Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 176 (2020) 78–87

Procedia Computer Science

www.elsevier.com/locate/procedia

24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

An interpretable model to measure fakeness and emotion in news

Guillaume Gadek[a], Paul Guélorget[a,b]

[a]Airbus, Advanced Information Processing, France
[b]Institut Polytechnique de Paris, Institut Mines-Télécom, Télécom SudParis

Abstract

Fake news and post-truth are everywhere. The huge number of online news outlets and the frequency of content creation underlines the demand for automatic information evaluation tools. Previous work usually focuses either on automatic fact-checking, or on fake-looking identification: the former tries to match a piece of content with trustable information, enough to confirm or infirm the claims. The latter gathers clues to help the reader's assessment of the piece of content. In this domain, there is no silver bullet: the reader desires verifiable information, thus the *fake news detector* should be interpretable or explainable.

In this article, we propose TC-CNN: an interpretable text classifier. We use it on two tasks: fake news detection and emotion classification. A second contribution relies on these two classifiers, and on a third-party hate detector, to perform a case study on this year real and fake-news press articles, in a comparison between mainstream and alt-right media.

Keywords: text classification; fake news detection; interpretable AI

1. Introduction

Since 2015, all media observers have become aware of the "fake news" problem, even though there has always been information manipulation campaigns throughout history. The easiness to create convincing news outlets and to diffuse media content without any field reporters leads to a massive amount of dubious contents [25]. These press- or news-looking content are then broadly diffused through the social media and trigger harmful consequences, notably observed in the form a discredit towards journalists and politics. Computer sciences and AI already proposed a few approaches to deal with this challenge. Automatic fact-checking is already partially addressed, but presents intrinsic limitations [12] and often relies on third-parties fact-checkers. A second approach to detect manipulations consists in evaluating the fakeness of every piece of information. Most fake news are persuasive, however they often do not resist against a well-formed mind. Manipulation clues detection in texts has long been a research topic, helped notably

E-mail address: guillaume.gadek@airbus.com

10.1016/j.procs.2020.08.009

by tools such as LIWC [22]. These expert-based approaches produce an insightful analysis on discourses and official announcements, but require rare competences to be established and updated.

News articles of today already mutated since the beginning of the century, with clickbait techniques evolving relatively fast: we believe that machine learning is more adapted to this task. Indeed, deep learning classification does not seem relevant to study whether a fact is false; however it brings some valuable hints about the writing *style* of a piece of text. In this article, we propose TC-CNN (Textual Class-activation-mappings Convolutional Neural Network), a new model for deep learning based text classification, that jointly produces an interpretation of the input, highlighting the patterns supporting or opposing a label prediction. Indeed, the structure of a fully convolutional network defines receptive fields that are naturally adapted to automatically detecting groups of successive words that reveal local, contextual and salient knowledge [17, 29]. These local patterns are then used as rationale for identifying the corresponding category in a classification task.

We apply this model on a fake news related task: the detection of biased press articles, written in English. This topic is intrinsically complex and this sole classifier cannot be considered a silver bullet. We are aware of this and propose a broader framework, capable of evaluating emotion, bias and hatred levels in a text. This technical proposal is applied on real data: we perform a case study on recent pieces of news and "fake" news, comparing the writing style of major English language news outlets.

The following sections are structured as follows: Section 2 gives some details about the challenging fake news problem and describes previous approaches to solve it. The we propose an interpretable model in Section 3, which usage is illustrated and validated in Section 4. A comparative case study is performed, comparing articles stemming from three main sources or aggregators, enabling the reader to grasp the benefit to reveal the complexity of today's media landscape. Finally, Section 6 concludes this paper.

2. Related Work

2.1. The fake news context

Various definitions and categorisations have been proposed in the domain of fake news analysis; basically, a distinction is done between "serious fabrications", when news articles are forged, mentioning events that never happened, "hoaxes" or rumours that only aim to be spread (and are sometimes referred to as *bullshit*), and "satire": fabrications with an obvious humoristic goal (e.g., *The Onion*[1]) [25]. Overall, the term "fake news" refers both to the globally speaking post-truth informational space, and to pieces of information that are intentionally diffused, while knowing they are false.

Beyond this term, the real problem deals with information manipulation, and the many ways to mix intent, information (e.g. messages) and knowledge (e.g. facts). An exhaustive formalisation of this problem has been recently proposed, with a special focus on the act of *lying* [13]: everything is not only based on content falsehood, but also on content perception based on its source, and its propagation.

Other recent initiatives are focuses on the automation (or at least, the up-scaling) of fact-checking, such as the CrossCheck[2] project launched in 2017 with Google News Lab in partnership with more than 20 media outlets. In parallel, Facebook associated with eight French media to reduce the amount of false information on its website. A similar project had already been launched in the United States with the support of ABC News, AP, FactCheck.org, Politifact and Snopes. CrossCheck is a collaborative journalism project that brings together editorial teams from all over the world to accurately deal with false, misleading or confusing statements circulating online, studying topics, comments, images and videos.

