Information Processing and Management 58 (2021) 102653

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

When classification accuracy is not enough: Explaining news credibility assessment

Piotr Przybyła [a],[*], Axel J. Soto [b],[c]

[a] Institute of Computer Science, Polish Academy of Sciences, 5 Jana Kazimierza Str., 01-248 Warsaw, Poland
[b] Institute for Computer Science and Engineering (UNS–CONICET), San Andrés 800, Bahía Blanca, Argentina
[c] Department for Computer Science and Engineering, Universidad Nacional del Sur, San Andrés 800, Bahía Blanca, Argentina

ARTICLE INFO

Keywords:
Visual analytics
Credibility
Text classification
Fake news
Natural language processing

ABSTRACT

Dubious credibility of online news has become a major problem with negative consequences for both readers and the whole society. Despite several efforts in the development of automatic methods for measuring credibility in news stories, there has been little previous work focusing on providing explanations that go beyond a black-box decision or score. In this work, we use two machine learning approaches for computing a credibility score for any given news story: one is a linear method trained on stylometric features and the other one is a recurrent neural network. Our goal is to study whether we can explain the rationale behind these automatic methods and improve a reader's confidence in their credibility assessment. Therefore, we first adapted the classifiers to the constraints of a browser extension so that the text can be analysed while browsing online news. We also propose a set of interactive visualisations to explain to the user the rationale behind the automatic credibility assessment. We evaluated our adapted methods by means of standard machine learning performance metrics and through two user studies. The adapted neural classifier showed better performance on the test data than the stylometric classifier, despite the latter appearing to be easier to interpret by the participants. Also, users were significantly more accurate in their assessment after they interacted with the tool as well as more confident with their decisions.

1. Introduction

Misinformation is widely considered a crucial challenge for modern societies. One of the problems behind its spread is the difficulty of assessing credibility of information, especially news, published and shared through online media. Artificial intelligence (AI) has been proposed as a solution to this problem (Ciampaglia et al., 2018) and, indeed, some positive results have been reported.

Unfortunately, even if such methods were to achieve near-perfect accuracy, it remains unclear how to use their output to inform credibility, raise readers' awareness and ultimately influence their behaviour. While automatic content filtering is considered equivalent to censorship (Llansó, 2020), explicit warnings about non-credible content used by social media services have small to moderate effect on users' perceptions (Clayton et al., 2019; Pennycook et al., 2020) and actions (Mena, 2020). Political beliefs are particularly challenging, as providing corrections to subjects' misperceptions can even strengthen confidence in them (Nyhan & Reifler, 2010). However, it may be effective to provide alternative explanations for a disputed claim (Nyhan & Reifler, 2015) or graphical information (Nyhan & Reifler, 2019). Research has also shown that providing visual explanations for AI decisions can increase user trust in automatic assessments (Ribeiro et al., 2016; Yang et al., 2020).

[*] Corresponding author.
E-mail address: piotr.przybyla@ipipan.waw.pl (P. Przybyła).

https://doi.org/10.1016/j.ipm.2021.102653
Received 19 January 2021; Received in revised form 30 April 2021; Accepted 25 May 2021
Available online 12 June 2021

The main goal of this work is to study how a text classifier can be adapted to influence users' assessment of credibility in online news. We rely on previous work regarding credibility classification using a style-based approach (stylometric and neural) that shows solid performance (above 80% accuracy), even for documents from previously unseen sources or about unseen topics (Przybyła, 2020). Within the present study, we have re-designed these methods so that they can work in a real time setting in the constrained context of a browser extension. We have also recreated the classifiers as a visual analytics system in order to assist users towards assessing news credibility in an interactive manner. This tool, which we called *Credibilator*, is then evaluated using standard classification metrics and through two user studies. Our evaluation allows us to investigate the following questions:

- How can resource-hungry machine learning (ML) models be tailored to restricted computational environments without sacrificing accuracy?
- What interactive visual means can be used to explain the decision of classifiers in a particular case?
- Do users' assessments of text credibility become more accurate and confident after interacting with our tool?

To encourage further research on these issues, the text corpus, the classification models and Credibilator code are made publicly available.[1]

2. Related work

The present study builds on three existing research areas. Firstly, on methods seeking to assess credibility of any given news text automatically, without any human assistance. Secondly, on expanding text classification by means of visual analytics in order to help a user understand how a certain classification result was obtained. Finally, on building tools that can help web users to understand the trustworthiness of online content. Here we present previous research in each of these areas.

2.1. Credibility assessment

Research in AI solutions for misinformation often focuses on detecting instances of *fake news*, as the most obvious target (Sharma et al., 2019; Zhang & Ghorbani, 2020). This term, however, is used to describe many phenomena and its definitions differ even in the most basic aspects, e.g. whether falsehood of content is an essential or optional element (Gelfert, 2018; Tandoc et al., 2017). Instead, we aim to measure *credibility*, which captures how worthy a document is to be believed in, based on a large variety of factors (Choi & Stvilia, 2015; Hilligoss & Rieh, 2008; Kakol et al., 2017).

Research on automatic assessment of credibility for online news can be separated in three main groups according to the signals they use. Some attempt to extract and verify claims made in text (i.e. *fact-checking*) with respect to knowledge bases (Atanasova et al., 2019; Ciampaglia et al., 2015; Thorne et al., 2018). Others concentrate on quantifying trustworthiness through the social media context of the news content and its author (Shu et al., 2017; Tacchini et al., 2017; Zubiaga et al., 2016).

The third possibility is relying on overall linguistic properties of text, i.e. writing *style*. Low credibility text, including fake news, is prepared in a way that maximises emotional response (Bakir & McStay, 2017), which translates to certain stylistic techniques. Such credibility indicators could be quantified by measuring language complexity, detecting syntactic patterns through n-grams of part of speech tags or counting words belonging to certain categories (Horne & Adali, 2017; Pérez-Rosas et al., 2018; Potthast et al., 2018; Przybyła, 2020; Rashkin et al., 2017; Reis et al., 2019).

We choose to base our study on style-based classifiers proposed in a study by Przybyła (2020) because of the following reasons:

- The style-base approach does not rely on external resources, such as social media context or knowledge bases, which makes it easier to scale, to extend to other languages or to implement in a restricted environment.
- The study involved evaluation with respect to previously unseen sources and topics, which is crucial for the considered use case,
- The proposed models are relatively light-weight (based on simple features or a neural network with few layers), making them well-suited for restricted environments.

2.2. Visualisation in text classification

Methods that incorporate interactive visualisation to augment understanding of machine learning models have been explored for several years now (Liu et al., 2017). Such visual analytics methods have been also applied to text data, where two main strategies stand out. The first strategy is to identify and highlight salient features that impact on the result of the classification. For instance, *Minerva* (Stoffel et al., 2015) analyses different feature types to gain insights into tasks such as sentiment classification. Brooks et al. (2015) go a step further as their system provides support for *ideation* (i.e. creation) of new features as a composition of original features in such a way that better classification results can be obtained through user involvement.

The second strategy is to leverage feature similarity or text similarity against specific ground truth to illustrate how seemingly obscure models work. One pioneer work in this line is *iVisClassifier* (Choo et al., 2010), where dimensionality reduction is used to visualise data instances and the decision boundary, and hence shed light on how the proposed linear discriminant analysis

[1] https://github.com/piotrmp/credibilator.

method works. A more recent approach is the study by Rauber et al. (2017) where the embedded representation used by a neural network is visualised by means of dimensionality reduction methods to show potentially similar samples, and hence to describe what the internal representation of a neural network looks like. Another example of similarity at the feature level is relative n-gram signatures (Jankowska et al., 2012), which visualises commonalities and discrepancies in character n-gram signatures for each book to understand subtle differences among authors in their choice of words. Our work explores both strategies as the most discriminative features are highlighted as well as a 2D scatter plot is presented to detect similar news in our labelled corpus, which reflects the stylistic-based internal text representation of Credibilator.

Interactive visualisation in text classification appears to be effective for tasks where class boundaries are not clear cut or in active learning scenarios. One such case is stance classification (Kucher et al., 2017), where visualisation helps find interesting samples for manual annotation and help users better understand stance phenomena. This work was extended to also encode sentiment-derived information in the visualisation (Kucher et al., 2020). A similar scenario is described for classifying questions based on stylistic features (Sevastjanova et al., 2018) by means of a visual interface. Our work also goes in the direction of making use of a visual interface to assist the user in a classification task whose correct class (i.e. credible or not) may not be always apparent.

2.3. Credibility tools

Finally, we review other tools that were used for credibility assessment of news. Research efforts described in Section 2.1 focus on classification performance rather than on explainability, with one exception (Reis et al., 2019), where Shapley Additive Explanations (SHAP) are used for feature importance visualisation. Separately, several tools (usually browser extensions) have been made available to help internet users assess credibility of various types of content. This includes *Reality Defender*[2] for detecting manipulated images, *InVID WeVerify*[3] for analysing videos and an automatic assistant to detect visual bias (Narwal et al., 2017). Text analysis is performed by *Fake News Detector*,[4] which flags dubious stories in social media, and *TrustedNews*,[5] which measures objectivity. However, no evaluation or user study has been published for either of them.

In terms of visual analytics systems, *VINCENT* (Ninkov & Sedig, 2020) is a tool for exploring online content related to the vaccine debate. It visualises relevant websites through several features, such as sentiment, most frequent words or geographical location. This helps users gain insights into the discussion, but it does not directly address the problem of credibility. *FakeNewsTracker* (Shu et al., 2019) is a neural-based tool for identification of fake news. While it provides a visual summary of common topics and patterns for the identified content, it does not aim at providing an explanation for the classification prediction.

The *Tweet Verification Assistant* (Boididou et al., 2018), *BRENDA* (Botnevik et al., 2020) and *XFake* (Yang et al., 2019) are the closest to our work. Similar to ours, the *Tweet Verification Assistant* provides a list of different features found to be relevant for the classification of a post using histograms. However, the tool is constrained for Twitter content, the explanation is limited to the presentation of histograms and there has been no evaluation on user perception. *BRENDA* is also a web browser extension that allows verifying veracity *in situ*. It differs to our study in the sense that the tool is focused on fact checking and on providing snippets of evidence, as opposed to identifying stylistic cues that denote lack of credibility. *XFake* uses document metadata and content to detect fake news and it provides visual explanation using word clouds and decision trees. However, synthetic evaluation shows low accuracy for a binary problem (53%–67%) and the results of a user study were not published.

To summarise, while some work has been carried out on automatic credibility prediction for online news and some of it takes into account explainability, no one has so far verified whether this kind of method can improve accuracy or confidence of human judgement.

3. Methods

In this section we describe our approach to make credibility assessment accessible online. We implemented two style-based machine learning models and packed them as a browser extension. To ensure users' privacy, the analysed content is processed in the user's browser. As a result, the classification models need to be suited to a constrained environment. The rationale behind the credibility score is presented using different interactive visualisations, which explain what aspects of the text contribute to the assigned score for each classifier.

3.1. Credibility assessment

Text scoring in Credibilator is based on two previously published classifiers (Przybyła, 2020), which were designed to detect low-credibility text by assessing its style: **Stylometric** and **Neural**. These classifiers, further described below, are trained using a large corpus created by web-scraping news websites with class labels coming from expert knowledge. We use documents identifiers (URLs) collected in previous work (Przybyła, 2020) through the following process:

[2] https://rd2020.org/.

[3] https://weverify.eu/verification-plugin/.

[4] https://fakenewsdetector.org/en.

[5] https://trusted-news.com/.

1. Obtaining the list of credible sources by choosing websites of news outlets that are more commonly trusted than distrusted by the US population according to a *Pew Research Centre* study (Mitchell et al., 2014) and excluding news aggregators (*Google News*, *Yahoo News*) and video-based outlets (*MSNBC*).
2. Obtaining the list of non-credible sources by choosing websites categorised as *Fake news* or *Imposter site* by *PolitiFact* (Gillin, 2017) and manually excluding those based on discussions, prank content and advise articles.
3. Crawling the websites, starting from the main page, up to the depth of five links and 10,000 URLs per website. The *WayBackMachine*[6] archives were used to obtain the version of the pages available at the time of the *PolitiFact* classification (2017).
4. Excluding web pages that are duplicates or do not contain continuous text (average paragraph length below 15 words).

