Capturing the Style of Fake News

Piotr Przybyła
Institute of Computer Science, Polish Academy of Sciences
Warsaw, Poland
piotr.przybyla@ipipan.waw.pl

Abstract

In this study we aim to explore automatic methods that can detect online documents of low credibility, especially fake news, based on the style they are written in. We show that general-purpose text classifiers, despite seemingly good performance when evaluated simplistically, in fact overfit to sources of documents in training data. In order to achieve a truly style-based prediction, we gather a corpus of 103,219 documents from 223 online sources labelled by media experts, devise realistic evaluation scenarios and design two new classifiers: a neural network and a model based on stylometric features. The evaluation shows that the proposed classifiers maintain high accuracy in case of documents on previously unseen topics (e.g. new events) and from previously unseen sources (e.g. emerging news websites). An analysis of the stylometric model indicates it indeed focuses on sensational and affective vocabulary, known to be typical for fake news.

Introduction

The problem of fake news and the wider issue of credibility in online media continue to attract considerable attention not only of their consumers and creators, but also of policy makers and the digital industry. One of the responses of social media sites is to signal untrustworthiness of certain content, e.g. as *disputed* in *Facebook* (Clayton et al. 2019). Unfortunately, the manual fact-checking involved is too laborious to be applied to every post published on these platforms and elsewhere. That is why we consider text analysis techniques that could automatically assess the credibility of documents published online, which could also be useful for other stakeholders intending to reduce the impact of misinformation, including journalists (Chen, Conroy, and Rubin 2015) and web users (Berghel 2017). The most straightforward approach to the problem is automatic verification of each claim included in a document. This task, however, has many challenges, namely insufficient level of text understanding methods and limited coverage and currency of knowledge bases.

It may seem easier to train a machine learning (ML) model on a collection of online documents, accompanied by expert-assigned labels indicating their credibility, using one of general-purpose text classification algorithms. As we show in the next section, this has indeed been done, sometimes leading to impressive classification accuracy. The disadvantage of such solutions is that we have no direct control over which features of the document (e.g. word occurrences in bag of words representation) the credibility assessment is based on. Since some features that provide good performance on the training or test data might not be desirable in real-life application scenario, it does not suffice to know that a classifier makes the right decision in most cases – we would like it to do so for the right reasons. This includes knowing what features are important for a particular decision (interpretability) and making sure they are not specific to the training data (generalisability).

For example, an ML model might learn to recognise the source a given document comes from (using its name appearing in text) and assign credibility label based on other documents from the same source, i.e. other articles from the the website, seen in the training data. While taking into account the reputation of a source is a heuristic heavily used by humans when assessing information online (Metzger, Flanagin, and Medders 2010) and commonly advised for fake news spotting (Hunt 2016), it may be misleading in an ML context. The fake news websites tend to be short-lived (Allcott and Gentzkow 2017), and such a model would be helpless when new sources replace them. The document topic could be another easily accessible, yet misleading feature. While fake news outlets indeed concentrate around a few current themes that are guaranteed to engage the target audience (Bakir and McStay 2017), these topics will be replaced over time, making a classifier obsolete.

In this study we focus on the *style* of writing, i.e. the form of text rather than its meaning (Ray 2015). Since fake news sources usually attempt to attract attention for a short-term financial or political goal (Allcott and Gentzkow 2017) rather than to build a long-term relationship with the reader, they favour informal, sensational, affective language (Bakir and McStay 2017). This indicator of low credibility could be used to build a reliable classifier.

Several directions could be pursued to avoid the model being biased by the sources or topics available in the training data. In this study we provide the following contribution:

- We present a textual corpus with 103,219 documents, covering a wide range of topics, from 223 sources labelled for credibility based on studies performed at *PolitiFact* and *Pew Research Center*, which is a useful resource in building unbiased classifiers.
- We use the corpus to construct evaluation scenarios measuring performance of credibility estimation methods more realistically by applying them to documents from sources and topics that were not available at training time.
- We propose two classifiers: a neural network and a model based on features used in stylometric analysis and demonstrate that the latter indeed captures the affective language elements.

In order to encourage and facilitate further research, we make the corpus, the evaluation scenarios and the code (for the stylometric and neural classifiers) available online[1].

Related work

The problem of fake news has been attracting major attention since the 2016 presidential elections in the US (Allcott and Gentzkow 2017). It has been a subject of research in journalism and political sciences, but much more is needed, especially to assess the widely discussed connections with social media and political polarisation (Tucker et al. 2018).

Annotated corpora

In the challenge of automatic detection of fake content, textual data annotated with respect to credibility, veracity or related qualities play a crucial role (Torabi Asr and Taboada 2019). One of the most commonly used resources of this kind is *OpenSources*[2], a publicly-available list of around 1000 web sources with human-assigned credibility labels. This list was used by a web browser plugin *B.S. detector*, whose decision in turn were used to generate a corpus of 12,999 posts from 244 websites it labelled as fake. The corpus was made available as a *Kaggle* dataset[3]. Another collection[4] was collected by automatically scraping the domains from *OpenSources*. Pathak and Srihari (2019) manually selected around 700 documents from the dataset and labelled them based on the type of misinformation they use.