A very relevant survey on fake news detection proposed to split the problem into four challenging tasks, evaluating: the fact, the style, the propagation and the credibility of the emitters of a piece of news [30]. Within the *style* challenge, "clickbait" articles detection is specifically mentioned. The different kinds of *fakes* is still an open debate. Many papers propose their own taxonomies, such as *malicious* / *hoax* / *satire* [24], on social media texts; *fake* / *true* for

[1] https://www.theonion.com/

[2] http://www.lemonde.fr/les-decodeurs/article/2017/02/28/lutte-contre-les-fausses-informations-le-monde-partenaire-du-projet-crosscheck_5086731_4355770.html

posts combining text, image and propagation data [20] or even to evaluate whether the title of a press article is related, or opposed, to its own text [15]. Always more information can be combined, such as the system XFake, which independently analyses semantic, linguistic and contextual (e.g., the attributes of the source: author, media outlet...) data [27]. A more classical approach, that we follow, is focused on classifying press articles along three classes: *biased* (actively promoting a point of view), *bullshit* (when the article only aims at attracting attention), and *legit* (the closer possible of a "true" label for a press article) [16].

2.2. Machine learning on related text classification problems

Text classification tasks have been widely addressed in numerous real applications over the last few years, such as emotion, sentiment and opinion classification [1, 5, 10], language identification [4] and even irony detection [6]. For the exploitation of raw textual data, the classification task is addressed by two main processes: feature extraction and the classification itself. Common feature extraction methods (word embeddings) include GloVe [23], FastText [14], BERT [9]. On the classification side, Convolutional Neural Networks (CNN) [17] and Recurrent Neural Networks (RNN) [7] have recently become widespread.

Among text classification tasks, emotion detection has known a quick recent evolution, from dictionary-based [21] to machine learning based approaches: empowered by a 7-class, 40,000 tweet dataset, accuracies have increased, from 60% [3] to 70% [19]. Powered by bigger social media datasets, the project *Text-emotion*[3] limited itself to 5 emotions (including neutral): various scientific opinions cohabit on the number of classes to keep, with regard to annotators agreement. This project reached a global accuracy of 62% on social media posts.

Entirely dedicated to the web content moderation, the detection of hatred has recently benefited from both theoretical and practical advances. Among the many contributions on this domain, *hatesonar*, a classifier embedded in a Python library (thus, very easy to integrate in a larger system), is grounded on a 3-classes hate detection: "hate" (directed towards a group of people), "offence" (directed towards an individual) or "neither" [8]. A widespread claim consists in combining different indicators to analyse Web2.0 contents: as an example, the system *Oasis* aggregates topic detection, and two CNN-LSTM based classifiers for sentiment (3 classes) and emotion (4 emotions, and neutral) detection [19].

2.3. Explainable AI and interpretable fake news

The idea to detect fake news with a useful explanation is relatively new, with initiatives such as dEFEND [26] relying on user comments to automatically spot controversial claims in news articles. This approach heavily relies on a sane community of readers-commenters, which may not be the case for every media outlet.

More generally, in a classification setup, obtaining comprehensive class-membership cues is a highly tedious task, significantly more complex than a simple labelling one. To this purpose, two techniques are often considered to characterise sub-documents segments. A first one consists in observing the behaviour of an attention mechanism [2]. The second approach is based on the extraction of class activation maps [29].

A self-attention mechanism was introduced in [18], where high attention scores assigned to crucial words are observed when performing text classification. Also, it has been shown that attention is transferable [28]. However, attention scores are flawed when considering the purposes of this paper. They can be interpreted as the salience of a sub-part of the document relatively to a generic classification task, and not relatively to a precise label. They are ranging from zero (no interest) to one (high interest) without carrying information whereas the considered sub-parts play in favour or against class membership.

A different approach concerns the so-called class activation maps (CAMs), introduced in [29]. CAMs are able to define the contribution of each sub-part of the document to each considered class. They can be seen as the penultimate neuron layer; they however require specific architecture conditions in order to conserve a logical matching with the inputs.

[3] https://github.com/tlkh/text-emotion-classification

2.4. Discussion

Past work paved the way for better algorithm architectures, and for better taxonomies. The former enables us to propose an accurate prediction, combined with a mechanism for interpretation of the result: we decide to follow the class activation maps approach, and to apply it on textual input, following recent work [11]. The latter is focused on the application on real-world data: *fake news* cannot be reduced to a simple yes/no classification problem. Pseudo-fakes, ideological articles, hatred and emotion-loaded content are too often mixed-up. They all do participate to the *fake news* phenomenon: they have to be measured separately; then combined to build a complete analysis of the content of press-looking articles, avoiding black-box predictions.

Also, this combination cannot be seen uniquely from the domain of machine learning. Fake news corpora exist and are invaluable to train our methods; yet they are by nature out-dated, topic-specific and susceptible of bias. Such systems have to be illustrated and exploited with two goals: to show their relevance, and to update our awareness about today's media.

To tackle this intricate challenge, we introduce an interpretable text classifier, grounded on a CNN architecture while exploiting the CAMs. We propose to train two models, to predict emotions and style, and re-exploit a previous hate detection library. With these tools, we propose a case study on this year press articles, either real or fake.

3. TC-CNN: an interpretable model for biased news article classification

3.1. Spatially interpretable architecture

Initially used for image classification purposes used [29], a recent work on text classification propose to adapt the CAM technique to textual data [11]. The method relies on the observation that spatiality is preserved across convolutional layers, whereas it is lost in the last fully-connected layers used by some CNNs. Hence, it only concerns fully-convolutional networks where global average pooling (GAP) or global max pooling (GMP) is applied to squash the T spatial features vectors associated to the deepest feature maps $F = \{F_1, ..., F_T\} \in \mathbb{R}^{T\times K}$ into a single, global feature vector $F_g \in \mathbb{R}^K$ in which all spatiality is lost. Here, K denotes the size of the considered feature maps.