As a result of this process, 103,219 document URLs were collected. Within the present work, we further refine the corpus by applying browser-based content extraction methods (see Section 3.2).

3.1.1. Stylometric model

The first classifier is built based on research in *stylometry*, which seeks to describe a given text with respect to its style and independently from its meaning. Stylometric features have proved successful in discovering authors' traits from their text (Argamon et al., 2009; Diermeier et al., 2011; Koppel et al., 2002; Przybyła & Teisseyre, 2014) and similar methods have also been applied to text credibility assessment (Horne & Adali, 2017; Pérez-Rosas et al., 2018; Rashkin et al., 2017).

The most important consideration here is avoiding features that could betray the document's topic or source. For example, if we used a regular bag of words implementation, the word *Obama* could be an indicator of low credibility (fake news sources often focus on contentious political topics) and *BBC* of high credibility (media names often appear in their articles). This is confirmed by results obtained with solutions using such features in classification (Ahmed et al., 2017; Rashkin et al., 2017) or visualisation (Shu et al., 2019), finding topical words, such as *Syria* (Rashkin et al., 2017) or *Trump* (Shu et al., 2019) to be the most important. While this may lead to good performance within the dataset, it can also cause accuracy drop in the long term, when both current topics and sources change. Instead, we use parts of speech (POS), which are less prone to be affected by these factors. Specifically, we compute frequencies of all POS unigrams, bigrams and trigrams occurring in at least 5 documents, normalised by the document length (number of POS-tagged words).

Another type of features used for stylometric purposes are those based on dictionaries, such as *Linguistic Inquiry and Word Count* (LIWC) (Tausczik & Pennebaker, 2009) or *General Inquirer* (GI) (Stone et al., 1962). They group words in broad meaning categories (e.g. expressions related to power and authority), which help in style representation, but have insufficient coverage. To solve this problem, we seek to automatically extend the 182 categories in GI with new words. Firstly, for each category C of original size n, we define a classification task, where each word is represented through a word embeddings (300 dimensions of *word2vec* (Mikolov et al., 2013)) and a class label set to 1, if the word belongs to C; and set to 0 otherwise. Then, we train a logistic regression model using the embedding of each word in the dictionary to predict the probability of its membership in C. Subsequently, C is extended with $4 \times n$ words with the highest positive class probability.

The new dictionary covers 34,293 different words, where the old one contained only 8640. On average, each of the 182 categories has 898 words, but their sizes differ significantly. The smallest category `say` (*words for say and tell*) has just 16 words, while the largest one `negativ` (*words of negative outlook*) includes 8020 words. The number of categories a word belongs to can be more than one; on average this value equals 4.43, but it achieves a maximum of 67 categories. For example, the word *piano* belongs to `sklasth` (*skill aesthetic, mostly arts*), `exprsv` (*associated with the arts, sports, and self-expression*), `object` (*references to objects*) and `tool` (*references to tools*). Frequencies of words belonging to each of the categories, normalised by the document length (number of words assigned to any GI category), constitute our second group of features.

Finally, we include some classic stylometric features describing text complexity (i.e. number of sentences, average number of words in a sentence, average word length) and letter casing (i.e. fraction of words in lower case, upper case, title case and other case schemes).

The total number of generated features is 21,651 (182 category-based, 21,461 POS n-grams and 8 other), which necessitates filtering before feeding them into a classifier. Filtering is performed through Pearson correlation of each feature with the class label in the training data. Specifically, we define variables $b_{i,j}$ that indicate if the value of feature j in document i (denoted by $x_{i,j}$) is non-zero:

$$b_{i,j} = \begin{cases} 1 & \text{if } x_{i,j} \neq 0 \\ 0 & \text{if } x_{i,j} = 0. \end{cases}$$

In similar solutions, features are filtered based on the number of positive entries in columns of this matrix, i.e. $\sum_i b_{i,j}$ (Potthast et al., 2018). Instead, we take into account the class label y and preserve feature j if:

$$|\mathrm{cor}(\overrightarrow{b_{\cdot,j}}, \overrightarrow{y})| > 0.05,$$

where $\overrightarrow{b_{\cdot,j}}$ is the jth column vector of the indicator matrix $[b_{i,j}]$. After the filtering process, 775 features are preserved (106 category-based, 664 POS n-grams and 5 other).[7]

[6] https://archive.org/web/.

[7] Provided values refer to the model built on the whole dataset for deployment. In cross-validation evaluation, different number of features are kept for every data split.

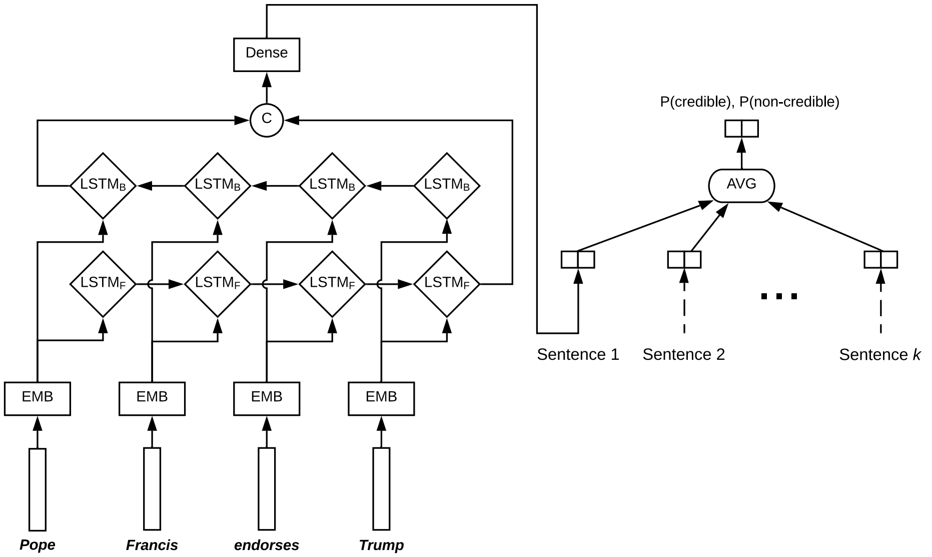


Fig. 1. Architecture of the BiLSTMAvg neural model.

Based on the retained features, an L_1-regularised (LASSO) logistic regression model using `glmnet` (Tibshirani, 1996) in *R* (R Core Team, 2013) is built. In order to choose the value of the regularisation penalty λ, we use the cross-validation procedure in the `cv.glmnet()` function. It randomly splits the training data into 10 folds and computes the predictions in a cross-validation scenario, returning the λ value that results in the lowest error ($\lambda = 0.0001578$ for our data). This value is then used to compute the final model on the full dataset, having non-zero coefficients for 616 features.[7]

The output of the classifier (i.e. probability of positive class or the non-credible news) is used directly as a non-credibility score. For the purposes of the quantitative evaluation, it is discretised using a 0.5 threshold.

3.1.2. Neural model

The architecture of our neural classifier, called *BiLSTMAvg*, is shown in Fig. 1. It is a deep neural network using well-established techniques in text classification. Specifically, the following layers are included:

- an embedding layer (300-dimensional) applied to each word in a sentence,
- two LSTM layers: backward and forward, representing the whole sentence through two 100-dimensional vectors,
- a dense layer, reducing dimensionality to 2 and applying softmax to obtain a sentence score,
- an averaging layer, computing the average score for all sentences and returning it as a document score.

The dense layer used to reduce dimensionality takes the concatenated outputs of the last elements of both LSTM layers as a single 200-dimensional input vector (v) and performs a linear transformation using a 2×200 matrix (A) to obtain a 2-dimensional output vector ($w = Av$). The layer uses softmax activation to return a pair of scores that sum up to 1 and express the likelihood of the credible and the non-credible class label.

While LSTM layers (Hochreiter & Schmidhuber, 1997) are commonly used in text classification, the averaging layer has been added specifically for this task. Its role is to alleviate the problem of long-term dependencies in text classification when applying LSTM to very long sequences, such as full article text. Among the proposed solutions were multi-level memory mechanisms (Xu et al., 2016), skipping irrelevant words (Du et al., 2020) and multi-scale (Liu et al., 2015) or hierarchical LSTM layers (Tai et al., 2015). Computing sentence credibility score and averaging over all sentences avoids long-term dependencies and encourages the classifier to find justification for a document label in each of its sentences. This helps the network to focus on stylistic properties that are present throughout the whole document.

Based on our observation of typical documents in the corpus, we set the maximum document length to 50 sentences and the maximum sentence length to 120 tokens. The neural model is trained in *TensorFlow* (Abadi et al., 2016) using the Adam optimiser (Kingma & Ba, 2015). We group data into batches of 32 documents each and train for 10 epochs with cross-entropy loss and uniform learning rate of 0.001.

3.2. Adaptation for constrained environments

Making the aforementioned classifiers available to internet users in a browser extension requires making certain adaptations to the basic design. Firstly, we consider the limitations of the procedure for extracting plain text from HTML documents. The ad-hoc heuristic used originally (Przybyła, 2020), while giving acceptable results for the original corpus, is not general enough to be applied to the layout of an arbitrary webpage that can be selected by a user. For that reason here we use *unfluff*,[8] a Node.js library for content extraction. Based on the output of the tool, we remove documents that do not contain enough content for style analysis (i.e. less than 500 characters). As a result, we obtain a slightly smaller, refined corpus with 95,900 documents (was 103,219) from 199 sources (was 205), which is used for subsequent training and evaluation. The corpus contains 47,869 non-credible documents (49.92%) and 48,031 credible ones (50.08%).

Secondly, the original implementation uses R, Java and Python (Przybyła, 2020), but a browser extension is limited to JavaScript (JS). This requires the transition to JS versions of libraries performing crucial linguistic and computational roles. We use *compromise*[9] for tokenisation and POS tagging, *javascript-lemmatizer*[10] for lemmatisation and *TensorFlow.js*[11] for performing neural network inference.

The final issue to consider is the size of the neural model. The fact that it will be run in a browser on a PC with unknown computing resources suggests that the model should be relatively compact. In our case, the most memory-consuming part of the network is the embedding matrix: 300-dimensional representation for each word from a dictionary of 929,019 unigrams available in the word2vec-based embedding dictionary.[12] Recent studies have shown that neural models for NLP can be greatly reduced in size with little cost in terms of performance (Cheng et al., 2018; Sanh et al., 2019). Hence, we only use the embeddings corresponding to the most frequent words and introduce a special token for remaining ones, `[UNK]`. Its embedding vector is part of the set of trainable weights of the *BiLSTMAvg* model and is therefore optimised during training of the whole network. Instead of word2vec, we use *fastText*, which is a method for computing vector representations that takes into account character n-grams to account for distributional similarities between words with common sub-words. The available fastText-based embedding collections[13] are computed based on larger corpora and obtain performance superior than word2vec in several tasks (Mikolov et al., 2017). Section 4.1 contains an evaluation of several models built with differing sizes of the embedding dictionary.

3.3. Visual and explainable credibility assessment

The goal of our visual text analytics tools is to make it possible for the user to understand why a certain score has been assigned to a given document. Thus, our efforts can be viewed as seeking explanation and increasing interpretability of the model (Lipton, 2016). The different challenges posed by the classification models can be described in terms of the explainability taxonomy proposed by Doran et al. (2017). The stylometric one can be considered *interpretable*, since both its inputs (number of words, POS n-grams, etc.) and the method for computing its output (linear classification) can be shown to the user. The neural model can be described as *comprehensible*, since some of its internal information can be displayed to the user (internal representation, sentence scores), but the rest remains hidden (LSTM computations). Note that many popular interpretability tools, e.g. LIME (Ribeiro et al., 2016), are designed for a different scenario: *black-box* or *opaque* systems, when explanation is sought without any information from the inner workings of the model.

