook, Twitter, YouTube, and Instagram are resources for all types of claims. The general website may not be a good predictor of false or true claims, but studying users and groups in those networks can be used to analyze many details of how claims start and circulate. In a previous paper [17], Alsmadi et al. proposed a model to rate OSN users based on the quality/credibility/publicity of the content they create. Claims originated from certain users can be given credibility ratings based on the credibility of the originating users.

5.2. Content-based feature extraction

How can we build models to judge an article's credibility automatically? Here our focus is on feature-based approaches. In our reference model, we described six entities that can be the source of features to be extracted about claims: hosting websites, authors, content, in-links or references cited, out-links (e.g., followers), and evaluators. In all models based on data analytics, the main factor that can ultimately judge model quality is the quality of the collected features. The typical cycle of extracting content-based features starts by using one of the known text-based feature extraction methods (e.g., BoWs, *n*-grams, top words, and named-entity recognition). In supervised models, feature quality is based on their ability to predict class labels with high accuracy.

We used existing datasets about credibility from websites such as Snopes, Politifact, and Emergent that have human-based claims rating. We also built our own datasets from those websites. Similar to other fact-checking websites, Snopes collects public subjects and summarizes the major claim in that subject. The content is also manually classified/categorized under different categories such as politics, news, "fauxtos," "inboxer rebellion," horror, questionable quotes, and media matters. Experts then classify a claim under one of the different "claim rating" categories (e.g., correct attribution, correctly attributed, incorrectly attributed, legend, misattributed, miscaptioned, mixture, mixture of true and false, information, mostly false, mostly true, multiple, outdated, probably false, research in progress, undetermined, unproven, false, and true).

In order to simplify data-analytic activities, we converted Snopes rating into three categories: false (for false and mostly false ratings), true (for true and mostly true ratings), and other (Figure 1). Table 3 shows the top categories with false rates. Here we are trying to answer the question, can we use the claim category (also called "topic" on some websites) as a significant feature to predict category? The answer to the question is that if the majority of the claims are in the false or true categories and not in the "other" category, then category/topic can be a good credibility predictor. This may vary, however, from one category to another.

Figure 1 shows that Snopes' dataset is somewhat imbalanced, where the volume of reported false claims is at least three times more than the volume of reported true claims. In order to improve model accuracy and as claims with label "other" can be broad, in most of our analysis, we focused only on true and false claims.

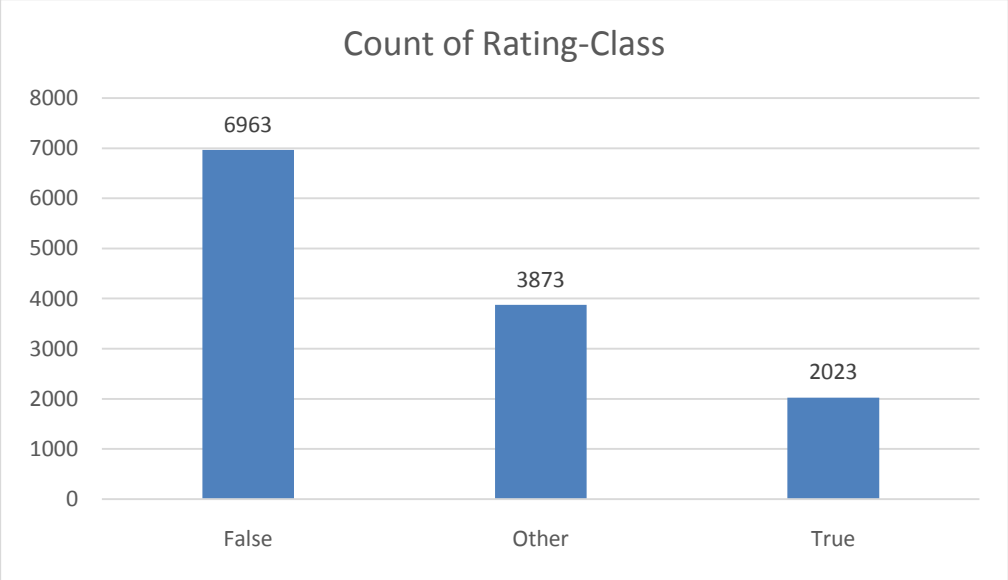

Figure 1. Rating class for our collected Snopes dataset.