The model that we have adopted is a fully-convolutional neural network made of 3 layers of 128 kernels of size 5 followed by a global average pooling and a softmax classification layer. We used the pre-trained FastText word embedding [14]: a context-aware embedding such as BERT means a much higher-complexity architecture and would degrade the interpretability of the final result because it already takes into account the role of the neighbouring words without explaining how. In our model however, we directly trace the importance of each word's contribution.

Because we need to preserve the temporality across layers, we use same padding for convolutions, so that there is exactly one output layer directly corresponding to each input token. Thus, the last convolutional layer presents a number of outputs that is equal to the number of input words. In this way, the t^{th} output of the last convolutional layer describes the t^{th} word of the input sentence, while taking into consideration its context within the convolutional receptive field.

The CAM extraction is explained hereafter. If there are C different labels, then the softmax input is defined as:

$$S = W^T F_g + b \tag{1}$$

where $W = \{w_c^k\} \in \mathbb{R}^{K\times C}$ and $b \in \mathbb{R}^C$ are the final weights and biases. For any input example x, the class activation map for label c at location t is obtained by summing the contribution $w_c^k F_t^k$ of each scalar feature F_t^k to the final score of label c, as described in the following equation:

$$CAM(x, c, t) = \sum_k w_c^k F_t^k \tag{2}$$

The CAM address the aforementioned limitations of the attention scores to fulfil our purpose: in a text classification context, it provides a signed contribution of each word to each class membership tackled by the model. Hence, we chose to use CAMs to extract dense, comprehensive knowledge from a network previously trained on a large set of documents, thus $R = CAM$. Our CAM-extracted model is illustrated in Figure 1.

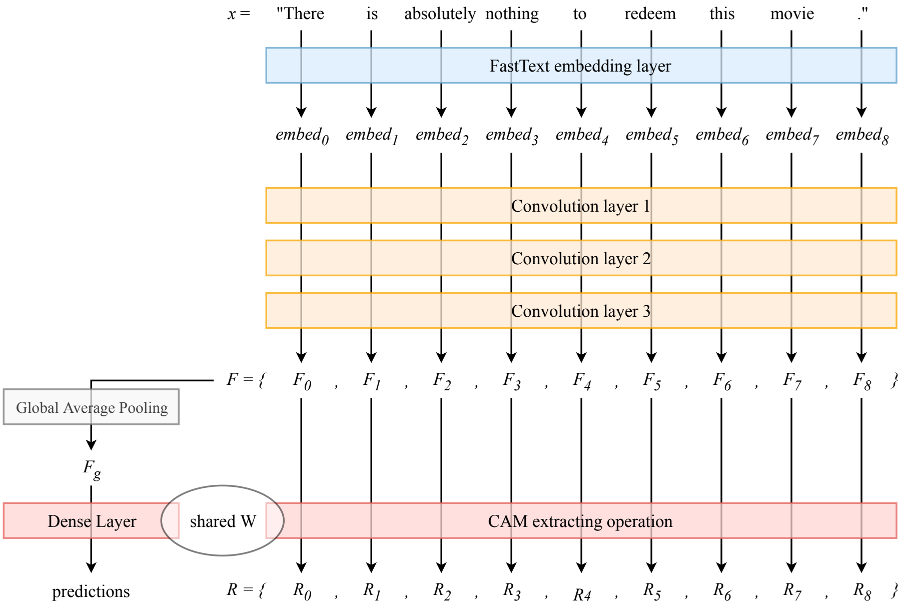

x = "There is absolutely nothing to redeem this movie ."
FastText embedding layer
embed0 embed1 embed2 embed3 embed4 embed5 embed6 embed7 embed8
Convolution layer 1
Convolution layer 2
Convolution layer 3
F = { F0 , F1 , F2 , F3 , F4 , F5 , F6 , F7 , F8 }
Global Average Pooling
Fg
Dense Layer
shared W
CAM extracting operation
predictions
R = { R0 , R1 , R2 , R3 , R4 , R5 , R6 , R7 , R8 }

Fig. 1. Extraction of Class Activation Maps from a tokenized sentence.

3.2. Predicting fake-like articles

We used this architecture on two classification problems. To begin against fake news, we rely on two well-known press articles datasets to train the model:

- Kaggle fake news[4] is composed of 13,000 documents in English, tagged as “fake”: either *biased* or *bullshit*.
- Signal-Media[5] gathers a huge corpora of press articles (in English), deemed to be legitimate. Following [16], we select a random sample so as to balance the classes.

The use of contextualised embeddings such as Bert or Elmo would be a return to the “black box” phenomenon, because they exploit the whole text to embed each token; instead we use the *fasttext* pre-trained embeddings, of dimension 300. The texts are only tokenised (including the punctuation), through the python library NLTK[6]. This pre-trained embeddings imply that the model has to grasp the context. A random train/test split of 20% is used to be able to train compute accuracy scores. We adopted an early stop strategy, resulting in a total of 41 epochs.

3.3. Secondary task: predicting emotion

Fake news and biased content are often intricate with high emotional values: the intuition is that the emotive impact would replace rational facts to persuade the readers. While emotion detection is no new task, we propose to train a second model, based on the same architecture, to enrich the analysis and combine “fakeness” predictions with emotive labels. The only modification to the model architecture is to output 5 class labels, to match the task labels. As a training corpus, we retain the dataset Text-emotion[7], which consists of over 40,000 tweets in English. Even though the document size varies between press articles and tweets, we are confident that our model is sufficiently focused on the sentence patterns, to be able to tackle this second classification task.

