FakeFlow: Fake News Detection by Modeling the Flow of Affective Information

Bilal Ghanem[1], **Simone Paolo Ponzetto**[2], **Paolo Rosso**[1], **Francisco Rangel**[3]
[1]Universitat Politècnica de València, Spain
[2]University of Mannheim, Germany
[3]Symanto Research, Germany
bigha@doctor.upv.es, simone@informatik.uni-mannheim.de,
prosso@dsic.upv.es, francisco.rangel@symanto.com

Abstract

Fake news articles often stir the readers' attention by means of emotional appeals that arouse their feelings. Unlike in short news texts, authors of longer articles can exploit such affective factors to manipulate readers by adding exaggerations or fabricating events, in order to affect the readers' emotions. To capture this, we propose in this paper to model the flow of affective information in fake news articles using a neural architecture. The proposed model, FakeFlow, learns this flow by combining topic and affective information extracted from text. We evaluate the model's performance with several experiments on four real-world datasets. The results show that FakeFlow achieves superior results when compared against state-of-the-art methods, thus confirming the importance of capturing the flow of the affective information in news articles.

1 Introduction

In today's information landscape, fake news are used to manipulate public opinion (Zhou and Zafarani, 2018) by reshaping readers' opinions regarding some issues. In order to achieve this goal, authors of fake news' narratives need to capture the interest of the reader. Thus, they are putting efforts to make their news articles look more objective and realistic. This is usually done by adding misleading terms or events that can have a negative or positive impact on the readers' emotions.

Short text false information, e.g., fake claims or misleading headlines, might be less harmful than news articles. They may have some eye-catching terms that aim to manipulate the readers' emotions (Chakraborty et al., 2016). In many cases, the identification of this kind of exaggeration in short statements can unmask the fabrication. On the other hand, in fake news articles the authors exploit the length of the news to conceal their fabricated story. This fact exposes the readers to be emotionally manipulated while reading longer texts that have several imprecise or fabricated plots. The flow of information has been investigated for different tasks: Reagan et al. (2016) studied the emotional arcs in stories in order to understand complex emotional trajectories; Maharjan et al. (2018) model the flow of emotions over a book and quantify its usefulness for predicting success in books; Kar et al. (2018) explore the problem of creating tags for movies from plot synopses using emotions.

Unlike previous works (Rashkin et al., 2017; Shu et al., 2018; Castelo et al., 2019; Ghanem et al., 2020) that discarded the chronological order of events in news articles, in this work we propose a model that takes into account the affective changes in texts to detect fake news. We hypothesize that fake news has a different distribution of affective information across the text compared to real news, e.g. more fear emotion in the first part of the article or more overall offensive terms, etc. Therefore, modeling the flow of such information may help discriminating fake from real news. Our model consists of two main sub-modules, topic-based and affective information detection. We combine these two sub-modules since a news article's topic may have a correlation with its affective information. For example, a fake news article about Islam or Black people is likely to provoke fear and express negative sentiment while another fake news that is in favor of a particular politician might try to evoke more positive emotions and also express some exaggerations.

The contributions of our work are as follows:

- We design a model that detects fake news articles by taking into account the flow of affective information[1].

[1] Available at https://github.com/bilalghanem/fake_flow

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 679–689
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

- Extensive experiments on four standard datasets demonstrate the effectiveness of our model over state-of-the-art alternatives.
- We build a novel fake news dataset, called MultiSourceFake, that is collected from a large set of websites and annotated on the basis of the joint agreement of a set of news sources.