Credibilator offers two levels of user interaction. In the first level, the user can trigger the extension while browsing any news webpage, which makes the tool extract the news text from the document HTML, and compute the stylometric-based score in a small pop-up window. If the user wants to know the reasons behind the computed score, he or she can interact with Credibilator in a second level, where visualisations are presented for both classifiers, i.e. neural and stylometric models, to shed light on their inner workings. Throughout the tool, shades of green are used to denote credibility, while shades of red are used to denote non-credibility. Credibilator uses multiple coordinated views (Scherr, 2008), as this feature enables users to more effectively analyse data from different perspectives. We use the D3.js library to implement the interactive visualisations (Bostock et al., 2011).

3.3.1. Neural mode

In the neural mode, three coordinated panels are displayed: the text panel, the sentence similarity panel and the sentence map, as it can be seen in Fig. 2. Since the document credibility score is computed as an average over its sentences, each sentence in the text panel is colour-coded according to its level of non-credibility (Fig. 2A). In this way the user can inspect every sentence (especially those highlighted as likely non-credible) and make an informed decision beyond a single value for the whole document. Text highlighting using a background colour has been proved to be an effective means to attract user attention and it also resembles what people do on paper for identifying salient sentences (Self et al., 2013; Strobelt et al., 2015), including in the context of credibility visualisation (Botnevik et al., 2020).

As described in Section 3.2, the size of the neural model embedding dictionary is reduced according to token frequency in general English. Less frequent words, including most proper names, numerical expressions and uncommon vocabulary, are all represented internally by the same (`[UNK]`) token. This is visualised through the *machine view* in the text panel by blurring all such tokens (Fig. 3, right panel). This view enables the user to understand whether a word is processed or ignored, and in the latter case the user can tell that these words do not influence the credibility score.

[8] https://www.npmjs.com/package/unfluff.

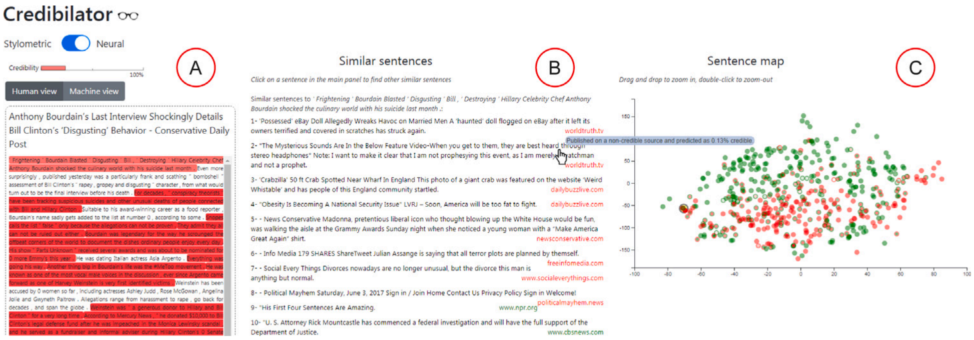
[9] https://github.com/spencermountain/compromise/.

[10] https://www.npmjs.com/package/javascript-lemmatizer.

[11] https://www.tensorflow.org/js.

[12] https://code.google.com/archive/p/word2vec/.

[13] https://fasttext.cc/docs/en/english-vectors.html.


Credibilator
Stylometric
Neural
Credibility
100%
Human view
Machine view
Anthony Bourdain's Last Interview Shockingly Details Bill Clinton's 'Disgusting' Behavior - Conservative Daily Post
A
B
C
Sentence map
Drag and drop to zoom in, double-click to zoom-out
Published on a non-credible source and predicted as 0.13% credible
150
50
0
-50
-100
-150
-100
-80
-60
-40
-20
0
20
40
60
80
100

Similar sentences

Click on a sentence in the main panel to find other similar sentences

Similar sentences to ' *Frightening ' Bourdain Blasted ' Disgusting ' Bill , ' Destroying ' Hillary Celebrity Chef Anthony Bourdain shocked the culinary world with his suicide last month .*:

1- 'Possessed' eBay Doll Allegedly Wreaks Havoc on Married Men A 'haunted' doll flogged on eBay after it left its owners terrified and covered in scratches has struck again. worldtruth.tv

2- "The Mysterious Sounds Are In the Below Feature Video-When you get to them, they are best heard through stereo headphones" Note: I want to make it clear that I am not prophesying this event, as I am merely watchman and not a prophet. worldtruth.tv

3- 'Crabzilla' 50 ft Crab Spotted Near Wharf In England This photo of a giant crab was featured on the website 'Weird Whistable' and has people of this England community startled. dailybuzzlive.com

4- "Obesity Is Becoming A National Security Issue" LVRJ – Soon, America will be too fat to fight. dailybuzzlive.com

5- - News Conservative Madonna, pretentious liberal icon who thought blowing up the White House would be fun, was walking the aisle at the Grammy Awards Sunday night when she noticed a young woman with a "Make America Great Again" shirt. newsconservative.com

6- - Info Media 179 SHARES ShareTweet Julian Assange is saying that all terror plots are planned by themself. freeinfomedia.com

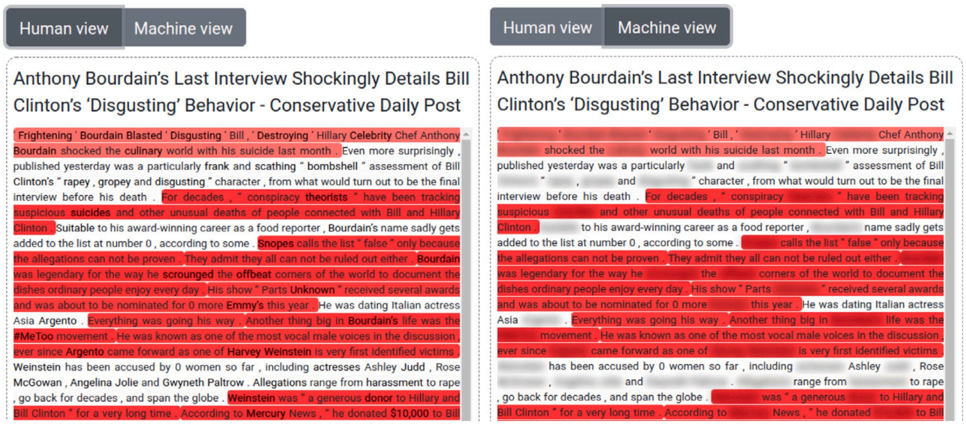
7- - Social Every Things Divorces nowadays are no longer unusual, but the divorce this man is anything but normal. www.socialeverythings.com

8- - Political Mayhem Saturday, June 3, 2017 Sign in / Join Home Contact Us Privacy Policy Sign in Welcome! politicalmayhem.news

9- "His First Four Sentences Are Amazing. www.npr.org

10- "U. S. Attorney Rick Mountcastle has commenced a federal investigation and will have the full support of the Department of Justice. www.cbsnews.com

Fig. 2. Credibilator in neural mode. (A) Text panel, (B) Sentence similarity panel, (C) Sentence map.



Anthony Bourdain's Last Interview Shockingly Details Bill Clinton's 'Disgusting' Behavior - Conservative Daily Post

' Frightening ' Bourdain Blasted ' Disgusting ' Bill , ' Destroying ' Hillary Celebrity Chef Anthony Bourdain shocked the culinary world with his suicide last month . Even more surprisingly , published yesterday was a particularly frank and scathing " bombshell " assessment of Bill Clinton's " rapey , gropey and disgusting " character , from what would turn out to be the final interview before his death . For decades , " conspiracy theorists " have been tracking suspicious suicides and other unusual deaths of people connected with Bill and Hillary Clinton . Suitable to his award-winning career as a food reporter , Bourdain's name sadly gets added to the list at number 0 , according to some . Snopes calls the list " false " only because the allegations can not be proven . They admit they all can not be ruled out either . Bourdain was legendary for the way he scrounged the offbeat corners of the world to document the dishes ordinary people enjoy every day . His show " Parts Unknown " received several awards and was about to be nominated for 0 more Emmy's this year . He was dating Italian actress Asia Argento . Everything was going his way . Another thing big in Bourdain's life was the #MeToo movement . He was known as one of the most vocal male voices in the discussion , ever since Argento came forward as one of Harvey Weinstein is very first identified victims . Weinstein has been accused by 0 women so far , including actresses Ashley Judd , Rose McGowan , Angelina Jolie and Gwyneth Paltrow . Allegations range from harassment to rape , go back for decades , and span the globe . Weinstein was " a generous donor to Hillary and Bill Clinton " for a very long time . According to Mercury News , " he donated $10,000 to Bill

Anthony Bourdain's Last Interview Shockingly Details Bill Clinton's 'Disgusting' Behavior - Conservative Daily Post

' [illegible] ' [illegible] ' [illegible] ' Bill , ' [illegible] ' Hillary [illegible] Chef Anthony [illegible] shocked the [illegible] world with his suicide last month . Even more surprisingly , published yesterday was a particularly [illegible] and [illegible] " [illegible] " assessment of Bill [illegible] " [illegible] , [illegible] and [illegible] " character , from what would turn out to be the final interview before his death . For decades , " conspiracy [illegible] " have been tracking suspicious [illegible] and other unusual deaths of people connected with Bill and Hillary Clinton . [illegible] to his award-winning career as a food reporter , [illegible] name sadly gets added to the list at number 0 , according to some . [illegible] calls the list " false " only because the allegations can not be proven . They admit they all can not be ruled out either . [illegible] was legendary for the way he [illegible] the [illegible] corners of the world to document the dishes ordinary people enjoy every day . His show " Parts [illegible] " received several awards and was about to be nominated for 0 more [illegible] this year . He was dating Italian actress Asia [illegible] . Everything was going his way . Another thing big in [illegible] life was the [illegible] movement . He was known as one of the most vocal male voices in the discussion , ever since [illegible] came forward as one of [illegible] is very first identified victims . [illegible] has been accused by 0 women so far , including [illegible] Ashley [illegible] , Rose [illegible] , [illegible] and [illegible] . [illegible] range from [illegible] to rape , go back for decades , and span the globe . [illegible] was " a generous [illegible] to Hillary and Bill Clinton " for a very long time . According to [illegible] News , " he donated [illegible] to Bill

Fig. 3. Text panel in neural mode in two variants: human view (all words) and machine view (out-of-dictionary words blurred). Sentences are highlighted with different intensities of red based on non-credibility scores.

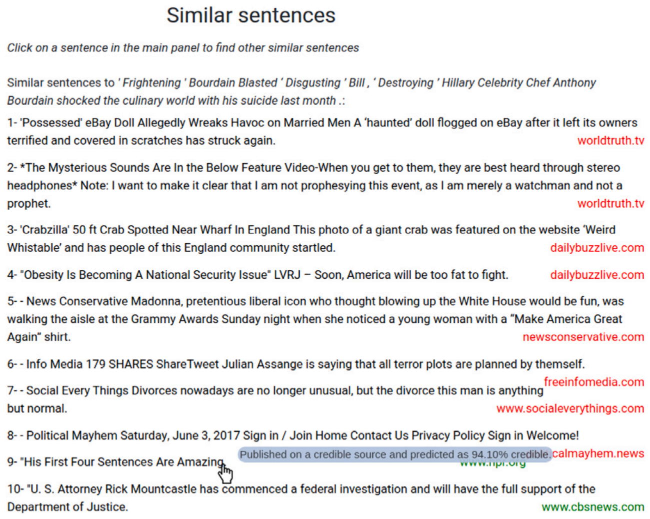
When the user clicks on a sentence, the sentence similarity panel retrieves other sentences from the training corpus that are similar according to the hidden representation drawn from the LSTM layers (Fig. 2B). The retrieved sentences from the corpus contain the predicted scores as well as the human assessment denoting the source credibility (Fig. 4). This may help a user understand why the sentence has been assigned such a score based on other sentences from the training corpus considered to be similar by the model. The sentence map depicts a 2-dimensional scatterplot with sentences from our corpus projected using a tSNE-based optimisation (van der Maaten & Hinton, 2008) of the neural-based embedded sentences (Fig. 2C). This panel is coordinated with the sentence similarity panel in such a way that the user can identify an area of the map where he or she can focus on to find common patterns among sentences similar to the one clicked. Since the number of items to be shown is considerably large (2,845,002 sentences), recommended guidelines for visual information seeking are implemented (Shneiderman, 2003), where an overview is shown first, and the user can zoom in to get more sentences on demand. The inner colour of the dots in the scatterplot encodes the credibility score predicted by the neural method, while the colour of the outline indicates the actual class label for its document. In this way, a user can identify areas of the map where documents tend to be misclassified or areas where the method is giving unconfident predictions. For instance, in Fig. 5 note the low confidence (light green/light red fill colour) of the predictions for sentences between those labelled as credible (top) and those labelled as non-credible (bottom). Visualisation of the internal representation of the model and its similarity to ground truth data instances has been explored by other methods seeking explainability (Choo et al., 2010; Rauber et al., 2017).