[4] https://www.kaggle.com/mrisdal/fake-news
[5] http://research.signalmedia.co/newsir16/signal-dataset.html
[6] https://www.nltk.org/
[7] https://github.com/tlkh/text-emotion-classification

4. Experiments: TC-CNN at work

A previous approach in the literature compared various machine learning approaches, highlighting that excellent accuracies could be reached using deep learning and more specifically, a recurrent neural network fed with pre-trained GloVe embeddings [16]. In Table 1, we include the score proposed in their article; however as we may have discrepancies (including in the train/test split), we also include our 300d-GloVe-based recoded baseline, *CNN-prev*, which consists of three convolutional layers with kernel size 3, a dense layer, and a global max pooling. We extend the comparison to another broadly used classification framework, *fasttext* [14].

Table 1. Classification scores

Task	Classifier	Precision	Recall	Weighted F_1
Fake news	As in [16]	-	-	0.85
	CNN-prev	0.911	0.907	0.907
	fasttext	0.878	0.876	0.871
	TC-CNN-fake	0.918	0.919	0.918
Emotion	*Text-emotion*[8]	-	-	0.62
	fasttext	0.597	0.600	0.590
	TC-CNN-emo	0.686	0.688	0.683

As a discussion about Table 1, our model performs equivalently or relatively better than the previous CNN-based architecture and than *fasttext*, while conserving a reasonable network size and more importantly, while bringing a huge gain: interpretable results.

A recurrent objection on fake news classification is the risk to overfit the classifier on a specific dataset, overfocusing on a few words. This problem is particularly hard to analyse, because of the scarce amount of similarly-labelled, thematically different datasets. The introduced model presents an advantage to investigate this point, through the class activation mappings. We propose to compare an example present amongst the biased articles of the Kaggle dataset in Figure 2, typical of the 2016 fake news: it deals with the Clinton affairs. “Clinton” being a recurrent target of disinformation campaigns, is likely to be determining in the output of the classifier. Figure 2 (upper) shows the activation values, in blue (positive) and orange (negative) for the class *bias*, which is the predicted class for this text. The classifier provides an interpretable output, which is the weight of each token in the prediction; we propose some hints about the presence of such weights, using red boxes. To begin with, the model has learnt a vocabulary: James Comey and Hillary Clinton were present in a number of biased articles, thus highlighted in blue. The model also learnt patterns: the abusive use of quotes every few words suggest bias. Precise information about places and dates are highlighted in orange, suggesting non-bias (thus looking more legitimate).

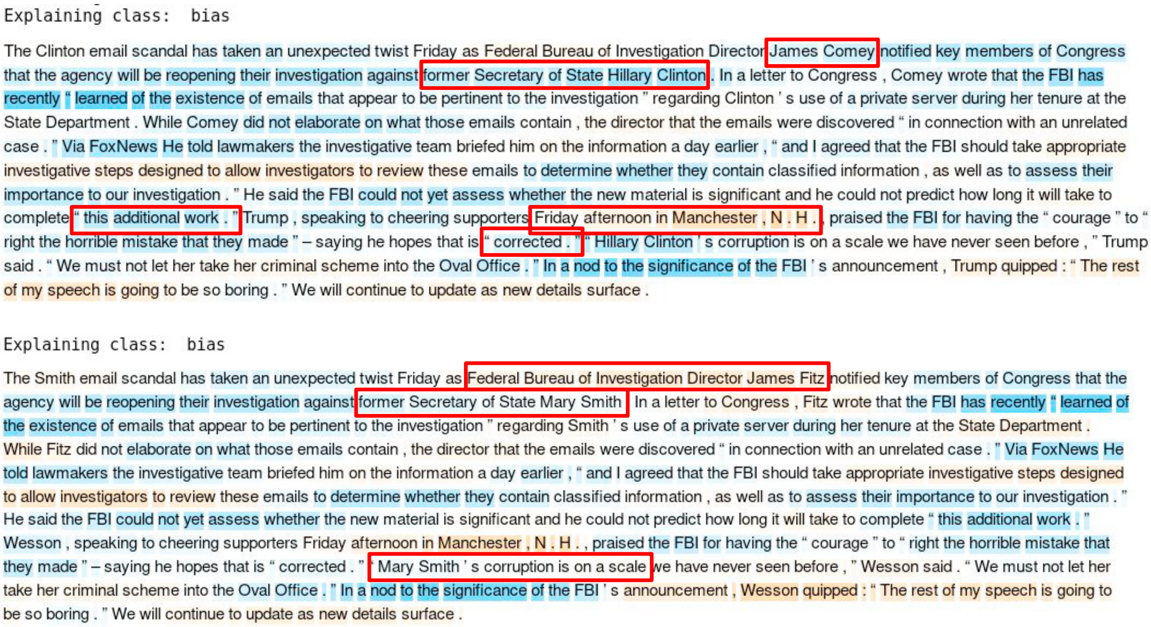
By replacing all family names by more neutral ones, we can see the differences on Figure 2 (lower): “Smith” is not as strongly linked to the *bias* class as “Clinton” (notably in the phrase “Hillary Clinton’s corruption”, in the last red box). However, the sentence structure itself is still determining, enabling the classifier to maintain its prediction and recognise a bias: as an example, the model appreciates as factual (here in orange) to fully give the title and name of a person.