2 Related Work

Previous work on fake news detection is mainly divided into two main lines, namely with a focus on social media (Zubiaga et al., 2015; Aker et al., 2017; Ghanem et al., 2019) or online news articles (Tausczik and Pennebaker, 2010; Horne and Adali, 2017; Rashkin et al., 2017; Barrón-Cedeno et al., 2019). In this work we focus on the latter one. Fact-checking (Karadzhov et al., 2017; Zlatkova et al., 2019; Shu et al., 2019a) is another closely related research topic. However, fact-checking targets only short texts (that is, claims) and focuses on using external resources (e.g. Web, knowledge sources) to verify the factuality of the news. The focus in previous work on fake news detection is mainly on proposing new feature sets. Horne and Adali (2017) present a set of content-based features, including readability (number of unique words, SMOG readability measure, etc.), stylistic (frequency of part-of-speech tags, number of stop words, etc.) and psycholinguistic features (i.e., several categories from the LIWC dictionary (Tausczik and Pennebaker, 2010)). When these features are fed into a Support Vector Machine (SVM) classifier and applied, for instance, to the task of distinguishing satire from real news, they obtain high accuracies. Using the same features for the task of fake news detection, however, results in somewhat lower scores. Pérez-Rosas et al. (2018) propose a model (FakeNewsDetector) that uses a feature set consisting of unigrams and bigrams, psycholinguistic, readability, punctuation and dependency-based syntactic features, and they evaluate the performance of their model in a cross-domain experiment. Rashkin et al. (2017) use a model based on ngram features with a MaxEntropy classifier and apply it to a dataset with different types of fake news articles (e.g., satire, hoax, propaganda, etc.). Similar to the previous work, the authors evaluate their system's performance on in-domain and out-of-domain test sets, respectively. News, and in particular fake news, are dynamic in nature and change constantly. In order to approach the dynamic nature of news, Castelo et al. (2019) propose a topic-agnostic model (TopicAgnostic) that is based on morphological (count of part-of-speech tags), psycholinguistic (personal concerns, affection, and perception categories from the LIWC dictionary), readability (Gunning Fog metric, etc.) and Web-Markup features to capture patterns of the Web pages' layout (frequency of advertisements, presence of an author name, etc.). All of the morphological, psycholinguistic and readability features in the TopicAgnostic model were extracted from headlines and texts of the news articles. The approach obtains a better performance than FakeNewsDetector on three different datasets using a SVM classifier. FakeNewsTracker (Shu et al., 2019b) is a deep neural network-based model that consists of two branches: one encodes news article texts and the other encodes social media engagements (e.g., tweets and their replies). A similar model called Emotionally Infused Network (EIN) is proposed in Ghanem et al. (2020). EIN encodes the text of the article and their affective content, based on several dictionaries, and then combines the two vector representations. The authors evaluate their model on a multi-class false information dataset and show the effectiveness of using emotion features extracted from the text. Despite the large variety of features and models that have been explored in previous work, none of these works considers the sequence of affective information in text; instead, they feed the entire news articles as one segment into their models. In contrast, the aim of our work is to evaluate this source of information, using a neural architecture.

3 The FakeFlow Model

Given an input document, the FakeFlow model first divides it into N segments. Then it uses both word embeddings and other affective features such as *emotions, hyperbolic words*, etc. in a way to catch the flow of emotions in the document. The model learns to pay attention to the flow of affective information throughout the document, in order to detect whether it is fake or real.

Figure 1 shows the architecture of the FakeFlow model. The neural architecture has two main modules: The first module uses a Convolutional Neural Network (CNN) to extract topic-based information from articles (left branch). The second module models the flow of the affective information within the articles via Bidirectional Gated Recurrent Units (Bi-GRUs) (right branch).

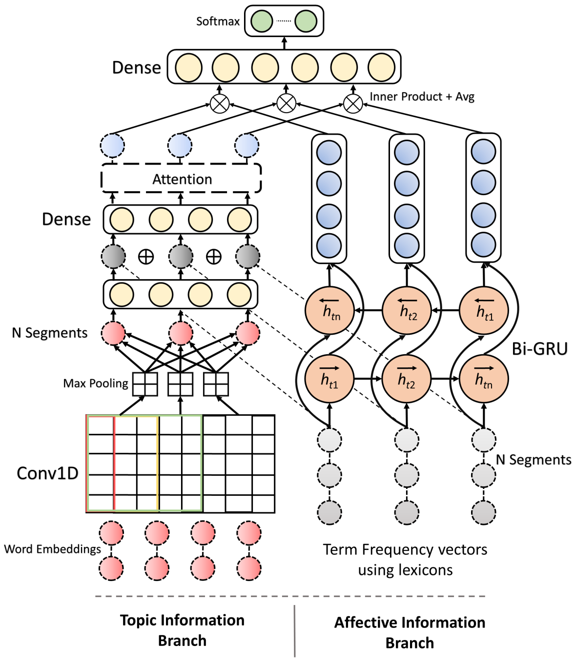


Figure 1: The architecture of the FakeFlow model.

3.1 Topic-based Information

Given a segment $n \in N$ of words, the model first embeds words to vectors through an embedding matrix. Then it uses a CNN that applies convolution processes and max pooling to get an abstractive representation of the input segment. This representation highlights important words, in which the topic information of the segment is summarized. Then it applies a fully connected layer on the output segments to get a smaller representation (v_{topic}) for later concatenation with the representation of affective information:

$$v_{topic} = f(W_a \, cnn_v + b_a)$$

where W_a and b_a are the corresponding weight matrix and bias terms, and f is an activation function such as ReLU, tanh, etc.