Journalists at *BuzzFeed News* contributed by assessing the veracity of 2,282 posts published before the 2016 elections by 9 Facebook pages (Silverman 2016). The data was later made available as a corpus[5]. A different approach to creating a corpus was taken by Shu et al. (2018), who explored the claims fact-checked by *PolitiFact* and *GossipCop* and automatically retrieved relevant webpages, obtaining 23,921 articles, with the vast majority covering celebrity gossip.

Credibility assessment

The first studies on recognition of fabricated news focused on machine-generated (Badaskar, Agarwal, and Arora 2008) or satirical (Burfoot and Baldwin 2009) articles. Approaches to what we currently call fake news were hampered by low amount of data available; e.g. Rubin, Conroy, and Chen (2015) worked on just 144 news items and the recognition performance was not significantly better than chance. Horne and Adali (2017) used datasets with 35 fake news items from a *BuzzFeed News* article (Silverman 2016) and another 75 items gathered by themselves manually and made interesting observations on the stylistic cues affecting credibility. However, the prediction performance may not be reliable due to small data size.

Pérez-Rosas et al. (2018) attempted to overcome the lack of data by artificially generating a fake news corpus through crowdsourcing, achieving a classification accuracy of 0.74 on a balanced test set. However, their classifier is trained on less than 1000 documents and uses word n-grams, making it prone to overfitting to sources or topics, which is confirmed by weaker results in a cross-topic evaluation scenario (around 0.50-0.60). Another way to collect a sufficient number of manually credibility-labelled documents with limited resources is active learning (Bhattacharjee, Talukder, and Balantrapu 2017). Rashkin et al. (2017) used a dataset including satire and *hoax* news stories and the evaluation, performed on previously unseen sources, showed an F-score of 0.65. The classifier is however unlikely to be topic-independent due to relying on word tri-grams, which is demonstrated by presence of keywords related to current topics among strong features (e.g. *syria*).

Ahmed, Traore, and Saad (2017) reported high accuracy (92% on a balanced test set) using only TF-IDF of word n-grams on the *Kaggle* dataset. Given how this type of features is prone to overfitting to particular news sources (e.g. through their names appearing in text) and that the evaluation was performed through ordinary cross-validation, it seems unlikely this accuracy would be upheld on new sources.

A recent study by Potthast et al. (2018), using the *BuzzFeed News* corpus, is similar to our work, as they ensured their classifier is style-based by building it on top of stylometric features. However, they argued that accurate identification of fake news was not possible, and instead focus on detecting *hyperpartisan* media, which consistently follow the right- or left-wing narrative, as opposed to mainstream outlets. Our study has a wider scope, since many sources included in our corpus lack such partisan consistency or even do not focus on politics at all.

To sum up, although there has been several attempts to classify credible and non-credible news publications, they were all limited by the amount of available data, resulting in either low performance or likely overfitting to topics or sources that were used during training. We aim to overcome this limitation by gathering a corpus much larger than any of the previously used for training such models, which allows us to achieve high classification performance while keeping the credibility assessment style-driven.

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

[2] https://github.com/BigMcLargeHuge/opensources

[3] https://www.kaggle.com/mrisdal/fake-news

[4] https://github.com/several27/FakeNewsCorpus

[5] https://zenodo.org/record/1239675

Corpus

Since the definition of fake news is still being discussed (Gelfert 2018), within this work we use the notion of credibility. We define a document as non-credible if the source it comes from has been assessed as such by experts. To gather non-credible documents, we use a list of websites (Gillin 2017) prepared by journalists at *PolitiFact*, a non-profit fact-checking centre. On the other hand, the most trusted news outlets from a study (Mitchell et al. 2014) by *Pew Research Center* (PRC), an independent public opinion research unit, are considered credible. Websites of both categories are crawled to obtain documents, which are then converted into plain text and treated as learning cases with the label denoting them as credible (0) or non-credible (1), according to the source they come from.

Collecting the documents

To obtain documents from non-credible sources we use the websites labelled in 2017 by *PolitiFact* as *fake news* (192 sources) and *impostor* (49) (Gillin 2017). Unfortunately, less than a quarter of them are still active in 2019, but most websites remain available in the *WayBackMachine* archives[6]. Since the list was last updated on 09.11.2017, the latest available snapshot of the main page of each website between 01.01.2017 and 09.11.2017 is selected for crawling. Six of the websites are excluded, since they do not contain any news pieces, but rather discussions, prank content and advise articles. The websites for which no documents except a front page are available in the archive are excluded, too.

As per the credible sources, we choose the 21 media outlets that were commonly trusted than distrusted, according to the survey report (Mitchell et al. 2014) by PRC. We exclude two news aggregators (*Google News*, *Yahoo News*), who do not create their own content, but link to external sources; and *MSNBC*, which contains only video materials.