Table 3. Top 12 categories with false ratings.

Category	False	Other	True	False Rate (%)
Junk news	1472	33	1	0.977
Fauxtography	1088	548	385	0.538
Politics	660	575	363	0.413
Uncategorized	560	270	122	0.588
Inboxer rebellion	312	233	121	0.468
Media matters	256	13	12	0.911
Politicians	254	146	84	0.525
Entertainment	241	73	74	0.621
Business	198	152	94	0.446
Medical	194	149	57	0.485
Crime	169	182	60	0.411
Humor	141	58	27	0.624

Table 4 shows the dataset from Snopes [18]. Although the time difference between our collected dataset and this dataset is 2–3 years, the percentages of false/true ratings in major categories are very similar.

Table 4. Top 12 categories in Snopes with false ratings.

Topic/Claim	False	Other	True	False Rate (%)
Politics	3000	1080	683	0.630
Fake news	2069	11	8	0.991
Fauxtography	1849	339	262	0.755
Medical	669	136	57	0.776
Political news	522	159	131	0.643
Crime	323	117	68	0.636
Business	213	145	50	0.522
Entertainment	198	20	61	0.710
Food	183	110	13	0.598
Science	171	57	88	0.541
History	131	38	44	0.615

For Politifact and Emergent, most false claims come from two main categories of websites, social networks, and news websites (Table 5).

Table 5. Claim versus claim-source domain for Politifact and Emergent.

Website	False	True	False Rate %
Daily mail	479	172	2.78
Snopes	236	24	9.83
The guardian	83	242	0.34
Mashable	13	77	0.17
USA Today	14	67	0.21
Business insider	26	74	0.35
Buzzfeed	70	25	2.80
Telegraph	21	62	0.34
Independent	49	86	0.57
The verge	5	32	0.16
Express	6	31	0.19
IBtimes	48	24	2.00
NBC news	12	36	0.33
Washington post	58	80	0.73
Inquisitr	39	18	2.17

6. Features based on named-entity recognition (NER)

Text-analysis strategies provide automated mechanisms to summarize texts and create certain

output constructs. Our goal is to evaluate some of those mechanisms to extract features from claims text. Named-entity recognition (NER) is an information-extraction process that seeks to extract and classify named entities in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, and percentages. As one example, we used spaCy 2.0.11 (https://spacy.io) (with the en_core_web_sm model) to annotate the part of speech (PoS) for each claim. The current spaCy model is a convolutional neural network trained on OntoNotes, a large corpus comprising news, conversational telephone speech, weblogs, Usenet newsgroups, broadcast, and talk-show texts. From our dataset of 12,860 entities, spaCy created 2864 entities based on its corpus. Table 6 shows the top entities based on how many times they occurred in different claims, and Table 7 shows the top entity types in the dataset.

Table 6. Top entities from spaCy library.

Entity Text	Entity Type	Count
Several dozen	Cardinal	397
Sixty	Cardinal	397
Thousands	Cardinal	397
Halloween percent	Percent	393
Evidence year	Date	391
Twentieth-century	Date	391
Democrats	NORP	381

Table 7. Top entity types from spaCy library.

Entity Type	Count	Entity Type	Count
Date	198	Quantity	13
Cardinal	74	Ordinal	11
Time	60	Product	10
NORP	23	Person	7
Money	18	Org	2

One interesting and relevant spaCy entity type that showed up frequently in claims' dataset is NORP (nationalities or religious or political groups). Ideally, this should be the focus and major entity type in the claims dataset. Additionally, it should be decomposed to cover finer classifications.

Figure 2 shows the number of entities per claim. Note that the majority of claims have very small numbers of derived entities based on spaCy NER. Libraries such as spaCy NER are designed to be generic. Our analysis triggers the need to generate more focused or domain-specific entities (e.g., for claims or fake-news domain).

