Disinformation: analysis and identification

Archita Pathak^{1,2} · Rohini K. Srihari¹ · Nihit Natu^{1,3}

Published online: 18 June 2021

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

We present an extensive study on disinformation, which is defined as information that is false and misleading and intentionally shared to cause harm. Through this work, we aim to answer the following questions:

- Can we automatically and accurately classify a news article as containing disinformation?
- What characteristics of disinformation differentiate it from other types of benign information?

We conduct this study in the context of two significant events: the US elections of 2016 and the 2020 COVID pandemic. We build a series of classifiers to (i) examine linguistic clues exhibited by different types of fake news articles, (ii) analyze “clickbaitiness” of disinformation headlines, and (iii) finally, perform fine-grained, veracity-based article classification through a natural language inference (NLI) module for automated disinformation verification; this utilizes a manually curated set of evidence sources. For the latter, we built a new dataset that is annotated with generic, veracity-based labels and ground truth evidence supporting each label. The veracity labels were formulated based on examining standards used by reputable fact-checking organizations. We show that disinformation derives features from both propaganda and mainstream news, making it more challenging to detect. However, there is significant potential for automating the fact-checking process to incorporate the degree of veracity. We provide error analysis that illustrates the challenges involved in the automated fact-checking task and identifies factors that may improve this process in future work. Finally, we also describe the implementation of a web app that extracts important entities and actions from a given article and searches the web to gather evidence from credible sources. The evidence articles are then used to generate a veracity label that can assist manual fact-checkers engaged in combating disinformation.

✉ Archita Pathak
architap@buffalo.edu

Extended author information available on the last page of the article

1 Introduction

Fake news is an increasing threat to society that has resulted in social unrest worldwide; hence, there is a desperate need for scalable solutions (Ungku et al. 2019). According to a recent Pew Research Study, Americans rate fake news as a more serious problem than racism, climate change, or illegal immigration.¹ Today, as the world is suffering from a COVID pandemic, fake news and ‘infodemics’ are creating further chaos and subverting efforts to bring the pandemic under control.^{2,3} Since it is relatively easy and inexpensive to rapidly disseminate content through social media platforms, there is an urgent need for automated solutions to combat disinformation. Furthermore, AI solutions are currently justifiably being scrutinized for fairness and ethical considerations. Thus, any solution to this problem must not be perceived as censorship or violation of fundamental rights and data privacy (de Cock Buning 2018).

This study’s central focus is disinformation, which is defined as false and misleading information intentionally shared to cause harm. A common tactic is to nefariously combine truthful elements with lies or false conclusions (Wardle and Derakhshan 2017). While disinformation is typically inaccurate, it does not always *have to be* inaccurate. It just has to be misleading (Fallis 2009). Disinformation intentionally attempts to mislead people into believing a manipulated narrative to cause chaos or, in extreme cases, violence. For example, recently, there was a proliferation of stories linking coronavirus to 5G networks that caused people in the UK to burn cellular equipment; several well-known celebrities jumped on this fake news bandwagon.⁴

The paper is organized as follows: in Sect. 2, we build classifiers for (i) distinguishing disinformation from other types of articles, (ii) distinguishing clickbait disinformation headlines from other headlines, and (iii) performing fine-grained classification of disinformation into veracity-based labels. Fine-grained classification provides more nuanced assessments of the veracity of an article, which can be useful to fact-checking organizations who are on the front lines of combating disinformation. Section 3 provides an overview of the FactFinder web app, an automated fact-checking tool developed for evidence-based veracity prediction of an input news article. Finally, we summarize the findings and highlight future work in Sect. 4.

Motivation With the explosion of disinformation, primarily due to viral propagation through social media, there is an urgent need to dramatically scale up fact-checking efforts that are currently dependent on manual curators. There are only a few fact-checking organizations worldwide compared to the volume of daily content requiring verification. The speed at which information is disseminated online limits the ability of manual fact-checking organizations to scale up efforts. Automating more of the fact-checking process enables human fact-checkers to devote more

¹ <https://www.journalism.org/2019/06/05/many-americans-say-made-up-news-is-a-critical-problem-that-needs-to-be-fixed/>.

² <https://www.nature.com/articles/d41586-020-01409-2>.

³ <https://news.un.org/en/story/2020/04/1061592>.

⁴ <https://www.cnet.com/news/fake-5g-coronavirus-theories-have-real-world-consequences/>.

time to complex cases requiring human judgment. This work is motivated to support and accelerate the work of fact-checkers by building a system that can automatically generate supporting or refuting evidence for claims in a given article.