5. Measuring offences and biases in the press

5.1. Sources of 2020 data

We constituted three news article datasets: *news_outlets*, contains a random sample of 30 articles from each of seven world-level English press references: CNN, MSNBC, FoxNews, The Guardian, BreitbartNews, The Daily Express and The BBC. In a separated *the_onion* dataset, an eighth media outlet is also considered: The Onion. We keep it segregated, because of its satirical nature: this outlet does not aim to propagate news, but humour. The included articles have been redacted in January and February 2020.

The dataset *gab_trends* has been collected by downloading the proposed links from a news aggregator, “Gab Trends”. Gab is a microblogging platform (Twitter-like), US-based, that brands itself as championing free speech:



```
Explaining class:  bias
```

The Clinton email scandal has taken an unexpected twist Friday as Federal Bureau of Investigation Director James Comey notified key members of Congress that the agency will be reopening their investigation against former Secretary of State Hillary Clinton . In a letter to Congress , Comey wrote that the FBI has recently " learned of the existence of emails that appear to be pertinent to the investigation " regarding Clinton ' s use of a private server during her tenure at the State Department . While Comey did not elaborate on what those emails contain , the director that the emails were discovered " in connection with an unrelated case . " Via FoxNews He told lawmakers the investigative team briefed him on the information a day earlier , " and I agreed that the FBI should take appropriate investigative steps designed to allow investigators to review these emails to determine whether they contain classified information , as well as to assess their importance to our investigation . " He said the FBI could not yet assess whether the new material is significant and he could not predict how long it will take to complete " this additional work . " Trump , speaking to cheering supporters Friday afternoon in Manchester , N . H . , praised the FBI for having the " courage " to " right the horrible mistake that they made " – saying he hopes that is " corrected . " " Hillary Clinton ' s corruption is on a scale we have never seen before , " Trump said . " We must not let her take her criminal scheme into the Oval Office . " In a nod to the significance of the FBI ' s announcement , Trump quipped : " The rest of my speech is going to be so boring . " We will continue to update as new details surface .

```
Explaining class:  bias
```

The Smith email scandal has taken an unexpected twist Friday as Federal Bureau of Investigation Director James Fitz notified key members of Congress that the agency will be reopening their investigation against former Secretary of State Mary Smith . In a letter to Congress , Fitz wrote that the FBI has recently " learned of the existence of emails that appear to be pertinent to the investigation " regarding Smith ' s use of a private server during her tenure at the State Department . While Fitz did not elaborate on what those emails contain , the director that the emails were discovered " in connection with an unrelated case . " Via FoxNews He told lawmakers the investigative team briefed him on the information a day earlier , " and I agreed that the FBI should take appropriate investigative steps designed to allow investigators to review these emails to determine whether they contain classified information , as well as to assess their importance to our investigation . " He said the FBI could not yet assess whether the new material is significant and he could not predict how long it will take to complete " this additional work . " Wesson , speaking to cheering supporters Friday afternoon in Manchester , N . H . , praised the FBI for having the " courage " to " right the horrible mistake that they made " – saying he hopes that is " corrected . " " Mary Smith ' s corruption is on a scale we have never seen before , " Wesson said . " We must not let her take her criminal scheme into the Oval Office . " In a nod to the significance of the FBI ' s announcement , Wesson quipped : " The rest of my speech is going to be so boring . " We will continue to update as new details surface .

Fig. 2. Examples of CAMs explaining a bias about Clinton: (a) original, (b) with replacements.

Table 2. Most frequent words, excluding stopwords, on *news_outlets*

Prediction	Words	Nb Docs
bias	Democrats, House, time, –, com, Iran, like, president, U, 2020, President, also, one, would, people, —, said, Trump	861
bullshit	campaign, Court, v, appointee, Pelosi, American, also, Senate, Louisiana, one Republican, trial, two, News, CNN, people, Democrats, voted, would, witnesses, court, House president, case, Donald, vote, President, impeachment, abortion, said, Trump	30
factual	first, two, like, —, told, Trump, time, year, also, one, would, people, ”, “, said	484

some of the hosted content would be banned on mainstream platforms. This dataset contains 1135 articles from a total of 150 different web domains. We crawled it from the 15th of December 2019, to the 15th of February, 2020.

While *gab_trends* indeed contains well-known news outlets such as the BBC, the Daily Mail and FoxNews, it also refers to other websites, which do not enjoy the same journalistic quality. Breitbart is the most cited (over 200 times), followed by Youtube (which may refer to any type of content); ZeroHedge (150) and InfoWars[9] (cited less than 40 times) are at least controversial.