Key to FakeFlow is its ability to capture the relevance of the affective information with respect to the topics. For this, we concatenate the topic summarized vector v_{topic} with the representation vector v_{affect}, aimed at capturing the affective information extracted from each segment (Section 3.2).

$$v_{concat} = v_{topic} \oplus v_{affect}$$

To merge the different representations and capture their joint interaction in each segment, the model processes the produced concatenated vector v_{concat} with another fully connected layer:

$$v_{fc} = f(W_c \, v_{concat} + b_c)$$

In order to create an attention-focused representation of the segments to highlight important ones and to provide the model with the ability to weight segments differently according to the similarity of neighboring segments, the model applies a context-aware self-attention mechanism (Zheng et al., 2018) on v_{fc}. This is a crucial step, as the importance of a segment at timestep t is related to the other segments since they share the same context in the news article. Moreover, applying the attention layer can help us understand which features are most relevant by showing to which words the network attends to during learning. The output of the attention layer is an attention matrix l_t with scores for each token at each timestep.

3.2 Affective Flow of Information

To model the affective information flow in the news articles, we choose the following lexical features, under the assumption that they have a different distribution across the articles' segments. We use a term frequency representation weighted by the articles' length to extract the following features from each segment n:

- *Emotions*: We use emotions as features to detect their change among articles' segments. For that we use the NRC emotions lexicon (Mohammad and Turney, 2010) that contains ∼14K words labeled using the eight Plutchik's emotions (*8 Features*).

- *Sentiment*: We extract the sentiment from the text, *positive* and *negative*, again using the NRC lexicon (Mohammad and Turney, 2010) (*2 Features*).

- *Morality*: We consider cue words from the Moral Foundations Dictionary[2] (Graham et al., 2009) where words are assigned to one (or more) of the following categories: *care, harm, fairness, unfairness (cheating), loyalty, betrayal, authority, subversion, sanctity* and *degradation* (*10 Features*).

- *Imageability*: We use a list of words rated by their degree of abstractness and imageability[3]. These words have been extracted from the MRC psycholinguistic database (Wilson, 1988) and then using a supervised learning algorithm, the

[2] https://moralfoundations.org/other-materials/

[3] https://github.com/ytsvetko/metaphor/tree/master/resources/imageability

words have been annotated by the degrees of abstractness and imageability. The list contains 4,295 and 1,156 words rated by their degree of abstractness and imageability, respectively (*2 Features*).

- *Hyperbolic*: We use a list of ∼350 hyperbolic words (Chakraborty et al., 2016), i.e., words with high positive or negative sentiment (e.g., terrifying, breathtakingly, soul-stirring, etc.). The authors extracted these eye-catching words from clickbaits news headlines (*1 Feature*).

To model the flow of the above features, we represent each segment of an article by a vector v_{affect} capturing all 23 features listed above. Then we feed the document's vectors to a Bi-GRU network to summarize the contextual flow of the features from both directions[4] to obtain v_{flow}.

Given the segments' flow representation (v_{flow}) of an article and their relevance to the topics (l_t), FakeFlow applies a dot product operation and then averages the output matrix across the segments to get a compact representation $v_{compact}$, which is then fed into a fully connected layer:

$$v_{final} = f(W_d\, v_{compact} + b_d)$$

Finally, to generate the overall factuality label of an article, a softmax layer is applied to the output of the fully connected layer.

4 Fake News Datasets

Despite the recent efforts for debunking online fake news, there is a dearth of publicly available datasets. Most of the available datasets are small in size (e.g., the Politifact[5] dataset in (Shu et al., 2018) has ∼700 available articles, the Celebrity dataset in (Pérez-Rosas et al., 2018) has ∼500 articles, etc.), their test parts have not been manually annotated, or have been collected from a very small number of news sources. Nonetheless, we evaluate FakeFlow on three different available datasets to demonstrate its performance. In addition, we create our own dataset. Table 1 gives an overview of the datasets that we used in our work.

MultiSourceFake: We rely on different resources for creating the training and test portions of the dataset, so as to provide a challenging benchmark.

For the training part, we use *OpenSources.co* (OS), *MediaBiasFactCheck.com* (MBFC), and PolitiFact[6] news websites' lists. OS list contains 560 domains, MBFC list has 548 domains, and the PolitiFact list has 227 domains. These lists have been annotated by professional journalists. The lists contain domains of online news websites annotated based on the content type (as in the OS news list: *satire, reliable*, etc.; and in the PolitiFact news list: *imposter, parody, fake news*, etc.) or from a factuality perspective (as in the MBFC news list: low, medium, and high factuality). From the OS list, we select domains that are in one of the following categories: *fake, bias, reliable, hate, satire,* or *conspiracy*. We consider domains under the *reliable* category as real news sources, and the rest as fake. The PolitiFact list is different from the OS list since it has only labels for domains that are either fake or with mixed content. We discard the mixed ones[7] and map the remaining ones to the fake news label. Finally, we select from the MBFC list those domains that are annotated either as high or low factual news and we map them to real and fake labels, respectively. Out of these three final lists, we select only those domains for our dataset that are annotated in all lists in a consistent way; for example, we discard those domains that are annotated as real in the OS list but their label in the MBFC list is fake (low factuality). The final list contains 85 news websites. We now proceed by projecting the domain-level ground truth onto the content of those domains and randomly sample articles, with a maximum of 100 news articles per domain.[8]