2 Identifying disinformation

2.1 Disinformation and other types of news

Our first study focuses on segregating disinformation from other types of news.⁵ The research in this domain recognizes the type of fake news based on the intent of the author (desire to deceive). In this study, we focus on the following news types:

- *Hoax*: convinces readers of the validity of a paranoia-fueled story (Rashkin et al. 2017)
- *Propaganda*: aims to influence people's mindset to advance a specific agenda (Da San Martino et al. 2019)
- *Satire*: mimics real news but still cues the reader that it is not meant to be taken seriously (Rashkin et al. 2017)

Based on these definitions, Rashkin et al. (2017) presented an analytical study on the language of news media by comparing the language of real news with that of satire, hoaxes, and propaganda. Following this work, we also included disinformation articles for language comparison with hoaxes, propaganda, and real news to find linguistic characteristics of disinformation compared to other types of news.

We focus on two significant events in recent history to analyze disinformation across years: (1) US Elections-2016 and (2) COVID-2020. Both of these events impacted society and were associated with widespread fake news campaigns. For disinformation on the 2016 US presidential elections, we extract articles from two datasets: (i) a Stanford dataset (Allcott and Gentzkow 2017) containing web links of viral fake news articles during the election period. Although many of these web pages are expired now, we found 26 articles still accessible on the web; (ii) a Kaggle dataset created by using the BS detector,⁶ a tool to identify fake or satirical news websites. Since the BS detector tool labels an article based only on its source, there are many anomalies in the dataset; for example, a movie or food review flagged as fake. To overcome this issue, we used PolitiFact's list of fake news websites⁷ to further winnow down to articles debunked by fact-checking sites such as Snopes, PolitiFact, or fact-checking groups comprised of journalists.

⁵ The code and data used for the study in this section are available at <https://github.com/architapathak/Disinformation-Analysis-and-Identification>.

⁶ <https://www.kaggle.com/mrisdal/fake-news>.

⁷ <https://www.politifact.com/article/2017/apr/20/politifact-guide-fake-news-websites-and-what-they/>.

Table 1 Distribution of articles used for analysis in this paper

	Time frame	Total	Avg. tokens
Election2016 disinfo	Aug–Dec	301	505.9
Election2016 mainstream	Sept–Dec	3017	533.7
COVID2020 disinfo	Jan–April	556	774.5
COVID2020 mainstream	March	3499	458.7

Fig. 1 Salient words in each type of article

COVID-2020 disinformation articles were taken from the EU StratCom Task Force,⁸ which has actively debunked disinformation generated and circulated by Kremlin/pro-Kremlin media. We also extracted articles from NewsPunch (formerly YourNewsWire.com), a disinformation website active since 2016, and published a mixture of true and false news. It is important to note that the articles in this study are labeled as disinformation based on two conditions: (i) the articles are debunked as fake by credible fact-checkers, or (ii) the articles are published by popular fake news websites such as NewsPunch and Infowars. We did not conduct a manual evaluation of the article's claims for the study presented in this section.

Mainstream news articles are sampled from the AllTheNews.⁹ corpus. The selected mainstream sources belong to a broad spectrum of ideological biases (evenly distributed from left to right). Figure 1 shows salient words appearing in the disinformation and mainstream articles used in this study. Table 1 presents the data distribution.

⁸ <https://euvsdisinfo.eu/about/>.

⁹ <https://components.one/datasets/all-the-news-2-news-articles-dataset/>.

We used 887 disinformation articles on Election2016 and COVID2020 from the news corpus as defined above. To maintain consistency in the distribution, we sampled 1,000 propaganda, hoax, and mainstream articles from the dataset released by Rashkin et al. (2017). This dataset was used to make news reliability predictions on satire, propaganda, hoax, and mainstream news. For simplicity, we call this dataset “TOVS”, which is the abbreviation of the first four words of the authors’ paper “**T**ru**T**h **O**f **V**arying **S**hades”.

The final dataset for our experiments was then split into 80% training and 20% testing sets.

Methodology The following models were created and trained for classification-based experiments:

1. *Max-Ent* A baseline max-entropy classifier with L2 regularization on n-gram TF-IDF feature vectors (up to trigrams). This model was trained for 500 iterations on our training data.
2. *DNN* Deep neural networks consisting of two dense layers. Since fake news articles, especially hoaxes and propaganda, may contain misspellings, these words are removed from training as out of vocabulary (OOV) words by pretrained vectors such as GloVe or word2vec, subsequently leading to loss of information. To overcome this issue, we made the word embedding layer of the DNN model trainable to generate a corpus-specific word representation. We used a vocabulary size of 300 K and 300-dimensional vectors to create each word’s embedding for this task. Furthermore, the first dense layer of the model consists of 1,024 units, and the second consists of 512 units. A dropout of 0.2 and batch normalization after every dense layer were applied to generalize the model.
3. *Bi-LSTM (WE)* Bidirectional LSTM with trainable word embedding. LSTM (Hochreiter and Schmidhuber 1997) is a popular model, especially when dealing with text data, because of its ability to learn long-term dependencies and preserve contextual information. The bidirectional nature of the model helps preserve sequence information from beginning to end and end to beginning simultaneously. We used 512 LSTM units in both forward and backward layers with a dropout of 0.2 for training.