The video provided as part of the supplementary material highlights other relevant features and interactions available in the tool.

3.3.2. Stylometric mode

When Credibilator is in stylometric mode, it contains equivalent panels to the ones described for the neural mode, although they focus on overall document style rather than on individual sentences. This means that the document similarity panel contains a list of articles similar to the document under analysis in terms of the stylometric features described in Section 3.1.1. Analogously, the document map projects the documents in the corpus based on the tSNE projection of these features.

In addition, the stylometric mode includes a feature contribution panel, which shows the top-10 most meaningful features that contributed to the predicted score (Fig. 6A). When one feature is clicked, it is possible to explore in the feature distribution panel



Similar sentences

Click on a sentence in the main panel to find other similar sentences

Similar sentences to *' Frightening ' Bourdain Blasted ' Disgusting ' Bill , ' Destroying ' Hillary Celebrity Chef Anthony Bourdain shocked the culinary world with his suicide last month .:*

1- 'Possessed' eBay Doll Allegedly Wreaks Havoc on Married Men A 'haunted' doll flogged on eBay after it left its owners terrified and covered in scratches has struck again. worldtruth.tv

2- *The Mysterious Sounds Are In the Below Feature Video-When you get to them, they are best heard through stereo headphones* Note: I want to make it clear that I am not prophesying this event, as I am merely a watchman and not a prophet. worldtruth.tv

3- 'Crabzilla' 50 ft Crab Spotted Near Wharf In England This photo of a giant crab was featured on the website 'Weird Whistable' and has people of this England community startled. dailybuzzlive.com

4- "Obesity Is Becoming A National Security Issue" LVRJ – Soon, America will be too fat to fight. dailybuzzlive.com

5- - News Conservative Madonna, pretentious liberal icon who thought blowing up the White House would be fun, was walking the aisle at the Grammy Awards Sunday night when she noticed a young woman with a "Make America Great Again" shirt. newsconservative.com

6- - Info Media 179 SHARES ShareTweet Julian Assange is saying that all terror plots are planned by themself. freeinfomedia.com

7- - Social Every Things Divorces nowadays are no longer unusual, but the divorce this man is anything but normal. www.socialeverythings.com

8- - Political Mayhem Saturday, June 3, 2017 Sign in / Join Home Contact Us Privacy Policy Sign in Welcome! calmayhem.news

9- "His First Four Sentences Are Amazing. www.npr.org

Published on a credible source and predicted as 94.10% credible.

10- "U. S. Attorney Rick Mountcastle has commenced a federal investigation and will have the full support of the Department of Justice. www.cbsnews.com

Fig. 4. Sentence similarity panel. A ranked list of sentences is shown based on neural representation similarity. Similarly to the clicked sentence in this example, most sentences start with a quoted phrase.

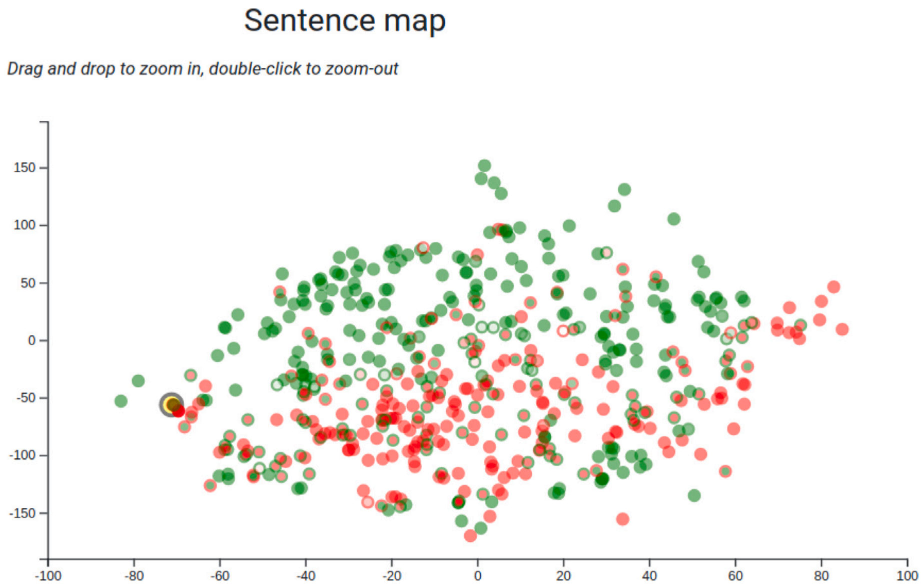


Fig. 5. Sentence map panel. Sentences from the corpus are colour-coded according to credibility label (outline) and predicted score (fill).

(Fig. 6B) how this feature value is distributed in our corpus of credible and non-credible news in terms of feature value deciles. These two panels are also coordinated with the main text panel, as words and phrases connected to the clicked feature are highlighted in the text panel to indicate their contribution to the feature value. The number of features to be shown can be increased or decreased on demand. The design choice for histograms has been also inspired by other techniques seeking explainability (Yang et al., 2019).

The *contribution* of a feature ($c(i,j)$) is a quantity expressing the degree to which the jth feature affects the non-credibility score of the ith document. In logistic regression, the score s for a document $\vec{x}_i$ is computed as:

$$s(\vec{x}_i) = \frac{1}{1+e^{-t(\vec{x}_i)}}, \quad t(\vec{x}_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j},$$

where β_j and $x_{\cdot,j}$ are the model coefficient and the value for the jth feature, respectively.

The above could suggest $c(i,j) = \beta_j x_{i,j}$ as the natural choice, since the higher this value, the higher the non-credibility score. However, we would also like the sign of c to be meaningful, i.e. positive values meaning that the feature indicates low credibility, and negative values, the opposite. For this reason we compute contribution with respect to the mean value of the feature in training

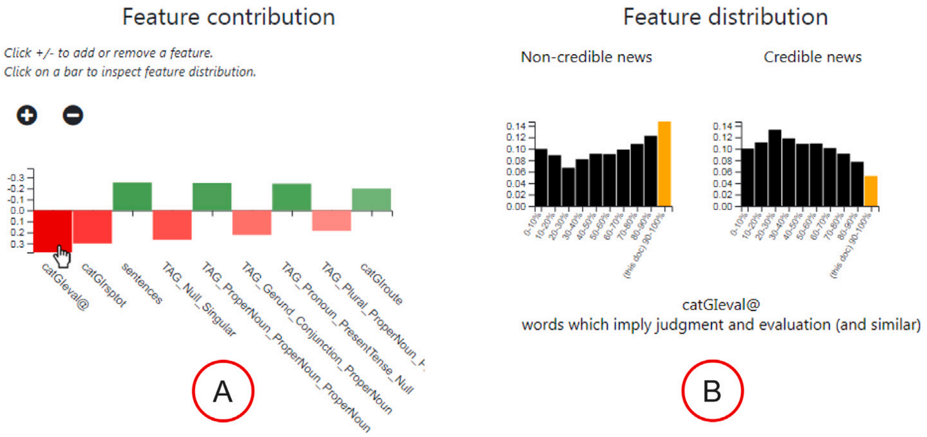
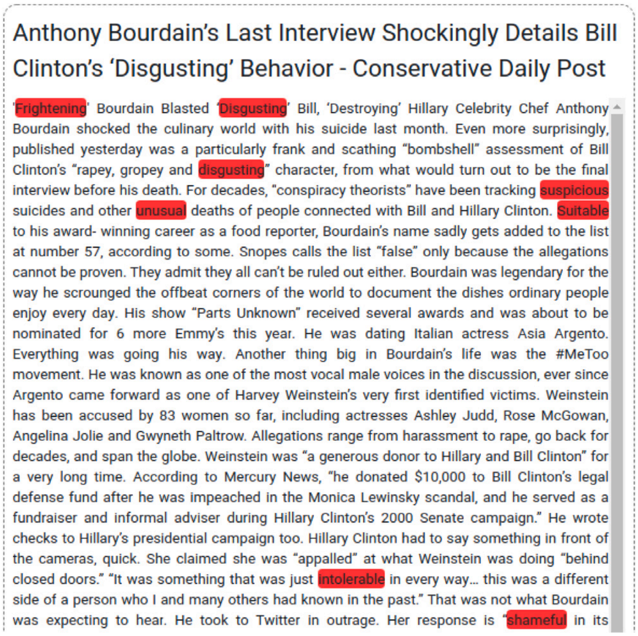


Fig. 6. Credibilator in stylometric mode. (A) Feature contribution panel. (B) Feature distribution panel.



Anthony Bourdain's Last Interview Shockingly Details Bill Clinton's 'Disgusting' Behavior - Conservative Daily Post

'Frightening' Bourdain Blasted 'Disgusting' Bill, 'Destroying' Hillary Celebrity Chef Anthony Bourdain shocked the culinary world with his suicide last month. Even more surprisingly, published yesterday was a particularly frank and scathing "bombshell" assessment of Bill Clinton's "rapey, gropey and disgusting" character, from what would turn out to be the final interview before his death. For decades, "conspiracy theorists" have been tracking suspicious suicides and other unusual deaths of people connected with Bill and Hillary Clinton. Suitable to his award- winning career as a food reporter, Bourdain's name sadly gets added to the list at number 57, according to some. Snopes calls the list "false" only because the allegations cannot be proven. They admit they all can't be ruled out either. Bourdain was legendary for the way he scrounged the offbeat corners of the world to document the dishes ordinary people enjoy every day. His show "Parts Unknown" received several awards and was about to be nominated for 6 more Emmy's this year. He was dating Italian actress Asia Argento. Everything was going his way. Another thing big in Bourdain's life was the #MeToo movement. He was known as one of the most vocal male voices in the discussion, ever since Argento came forward as one of Harvey Weinstein's very first identified victims. Weinstein has been accused by 83 women so far, including actresses Ashley Judd, Rose McGowan, Angelina Jolie and Gwyneth Paltrow. Allegations range from harassment to rape, go back for decades, and span the globe. Weinstein was "a generous donor to Hillary and Bill Clinton" for a very long time. According to Mercury News, "he donated $10,000 to Bill Clinton's legal defense fund after he was impeached in the Monica Lewinsky scandal, and he served as a fundraiser and informal adviser during Hillary Clinton's 2000 Senate campaign." He wrote checks to Hillary's presidential campaign too. Hillary Clinton had to say something in front of the cameras, quick. She claimed she was "appalled" at what Weinstein was doing "behind closed doors." "It was something that was just intolerable in every way… this was a different side of a person who I and many others had known in the past." That was not what Bourdain was expecting to hear. He took to Twitter in outrage. Her response is "shameful" in its

Fig. 7. Text panel in stylometric mode. Words are highlighted to make evident the connection with the interacted features.

data ($\overline{x_j}$):

$$c(i,j) = \beta_j (x_{i,j} - \overline{x_j}).$$

This formulation is equivalent to feature importance in the SHAP interpretability framework in the case of linear models (Lundberg & Lee, 2017).

For example, the feature computed as percentage of lowercase words in the text has coefficient $\beta_j = -2.98$ (such words indicate higher credibility) and mean value $\overline{x_j} = 0.83$. As a result, c will be positive (increased non-credibility) for documents with less than 83% of lowercase words, c will be negative (increased credibility) for documents with more.