In the second spaCy experiment, we divided claims into those rated as false and those rated as true. Our goal was to see if unique entities could be distinguished between the two rating classes. Using the spaCy library, no significant distinction between false and true ratings was noticed. This is largely because the stop-words removal process in the library missed many of the words that looked generic and hence should be eliminated in this preprocessing stage.

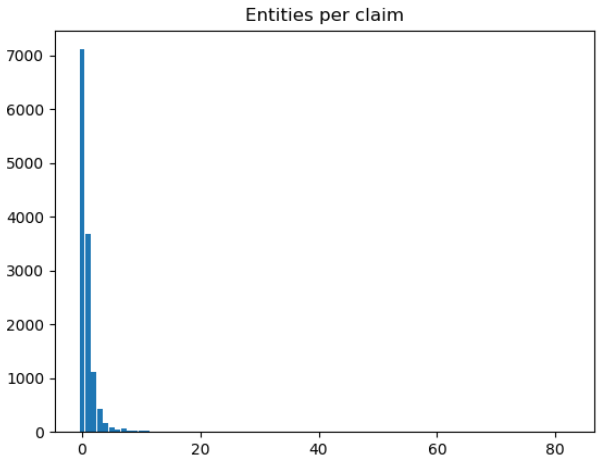

Figure 2. Entities per claim using spaCy NER.

We learned from those two spaCy experiments and some other NER libraries that they extract entities of general types or nature. Such entities may not be able to uniquely distinguish claims for different goals, such as credibility assessment. As a result, some domain-specific research efforts focus on extending those general entity types to include domain-specific entity types.

Our future effort will focus on creating a customized NER for the different domains or categories of domains. The followings are the main steps for building those domain-specific NERs:

- Collect claims from different fact-checking websites along with their categories (examples of Snopes categories include politics, war/anti-war, crime, medical, uncategorized, entertainment, horrors, and guns).
- Extract top N entities in each category and use the category name as the entity type (referred to as "labels" in the spaCy library).
- Train the library with different dataset instances to learn how to recognize the new labels.
- Build a claim-related NER and add it to public libraries such as spaCy.
- Build classification models using NERs as features.

We also evaluated the usage of another library, Algorithmia, which classifies entities under categories similar to what is in spaCy. We used three functions from Algorithmia:

- NER based on the predefined categories; we extracted the number of those entities in each claim.
- Sentiment score.
- Profanity score: Algorithmia employs a simple algorithm to create a profanity score based on certain keywords. Our goal was to evaluate whether false claims contain more profanity terms than true claims.

6.1. Lessons learned

Most of the NER-based features returned small counts due to the small size of most claims. However, we noticed that the profanity score-matched highly with false claims.

7. Politeness analysis

In evaluating claims' politeness levels, we want to evaluate the hypothesis that false claims tend to be less polite than true claims. In other words, can we distinguish claims ratings using politeness metrics or attributes? Politeness library [19] can analyze input text and extract features such as hedges, positive emotion, negative emotion, impersonal pronouns, swearing, negation, and title. Table 8 shows politeness features that have nonzero values, based on the following (excerpted) claim: "Due to budget cuts, U.S. troops deployed overseas are no longer provided with breakfast. So my lil brother is in Afghanistan. he skypes us yesterday and proceeds to tell us the army is cutting their meals to two a day." The unexcepted claim text is very emotional and involves anger, cursing, and the like, which explains why so many features are relatively high.

Table 8. A sample of politeness features.

Hedges	Positive Emotion	Negative Emotion	Impersonal Pronoun	Swearing	Negation	For You	Informal Title
2	16	15	41	1	24	1	2
Reasoning	Actually	First Person Single	Second Person	First Person Plural	Questions	Gratitude	
4	2	12	4	19	3	1	

As text size in most Snopes claims is small, we included the same politeness features for the credibility-evaluator content that annotate the claim and justify its final credibility classification. We distinguished those features coming from the claim annotation content (not the content of the claim itself) by the letter (C) at the end of the feature name. Table 9 shows top features (using SelectKBest method from the SKLearn library) that correlate with rating class. We evaluated the elimination of the (other) class label in credibility, which shows a significant improvement in prediction accuracy.