We used Keras with the TensorFlow backend to build the DNN and Bi-LSTM models. The Adam optimizer was used to optimize the network, and categorical cross-entropy was used as the loss function to perform softmax classification into four labels. The models were trained for 10 epochs with a batch size of 32.

As another baseline, we also trained our best performing model on the TOVS dataset, which contains satire but not disinformation. Figure 2 summarizes the results on a weighted F1 score. On the TOVS dataset, DNN outperforms the previous model on this dataset by achieving an F1 score of 95.95%. The score is reduced to 90.44% for the classification task on the dataset containing disinformation, hoax, propaganda, and mainstream news. This indicates that the linguistic clues in disinformation are more challenging for differentiation from other types of news than satire. Our experiments also show that a simple model such as DNN

Fig. 2 Weighted F1 scores for 4-way classification on type of news

can outperform Bi-LSTM for this task, which signals that preserving sequence information is not required when learning different news types.

Although the accuracy of classification is more than 90%, manual error analysis shows that some of the articles retrieved from the disinformation website (News-punch.com) often contained true news. In such cases, the model that learns the surface-level linguistic features of a specific domain's articles may incorrectly predict accurate news as disinformation, hoax, or propaganda. Therefore, detection based on pattern learning methods alone is helpful but not sufficient for scaling to real-world situations.

2.2 Disinformation and clickbait headline detection

As another signal for disinformation detection, we also analyzed the “clickbait-yness” of the headlines in disinformation articles. Clickbait is a term commonly used to describe eye-catching and teaser headlines in online media (Shu et al. 2017). Although mainstream news media sometimes use clickbait headlines, their presence is much more prominent in fake news and sometimes is a direct giveaway of an article being fake.¹⁰ (For example: “Trump Supporters Didn’t Vote Because Of The Economy; It Was Racism, And Here’s PROOF”).

Dataset For this task, we used the dataset released by Chakraborty et al. (2016), which contains an even distribution of 16,000 clickbait headlines and 16,000 non-clickbait headlines. The nonclickbait headlines in the dataset were sourced from Wikinews, and clickbait headlines were sourced from BuzzFeed, Upworthy, Viral-Nova, Scoopwhoop and ViralStories. The clickbait headlines have features such as usage of hyperbolic words (e.g., *awe-inspiring* and *gut-wrenching*), internet slang (e.g., *WOW*, *LOL*, *LMAO*) and common bait phrases (e.g., “*You Won’t Believe*”).

¹⁰ <https://www.cits.ucsb.edu/fake-news/what-is-fake-news>.

Fig. 3 Core elements of automated fact-checking (Graves 2018)

Methodology We built a Bi-LSTM model that uses a combination of word and character embeddings based on the work presented in Anand et al. (2017). The model presented in this paper achieves a state-of-the-art result with an F1 score of $\approx 99\%$. The trained model is then used to make predictions on our disinformation dataset containing 887 articles. Interestingly, only $\approx 36\%$ of the disinformation articles were predicted to have clickbait headlines, which signals that disinformation may not always use flashy, obvious attention-seeking headlines. However, manual analysis shows that some of these headlines still have subtle indicators that make them sensational. For example, the headline “Soros Sees Big Opportunities After Coronavirus Pandemic” does not have stylistic features of being clickbait, but certain demographics may find it sensational.

2.3 Disinformation and degree of veracity

Analyses presented in the previous sections are good indicators but not enough to scale to real-world situations. This is mainly due to the dataset and model-induced biases when classifying news articles without fact-checking. Dataset bias is introduced when articles are labeled based on their domain (disinformation websites can also publish true news.) Model bias is introduced when learning stylistic features without analyzing the text’s meaning. Real-world situations also include issues that require attention while building automated models for disinformation identification. These issues include but are not limited to freedom of expression, subjective and confirmation biases, and censorship. Further, persuading readers requires a deeper investigation of claims made in news articles by leveraging additional world

knowledge. Therefore, there is a need to go beyond predictions using broad labels and focus on nuanced judgments of the content's veracity, for example, false, partially true, and true.

These judgments are dependent upon automated fact-checking, which is a task of automatically assessing the truthfulness of a given text (Vlachos and Riedel 2014). (Graves 2018) specifies three core elements of an end-to-end automated fact-checking system: identification, verification, and correction (as shown in Fig. 3). This section focuses on predicting a given article's degree of veracity by identifying factual statements followed by fact verification against authoritative sources. Assessing the article's truthfulness by using authoritative sources may also help overcome biases introduced based on time, space, topics, and sources.