In total, this procedure retains 205 non-credible and 18 credible websites. The sites are crawled by following HTML links, starting from the main page and limiting the path length to 5 and maximum number of visited links to 10,000. We ignore pages not archived in 2017, duplicates and subpages with text of average line length less than 15 words. This process results in a corpus of 52,790 pages from non-credible and 50,429 from credible sources. Finally, after conversion from HTML to plain text using manually designed heuristic rules, we obtain a textual corpus of 103,219 documents and 117M tokens.

Corpus exploration

The numbers of documents coming from respective sources in the corpus differ greatly. This is especially true for the non-credible ones, which span from a few large websites (13 have more than 1000 documents) to plenty of very small ones (77 have less than 50 documents). The fact that fake news is published by numerous small sources, three quarters of which are unavailable after two years, illustrates the importance of the credibility assessment not relying on a particular source. The credible outlets are larger and closer in size: from 1227 to 6019 documents. While the corpus is gathered from US media, the distribution of non-credible content over numerous outlets with much smaller size than the credible sources was also confirmed in an analysis of online media reach in Europe (Fletcher et al. 2018).

In order to model topical differences between sources, we compute a model of 100 topics using LDA (Latent Dirichlet Allocation) (Blei, Ng, and Jordan 2003) implemented in *Mallet* (McCallum 2002). Next, each document is assigned to the topic it has the strongest association with. Figure 1 shows how many of the documents from credible and non-credible sources are assigned to the largest 15 topics, described by associated keywords. We can see that some themes are far more popular in the non-credible part: the comparisons between the current president and his predecessor and election rival (topics #19 and #70), media coverage (#85), Muslims and immigration (#23 and #11) and health/nutrition (#76). The areas that are more commonly covered by credible sources include cinema (#50) and sports (#5). Some issues popular in both classes are the Russia investigations (#62), crime (#55) and international conflicts in the middle east and Korea (#17 and #2).

This analysis further illustrates the need for a credibility classifier to avoid relying on topics in its decisions. Many of the differences in vocabulary between the source types come from interest in very specific and current themes, such as Hillary Clinton's e-mails, Donald Trump's presidency or illegal immigration through Mexico. While these features may seem discriminatory at a certain period of time, a classifier based on it may be unable to perform well in future, when the media attention turns elsewhere.

Stylometric classifier

In terms of general architecture of the stylometric classification, we use a collection of stylystic features followed by linear modeling. While similar approaches were applied to credibility assessment (Burfoot and Baldwin 2009; Ahmed, Traore, and Saad 2017; Horne and Adali 2017; Rashkin et al. 2017; Pérez-Rosas et al. 2018), in this study we take special care to avoid using features that would allow a classifier to overfit to particular sources and topics. That is why instead of popular n-grams of words, we use n-grams of Part of Speech (POS) tags.

Another group of tools frequently employed in stylistic analysis are dictionaries, e.g. *Linguistic Inquiry and Word Count* (LIWC) (Tausczik and Pennebaker 2009), used in fake news detection (Horne and Adali 2017; Rashkin et al. 2017; Pérez-Rosas et al. 2018), or *General Inquirer* (GI) (Stone et al. 1962), used for hyperpartisan news recognition (Potthast et al. 2018). The weakness of these resources lies in the limited dictionary size, e.g. GI contains 8640 words in 182 categories[7]. We therefore increase its size by expanding each category with words similar according to *word2vec* (Mikolov et al. 2013) representation. Firstly, for each category of size n, we build a logistic regression model of belonging to this category using all words represented by vec-