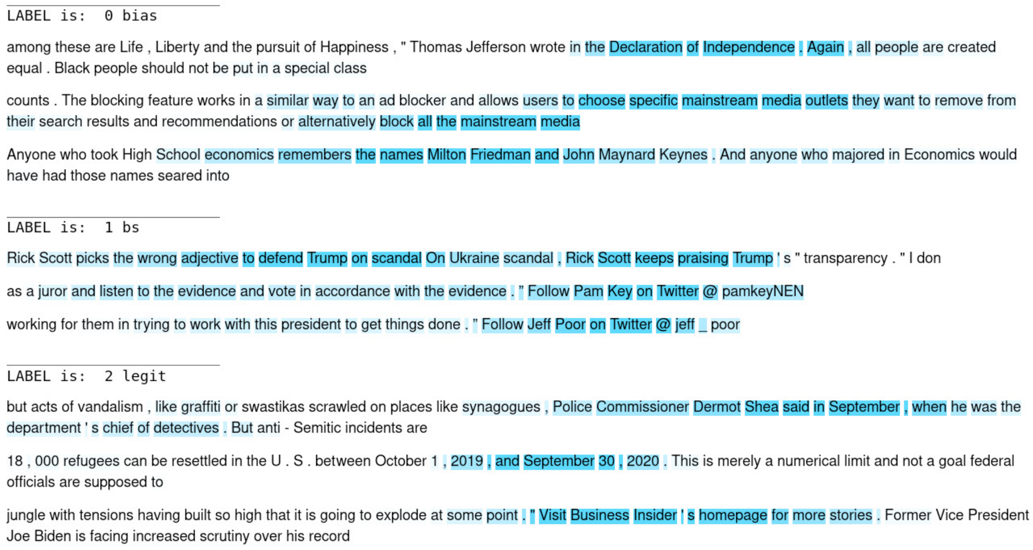
5.2. Qualitative analysis on a sample

A first deep dive in both the *news_outlets* and the *gab_trends* corpora consists in genuinely looking at the most frequent words for each predicted label, as displayed in Table 2. As a classic frequency count retrieves generic words, we decided to hide the stopwords. A first observation is the dominance of politics related terms: “Trump” is the most frequent person in all three classes; the bullshit class does not have the same importance than the two others, because of the scarce amount of articles that were tagged this way.

A second opportunity is offered thanks to the CAMs, to display the words that trigger each class of our classifier. In Figure 3, we retain the words with class-relative-CAM highest values (coloured words in the middle) and their immediate text for three documents predicted in this class. The second biased example is typical, with the reference to

[9] “InfoWars is a far-right American conspiracy theory and fake news website”, https://en.wikipedia.org/wiki/InfoWars

“mainstream media”; the bullshit class seems attracted by Twitter users *follow me* requests. On the legitimate articles side, the first example is very well documented (title, name and date); the second illustrate the focus on precise dates; the last suggest the presence of a given media outlet in the training dataset.



```
LABEL is:  0 bias
```

among these are Life , Liberty and the pursuit of Happiness , " Thomas Jefferson wrote in the Declaration of Independence . Again , all people are created equal . Black people should not be put in a special class

counts . The blocking feature works in a similar way to an ad blocker and allows users to choose specific mainstream media outlets they want to remove from their search results and recommendations or alternatively block all the mainstream media

Anyone who took High School economics remembers the names Milton Friedman and John Maynard Keynes . And anyone who majored in Economics would have had those names seared into

```
LABEL is:  1 bs
```

Rick Scott picks the wrong adjective to defend Trump on scandal On Ukraine scandal , Rick Scott keeps praising Trump ' s " transparency . " I don

as a juror and listen to the evidence and vote in accordance with the evidence . " Follow Pam Key on Twitter @ pamkeyNEN

working for them in trying to work with this president to get things done . " Follow Jeff Poor on Twitter @ jeff _ poor

```
LABEL is:  2 legit
```

but acts of vandalism , like graffiti or swastikas scrawled on places like synagogues , Police Commissioner Dermot Shea said in September , when he was the department ' s chief of detectives . But anti - Semitic incidents are

18 , 000 refugees can be resettled in the U . S . between October 1 , 2019 , and September 30 , 2020 . This is merely a numerical limit and not a goal federal officials are supposed to

jungle with tensions having built so high that it is going to explode at some point . " Visit Business Insider ' s homepage for more stories . Former Vice President Joe Biden is facing increased scrutiny over his record

Fig. 3. Words with highest CAMs, in their text

5.3. *Statistical overview*

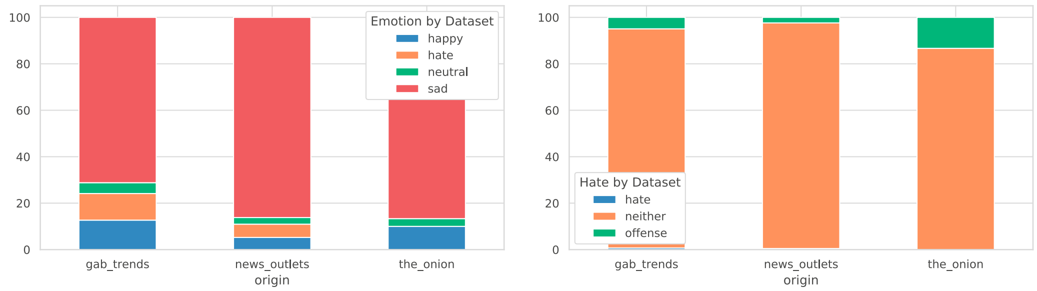


Fig. 4. Comparing the datasets, by emotion (a), by presence of hate (b)

For the three datasets, the proportion of each emotion and hate presence is shown in Figure 4. On the emotion side, the majority is sad: while “neutral” would have been expected, we believe that the semantic fields around politics, geopolitics and economics are globally closer to sadness. We conserved the original taxonomy for the emotions, thus the emotion “hate” refers to anger, while the hate label “Hate” refers to the presence of at least one hateful or offensive sentence in the article: a continuously high level of hate in a whole article is somehow improbable. Gab Trends proposes higher presence rates of either happy or hateful articles. The Onion somehow manages its goal with a sadness level comparable to serious newspapers: we confess that our emotion detector does not spot humour. On the hate side, only a handful of articles are marked as “hateful”, all in our gab_trends dataset. They do not go beyond clichés about demographics (white people share among the US population), or about feminism.