For instance, in Fig. 6A we see that the most discriminative feature is `catGIEval@` (i.e. number of words which imply judgement and evaluation), which is currently selected. We can then see in the feature distribution panel (Fig. 6B) that our document has a value for this feature that is in the last decile of the feature range (orange bar: top 10% of values observed in training data), and that such a high ratio of these words is more common in non-credible news than in credible news, hence this feature contributes towards non-credibility (red bar). When feature `@catGI` is clicked, every occurrence of words belonging to this category, e.g. *frightening*, *disgusting*, *suspicious*, etc., is highlighted in the text panel (Fig. 7).

Finally, a back-end service over the web is run for the retrieval of stylistically similar documents and sentences in our corpus. While fast retrieval of similar text is needed to facilitate proper interactivity, it is important not to transfer plain text to account for users' privacy. These aspects are achieved by implementing a two-level retrieval mechanism using the classification features (either the stylistic features or the neural sentence embeddings). At the first level, locality-sensitive hashing (Indyk et al., 1997; Ole Krause-Sparmann, 2018) is applied for constant time retrieval of candidate sentences or documents. Hashes of 10 bits are used, which gives a collision probability close to 0.1%. This means that about 2000–3000 sentences are selected as candidates for pairwise comparison. At the second level, cosine similarity between the query features and the retrieved candidate features is performed for final ranking. The top ten sentences or documents are retrieved after each search. Therefore, no plain text is transmitted and a constant-time search on a large corpus is achieved. The back-end is also used for resampling new documents as the user zooms in a given region of the sentence or document map.

4. Evaluation

Credibilator is evaluated in two ways. First, we measure the performance of the classification models through cross-validation to learn the optimal model size and compare with previous work. Second, we conduct user-based evaluations to check how the interaction with Credibilator affects the accuracy and confidence of human credibility assessment.

4.1. Model evaluation

The purpose of the model evaluation is to check how the adaptations (described in Section 3.2), which were motivated by the limited resources, affect the classification accuracy. Preferably, we would like the reduced model to perform as well as the original one.

For evaluation we use the data published with the classifiers we base our work on (Przybyła, 2020),[14] split in two five-fold cross-validation (CV) scenarios. In *document CV*, each document is independently and randomly assigned to a CV fold. In *source CV*, all documents from a given source (news website) are assigned to the same fold. This scenario is more challenging, as it means the classifier is tested on documents from sources unseen in training. To enable fair comparison with previous work evaluated on the same dataset (Przybyła, 2020), we use the same CV folds. We use accuracy instead of precision or recall since the dataset is fairly balanced (49.92% documents come from non-credible sources).

We perform two experiments of this type. Firstly, we check how the size of the embedding dictionary (2 million; 1 million; 100,000; 10,000; 1,000 or 100) affects the performance of the neural classifier. Secondly, the accuracy obtained by both classifiers in both CV scenarios is compared to previous work (Przybyła, 2020) and a commonly used general-purpose text classification approach: tuned BERT.[15]

4.2. User study A

In order to evaluate how helpful Credibilator is in assessing trustworthiness of news articles, we recruited 14 participants for the first user study. Each participant is presented with five articles related to five current topics. Each topic has a credible and a non-credible news story and the user's task is to judge their credibility using either the stylometric or the neural-based interface. To ensure the users focus on text style rather than source name or webpage appearance, we reformat the articles to plain HTML files. The users rate each article on a scale from 1 (*Not credible at all*) to 5 (*Very credible*): first by reading them without any automated assistance (i.e. unaided) and then again after interacting with Credibilator (i.e. aided). Finally, we ask the participants to answer some general questions about their experience with the system.

We selected five topics recently (as of September 2020) covered by sources included in the study. Then, for each of these topics we found one article from a non-credible (NC) and a credible (C) source. While the articles describe current matters and were not used before, the sources they come from (and their credibility labels) are taken from the training corpus (Przybyła, 2020). The selected articles are shown in Table 1.

Each participant was given five *tasks* corresponding to the topics above. A task has two parameters:

- *model*: stylometric (S) or neural (N), i.e. whether the user is expected to use the stylometric or neural mode while using Credibilator.
- *credibility*: credible (C) or non-credible (NC), i.e. the true credibility label of the article displayed for a given topic. This parameter is hidden from the user.

Table 2 shows the parameters of the tasks assigned to users. These assignments were designed so that each user can interact with both models and both credibility levels. In each task, a user was expected to perform the following actions:

1. Read the article provided,
2. Rate its credibility (*unaided*),
3. Analyse the credibility of the article using the specified model in Credibilator,
4. Rate the credibility again (*aided*).

[14] The documents discarded because of containing insufficient amount of text following the improved text extraction (see Section 3.2) were not taken into account in the evaluation.

[15] BERT Base tuned for 3 epochs in document classification setup, with each article limited to the first 256 tokens.

Table 1
Articles provided to the users for credibility assessment.

Topic	Source	Title	URL	Cred.
A: Nomination of Donald Trump for Nobel peace prize	Conservative Daily Post	Trump Finally Nominated for Nobel Peace Prize After President Actually Delivers Peace	https://conservativedailypost.com/trump-finally-nominated-for-nobel-peace-prize-after-president-actually-delivers-peace/	NC
	BBC News	Trump Nobel Peace Prize nomination - what you need to know	https://www.bbc.com/news/world-us-canada-54092960	C
B: State of Virginia passing new legislation on policing	enVolve	DANGER: Virginia's Radically Left Democrats Pass SICKENING Bill That Opens Up Police To Assault!	https://en-volve.com/2020/08/27/danger-virginias-radically-left-democrats-pass-sickening-bill-that-opens-up-police-to-assault/	NC
	The Washington Post	Virginia Senate passes sweeping police overhaul bill	https://www.washingtonpost.com/local/virginia-politics/virginia-senate-passes-sweeping-police-overhaul-bill/2020/09/10/1f85a0b2-f36e-11ea-999c-67ff7bf6a9d2_story.html	C
C: Water contamination in Flint, Michigan	News 4	22 Dead Bodies Discovered In Flint River Found To Be The Source Of Water Contamination	http://news4ktla.com/22-dead-bodies-discovered-flint-river-found-source-water-contamination/	NC
	CBS News	Charges dropped against 8 people in Flint water scandal	https://www.cbsnews.com/news/flint-michigan-water-crisis-charges-dropped-against-several-people-today-2019-06-13/	C
D: NFL opening the season with a minute of silence	Mad World News	NFL Gets Reality Check From Fans During 'Social Justice' Moment Of Silence	https://madworldnews.com/nfl-social-justice-moment-silence/	NC
	ABC News	Gabrielle Union reacts to NFL fans booing moment of silence: 'How do you boo unity?'	https://abcnews.go.com/Entertainment/gabrielle-union-reacts-nfl-fans-booing-moment-silence/story?id=73052831	C
E: Kanye West releasing a new album	Neon Nettle	Kanye West Credited With Spike in People Googling About Christianity and Jesus	https://neonnettle.com/features/1734-kanye-west-credited-with-spike-in-people-googling-about-christianity-and-jesus	NC
	NPR	Kanye West's 'Jesus Is King,' Like Its Creator, Asks A Little Too Much Of Us	https://www.npr.org/2019/10/28/774137186/kanye-wests-jesus-is-king-like-its-creator-asks-a-little-too-much-of-us?t=1600437946200&t=1602247618502	C

We collected 140 (14 participants × 5 topics × 2 ratings) credibility ratings. Each of them was expressed on a scale 1–5 with the following explanation:

- 1: Not credible at all,
- 2: Rather not credible,
- 3: I don't know,
- 4: Fairly credible,
- 5: Very credible.

Due to the pandemic, the study was performed remotely. Each participant received a one-page document describing the research goals plus a five-minutes video to showcase the tool. This video is included in the supplementary material.

Finally, after finishing all the tasks, each user answered a questionnaire regarding their general experience with Credibilator. The questions and possible answers are provided in Table 3. Further details about the design of this user study can be found in the Supplementary material.

Table 2
Design of user study A in terms of model used and credibility assigned to each participant.

User	Topic									
	Model					Credibility				
	A	B	C	D	E	A	B	C	D	E
1	S	N	S	N	S	C	C	NC	NC	C
2	S	N	S	N	S	NC	NC	C	C	C
3	N	S	N	S	N	NC	C	C	C	NC
4	S	N	S	N	S	C	C	C	NC	NC
5	N	S	N	S	N	NC	NC	C	C	NC
6	S	N	S	N	S	NC	C	C	NC	NC
7	N	S	N	S	N	C	C	NC	NC	NC
8	S	N	S	N	S	C	NC	NC	NC	C
9	N	S	N	S	N	NC	NC	NC	C	C
10	S	N	S	N	S	C	C	NC	NC	C
11	N	S	N	S	N	C	NC	NC	C	C
12	S	N	S	N	S	NC	NC	C	C	C
13	S	N	S	N	S	C	C	C	NC	NC
14	S	N	S	N	S	NC	C	C	NC	NC

Table 3
Questions and answers included in the questionnaire for the study participants.

ID	Question		Answers
Q1	How much do you agree with the following statement:	It was easier to assess the text credibility thanks to knowing the credibility scores.	A1: Strongly agree A2: Agree A3: Neither agree nor disagree A4: Disagree A5: Strongly disagree
Q2	*as above*	It was easier to assess the text credibility thanks to interacting with the visual interface.	*as above*
Q3	*as above*	I could understand how the visualised factors influenced the final score.	*as above*
Q4	*as above*	The interface was easy to use.	*as above*
Q5	*as above*	Such a tool could be helpful in my web browsing.	*as above*
Q6	Which variant of the tool would you prefer?		A1: Stylometric classifier A2: Neural classifier A3: I'm not sure
Q7	Please share any other comments you have below:		(Text field)

4.3. User study B

It is important to acknowledge that there might be a variety of additional factors contributing to the results of the user study A, such as the participants' unconscious bias towards relying on the automated tool or simply closer inspection of the text while reading the article for a second time. To filter these factors out, we prepared a modified version of our tool, with the same interface, yet displaying inaccurate results and visualisations. We make sure that results look consistent with the corresponding explanations, as otherwise users may notice an abnormal behaviour instead of an incorrect automatic assessment.

The faulty version of Credibilator was implemented differently depending on whether the stylometric or the neural versions are used. For the stylometric classifier, we used a logistic regression model with randomly perturbed coefficients. Searches on the similarity map were performed by randomly permuting the feature positions. This leads to documents in the corpus which are likely not stylistically similar to the target document, but at least similar to each other. For the neural-based classifier, the faulty model used random weights. This makes the highlighting of sentences consistent, although incorrect. The search for similar sentences was done using the activations of this last layer, and, similarly to the stylometric classifier, it yielded a cluster of similar documents associated to the random activations of the last layer.

We recruited nine new participants for this second user study. The setup was similar to the one used in the first user study: the same ten articles were used (five credible and five non-credible) and we randomised the credibility of the articles as well as the type of classifier participants had to use for each article. We used Table 2 to separate the tasks assigned to each user. In addition, we randomised the tasks where a user had to use the faulty version of Credibilator. Throughout the study, either version of the tool was used for half of the credibility assessments. Further details about the design of this user study can be found in the Supplementary material.

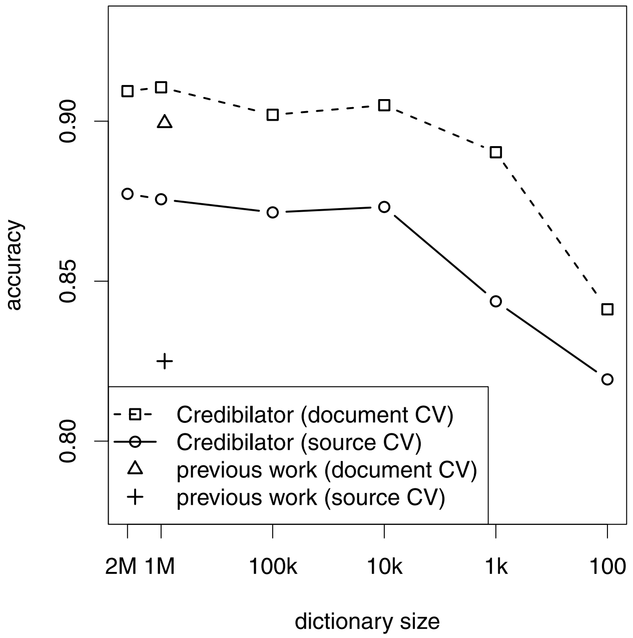


Fig. 8. Classification accuracy of Credibilator's neural model, measured in two evaluation scenarios (document cross-validation and source cross-validation) for different dictionary sizes (x axis, logarithmic), compared to previous results on the same corpus (Przybyła, 2020).