Dataset A seminal dataset for the task of fact verification against textual sources was introduced by Thorne et al. (2018) called the FEVER (Fact Extraction and VERification) dataset. It contains 185,445 claims generated by altering the sentences extracted from Wikipedia and subsequently verified without knowing the sentence they were derived from. The dataset is specific only to Wikipedia claims and evidence. It does not contain labels representing the claim's degree of veracity. Regarding news article veracity prediction, NELA-GT-2018 Nørregaard et al. (2019) contains articles with the degree of veracity labels retrieved from 8 different assessment sites. However, the label definitions are not generic and dependent on external organizations. Additionally, there is no evidence set supporting the veracity label of each article. Lack of evidence is also an issue present in other datasets for veracity predictions (Pathak and Srihari 2019; Rashkin et al. 2017; Wang 2017).

Due to the lack of datasets that contain all important pieces of information required for automated fact-checking, we introduce a novel dataset called DNF-300 (**DisiNFormation**) that can be used for this specific task. We used disinformation articles on the US elections 2016 to build this dataset. Since external world knowledge is an important aspect of fact-checking, each article in DNF-300 is associated with a veracity label and corresponding evidence. Two annotators participated in the process of annotating the articles with labels and evidence.

The annotation process involves identifying sentences from each article that can be considered "claims". In this work, we define "claim" as a sentence that is important to the point of the article but requires verification. These sentences were then used to query the web, and the top 10 results were used to gather evidence from credible sources. The list of credible sources, which contains approximately 400 sources, is retrieved from the Media Bias/Fact Check website.¹¹ This website rates the media sources based on factual accuracy and political bias. Based on the evidence found, the annotators labeled the entire article into five veracity labels: {(0) *false*; (1) *partial truth*; (2) *opinions stated as fact*; (3) *true*; (4) *NEI*}. Table 2 shows the description and distribution of these labels. The label comparison with two popular fact-checking websites is displayed in Fig. 4.

This dataset is also a key contribution of this paper since the articles are manually read and subsequently annotated. Additionally, the dataset contains two novel

¹¹ <https://mediabiasfactcheck.com/>.

Table 2 DNF-300 label description and distribution

Label	False	Partial true	Opinion stated as fact	True	NEI
Desc.	Found evidence to refute the entire article	Article about true event, however, found evidence refuting some of the claims	Article contains false/manipulated claims, however, it is an opinion article which cannot be labeled as fake	Found evidence supporting the entire article	No evidence could be found
Total	62	74	49	28	87

Claims are sentences which are important to the point of the article but that require verification. Interestingly, some of the articles, which were labeled as fake in other datasets due to the source of the article, turned out to be true

Fig. 4 Label comparison with Snopes and PolitiFact ratings

Table 3 Weighted F1-score for fine-grained classification into veracity label

Model	Configuration	Weighted F1 score
Random	—	20%
DNN (5-fold)	16 batch, 10 epochs, 2 dense layers	12.4%
Bi-LSTM (CE) (5-fold)	16 batch, 10 epochs, 64 hidden nodes	22.5%
Bi-LSTM (WE) (5-fold)	16 batch, 10 epochs, 64 hidden nodes	30%
BERT-base (80:20 split)	16 batch, output layer with 768 hidden nodes, 10 epochs	69.47%
BERT-base (90:10 split)	16 batch, output layer with 768 hidden nodes, 10 epochs	74.75%

features that are essential for the verification task: (i) generic veracity-based label set, independent of any external organization, and (ii) ground truth evidence corresponding to each label.

Methodology We first start by learning linguistic features of the content to perform fine-grained veracity classification. We used DNN, Bi-LSTM (WE), and Bi-LSTM (CE) models for this task. The Bi-LSTM (CE) model uses character embeddings, a concept introduced by Zhang et al. (2015), and has been shown to be very efficient in learning orthographic and morphological text features. To build character embedding using the content in the DNF-300 dataset, 3 layers of 1-D CNN with a filter size of [196, 196, 300], pool size 2, and kernel stride of 3 are added above the Bi-LSTM model. The CNN layer creates a one-hot vector of each character from our training data. Finally, we perform max-pooling across the sequence to identify features that produce strong sentiment. The dataset is split into 80% training and 20% testing sets. As shown in Table 3, the best performance for classification based on linguistic features is only 30%, which is better than random but still not sufficient. This also demonstrates the challenge posed by the DNF dataset in categorizing articles into the target veracity labels.

Fig. 5 Confusion matrix for BERT-based NLI task to predict veracity labels {(0) *false*; (1) *partial truth*; (2) *opinions stated as fact*; (3) *true*; (4) *NEI*}

Following this, we modeled this problem as a natural language inference (NLI) task (Bowman et al. 2015; MacCartney and Manning 2009; Williams et al. 2018; Welleck et al. 2019) and formally defined it as follows.