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

[7] http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm

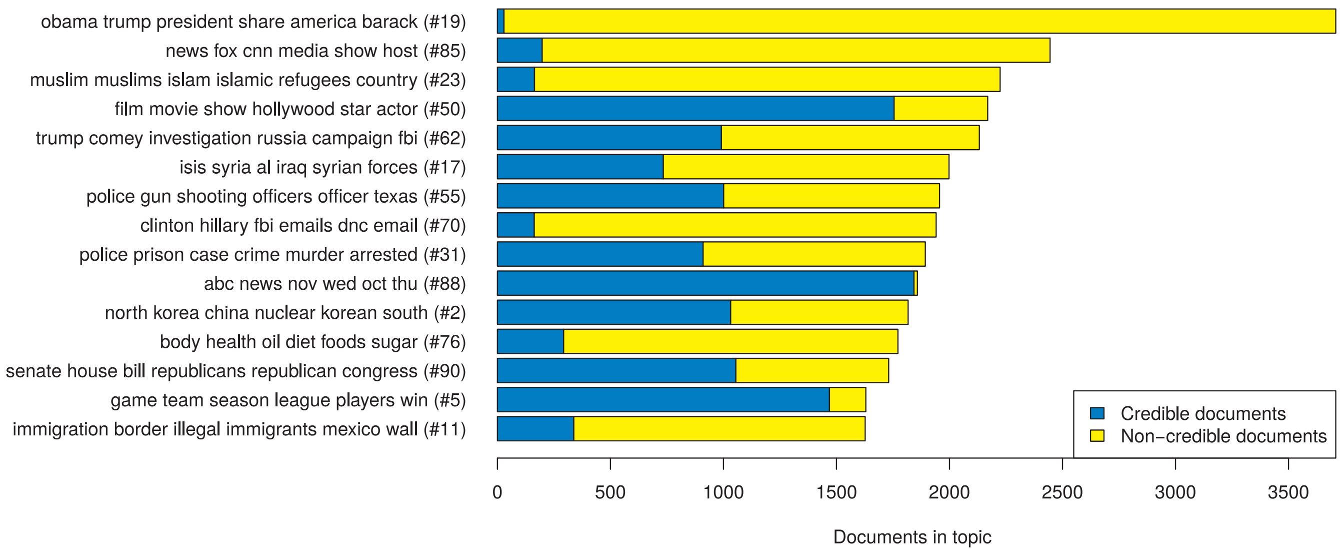


Figure 1: The largest 15 LDA topics in the corpus, each shown with the six most significant keywords, an identifier and bars illustrating number of credible and non-credible documents associated with it.

tors trained on *Google News* corpus[8]. Then, $4 \times n$ new words with the highest score are added to the category. Performing this procedure for all 182 categories yields a dictionary with a total size of 34,293 words.

Stylometric features

The documents are preprocessed by *Stanford CoreNLP* (Manning et al. 2014), including sentence segmentation, tokenisation and POS tagging. This annotation is used to generate the following document features:

- number of sentences, average sentence length (in words) and average word length (in characters),
- number of words matching different letter case schemes (all lower case, all upper case, just first letter upper case, other), represented as counts normalised by the document length,
- frequencies of POS unigrams, bigrams and trigrams, represented as counts normalised by the document length (if present in at least 5 documents),
- frequencies of words belonging to the 182 word categories in the expanded GI dictionary, represented as counts normalised by the document length.

Classifier

The dataset includes 103,219 instances described by 39,235 features. We apply a two-stage approach for selecting relevant features: first preliminary filtering, then building a regularised classifier.

At the filtering stage, we use the Pearson correlation with the output variable, which is a common technique for linear classifiers (Guyon and Elisseeff 2003). First, we check whether feature j is present in document i by computing a binary matrix with elements $b_{i,j} = \mathbf{1}[x_{i,j} \neq 0]$. Potthast et al. (2018) perform filtering by removing features that occur in less than 2.5% or 10% of documents. We argue this could lead to a loss of information, since a less frequent feature could still be significant, as long as a large majority of the documents it occurs in belongs to the same class. Therefore, we take into account class label y by computing the correlation coefficient and including each feature j such that $\mathbf{b}_j = [b_{1,j}, b_{2,j}, \ldots]$ satisfies $|\text{cor}(\mathbf{b}_j, \mathbf{y})| > 0.05$. The number of retained features depends on the training-test split, but we observe it to be always below 5%.

To assess the probability of a document belonging to the non-credible category, a logistic regression model is built. The vastness of the feature space implies a need for regularisation, so we apply the L_1 version (LASSO), as implemented in `glmnet` (Friedman, Hastie, and Tibshirani 2010) package in *R* (R Core Team 2013) with penalty parameter λ selected through cross-validation over training set.

The classifier output is used directly as a non-credibility score taking values between 0 (credible) and 1 (non-credible). When evaluation demands a discrete output (to compute accuracy), a threshold of 0.5 is applied.

BiLSTMAvg

The second of the applied solutions, called BiLSTMAvg, is a neural network with architecture based on elements used in natural language processing, i.e. word embeddings (Mikolov et al. 2013) and bidirectional LSTM (Hochreiter and Schmidhuber 1997). Since LSTM is most commonly employed to represent the meaning of short text fragments, esp. sentences, we have decided to add an additional layer that computes the credibility scores (probabilities of *credible* and *non-credible* classes) of an article by averaging the

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

scores of all its sentences. This should also encourage the classifier to seek credibility clues in every sentence rather than just focus on the easy ones (e.g. mentioning source name). Specifically, the following layers are included:

- An embedding layer, representing each token using a 300-dimensional word2vec vector trained on *Google News*,
- Two LSTM layers, forward and backward, representing each sentence by two 100-dimensional vectors (output of the last cell in a sequence),
- A densely-connected layer, reducing the dimensionality to 2 and applying softmax to compute class probability,
- An averaging layer, representing each document's class probability scores by averaging the scores for all its sentences.

The neural network is implemented and trained in *TensorFlow* for 10 epochs with sentence length limited to 120 tokens and document length limited to 50 sentences.

Baseline classifiers

To understand if general-purpose text classifiers are able to capture document style without overfitting to features indicating a source or topic and to put the performance of our stylometric and neural solutions in perspective, we also evaluate two baseline models: bag of words and BERT.