The proportion of biased articles is grouped by dataset on Figure 5: for the articles shared on *Gab_trends* (a), and for our selection of news outlets (including *The Onion*) on (b). As the referred article sets are different, the scores of fakeness can vary for a same source present in both (a) and (b). On the left side, our model identifies the classic newspapers (Nypost, BBC, Daily Mail and Fox) even though they cover a variety of styles: as an example, the N.Y.

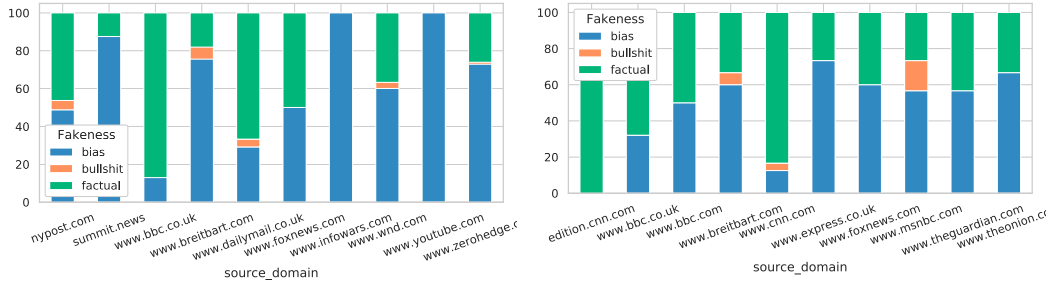


Fig. 5. Article quality by source for *gab_trends* (a), for *news_outlets* and *the_onion* (b)

Post is a tabloid. The model is however quite critical towards *summit*, *wnd* and *infowars*, whose content is predicted as biased. On the right, CNN and the BBC are split in two domain names: their writing styles are globally tagged as factual journalism. FoxNews is also present and reasonably factual. Our selection of Breitbart articles results more factual than Gab's selection.

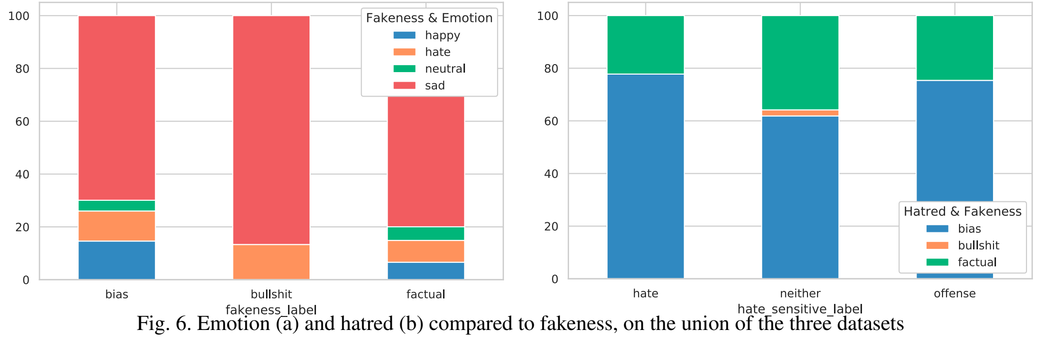


Fig. 6. Emotion (a) and hatred (b) compared to fakeness, on the union of the three datasets

An interesting point is shown in Figure 6, comparing the repartition of dominant emotion along the fakeness (a), and comparing hatred and fakeness (b). For this figure, we aggregated the three datasets. It appears that classic factual articles are expected to be sad; a second recurrent trend for biased articles is to bring either happiness or anger. On the right hand side, the figure shows the distribution of hatred with regard to the fakeness. Note that most of the articles are not hateful, represented in the "neither" column. We observe that hate and offense are most likely found in biased articles.

6. Conclusion

Fake news and information manipulation detection have long been a difficult, intricate problem, which can be decomposed in many scientific challenges. While automatic fact-checking always requires a set of trusted sources of truth, content quality can be learnt offline. In this article, we followed our intuition to build generic content evaluation tools, to help the readers understand what is aimed by the writers and see the lack of references and facts. We think that fake news cannot be seen only as a classification problem: how could one trust an algorithm when one does not even trust newspapers? The role of the algorithm and the weight to give to its prediction has to be well thought, as the aim is to detect bias, not to propagate one. In this regard, our method displays a great advantage, thanks to its interpretability: the reader sees why and what triggered the prediction. We combined this spectrum of analysis with related tasks: hate and emotion detection.

The case study conducted on recent news articles, redacted at the beginning of 2020, gives hints about the current mixing of factual and persuasive articles on the media landscape, on "mainstream" and "non-mainstream" news outlets. Notably, the studied "alternative media" mixes high- and poor-quality articles and outlets, equally presenting valuable sources (the BBC) and scam websites (infowars). As part of future works, we aim to aggregate more content evaluation modules to enable a holistic analysis of fake content propagation on social media.

References

[1] Abbasi, A., Chen, H., Salem, A., 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. ACM Transactions on Information Systems (TOIS) 26, 12.