Given an input $\{s_1, s_2\}$, where s_1 is the premise that contains factual statements from an article requiring fact-checking and s_2 is the hypothesis that has candidate snippets from an evidence article, and the target veracity labels $\{0, 1, 2, 3, 4\}$, the problem is then to learn the function

$$f_{NLI}(s_1, s_2) \longrightarrow \{0, 1, 2, 3, 4\} \quad (1)$$

We used the uncased version of BERT-base model (Devlin et al. 2018) to form the task as an attention-based function

$$f_{MLP}(g_{att}(s_1, s_2)) \quad (2)$$

We make the following two assumptions to identify the *premise* and *hypothesis* for this task:

- (i) first N words of an article that requires fact-checking contains a significant amount of factual content and
- (ii) a sufficient amount of relevant context can be found in the last N words of an evidence article.

To choose the value of N , we experimented with a varying number of words in the range $[100, 150, 200, 250]$ in the content and the evidence article. The best performance was obtained by using $N = 200$. As shown in Table 3, the NLI

Fig. 6 Auto-fact-checking pipeline

approach leads to significant and promising improvement in classification accuracy by $\approx 40\%$. Note that an article can be associated with multiple pieces of evidence that lead to improved veracity level classification. Since the dataset is small, we subsequently experimented with a train-test split of 90:10 to learn more features from the training set. On the 38 test articles, the model attains an F1 score of 74.75%.

Figure 5 shows the confusion matrix of BERT results on 38 test instances. Out of five labels, the NLI model performed well in predicting *true*, *partial truth* and *NEI* classes. However, many of the *opinion* articles are confused with *NEI*. This is because an opinion article may not always have an associated evidence article in the dataset. An additional approach to analyzing these types of articles is feature-based by identifying opinionated phrases/sentences.

Furthermore, in the current test set, 2 out of 7 *partial truth* articles are confused with *false* articles. This may be due to our approach for selecting the experiment premise and hypothesis - the first and last N words from disinformation and evidence articles, respectively. Since not every article may contain enough factual information in the first and last N words, one of the future directions of this work is to (i) identify factual claim sentences from a disinformation article and (ii) mine candidate snippets from evidence articles based on the relevancy with factual claims.

Claim identification also facilitates mining relevant evidence from the web, which is an important task in the fact-checking pipeline. Correctly identified evidence can be fed back to the verification module, creating an enriched dataset. Another limitation of this work that we would like to address in the future is to expand the dataset to include articles across various topics and languages. Additionally, we plan to perform user studies to analyze (i) which claims make more of an impact on readers and (ii) whether the evidence is compelling enough to categorize the article into a given veracity label. Expanded datasets and human evaluated ground truth will lead to more trust-worthy classifiers, assisting human fact-checkers.

3 Auto-fact-checking web application: FactFinder

Based on the work presented in the previous sections, we present a web app for an end-to-end fact-checking system called FactFinder. It uses a pipeline-based system to (i) sample important entities and associated actions from a given article, (ii) gather evidence from the web, and (iii) use the evidence to predict the degree of veracity of the article. The implementation of this web app can be found at <https://github.com/architapathak/FactFinder>.

Note that we are continuously adding more functionalities to this application. Currently, as shown in Fig. 6, the pipeline consists of the following stages:

1. *Query formulation* To retrieve credible news articles from the web related to the input news article, a named entity recognition (NER) tagger is used to identify important entities such as person, organization, and locations, which are then used to form a reasonable web search query. For this task, we trained a BERT-based model on an NER dataset provided by CoNLL 2003 (Sang and De Meulder 2003). The dataset consists of sentences with various types of entities of interest. BERT, which encompasses the encoder-decoder-based transformer model, achieves better results. The entities are predicted in the IOB format. For example, “Alex is going to Los Angeles” is predicted as [B-PER, O, O, O, B-LOC, I-LOC].
Since a news article can have numerous entities, the next step is to perform salience sampling of entities relevant to the point of the article. To do this, we count the number of verbs around the extracted entities. We hypothesize that the verb count around a specific entity highlights its importance to the point of the article. The set of nouns and verbs are then concatenated to form the query.
2. *Evidence gathering* To extract evidence articles, we queried the keywords generated by the query formulation task using the Bing Search API. We used the top 30 results and filtered them based on (i) whether they belong to credible sources and (ii) whether they were published around the same time the potential disinformation article was published. The list of credible sources contains names of news sources rated as “least biased” by Media Bias/Fact Check.¹² These sources have been acknowledged by the journalistic community as being factual and well sourced. We also added a few more left, left-center, right-center, and right news sources to this list based on our research. For the date filter, we used the range of 3 months before and after publishing the potentially disinforming article. Finally, the top 3 evidence articles were selected from the filtered URLs. Note that we call the article used for verification “potential” disinformation. Without an evidence-based verdict, it cannot be accurately identified as disinformation.
3. *Degree of veracity prediction* This is the NLI task for predicting the veracity label of an article based on the evidence. The best performing BERT model, as described in Sect. 2.3, was used for this task.