Execution of this user study was also similar to the first one: participants rated credibility of the article with and without using Credibilator. The main difference with the previous user study was that Credibilator, unbeknownst to the user, may be functioning in a faulty manner. There were also no questions about usability or the preference of one classifier towards the other as these issues were addressed in the first study. The study was also carried out remotely and the same five-minute video was provided to showcase the tool.

5. Results

In this section we provide the results of the experiments described in Section 4. Similarly, we divide these results into those using standard machine learning evaluation schemes and those about the user studies.

5.1. Model evaluation

The performance of the neural model with respect to the size of the embedding dictionary is shown in Fig. 8. Firstly, we can compare our adapted approach to the original implementation for a similar dictionary size. While the difference in document CV is quite small, in source CV the new approach achieves noticeably better results. It appears that overfitting to known sources is reduced in the new version, most likely due to better document representation: using *fastText* embeddings and improved text extraction through the *unfluff* library. Understandably, the accuracy drops as the dictionary size is reduced, but the accuracy loss remains negligible until we reach the level of 1000 or 100 words. Based on these observations, we proceed using a model with 10,000 words, which is 15 MB in size, compared to the 478 MB when using the full vocabulary.

In Table 4 we show how our adapted preprocessing and classification procedures affect the accuracy compared to the previous work (Przybyła, 2020). We can see that the only noticeable performance drop happens in the document CV scenario of the stylometric classifier, which could be interpreted as a reduction in overfitting. In the more realistic scenario (source CV: unseen sources in the test data), the accuracy remains virtually the same (stylometric) or improves greatly (neural) with respect to the reference work. The size of the neural model is reduced by tuning the embedding dictionary, but it remains much larger than the stylometric model, which may affect its interpretability. The BERT model appears to benefit even more thanks to the preprocessing adaptations, providing the best performance in both scenarios. Note however that this classifier is not implemented as part of Credibilator since making such a complex model both interpretable and usable in a restricted environment is outside the scope of our study.

5.2. User study A

Table 5 shows the credibility assessments provided by the study participants, before and after using Credibilator. To measure the correctness of these assessments, we compare them to the ground truth credibility labels. *Accuracy* is the percentage of the

Table 4
Comparison between the classification performance of different versions of the discussed models. Includes two variants of preprocessing: the original implementation (Przybyła, 2020) and our solution, adapted to browser environment. Regarding classifiers, we compare solutions based on style (stylometric and neural) with tuned BERT. The table shows classification accuracy in two CV scenarios and model size (number of parameters).

Preprocessing	Classifier	Accuracy		Size
		doc. CV	source CV	
Original	Stylometric (original)	0.9274	0.8097	939
	Neural (original)	0.8994	0.8250	85.18M
	BERT	0.9976	0.7960	110M
Adapted	Stylometric (Credibilator)	0.8680	0.8081	617
	Neural (Credibilator)	0.9050	0.8732	3.26M
	BERT	0.9629	0.9292	110M

Table 5
Credibility assessments provided by the users. The more accurate assessment for each user and task is underlined.

User	Task	Unaided	Aided	User	Task	Unaided	Aided
1	A	5	5	8	A	5	5
1	B	3	4	8	B	1	1
1	C	2	1	8	C	2	1
1	D	3	2	8	D	2	2
1	E	3	5	8	E	4	5
2	A	2	2	9	A	4	3
2	B	4	1	9	B	3	2
2	C	4	5	9	C	2	2
2	D	4	5	9	D	2	2
2	E	5	5	9	E	4	4
3	A	1	1	10	A	4	5
3	B	4	5	10	B	2	4
3	C	5	4	10	C	1	1
3	D	2	4	10	D	4	2
3	E	3	3	10	E	4	5
4	A	2	5	11	A	5	5
4	B	5	3	11	B	2	2
4	C	3	5	11	C	1	1
4	D	1	2	11	D	4	4
4	E	5	3	11	E	5	5
5	A	2	2	12	A	1	1
5	B	1	1	12	B	1	1
5	C	4	4	12	C	4	5
5	D	5	4	12	D	5	5
5	E	4	3	12	E	5	5
6	A	1	2	13	A	4	4
6	B	3	4	13	B	2	2
6	C	5	5	13	C	2	4
6	D	1	1	13	D	1	1
6	E	2	4	13	E	2	3
7	A	4	2	14	A	2	2
7	B	4	4	14	B	4	5
7	C	2	1	14	C	4	4
7	D	1	1	14	D	2	2
7	E	1	1	14	E	4	2

Table 6
Accuracy and confidence of credibility assessments made by users in user study A before using the tool (*unaided*) and after interacting with Credibilator (*aided*). Values shown with respect to classifier used and true credibility of the article.

	Accuracy		Confidence	
	Unaided	Aided	Unaided	Aided
Stylometric	75.68%	89.19%	62.16%	74.32%
Neural	69.70%	81.82%	63.64%	65.15%
Credible	72.22%	88.89%	59.72%	75.00%
Non-credible	73.53%	82.35%	66.18%	64.71%
All	72.86%	85.71%	62.86%	70.00%

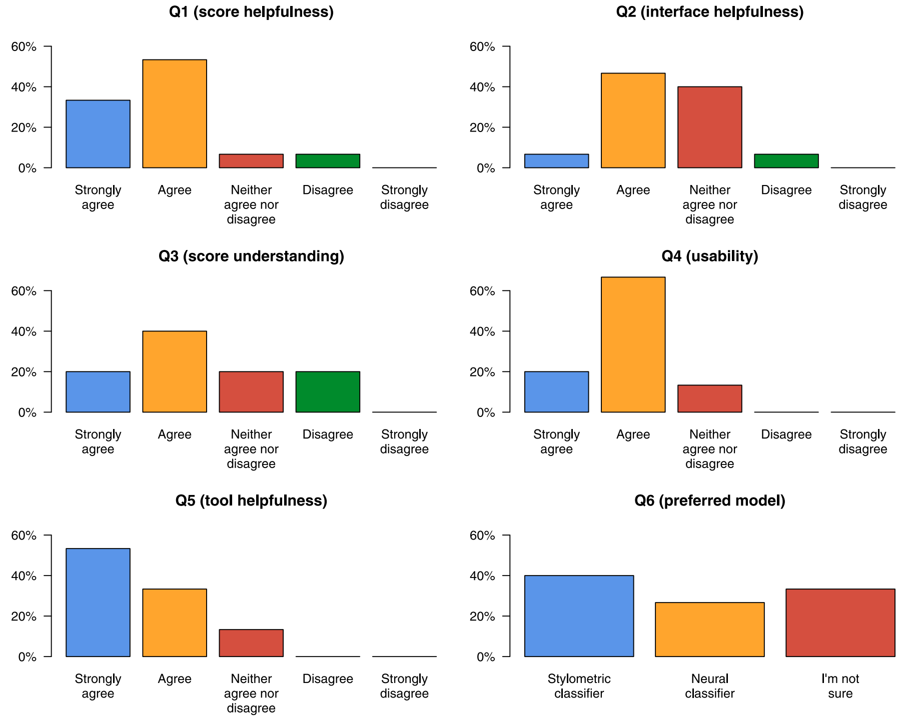


Fig. 9. Distribution of answers to the final survey.

participants leaning towards the correct label (1 or 2 for non-credible; 4 or 5 for credible). *Confidence* is computed as distance from the undecided answer (3, *I don't know*) and normalised to 0%–100%.

Table 6 shows the obtained results in terms of accuracy and confidence. The main conclusion is that using Credibilator increases the rate of correct assessments from 72.86% to 85.71% and this difference is significant ($p < 0.05$) according to the randomised permutation test (Morgan, 2006). While accuracy improves after using either classifier, the confidence score remains similar in the neural variant. Looking at the breakup with respect to true credibility, Credibilator helps spot non-credible documents, but it is even more effective in strengthening confidence in the credible ones.

Fig. 9 provides the results of the final survey as the percentage of participants choosing each answer (see Section 4.2). The results show that while the visual interface was rather helpful than unhelpful (Q2: 53% *[Strongly] agree*[16] vs 7% *[Strongly] disagree*) the single credibility score performed much better (Q1: 87% *[Strongly] agree* vs 7% *[Strongly] disagree*). This is despite the fact that only 20% could not understand entirely how the visualised factors influenced the final score (Q3) and the interface was widely considered easy to use (Q4: 87%). In accordance with better accuracy and confidence of users' judgements when using the stylometric mode, it was more often preferred (40%) than the neural one (27%). A summary of all the free-text comments have been manually organised for enhanced readability and shown in Fig. 10.

5.3. User study B

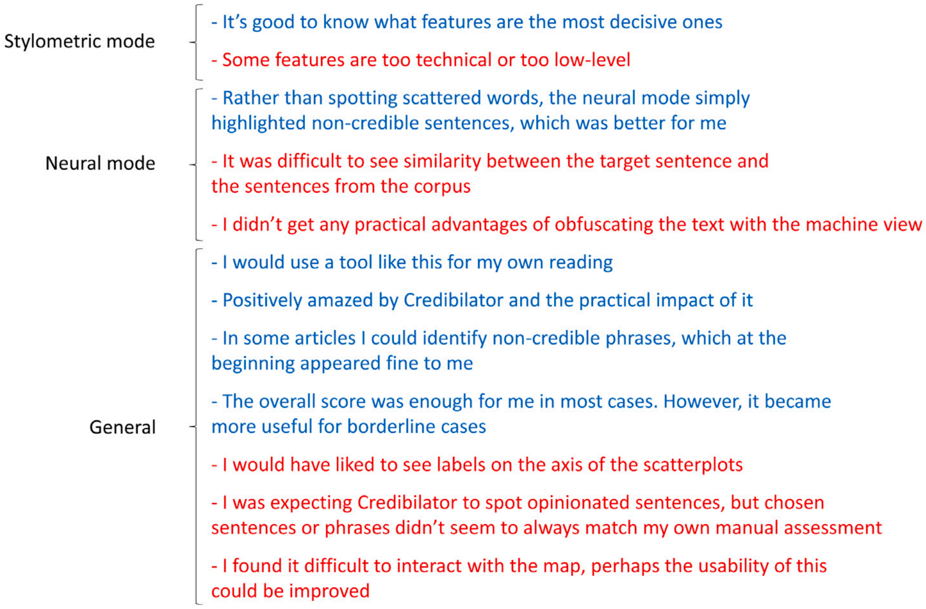
The accuracy and confidence of credibility assessments made by the participants in the second user study is shown in Table 7. Interestingly, accuracy increases by less than 5% when using the faulty Credibilator, likely due to closer inspection of article text while using the tool. However, a much larger difference in accuracy can be recognised (about 18%) when the real version of Credibilator is used.

The situation is different when evaluating confidence. While average participants' confidence does not change when the faulty Credibilator was used, there is a 14% increase in confidence when the real Credibilator was used. This result clearly suggests that the tool is virtually ineffective in influencing users if inaccurate explanations are provided.

6. Discussion

One of the questions we aimed to answer here was how much accuracy would decrease following the transition to the constrained browser environment. We noticed some loss in the case of the stylometric model, which is most likely because the NLP tools that

[16] I.e., *Strongly agree* or *Agree*.