Bag of words

This simple model represent documents through frequencies of unigrams, bigrams and trigrams of lemmata (base forms) of words, occurring in at least 200 documents. The feature filtering and logistic regression model construction is performed as in stylometric classification.

BERT

To employ BERT, a commonly used pre-trained language model (Devlin et al. 2018), we take the uncased base version and fine-tune it in a supervised text classification task using the recommended architecture (linear prediction over the output corresponding to the `[CLS]` element). We use the first 512 tokens of each document and the process is executed in each CV fold independently.

Evaluation

The main evaluation procedure involves running the model construction and prediction in a 5-fold cross validation (CV) scenario and comparing its output to true labels. Due to sufficiently balanced classes, we use accuracy instead of precision or recall. Three scenarios are considered:

- plain document-based CV, where folds include completely random documents from across the dataset.
- topic-based CV, where each of the LDA topics, generated as described previously, is assigned to one of the CV folds with all associated documents. This scenario simulates a situation when a test document belongs to previously unseen topic, e.g. corresponding to a new event.
- source-based CV, where each of the document sources is assigned to one of the CV folds with all its documents. This allows us to measure the performance expected for articles from previously unseen websites.

Method	**doc. CV**	**topic CV**	**source CV**
Stylometric	0.9274	0.9173	0.8097
BiLSTMAvg	0.8994	0.8921	0.8250
Bag of words	0.9913	0.9886	0.7078
BERT	0.9976	0.9965	0.7960

Table 1: Classification accuracy of our stylometric and neural classifiers compared to baselines in three evaluation scenarios, simulating, respectively, a new document from known sources and topics, a document from unknown topic and a document from unseen source.

Using CV, while increasing the computation time, helps to strengthen the evaluation by including documents from all sources in the test sets. To facilitate comparisons of classification performance with other approaches, the corpus download site includes the assignment of documents to CV folds.

Using table 1, showing classification accuracy, we can clearly distinguish two groups of methods. Firstly, the popular general-purpose classifiers (bag of words and BERT) perform extremely well in document CV, but lose 20-30% when applied in source CV, which indicates they overfit to sources seen in training. The stylometric method, although noticeably weaker in document CV, proves more resistant to new sources, almost reaching 81%, which could be considered a positive result compared to the most similar previous work (Potthast et al. 2018). Interestingly, even better results are provided by the BiLSTMAvg network, which despite the worst performance in document CV beats everything else in source CV. Topic CV is less of a challenge, as all of the tested methods loose at most 1% in this scenario.

For better understanding of stylistic differences between sources, figure 2 shows a boxplot of the non-credibility scores computed by the stylometric classifier in source-based CV, grouped by sources sorted by mean score, with colour corresponding to the true category. We can see the credible (blue) sources mostly below the 0.5 threshold and the non-credible ones (yellow) above it. Nevertheless, the wide range of scores in the large sources means some of documents are misclassified.

We can also notice cases where the mean score places a source in a wrong category. The most striking non-credible examples of that (two leftmost yellow bars) are *The Times Mexico* (times.com.mx) and *Before It's News* (beforeitsnews.com). The first one (currently unavailable) was labelled by *PolitiFact* as 'Imposter site', as it pretends to be a credible medium – a branch of *The Times*. It has relatively few pages (34 in the corpus), but it mixes made-up stories, e.g. *Leaked Audio: Mexican President Agrees to Pay For Wall*, with articles copied from reputable sources, e.g. *Snapchat's Physical Footprint Reveals Core Priority of the Brand*, originally from *Yahoo Finance*. Such instances will be challenging for any content- or style-based classifier. The second problematic source is a large (3,303 documents