¹² <https://mediabiasfactcheck.com/center/>.

FactFinder

Try these examples:

*Physician Confirms Hillary Clinton Has Parkinson's Disease ----- FBI Agent Suspected in
Hillary Email Leaks Found Dead in Apparent Murder-Suicide ----- President Obama
Confirms He Will Refuse to Leave Office If Trump Is Elected*

Headline:

Physician Confirms Hillary Clinton Has Parkinson's Disease

Article:

According to pharmaceutical executive Martin Shkreli, footage of Hillary Clinton exhibiting bizarre behavior is proof that the presidential candidate is suffering from Parkinson's disease. Citing his "15 year background in drug discovery and pharmaceuticals," Shkreli asserts that the videos of Hillary's strange facial movements and her difficulties with walking are "pretty unmistakable signs of Parkinson's disease." Explaining that a stroke or a concussion wouldn't explain the symptoms, Shkreli says that Hillary's walking

Approx. Month and Year of the Article:

September 2016

Fig. 7 First page of the FactFinder UI

3.1 User interface

Figure 7 shows the first page of the UI of the FactFinder app. Users can try different examples of fake news articles as provided on the top of the input page. Given the headline, the article's content, and approximate date of publication, this app provides different analyses, as shown in Fig. 8. The user can first see the probability of the headline being clickbait. For this, we use the clickbait detection model

Fig. 8 The analysis page of the FactFinder web app displaying the analysis on the clickbaitiness of the headline, and a word cloud of the article content

described in Sect. 2.2. In the example shown, there is a high chance of the headline being classified as nonclickbait since it does not contain any stylistic features of “clickbaitiness”. The UI also shows a word cloud of the most frequent keywords used in the article, which provides some insight into the article’s content.

The next section of the UI shows the content of the article (Fig. 9). Words that are highlighted in red are the query words extracted during the query formulation stage. As shown, the importance sampling step helps winnow down the important entities and verbs for query formulation. The yellow highlighted sentences provide a preview of future work, which is to select *factual claim sentences* as the premise for the NLI model.

Finally, the bottom of the page shows a list of evidence URLs and corresponding veracity label prediction (Fig. 10). The user can click on these links to refer

Article:

According to pharmaceutical executive Martin Shkreli, footage of **Hillary Clinton exhibiting** bizarre behavior **is** proof that the presidential candidate **is** suffering from Parkinson's disease. Citing his "15 year background in drug discovery and pharmaceuticals," **Shkreli asserts** that the videos of Hillary's strange facial movements and her difficulties with **walking** are "pretty unmistakable signs of Parkinson's disease." Explaining that a stroke or a concussion wouldn't explain the symptoms, **Shkreli says** that Hillary's **walking** difficulties are a "cardinal symptom of Parkinson's disease" and what **is** known as "freezing gait". **Shkreli adds** that he **helped develop** a drug to treat the problem. Hillary's "on-off episodes" are a result of dopamine depletion, according to Shkreli, which **is** impacting Hillary's movement. "We've all seen the videos of her kind of making these perplexed facial movements," **says** Shkreli, noting that Hillary's over-reaction to balloons at the DNC was "unusual" because it betrayed an "inability to control her facial movements". However, Hillary's seizure-like behavior in response to reporters asking questions in another clip was even more odd because she "made this involuntary movement for about ten seconds which was truly unusual," states Shkreli, identifying the condition as a form of Parkinson's-induced dyskinesia. "This ia classic symptom....if you showed someone that symptom and said 'name the disorder' it's not a seizure....that **is** simply involuntary movement, it's PD-LID, no doubt about it in my mind," **adds** Shkreli, concluding that when matched with Hillary's "freezing gait," it "explains everything." **Shkreli** emphasizes the fact the he **is** non-partisan and has given more money to Democrats than Republicans. Despite his controversial background

Fig. 9 The content of the news article as displayed on the FactFinder web app. Words highlighted in red are query words used for evidence mining. Sentences highlighted in yellow are the claim sentences that are salient to the point of the article

to the articles. This provides transparency into the model's decision that the users can evaluate.

4 Conclusion and future work

We presented a detailed study on disinformation by building classifiers that use stylistic, semantic, orthographic, and morphological features along with inference to make predictions on (i) the type of fake news, (ii) clickbait detection based on the disinformation headline, and (iii) fine-grained, veracity-based classification of disinformation. We show that style-based classifiers are not sufficient in overcoming various biases and accurately identifying disinformation; fact-checking is required for credible debunking.