Table 9. Best politeness features to predict credibility.

Best Features: Credibility 3 class labels		**Best Features: Credibility 2 class labels**	
Specs	**Score**	**Specs**	**Score**
Negative Emotion (C)	73.8	Negative Emotion (C)	49.8
Second Person	67.9	Impersonal Pronoun (C)	46.8
Impersonal Pronoun (C)	53.4	Positive Emotion (C)	44.8
Positive Emotion (C)	45.0	Second Person	32.8
Second Person (C)	40.0	Second Person (C)	18.6
Impersonal Pronoun	34.4	Hedges (C)	15.3
Positive Emotion	23.8	First Person Single	9.2
Informal Title (C)	17.7	Negation (C)	8.6
First Person Single (C)	16.7	Pause (C)	7.1
Hedges (C)	15.4	Gratitude	6.7

We evaluated different classifiers, and they all scored average accuracy (Table 10), which indicates the average ability of politeness features to predict credibility class. When the text of claim or claim annotation is large enough to produce significant values in those politeness features, such features then show a significant correlation with the claim-rating class label, i.e., whether a claim is false or true. False claims have significantly more negative politeness features. On the other hand, and due to the small size of most claim contents, most claims return null values for the majority of politeness features.

Table 10. Accuracy of classifiers on credibility based on claim politeness features.

Credibility classifiers accuracy: 3 class labels		**Credibility classifiers accuracy: 2 class labels**	
Classifier	**Accuracy Score**	**Classifier**	**Accuracy Score**
KNN	0.57	KNN	0.79
XGB Classifier	0.57	XGB Classifier	0.79
Support Vector Machines	0.57	Support Vector Machines	0.79
Ada Boost Classifier	0.57	Linear Discriminant Analysis	0.79
Linear Discriminant Analysis	0.56	Gradient Boosting Classifier	0.79
Gradient Boosting Classifier	0.56	Ada Boost Classifier	0.79
Perceptron	0.53	Stochastic Gradient Descent	0.77
Extra Trees Classifier	0.52	Extra Trees Classifier	0.77
Logistic Regression	0.49	Logistic Regression	0.73
Naive Bayes	0.48	Naive Bayes	0.73
Stochastic Gradient Descent	0.47	Perceptron	0.69
Linear SVC	0.44	Decision Tree	0.68
Decision Tree	0.44	Linear SVC	0.68
Random Forest	0.18	Random Forest	0.27

8. Bag of words (BoW) analysis

Feature extraction in text-analysis models based on frequent terms is common. It might not show very high accuracy in predicting target classes, but it can help us learn some of the relevant terms or subjects. As our datasets of claims are supervised by human experts, i.e., target classification for each claim is verified independently from the classification model and process, the value of the results of BoW experiments can be judged based on their ability to successfully and consistently predict target classes. We want to integrate BoW's frequent terms with features extracted from claim-cited websites. In the first part of this experiment, we conducted the following steps:

1. After all preprocessing steps (e.g., removal of special characters and stop words), extract the top N most-frequent words in the dataset.
2. Loop through the top N frequent words in a stepwise manner (e.g., m-steps in each loop) and find m-terms that show the highest information gain (using the different options in "Sklearn.feature_selection" library).
3. Use a list of classifiers (e.g., logistic regression, linear discriminant analysis, decision tree, and Gaussian methods) to calculate an ensemble accuracy for all classifiers.
4. Increase the number of features as far as accuracy is improving, or else after three cycles of

accuracy inclination, stop the model. Figure 3 shows accuracy, precision, and recall for the first few cycles.