Stylometric mode
- It's good to know what features are the most decisive ones
- Some features are too technical or too low-level

Neural mode
- Rather than spotting scattered words, the neural mode simply highlighted non-credible sentences, which was better for me
- It was difficult to see similarity between the target sentence and the sentences from the corpus
- I didn't get any practical advantages of obfuscating the text with the machine view

General
- I would use a tool like this for my own reading
- Positively amazed by Credibilator and the practical impact of it
- In some articles I could identify non-credible phrases, which at the beginning appeared fine to me
- The overall score was enough for me in most cases. However, it became more useful for borderline cases
- I would have liked to see labels on the axis of the scatterplots
- I was expecting Credibilator to spot opinionated sentences, but chosen sentences or phrases didn't seem to always match my own manual assessment
- I found it difficult to interact with the map, perhaps the usability of this could be improved

Fig. 10. Summary of free-text comments provided by the participants. Comments in blue and red reflect positive and negative aspects, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 7
Accuracy and confidence of credibility assessments made by users in user study B before using the tool (*unaided*) and after interacting with Credibilator (*aided*) in the faulty or real variant.

	Accuracy		Confidence	
	Unaided	Aided	Unaided	Aided
Faulty Credibilator	61.90%	66.67%	57.14%	57.14%
Real Credibilator	66.67%	85.71%	52.38%	66.67%

are available for this situation are not as accurate as the *Stanford CoreNLP* (Manning et al., 2014) used in previous experiments. More interestingly, the adapted version of the neural model performed better than the original. The gains are larger in the more challenging evaluation scenario (source CV), which suggests that the lack of representation of less common words is helpful for reducing overfitting. While a significant amount of work has been done in decreasing neural model size (Cheng et al., 2018), including reducing computation precision (Gupta et al., 2015), removing architectural elements (attention heads) (Michel et al., 2019) and knowledge distillation (Sanh et al., 2019), less attention has been given to the initial representation layer. One of the implications of our research is that frequency-based dictionary pruning can be a quick and easy method for compressing neural models in NLP. A possible avenue of further research is to explore other methods for decreasing the size of embedding layers, e.g. through dimensionality reduction (Raunak et al., 2019).

The best performance level observed in source CV (87% accuracy) suggests that there is still room for improvement in the task. One of the big challenges for research in credibility assessment has been the lack of commonly agreed benchmark datasets or evaluation scenarios taking into account the problem of overfitting to sources or topics. We hope our dataset will be helpful in that respect and our experiments will be followed by evaluation of other approaches. Since a classifier based on a pretrained language model provides better performance than style-based methods on this dataset (Przybyła, 2020), making such solutions interpretable and usable in constrained user environments is a promising avenue of research.

Another question that we investigated is whether and what interactive visualisations allow inspecting the rationale behind the decision of the classifier. Results showed how confidence increases after interaction with the tool, and this same sense of reassurance of the credibility results was reported in the free-form responses. Visual metaphors used in this tool (e.g. scatter plots, bar plots, text highlighting) are well established and previously used for similar efforts, which make them sensible design options. We also think that multiple coordinated views (Scherr, 2008) was an effective interactive strategy in this scenario, since it is also a well-studied strategy in the context of exploratory analysis and insight gaining.

While we observed a clear and statistically significant difference in credibility assessment accuracy, there are some limitations to this result. Most noticeably, the number of participants in the user studies was relatively small, which limited our ability to draw more fine-grained conclusions. For example, it would be beneficial to compare the influence of the explainable models we used to that of simple warning labels evaluated before (Clayton et al., 2019; Pennycook et al., 2020). Furthermore, we verified the users' perception of news credibility following the interaction with the tool, while others focused on their actions, such as sharing in social media (Mena, 2020).

The most interesting conclusion from our experiments is that while the neural approach performed better in terms of automatic accuracy (87% vs 81%), the feature-based classifier was more helpful to the users, which was shown in their accuracy, confidence and survey results. This is consistent with the observations that learning an explanation for ML result increases users' trust in it (Yang et al., 2020) and the fact that the linear models could be more thoroughly explained due to their (relative) simplicity. Moreover, we aimed at explaining the way the classifier has reached a given result, but generating explanations in the *black-box* scenario (e.g. as in LIME) remains a valid avenue for further research. Whether the correspondence between the generated explanation and the classifier inner workings is important for gaining user trust remains an open question.

The computational complexity of the underlying methods of Credibilator is well suited for real-time analysis of news. In the case of the neural model, each sentence is tokenised and fed to the neural architecture to output its score in constant time. This is repeated for all sentences. In the case of the stylometric analysis, it requires a full scan of the article to extract and count the different features, which are then combined in a single logistic function. Therefore, for both methods the processing time grows linearly with the length of the text.

In addition to the scenario described in the first user study, we envisioned other application scenarios where Credibilator can be used. One scenario is that of a linguist, who is interested in understanding language-related traits for credible or non-credible news articles. The focus of Credibilator in the stylometric analysis, and the visual and interactive presentation of results would allow them to investigate the stylistic features of each type of text. Another scenario would be a journalist who may want to check if their own text has features associated with low-credibility.

Finally, our user studies demonstrate that data-based classifiers could be translated into tools helping users in recognising low-credibility text. It opens up a new perspective towards performing a similar transformation of other misinformation-addressing solutions. For example, credibility assessment could also be performed through fact-checking, which poses a large challenge for visual analytics, since it requires taking into account the external sources used as a reference and explaining the reasons for matching sentences. In many use cases in the misinformation area, gaining users' trust requires taking into account several factors beyond accuracy, including interpretability and user experience.

7. Conclusion

To the best of our knowledge, this work presents the first evidence that automatic credibility assessment can not only perform well in terms of predictive accuracy, but also make the human perception of credibility significantly more accurate. Specifically, we have shown that carefully reducing the size of the embedding dictionary, we can achieve performance comparable, if not better, with the full model. Moreover, we have proposed a suite of interactive visualisations and shown how they can explain the automatic credibility scores, both for the neural and feature-based models. After using the interface, the credibility assessments by the participants of the user studies were significantly more likely to agree with the experts' judgement. Further research and development in this area is necessary to turn accurate machine learning models into user-centred tools that are helpful in addressing misinformation.

CRediT authorship contribution statement

Piotr Przybyła: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Project administration. **Axel J. Soto:** Methodology, Software, Validation, Investigation, Writing - original draft, Writing - review & editing, Visualization.

Acknowledgements

This work was supported by the *Polish National Agency for Academic Exchange* through a *Polish Returns* grant number PPN/PPO/2018/1/00006 and a computing grant number 447 at the *Poznan Supercomputing and Networking Center* in Poland. We also thank *Ministerio de Ciencia y Tecnología (Argentina)* for its grant PIDRI PRH-2017-0007, and *Google* for its *2020 Latin American Research Award*.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ipm.2021.102653.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., ... Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv:1603.04467.

Ahmed, H., Traore, I., & Saad, S. (2017). Detection of online fake news using N-gram analysis and machine learning techniques. In *Intelligent, secure, and dependable systems in distributed and cloud environments* (pp. 127–138). Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-69155-8_9, URL: http://link.springer.com/10.1007/978-3-319-69155-8_9.

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, *52*(2), 119. http://dx.doi.org/10.1145/1461928.1461959, URL: http://dl.acm.org/ft_gateway.cfm?id=1461959&type=html.

Atanasova, P., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Karadzhov, G., Mihaylova, T., Mohtarami, M., & Glass, J. (2019). Automatic fact-checking using context and discourse information. *Journal of Data and Information Quality*, *11*(3), 1–27. http://dx.doi.org/10.1145/3297722, arXiv:1908.01328, URL: https://dl.acm.org/doi/10.1145/3297722.

Bakir, V., & McStay, A. (2017). Fake news and the economy of emotions: Problems, causes, solutions. *Digital Journalism*, *6*(2), 154–175. http://dx.doi.org/10.1080/21670811.2017.1345645.

Boididou, C., Papadopoulos, S., Zampoglou, M., Apostolidis, L., Papadopoulou, O., & Kompatsiaris, Y. (2018). Detection and visualization of misleading content on Twitter. *International Journal of Multimedia Information Retrieval*, *7*(1), 71–86.

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2301–2309. http://dx.doi.org/10.1109/TVCG.2011.185.

Botnevik, B., Sakariassen, E., & Setty, V. (2020). BRENDA: Browser extension for fake news detection. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval* (pp. 2117–2120). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3397271.3401396, URL: https://doi.org/10.1145/3397271.3401396.

Brooks, M., Amershi, S., Lee, B., Drucker, S. M., Kapoor, A., & Simard, P. (2015). Featureinsight: Visual support for error-driven feature ideation in text classification. In *2015 IEEE conference on visual analytics science and technology* (pp. 105–112). http://dx.doi.org/10.1109/VAST.2015.7347637.

Cheng, Y., Wang, D., Zhou, P., & Zhang, T. (2018). Model compression and acceleration for deep neural networks: The principles, progress, and challenges. *IEEE Signal Processing Magazine*, *35*(1), 126–136. http://dx.doi.org/10.1109/MSP.2017.2765695.

Choi, W., & Stvilia, B. (2015). Web credibility assessment: Conceptualization, operationalization, variability, and models. *Journal of the Association for Information Science and Technology*, *66*(12), 2399–2414. http://dx.doi.org/10.1002/asi.23543, URL: https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/asi.23543, https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23543, https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.23543.

Choo, J., Lee, H., Kihm, J., & Park, H. (2010). Ivisclassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In *2010 IEEE symposium on visual analytics science and technology* (pp. 27–34). http://dx.doi.org/10.1109/VAST.2010.5652443.

Ciampaglia, G. L., Mantzarlis, A., Maus, G., & Menczer, F. (2018). Research challenges of digital misinformation: Toward a trustworthy web. *AI Magazine*, *39*(1), 65. http://dx.doi.org/10.1609/aimag.v39i1.2783, URL: https://144.208.67.177/ojs/index.php/aimagazine/article/view/2783.

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., & Flammini, A. (2015). Computational fact checking from knowledge networks. In A. Barrat (Ed.), *PLoS One*, *10*(6), Article e0128193. http://dx.doi.org/10.1371/journal.pone.0128193.

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., Sandhu, M., Sang, R., Scholz-Bright, R., Welch, A. T., Wolff, A. G., Zhou, A., & Nyhan, B. (2019). Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, http://dx.doi.org/10.1007/s11109-019-09533-0, URL: http://link.springer.com/10.1007/s11109-019-09533-0.

Diermeier, D., Godbout, J.-F., Yu, B., & Kaufmann, S. (2011). Language and ideology in congress. *British Journal of Political Science*, *42*(01), 31–55, URL: http://journals.cambridge.org/abstract_S0007123411000160.

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. In *Proceedings of the first international workshop on comprehensibility and explanation in AI and ML 2017 co-located with 16th international conference of the italian association for artificial intelligence*. Bari, Italy.

Du, J., Huang, Y., & Moilanen, K. (2020). Pointing to select: A fast pointer-LSTM for long text classification. In *Proceedings of the 28th international conference on computational linguistics* (pp. 6184–6193). Barcelona, Spain (Online): International Committee on Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.coling-main.544, URL: https://www.aclweb.org/anthology/2020.coling-main.544.

Gelfert, A. (2018). Fake news: A definition. *Informal Logic*, *38*(1), 84–117. http://dx.doi.org/10.22329/il.v38i1.5068, URL: https://ojs.uwindsor.ca/index.php/informal_logic/article/view/5068.

Gillin, J. (2017). Politifact's guide to fake news websites and what they peddle. *PolitiFact*, URL: http://www.politifact.com/punditfact/article/2017/apr/20/politifacts-guide-fake-news-websites-and-what-they/.

Gupta, S., Agrawal, A., Gopalakrishnan, K., & Narayanan, P. (2015). Deep learning with limited numerical precision. In *32nd international conference on machine learning*. arXiv:1502.02551.

Hilligoss, B., & Rieh, S. Y. (2008). Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. *Information Processing and Management*, *44*(4), 1467–1484. http://dx.doi.org/10.1016/j.ipm.2007.10.001.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Horne, B. D., & Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the 2nd international workshop on news and public opinion at ICWSM*. Association for the Advancement of Artificial Intelligence, arXiv:1703.09398.

Indyk, P., Motwani, R., Raghavan, P., & Vempala, S. (1997). Locality-preserving hashing in multidimensional spaces. In *Proceedings of the twenty-ninth annual ACM symposium on theory of computing* (pp. 618–625).