Our results on using NLI to automate/semiautomate the fact-checking process show promising improvement in veracity-based classification culminating in the development of a web app, FactFinder. This app identifies important entities and

Evidence :

1. <https://www.snopes.com/fact-check/hillary-clinton-has-parkinsons-disease/>
2. <https://www.truthorfiction.com/hillary-clinton-probably-parkinsons-disease/>
3. <https://www.washingtonpost.com/news/the-fix/wp/2016/08/23/a-guide-to-hillary-clintons-many-illnesses-as-diagnosed-in-the-conservative-media/>

Veracity Prediction :

URL	https://www.snopes.com/fact-check/hillary-clinton-has-parkinsons-disease/	https://www.truthorfiction.com/hillary-clinton-probably-parkinsons-disease/	htl fix cli the
Labels	False	False	Fal

Fig. 10 The evidence is selected based on the credibility of the sources and date of publication. Each piece of evidence is then used by the NLI module for the degree of veracity prediction

associated actions from the headline and content provided by the user and processes them for evidence gathering. The evidence is for veracity prediction based on generic labels. In the future, we aim to build upon this model and incorporate user studies in improving the dataset as well as the end-to-end fact-checking process. Another future direction is to perform claim identification on articles for effective evidence gathering and verification.

References

- Allcott H, Gentzkow M (2017) Social media and fake news in the 2016 election. *J Econ Perspect* 31(2):211–36
- Anand A, Chakraborty T, Park N (2017) We used neural networks to detect clickbaits: you won't believe what happened next! In: *European conference on information retrieval*. Springer, pp 541–547
- Bowman SR, Angeli G, Potts C, Manning CD (2015) A large annotated corpus for learning natural language inference. In: *Proceedings of the 2015 conference on empirical methods in natural language processing*. Association for Computational Linguistics, Lisbon, Portugal, pp 632–642. <https://doi.org/10.18653/v1/D15-1075>, <https://www.aclweb.org/anthology/D15-1075>
- Chakraborty A, Paranjape B, Kakarla S, Ganguly N (2016) Stop clickbait: Detecting and preventing clickbaits in online news media. In: *2016 IEEE/ACM international conference on advances in social networks analysis and mining*. IEEE, pp 9–16
- de Cock Buning M (2018) A multi-dimensional approach to disinformation: Report of the independent high level group on fake news and online disinformation. Publications Office of the European Union
- Da San Martino G, Barrón-Cedeño A, Nakov P (2019) Findings of the NLP4IF-2019 shared task on fine-grained propaganda detection. In: *Proceedings of the second workshop on natural language processing for internet freedom: censorship, disinformation, and propaganda*. Association for Computational Linguistics, Hong Kong, China, pp 162–170. <https://doi.org/10.18653/v1/D19-5024>, <https://www.aclweb.org/anthology/D19-5024>