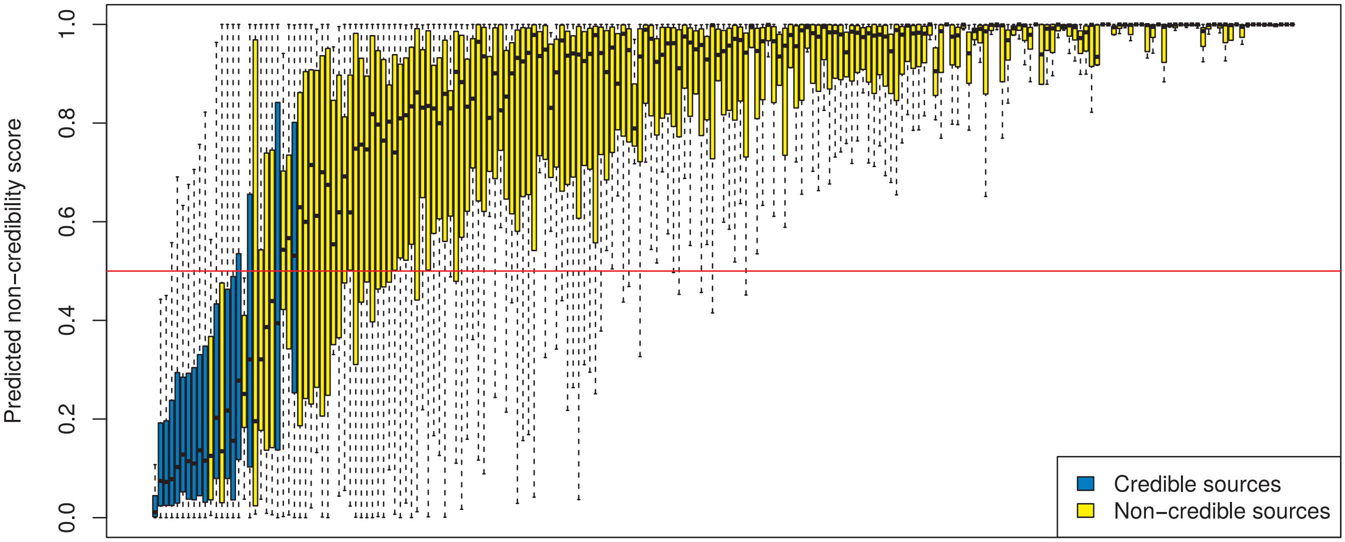


Figure 2: Predicted non-credibility scores assigned to documents by the stylometric classifier, grouped by sources (credible and non-credible), sorted by average score of all documents in each source.

in the corpus) citizen journalism portal, allowing anonymous users to post their content of various kind, frequently resembling discussion forum rather than a news outlet. This could render the classification difficult, but the portal includes obviously fake stories, too, e.g. *Worker Says Nanobot Mosquito Killed Woman!*.

The most challenging cases of credible sources (two rightmost blue bars), whose style resembles the non-credible ones, are *Fox News*, and *TheBlaze*, which have been regarded as the most distrusted and the least known of the credible sources included (Mitchell et al. 2014).

Finally, thanks to the choice of the features and the tendency of regularised regression to include a limited number of them in a model, we are able to visualise some of the motives behind a non-credibility score provided by the stylometric model. This can help us to make sure the classifier indeed relies on stylistic elements of a document. While a detailed analysis of the syntactical patterns remains beyond the scope of this work, the GI dictionary features have clear correspondence in text, which was used to highlight significant words in a fragment of an article[9] from the non-credible portion (figure 3). Specifically, we have highlighted in yellow the words for which GI features contribution was above 20. We can see that the words indicating lower credibility are indeed quite affective, e.g. *idiots*, *outrage*, *digrace*. Note how the highlighted words are not specific to the topic of the article (except one: *collusion*), which can explain good performance in topic-based evaluation. To compare with the bag of words model, the words covered by n-grams with coefficients above 20 are underlined. This baseline approach appears to be less style-driven: while there is some overlap (i.e. *idiots*), other words clearly indicate the focus on the document topic (i.e. *Barack Obama*, *Trump*, *Democrats*).

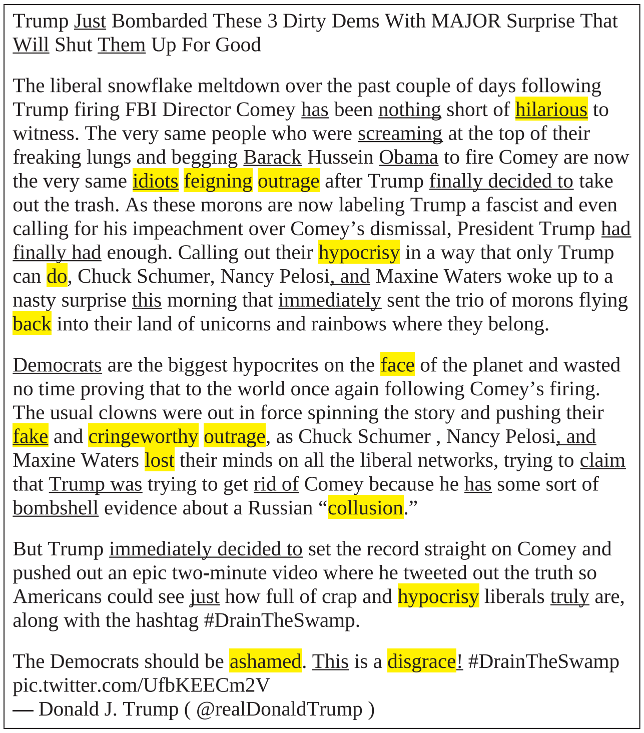
Discussion

The obtained results clearly emphasise the importance of designing realistic evaluation scenarios when measuring the performance of credibility assessment solutions. Naive evaluation (through random held-out subset) might suggest that a simple bag of words model can achieve near-perfect accuracy, while in confrontation with documents from unseen source it performs poorly. To the best of our knowledge, this is the first time this aspect of credibility assessment is observed and measured.