Jankowska, M., Kešelj, V., & Milios, E. (2012). Relative N-gram signatures: Document visualization at the level of character N-grams. In *2012 IEEE conference on visual analytics science and technology* (pp. 103–112). IEEE.

Kakol, M., Nielek, R., & Wierzbicki, A. (2017). Understanding and predicting web content credibility using the content credibility corpus. *Information Processing and Management*, *53*(5), 1043–1061. http://dx.doi.org/10.1016/j.ipm.2017.04.003.

Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. In *3rd international conference on learning representations, ICLR 2015 - conference track proceedings*. San Diego, USA: International Conference on Learning Representations, ICLR, arXiv:1412.6980.

Koppel, M., Argamon, S., & Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, *17*(4), 401–412. http://dx.doi.org/10.1093/llc/17.4.401, URL: http://llc.oxfordjournals.org/content/17/4/401.abstract.

Kucher, K., Martins, R. M., Paradis, C., & Kerren, A. (2020). Stancevis prime: visual analysis of sentiment and stance in social media texts. *Journal of Visualization*, *23*(6), 1015–1034.

Kucher, K., Paradis, C., Sahlgren, M., & Kerren, A. (2017). Active learning and visual analytics for stance classification with ALVA. *ACM Transactions on Interactive Intelligent Systme*, *7*(3), http://dx.doi.org/10.1145/3132169, https://doi.org/10.1145/3132169.

Lipton, Z. C. (2016). The mythos of model interpretability. In *Proceedings of the 2016 ICML workshop on human interpretability in machine learning*. New York, NY, USA: arXiv:1606.03490.

Liu, P., Qiu, X., Chen, X., Wu, S., & Huang, X. (2015). Multi-timescale long short-term memory neural network for modelling sentences and documents. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 2326–2335). Lisbon, Portugal: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D15-1280, URL: https://www.aclweb.org/anthology/D15-1280.

Liu, S., Wang, X., Liu, M., & Zhu, J. (2017). Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics*, *1*(1), 48–56.

Llansó, E. J. (2020). No amount of "AI" in content moderation will solve filtering's prior-restraint problem. *Big Data and Society*, *7*(1), http://dx.doi.org/10.1177/2053951720920686, URL: http://journals.sagepub.com/doi/10.1177/2053951720920686.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st conference on neural information processing systems*.

Manning, C. D., Bauer, J., Finkel, J., Bethard, S. J., Surdeanu, M., & McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: System demonstrations*. Association for Computational Linguistics, URL: http://aclweb.org/anthology/P14-5010.

Mena, P. (2020). Cleaning up social media: The effect of warning labels on likelihood of sharing false news on facebook. *Policy & Internet*, *12*(2), 165–183. http://dx.doi.org/10.1002/poi3.214, URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/poi3.214.

Michel, P., Levy, O., & Neubig, G. (2019). Are sixteen heads really better than one? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems (Vol. 32)* (pp. 14014–14024). Curran Associates, Inc., URL: http://papers.nips.cc/paper/9551-are-sixteen-heads-really-better-than-one.pdf.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). Advances in pre-training distributed word representations. In *Proceedings of the eleventh international conference on language resources and evaluation*. Miyazaki, Japan: European Language Resources Association (ELRA), arXiv:1712.09405.

Mitchell, A., Kiley, J., Gottfried, J., & Matsa, K. E. (2014). *Political polarization & media habits*: *Technical report*, Pew Research Center, URL: https://www.journalism.org/2014/10/21/political-polarization-media-habits/.

Morgan, W. (2006). Statistical hypothesis tests for NLP. URL: https://cs.stanford.edu/people/wmorgan/sigtest.pdf.

Narwal, V., Salih, M. H., Lopez, J. A., Ortega, A., O'Donovan, J., Höllerer, T., & Savage, S. (2017). Automated assistants to identify and prompt action on visual news bias. In *Conference on human factors in computing systems - proceedings (Vol. Part F1276)* (pp. 2796–2801). New York, New York, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3027063.3053227, arXiv:1702.06492. URL: http://dl.acm.org/citation.cfm?doid=3027063.3053227.

Ninkov, A., & Sedig, K. (2020). The online vaccine debate: Study of a visual analytics system. *Informatics*, *7*(1), 3. http://dx.doi.org/10.3390/informatics7010003, URL: https://www.mdpi.com/2227-9709/7/1/3.

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330. http://dx.doi.org/10.1007/s11109-010-9112-2.

Nyhan, B., & Reifler, J. (2015). Displacing misinformation about events: An experimental test of causal corrections. *Journal of Experimental Political Science*, *2*(1), 81–93. http://dx.doi.org/10.1017/XPS.2014.22, URL: https://www.cambridge.org/core/product/identifier/S2052263014000220/type/journal_article.

Nyhan, B., & Reifler, J. (2019). The roles of information deficits and identity threat in the prevalence of misperceptions. *Journal of Elections, Public Opinion and Parties*, *29*(2), 222–244. http://dx.doi.org/10.1080/17457289.2018.1465061, URL: https://www.tandfonline.com/doi/abs/10.1080/17457289.2018.1465061.

Ole Krause-Sparmann (2018). NearPY. URL: https://github.com/pixelogik/NearPy.

Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, 1–14. http://dx.doi.org/10.1287/mnsc.2019.3478, URL: http://pubsonline.informs.orghttp//www.informs.orgMANAGEMENTSCIENCEhttp://pubsonline.informs.org/journal/mnsc.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic detection of fake news. *Proceedings of the 27th international conference on computational linguistics*, 3391–3401, URL: https://aclanthology.info/papers/C18-1287/c18-1287.

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long papers)* (pp. 231–240). Association for Computational Linguistics, URL: https://www.aclweb.org/anthology/P18-1022.

Przybyła, P. (2020). Capturing the style of fake news. In *Proceedings of the thirty-fourth AAAI conference on artificial intelligence*: *Vol. 34*, (pp. 490–497). New York, USA: AAAI Press, http://dx.doi.org/10.1609/aaai.v34i01.5386, URL: https://aaai.org/ojs/index.php/AAAI/article/view/5386.

Przybyła, P., & Teisseyre, P. (2014). Analysing utterances in polish parliament to predict speaker's background. *Journal of Quantitative Linguistics*, *21*(4), 350–376. http://dx.doi.org/10.1080/09296174.2014.944330, URL: http://www.tandfonline.com/doi/abs/10.1080/09296174.2014.944330.

R Core Team (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, URL: http://www.r-project.org/.

Rashkin, H., Choi, E., Jang, J. Y., 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*. Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/d17-1317.

Rauber, P., Fadel, S. G., Falcão, A. X., & Telea, A. C. (2017). Visualizing the hidden activity of artificial neural networks. *IEEE Transactions on Visualization and Computer Graphics*, *23*(1), 101–110. http://dx.doi.org/10.1109/TVCG.2016.2598838.

Raunak, V., Gupta, V., & Metze, F. (2019). Effective dimensionality reduction for word embeddings. In *Proceedings of the 4th workshop on representation learning for NLP* (pp. 235–243). Florence, Italy: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/W19-4328, URL: https://www.aclweb.org/anthology/W19-4328.

Reis, J. C., Correia, A., Murai, F., Veloso, A., & Benevenuto, F. (2019). Explainable machine learning for fake news detection. In *WebSci 2019 - Proceedings of the 11th ACM conference on web science* (pp. 17–26). New York, New York, USA: Association for Computing Machinery, Inc, http://dx.doi.org/10.1145/3292522.3326027, URL: http://dl.acm.org/citation.cfm?doid=3292522.3326027.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should i trust you? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining - KDD '16*, http://dx.doi.org/10.1145/2939672.2939778, arXiv:1602.04938.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. In *Proceedings of the 5th workshop on energy efficient machine learning and cognitive computing - NeurIPS 2019*. Vancouver, Canada: URL: https://arxiv.org/abs/1910.01108.

Scherr, M. (2008). Multiple and coordinated views in information visualization. *Trends in Information Visualization*, *38*, 1–33.

Self, J. Z., Zeitz, R., North, C., & Breitler, A. L. (2013). Auto-highlighter: Identifying salient sentences in text. In *2013 IEEE international conference on intelligence and security informatics* (pp. 260–262). IEEE.

Sevastjanova, R., El-Assady, M., Hautli-Janisz, A., Kalouli, A.-L., Kehlbeck, R., Deussen, O., Keim, D. A., & Butt, M. (2018). Mixed-initiative active learning for generating linguistic insights in question classification. In *3rd workshop on data systems for interactive analysis (DSIA) at IEEE VIS*.

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(3), 1–42. http://dx.doi.org/10.1145/3305260, URL: http://dl.acm.org/citation.cfm?doid=3325195.3305260.

Shneiderman, B. (2003). The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization* (pp. 364–371). Elsevier.

Shu, K., Mahudeswaran, D., & Liu, H. (2019). Fakenewstracker: a tool for fake news collection, detection, and visualization. *Computational and Mathematical Organization Theory*, *25*(1), 60–71.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media. *ACM SIGKDD Explorations Newsletter*, http://dx.doi.org/10.1145/3137597.3137600, arXiv:1708.01967.

Stoffel, F., Flekova, L., Oelke, D., Gurevych, I., & Keim, D. A. (2015). Feature-based visual exploration of text classification. In *Symposium on visualization in data science (VDS) at IEEE VIS 2015*.

Stone, P. J., Bales, R. F., Namenwirth, J. Z., & Ogilvie, D. M. (1962). The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, *7*(4), 484–498. http://dx.doi.org/10.1002/bs.3830070412.

Strobelt, H., Oelke, D., Kwon, B. C., Schreck, T., & Pfister, H. (2015). Guidelines for effective usage of text highlighting techniques. *IEEE Transactions on Visualization and Computer Graphics*, *22*(1), 489–498.

Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., & de Alfaro, L. (2017). Some like it hoax: Automated fake news detection in social networks. In *CEUR workshop proceedings*: *Vol. 1960*, http://dx.doi.org/10.1257/jep.31.2.211, arXiv:1704.07506.

Tai, K. S., Socher, R., & Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing*, (pp. 1556–1566). http://dx.doi.org/10.1515/popets-2015-0023.

Tandoc, E. C., Lim, Z. W., & Ling, R. (2017). Defining "Fake News". *Digital Journalism*, *6*(2), 137–153. http://dx.doi.org/10.1080/21670811.2017.1360143.

Tausczik, Y. R., & Pennebaker, J. W. (2009). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. http://dx.doi.org/10.1177/0261927X09351676.

Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., & Mittal, A. (2018). The fact extraction and verification (FEVER) shared task. In *Proceedings of the first workshop on Fact Extraction and VERification (FEVER)*. arXiv:1811.10971v1.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 267–288.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605, URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

Xu, J., Chen, D., Qiu, X., & Huang, X. (2016). Cached long short-term memory neural networks for document-level sentiment classification. In *Proceedings of the 2016 Conference on empirical methods in natural language processing* (pp. 1660–1669). Austin, Texas: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D16-1172, URL: https://www.aclweb.org/anthology/D16-1172.

Yang, F., Du, M., Ragan, E. D., Pentyala, S. K., Yuan, H., Ji, S., Mohseni, S., Linder, R., & Hu, X. (2019). Xfake: Explainable fake news detector with visualizations. In *The web conference 2019 - Proceedings of the world wide web conference* (pp. 3600–3604). New York, New York, USA: Association for Computing Machinery, Inc, http://dx.doi.org/10.1145/3308558.3314119, arXiv:1907.07757. URL: http://dl.acm.org/citation.cfm?doid=3308558.3314119.

Yang, F., Huang, Z., Scholtz, J., & Arendt, D. L. (2020). How do visual explanations foster end users' appropriate trust in machine learning? In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (pp. 189–201). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3377325.3377480, URL: https://doi.org/10.1145/3377325.3377480.

Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing and Management*, *57*(2), Article 102025. http://dx.doi.org/10.1016/j.ipm.2019.03.004.

Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G., & Tolmie, P. (2016). Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS One*, *11*(3), http://dx.doi.org/10.1371/journal.pone.0150989, arXiv:1511.07487.