- Devlin J, Chang M, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805**, [arXiv:1810.04805](https://arxiv.org/abs/1810.04805)
- Fallis D (2009) A conceptual analysis of disinformation. Proceedings of iConference, <http://hdl.handle.net/2142/15205>
- Graves D (2018) Understanding the promise and limits of automated fact-checking
- Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput* 9(8):1735–1780
- MacCartney B, Manning CD (2009) An extended model of natural logic. In: Proceedings of the eight international conference on computational semantics, pp 140–156
- Nørregaard J, Horne BD, Adalı S (2019) Nela-gt-2018: a large multi-labelled news dataset for the study of misinformation in news articles. In: Proceedings of the international AAAI conference on web and social media, vol 13, pp 630–638
- Pathak A, Srihari R (2019) BREAKING! presenting fake news corpus for automated fact checking. In: Proceedings of the 57th annual meeting of the association for computational linguistics: student research workshop. Association for Computational Linguistics, Florence, Italy, pp 357–362. <https://doi.org/10.18653/v1/P19-2050>, <https://www.aclweb.org/anthology/P19-2050>
- Rashkin H, Choi E, Jang JY, Volkova S, Choi Y (2017) Truth of varying shades: analyzing language in fake news and political fact-checking. In: Proceedings of the 2017 conference on empirical methods in natural language processing, pp 2931–2937
- Sang EF, De Meulder F (2003) Introduction to the conll-2003 shared task: language-independent named entity recognition. arXiv preprint [cs/0306050](https://arxiv.org/abs/cs/0306050)
- Shu K, Sliva A, Wang S, Tang J, Liu H (2017) Fake news detection on social media: a data mining perspective. *ACM SIGKDD Explor Newsl* 19(1):22–36
- Thorne J, Vlachos A, Christodoulopoulos C, Mittal A (2018) FEVER: a large-scale dataset for fact extraction and VERification. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: Human Language Technologies, vol 1 (Long Papers). Association for Computational Linguistics, New Orleans, Louisiana, pp 809–819. <https://doi.org/10.18653/v1/N18-1074>, <https://www.aclweb.org/anthology/N18-1074>
- Ungku F, Fernandez C, Brock J (2019) Factbox: ‘fake news’ laws around the world. Reuters
- Vlachos A, Riedel S (2014) Fact checking: task definition and dataset construction. In: Proceedings of the ACL 2014 workshop on language technologies and computational social science. Association for Computational Linguistics, Baltimore, MD, USA, pp 18–22. <https://doi.org/10.3115/v1/W14-2508>, <https://www.aclweb.org/anthology/W14-2508>
- Wang WY (2017) “liar, liar pants on fire”: a new benchmark dataset for fake news detection. In: Proceedings of the 55th annual meeting of the association for computational linguistics, vol. 2: Short Papers. Association for Computational Linguistics, Vancouver, Canada, pp. 422–426. <https://doi.org/10.18653/v1/P17-2067>, <https://www.aclweb.org/anthology/P17-2067>
- Wardle C, Derakhshan H (2017) Information disorder: toward an interdisciplinary framework for research and policy making. Council of Europe report, DGI (2017) 9
- Welleck S, Weston J, Szlam A, Cho K (2019) Dialogue natural language inference. In: Proceedings of the 57th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Florence, Italy, pp 3731–3741 <https://doi.org/10.18653/v1/P19-1363>, <https://www.aclweb.org/anthology/P19-1363>
- Williams A, Nangia N, Bowman S (2018) A broad-coverage challenge corpus for sentence understanding through inference. In: Proceedings of the 2018 conference of the North American Chapter of the association for computational Linguistics: Human Language Technologies, vol 1 (Long Papers). Association for computational linguistics, New Orleans, Louisiana, pp 1112–1122. <https://doi.org/10.18653/v1/N18-1101>, <https://www.aclweb.org/anthology/N18-1101>
- Zhang X, Zhao J, LeCun Y (2015) Character-level convolutional networks for text classification. In: Advances in neural information processing systems, pp 649–657

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Archita Pathak Archita is a doctoral candidate in the CSE department at the University at Buffalo (SUNY), working with her advisor Dr. Rohini K. Srihari. Archita's research focus is on using Deep Learning and NLP for Fake News Detection. Her research seeks to advance simple fake news detection based on linguistic clues to robust, evidence-based automated fact-checking that can be deployed at a large scale. This involves challenging tasks like identification of key claims in news articles that are written to deliberately mislead people and evidence-based claim verification to tackle the menace of widespread online disinformation. Archita's work on novel dataset creation for automated fact-checking and claim identification has been published at ACL-SRW 2019 workshop and ICON 2020 conference, respectively. Recently, her work on disinformation analysis and identification won the Disinformation Challenge organized by SBP-BRiMS 2020. More information can be found at: <http://www.acsu.buffalo.edu/~architap/>.

Rohini K. Srihari Rohini Srihari is a scientist, educator and entrepreneur. Dr. Srihari is a professor in the Dept. of Computer Science and Engineering at the University at Buffalo (SUNY) where her research focuses on natural language processing, artificial intelligence, machine learning and information retrieval. She teaches graduate courses in information retrieval and natural language processing and offers seminars on topics including conversational AI, deep learning in NLP and AI for social good. Her work has been funded by agencies such as the National Science Foundation, DARPA and IARPA. She has published over 100 papers in leading journals and conferences. Her Google Scholar h-index is 39. She is the author of two US patents, one on multilingual text mining. A selected list of publications can be found at <http://www.buffalo.edu/~rohini>. Dr. Srihari received a B. Math degree from the University of Waterloo (Canada) and a PhD in Computer Science from the University at Buffalo.

Nihit Natu is a graduate student in the CSE department at University of Buffalo (SUNY). His research focus lies in Natural Language Processing and Deep Learning domains. His work on Deep Learning based Recommendation System has been published in Advanced Computing Technologies and Applications Journal in 2019. Currently, he is working with Dr. Rohini K. Srihari on entity extraction, event detection and web information retrieval for the task of fake news detection. He has been a part of the work that won the Disinformation Challenge organized by SBP-BRiMS 2020. Apart from this, Nihit also have an experience and expertise working on recommendation systems, text summarization and sentiment analysis tasks.

Authors and Affiliations

Archita Pathak^{1,2} · Rohini K. Srihari¹ · Nihit Natu^{1,3}

Rohini K. Srihari
rohini@buffalo.edu

Nihit Natu
nihitume@buffalo.edu

¹ University at Buffalo (SUNY), 338D Davis Hall, Buffalo, NY 14260-2500, USA

² Farmington Hills, USA

³ Buffalo, USA