The proposed stylometric classifier, while showing more consistent performance over evaluation scenarios, still loses over 10% of the accuracy on unseen sources. The most straighforward explanation might be that instead of the general style of fake news, the model captures styles of individual sources. Another possible reason for that is imperfection of text extraction procedure, which sometimes produces a plain text version that includes not only the actual news description, but also some standard website elements (e.g. encouraging social media sharing or commenting), that are converted to POS n-grams and could be picked up by the classifier. Unfortunately, these fragments differentiate only specific websites, not the general categories we seek to recognise.

The interpretability of the stylometric classifier allows us to verify that it indeed takes into account the affective words. Such interpretability is helpful in making sure a model generalises well (Lipton 2016), but also plays a role in obtaining users' trust (Pieters 2011). Being able to explain why a model has made certain decision is crucial in the application scenarios relevant for this work. Take political misperceptions as a prominent example (Flynn, Nyhan, and Reifler 2017) – providing an alternative explanation for observed events has been shown to be significantly more effective than a simple contradiction (Nyhan and Reifler 2015).

[9] http://freedomdaily.com/trump-just-bombarded-3-dirty-dems-major-surprise-will-shut-good/



Trump Just Bombarded These 3 Dirty Dems With MAJOR Surprise That Will Shut Them Up For Good

The liberal snowflake meltdown over the past couple of days following Trump firing FBI Director Comey has been nothing short of hilarious to witness. The very same people who were screaming at the top of their freaking lungs and begging Barack Hussein Obama to fire Comey are now the very same idiots feigning outrage after Trump finally decided to take out the trash. As these morons are now labeling Trump a fascist and even calling for his impeachment over Comey's dismissal, President Trump had finally had enough. Calling out their hypocrisy in a way that only Trump can do, Chuck Schumer, Nancy Pelosi, and Maxine Waters woke up to a nasty surprise this morning that immediately sent the trio of morons flying back into their land of unicorns and rainbows where they belong.

Democrats are the biggest hypocrites on the face of the planet and wasted no time proving that to the world once again following Comey's firing. The usual clowns were out in force spinning the story and pushing their fake and cringeworthy outrage, as Chuck Schumer , Nancy Pelosi, and Maxine Waters lost their minds on all the liberal networks, trying to claim that Trump was trying to get rid of Comey because he has some sort of bombshell evidence about a Russian "collusion."

But Trump immediately decided to set the record straight on Comey and pushed out an epic two-minute video where he tweeted out the truth so Americans could see just how full of crap and hypocrisy liberals truly are, along with the hashtag #DrainTheSwamp.

The Democrats should be ashamed. This is a disgrace! #DrainTheSwamp
pic.twitter.com/UfbKEECm2V
— Donald J. Trump (@realDonaldTrump)

Figure 3: Sample document fragment with words indicating low credibility according to the GI dictionary features of the stylometric classifier (highlighted in yellow) and the baseline bag of word model (underlined).

Unfortunately, the sentence-level credibility scores provided by BiLSTMAvg are too coarse-grained to be informative, so additional work is necessary to obtain word-level importance measure. The mechanism of attention (Vaswani et al. 2017) is commonly used in this role, but the validity of its use as explanation is debatable (Jain and Wallace 2019). Despite the interpretability issues associated with neural networks in general, the obtained results show they are worth considering in this scenario. BiLSTMAvg, despite lacking any explicit style-focused features, avoids overfitting and delivers the best performance on new sources.

The most important limitations of the study come from the basic assumptions we make about credibility. Firstly, that its assessment at the level of a source is inherited by all documents within it. We can expect that not every document from a non-credible source contains false information, just as not every news item from the trusted outlets is perfectly accurate. Whether this understanding of credibility reflects the concept of *fake news* will depend on which of its definitions we apply– while some rely on the veracity of provided information, others emphasise being *misleading by design* (Gelfert 2018). We aim to address this problem in future by extending the corpus with document-level credibility assessment.

Secondly, the dependency between writing style of a document and its credibility observed in our dataset might not be universal or permanent. While the current misinformation landscape is dominated by obvious profit-driven websites, there may exist some (possibly more in future) that are on par with real news outlets in terms of quality and style, yet provide misleading content.

Conclusions

To sum up, the credibility of news articles in our corpus can indeed be estimated based on the style they are written in. However, given how subtle the manifestation of style might be compared to more prominent traits, such as source or topic, special care is needed when collecting a learning sample and designing classification and evaluation procedures to make sure its theoretical performance translates to a benefit in a social context. The high classification accuracy obtained in the experiments indicates that despite previous claims that automatic fake news detection based on style does not work in general (Potthast et al. 2018) or may never be possible (Tucker et al. 2018), it is a worthwhile direction of research. We hope that future work in this field will be facilitated by the contributed corpus and evaluation scenarios.

Acknowledgments

This work was supported by the *Polish National Agency for Academic Exchange* through a *Polish Returns* grant number PPN/PPO/2018/1/00006 and by the *Google Cloud Platform* through research credits.