[2] Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv:1409.0473 [cs, stat] URL: http://arxiv.org/abs/1409.0473. arXiv: 1409.0473.

[3] Bouazizi, M., Ohtsuki, T., 2017. A pattern-based approach for multi-class sentiment analysis in twitter. IEEE Access 5, 20617–20639.

[4] Castro, D.W., Souza, E., Vitório, D., Santos, D., Oliveira, A.L., 2017. Smoothed n-gram based models for tweet language identification: A case study of the Brazilian and European Portuguese national varieties. Applied Soft Computing 61, 1160–1172.

[5] Catal, C., Nangir, M., 2017. A sentiment classification model based on multiple classifiers. Applied Soft Computing 50, 135–141.

[6] Charalampakis, B., Spathis, D., Kouslis, E., Kermanidis, K., 2016. A comparison between semi-supervised and supervised text mining techniques on detecting irony in greek political tweets. Engineering Applications of Artificial Intelligence 51, 50–57.

[7] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation URL: https://arxiv.org/abs/1406.1078.

[8] Davidson, T., Warmsley, D., Macy, M., Weber, I., 2017. Automated hate speech detection and the problem of offensive language, in: Eleventh international aaai conference on web and social media.

[9] Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs] URL: http://arxiv.org/abs/1810.04805. arXiv: 1810.04805.

[10] Giatsoglou, M., Vozalis, M.G., Diamantaras, K., Vakali, A., Sarigiannidis, G., Chatzisavvas, K.C., 2017. Sentiment analysis leveraging emotions and word embeddings. Expert Systems with Applications 69, 214–224.

[11] Guélorget, P., Grilheres, B., Zaharia, T., 2020. Deep active learning with simulated rationales for text classification, in: Proceedings of the 2020 International Conference on Pattern Recognition and Artificial Intelligence.

[12] Hassan, N., Adair, B., Hamilton, J.T., Li, C., Tremayne, M., Yang, J., Yu, C., 2015. The quest to automate fact-checking, in: Proceedings of the 2015 Computation+ Journalism Symposium.

[13] Icard, B., 2019. Lying, deception and strategic omission: definition et evaluation. Ph.D. thesis. Paris Sciences et Lettres.

[14] Joulin, A., Grave, E., Bojanowski, P., Mikolov, T., 2016. Bag of Tricks for Efficient Text Classification. arXiv:1607.01759 [cs] URL: http://arxiv.org/abs/1607.01759. arXiv: 1607.01759.

[15] Jwa, H., Oh, D., Park, K., Kang, J.M., Lim, H., 2019. exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert). Applied Sciences 9, 4062.

[16] Katsaros, D., Stavropoulos, G., Papakostas, D., 2019. Which machine learning paradigm for fake news detection?, in: 2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI), IEEE. pp. 383–387.

[17] Kim, Y., 2014. Convolutional Neural Networks for Sentence Classification. arXiv:1408.5882 [cs] URL: http://arxiv.org/abs/1408.5882. arXiv: 1408.5882.

[18] Lin, Z., Feng, M., Santos, C.N.d., Yu, M., Xiang, B., Zhou, B., Bengio, Y., 2017. A Structured Self-attentive Sentence Embedding. arXiv:1703.03130 [cs] URL: http://arxiv.org/abs/1703.03130. arXiv: 1703.03130.

[19] Liu, L., Huang, X., Xu, J., Song, Y., 2019. Oasis: Online analytic system for incivility detection and sentiment classification, in: 2019 International Conference on Data Mining Workshops (ICDMW), IEEE. pp. 1098–1101.

[20] Maigrot, C., Claveau, V., Kijak, E., 2018. Fusion-based multimodal detection of hoaxes in social networks, in: 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), IEEE. pp. 222–229.

[21] Munezero, M., Montero, C.S., Mozgovoy, M., Sutinen, E., 2015. Emotwitter–a fine-grained visualization system for identifying enduring sentiments in tweets, in: Computational Linguistics and Intelligent Text Processing. Springer, pp. 78–91.

[22] Pennebaker, J.W., Francis, M.E., Booth, R.J., 2001. Linguistic inquiry and word count: Liwc 2001. Mahway: Lawrence Erlbaum Associates 71, 2001.

[23] Pennington, J., Socher, R., Manning, C., 2014. Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543.

[24] 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.

[25] Rubin, V.L., Chen, Y., Conroy, N.J., 2015. Deception detection for news: three types of fakes, in: Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community, American Society for Information Science. p. 83.

[26] Shu, K., Cui, L., Wang, S., Lee, D., Liu, H., 2019. defend: Explainable fake news detection, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 395–405.

[27] Yang, F., Pentyala, S.K., Mohseni, S., Du, M., Yuan, H., Linder, R., Ragan, E.D., Ji, S., Hu, X., 2019. Xfake: explainable fake news detector with visualizations, in: The World Wide Web Conference, pp. 3600–3604.

[28] Zagoruyko, S., Komodakis, N., 2017. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. arXiv:1612.03928 [cs] URL: http://arxiv.org/abs/1612.03928. arXiv: 1612.03928.

[29] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning Deep Features for Discriminative Localization, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA. pp. 2921–2929. URL: http://ieeexplore.ieee.org/document/7780688/, doi:10.1109/CVPR.2016.319.

[30] Zhou, X., Zafarani, R., 2018. Fake news: A survey of research, detection methods, and opportunities. arXiv preprint arXiv:1812.00315 .