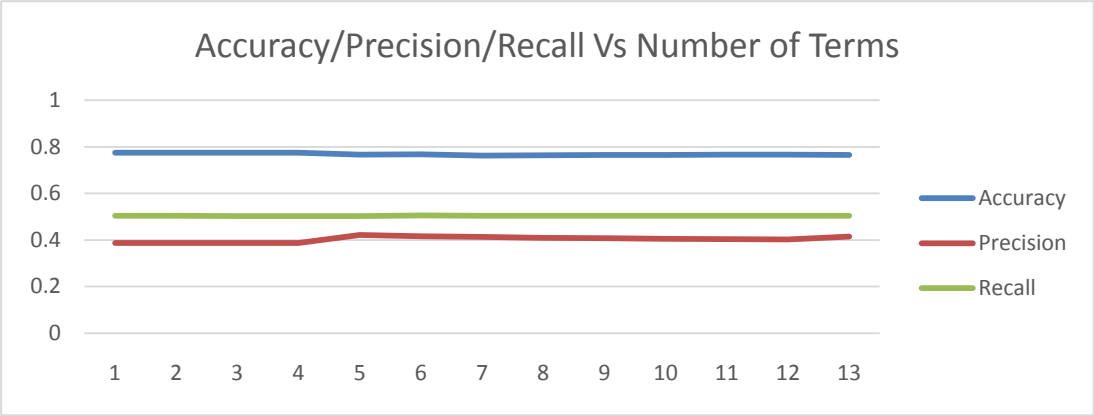

Figure 3. Accuracy, precision, and recall versus terms in the BoWs experiment.

Figure 3 shows that accuracy is stable and does not increase significantly as more of the top frequent words are added. This indicates that those different top words have an independent impact on each other. We realize that the process of discovering top frequent words that can be good false or true claims predictors is experimental and dynamic. An ideal design for automated-based claims rating is to be semi-supervised, where the results of classifying earlier claims can be used to improve accuracy for predicting the rating of future unsupervised claims.

9. Conclusion

In the current Internet-information world, where all humans can be content creators, information accuracy or credibility becomes a serious problem. Practically, many challenges exist in any effort to improve online information credibility or enforce any possible regulations. To this end, our main goal here was to propose and evaluate methods to programmatically extract relevant features that can be good predictors of claim classification or rating as false or true.

There is an urgent need for systems that can automatically label claims through the Internet, as the current human-based effort accomplished by few fact-checking websites is limited and impractical. In those programmable credibility-assessment models, the process will focus on extracting features about the claim (e.g., its content, originator, and hosting website) and use those features to predict the claim rating.

Features of such credibility-assessment models can be extracted from different sources. We focused our experiments and analysis on two sources: claim content and websites that cite those claims. We conducted experiments to evaluate features such as content-named entities, politeness, and profanity. We also evaluated patterns of websites that cite claims and whether there is a consistent trend in some websites to rate largely false or largely true claims.

We experienced different limitations in the different libraries that we used to extract content-based features. For example, most NER libraries are designed for general purposes or domains, limits the possibility of using such entities as features to predict claim rating with significant prediction accuracy.

We also experienced challenges related to the limited text size of claims and the fact that this limited size affects the number of features that can be extracted from the claim content. As an alternative, and to extend the text size of claims, we evaluated adding to the claims, the assessments provided by fact-checking websites.

In the second category of our experiments, evaluating claims' citing websites as features, we used two major search engines, Google and Bing, to crawl websites citing those claims. This is part of an effort to associate claims rating with the different websites and discover whether certain websites or website categories could be used to judge the credibility or rating of a claim.

We noticed that news websites (e.g., CNN and the Guardian) tend to cite more true claims than false claims. On the other hand, OSNs that enable regular users to create online contents (e.g., Wikipedia and YouTube) tend to cite more false claims than true claims. Ultimately, we can use the websites citing a claim to predict its claim rating—for example, a claim that is cited by all news websites will be more likely a true claim, as an alternative to a claim cited by most OSN websites.