References

Ahmed, H.; Traore, I.; and 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*, 127–138. Springer International Publishing.

Allcott, H., and Gentzkow, M. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* 31(2):211–236.

Badaskar, S.; Agarwal, S.; and Arora, S. 2008. Identifying Real or Fake Articles: Towards better Language Modeling. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*. Asian Federation of Natural Language Processing.

Bakir, V., and McStay, A. 2017. Fake News and The Economy of Emotions: Problems, causes, solutions. *Digital Journalism* 6(2):154–175.

Berghel, H. 2017. Lies, Damn Lies, and Fake News. *Computer* 50(2):80–85.

Bhattacharjee, S. D.; Talukder, A.; and Balantrapu, B. V. 2017. Active learning based news veracity detection with feature weighting and deep-shallow fusion. In *Proceedings of the IEEE International Conference on Big Data*. IEEE.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3:993–1022.

Burfoot, C., and Baldwin, T. 2009. Automatic satire detection: are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 161–164. Association for Computational Linguistics.

Chen, Y.; Conroy, N. J.; and Rubin, V. L. 2015. News in an online world: The need for an "automatic crap detector". *Proceedings of the Association for Information Science and Technology* 52(1).

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.; and 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*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186. Association for Computational Linguistics.

Fletcher, R.; Cornia, A.; Graves, L.; and Nielsen, R. K. 2018. Measuring the reach of ”fake news” and online disinformation in Europe. Technical report, Reuters Institute for the Study of Journalism.

Flynn, D. J.; Nyhan, B.; and Reifler, J. 2017. The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs About Politics. *Political Psychology* 38:127–150.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 33(1).

Gelfert, A. 2018. Fake News: A Definition. *Informal Logic* 38(1):84–117.

Gillin, J. 2017. PolitiFact's guide to fake news websites and what they peddle. *PolitiFact*.

Guyon, I., and Elisseeff, A. 2003. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research* 3:1157–1182.

Hochreiter, S., and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation* 9(8):1735–1780.

Horne, B. D., and 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.

Hunt, E. 2016. What is fake news? How to spot it and what you can do to stop it. *The Guardian*.

Jain, S., and Wallace, B. C. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3543–3556. Minneapolis, Minnesota: Association for Computational Linguistics.

Lipton, Z. C. 2016. The Mythos of Model Interpretability. In *Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*.

Manning, C. D.; Bauer, J.; Finkel, J.; Bethard, S. J.; Surdeanu, M.; and 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.

McCallum, A. K. 2002. MALLET: A machine learning for language toolkit.

Metzger, M. J.; Flanagin, A. J.; and Medders, R. B. 2010. Social and Heuristic Approaches to Credibility Evaluation Online. *Journal of Communication* 60(3):413–439.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781*.

Mitchell, A.; Kiley, J.; Gottfried, J.; and Matsa, K. E. 2014. Political Polarization & Media Habits. Technical report, Pew Research Center.

Nyhan, B., and Reifler, J. 2015. Displacing Misinformation about Events: An Experimental Test of Causal Corrections. *Journal of Experimental Political Science* 2(1):81–93.

Pathak, A., and Srihari, R. 2019. BREAKING! Presenting Fake News Corpus for Automated Fact Checking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 357–362. Florence, Italy: Association for Computational Linguistics.

Pérez-Rosas, V.; Kleinberg, B.; Lefevre, A.; and Mihalcea, R. 2018. Automatic Detection of Fake News. *Proceedings of the 27th International Conference on Computational Linguistics* 3391–3401.

Pieters, W. 2011. Explanation and trust: what to tell the user in security and AI? *Ethics and Information Technology* 13(1):53–64.

Potthast, M.; Kiesel, J.; Reinartz, K.; Bevendorff, J.; and 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)*, 231–240. Association for Computational Linguistics.

R Core Team. 2013. R: A Language and Environment for Statistical Computing.

Rashkin, H.; Choi, E.; Jang, J. Y.; Volkova, S.; and 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.

Ray, B. 2015. *Style: An Introduction to History, Theory, Research, and Pedagogy*. Parlor Press, The WAC Clearinghouse.

Rubin, V. L.; Conroy, N. J.; and Chen, Y. 2015. Towards news verification: Deception detection methods for news discourse. *Proceedings of the Hawaii International Conference on System Sciences*.

Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; and Liu, H. 2018. FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media. *arXiv:1809.01286*.

Silverman, C. 2016. This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook. *BuzzFeed News*.

Stone, P. J.; Bales, R. F.; Namenwirth, J. Z.; and 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.

Tausczik, Y. R., and 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.

Torabi Asr, F., and Taboada, M. 2019. Big Data and quality data for fake news and misinformation detection. *Big Data & Society* 6(1).

Tucker, J. A.; Guess, A.; Barberá, P.; Vaccari, C.; Siegel, A.; Sanovich, S.; Stukal, D.; and Nyhan, B. 2018. Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. Technical report, Hewlett Foundation.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc. 5998–6008.