In future extensions of this work, we plan to evaluate online collaborating models in which mass Internet users can be part of claims' credibility assessment. Such models can solve the problem of scarce resources in order to label the large volume of data and claims through OSNs. Additionally, such claims will be rated based on the peers/users of the website/person that created such content. For such models, we propose that OSN websites should enable the extension of options in which users can annotate and rate created content based on credibility.

Acknowledgments

The work reported here was funded by Texas A&M–San Antonio University grant no. 218056-20014.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. J. E. Alexander, M. A. Tate, *Web Wisdom: How to Evaluate and Create Information Quality on the Web*. Erlbaum, Hillsdale, NJ (1999).
2. D. S. Brandt, Evaluating information on the Internet, *Comp. Lib.*, **16** (1996), 44–46.
3. J. W. Fritch, R. L. Cromwell, Evaluating Internet resources: Identity, affiliation, and cognitive authority in a networked world, *J. Am. Soc. Inf. Sci. Tec.*, **52** (2001), 499–507.
4. M. Meola, Chucking the checklist: A contextual approach to teaching undergraduates Web-site evaluation, *portal: Lib. Acad.*, **4** (2004), 331–344.
5. L. A. Tran, Evaluation of community web sites: A case study of the Community Social Planning Council of Toronto web site, *Online Inform. Rev.*, **33** (2009), 96–116.
6. U. K. H. Ecker, J. L. Hogan, S. Lewandowsky, Reminders and repetition of misinformation: Helping or hindering its retraction? *J. App. Res. Mem. Cogn.* **6** (2017), 185–192.
7. X. Wang, C. Yu, S. Baumgartner, F. Korn, Relevant document discovery for fact-checking articles, *WWW '18 Companion Proc. Web Conf.*, (2018), 525–533.

8. C. Shao, G. L. Ciampaglia, A. Flammini, F. Menczer, Hoaxy: A platform for tracking online misinformation, *WWW '16 Companion Proc. Web Conf.*, (2016), 745–750.
9. C. Shao, G. L. Ciampaglia, O. Varol, K. C. Yang, A. Flammini, F. Menczer, The spread of low-credibility content by social bots, *Nat. Comm.*, **9** (2018), 4787.
10. Z. Jin, J. Cao, H. Guo, Y. Zhang, Y. Wang, J. Luo, Detection and analysis of 2016 U.S. presidential election-related rumors on Twitter, in *Social, Cultural, and Behavioral Modeling* (eds. D. Lee, Y. R. Lin, N. Osgood, and R. Thomson) (2017), 14–24, Springer, Cham, Switzerland.
11. J. Kim, B. Tabibian, A. Oh, B. Schölkopf, M. Gomez-Rodriguez, Leveraging the crowd to detect and reduce the spread of fake news and misinformation, *Proc. Eleventh ACM Int. Conf. Web Search Data Mining*, (2018), 324–332.
12. S. Androniki, K. E. Psannis, Social networking data analysis tools & challenges, *Future Generat. Comput.r Syst.*, **86** (2018): 893–913.
13. S. Mizzaro, How many relevances in information retrieval?, *Interact. Comp.*, **10** (1998), 303–320.
14. D. Wilson, *Web Site Evaluation*. Market Difference Communications Group, Rocklin, CA (2010).
15. C. Castillo, M. Mendoza, B. Poblete, Information credibility on Twitter, *Proc. 20th Int. Conf. World Wide Web*, (2011), 675–684.
16. S. Sikdar, B. Kang, J. O'Donovan, T. Hollerer, S. Adal, Cutting through the noise: Defining ground truth in information credibility on Twitter, *Human*, (2013), 151–167.
17. I. Alsmadi, X. Dianxiang, J. H. Cho, Interaction-based reputation model in online social networks, in *Proceedings of the Second International Conference on Information Systems Security and Privacy* (2016), 265–272, Science and Technology Publishers, Setúbal, Portugal.
18. A. Nourbakhsh, Who starts and who debunks rumors, https://www.kaggle.com/arminehn/rumor-citation, 2017.
19. M. Yeomans, A. Kantor, D. Tingley, The politeness package: Detecting politeness in natural language, *R J.*, (2018), 489–502.

AIMS Press