For the test part, we use the *leadstories.com* fact checking website for which professional journalists annotated online news articles on the article level as fake or real. We do not follow the way we annotate the training part since the projection of the domain-level ground truth inevitably introduces noise. The journalists that annotated *leadstories.com* assigned a set of labels to the fake news articles like, e.g., *false, no evidence, satire, misleading*, etc.; we map them all to the *fake* label. In addition, we discard all articles that are multimedia-based. After collecting the news articles, we postprocess them by discarding very short articles (less than 30 words). The test part includes 689 fake news articles. We complement the set with a sample of 1,000 real

[4] During prototyping, GRU produced better overall results than LSTM.

[5] https://www.politifact.com/

[6] https://www.politifact.com/article/2017/apr/20/politifacts-guide-fake-news-websites-and-what-they/

[7] The discarded label is "Some fake stories".

[8] Some of the websites included less than 100 news articles.

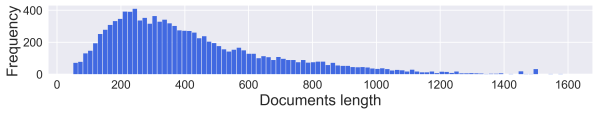


Figure 2: The distribution of the documents' length in the MultiSourceFake dataset.

Name	Total	Training	Test
MultiSourceFake	11,397	9,708	1,689
TruthShades	23,000	16,000	4,000 - 3,000
PoliticalNews	14,240	11,392	2,848
FakeNewsNet	20,208	16,156	4,039

Table 1: Number of articles in the datasets.

news articles from the training part. The overall dataset consists of 5,994 real and 5,403 fake news articles. The average document length (number of words) in the MultiSourceFake dataset is 422 words, and the 95th percentile value is 942. Figure 2 shows the distribution of the documents' length in the dataset.

TruthShades: This dataset has been proposed in Rashkin et al. (2017). The dataset was crawled from a set of domains that are annotated by professional journalists as either *propaganda, hoax, satire*, or *real*. The dataset has been built from the English Gigaword corpus for real news, and other seven unreliable domains that annotated in one of the three previous false information labels.

PoliticalNews: Due to the fact that: "a classifier trained using content from articles published at a given time is likely to become ineffective in the future" (Castelo et al., 2019), the authors of this work collected a dataset by crawling news websites in between the years 2013 to 2018 in order to evaluate their model's performance on different years.

FakeNewsNet: is a fake news repository that consists of two comprehensive datasets, one collected using claims from PolitiFact and the other from the GossipCop fact checking website. Given the large number of true and false claims from these two fact checking websites, Shu et al. (2018) built news datasets that contain visual and textual news articles content and social media information by searching Twitter for users who shared news. Out of the whole collected information, we use only the textual information of news articles, which is the part we are interested in.

5 Experiments

Experimental setup. We split the articles' text into N segments and set the maximum length of segments to 800 words, applying zero padding to the ones shorter than 800 words. Concerning the FakeFlow hyper-parameters, we tune various parameters (*dropout, the size of the dense layers, activation functions, CNN filter sizes and their numbers, pooling size, size of the GRU layer, and the optimization function*) (see Appendix A for the search space) using early stopping on the validation set. In addition to these hyper-parameters, we also use the validation set to pick the best number of segments (N). Regarding the MultiSourceFake dataset, we use 20% of the training part for validation. We represent words using pre-trained word2vec *Google-News-300* embeddings[9]. For evaluation, we follow the setup from related work. We report accuracy and weighted precision, recall and F1 score, and macro F1 for some datasets where the classes are imbalanced.

Baselines. To evaluate the performance of our model, we use a combination of fake news detection models and deep neural network architectures:

- **CNN, LSTM**: We use CNN and LSTM models and validate their performance when treating each document as one fragment. We experiment with different hyper-parameters and report results for the ones that performed best on the validation set.
- **HAN**: The authors of (Yang et al., 2016) proposed a Hierarchical Attention Networks (HAN) model for long document classification. The proposed model consists of two levels of attention mechanisms, i.e., word and sentence attention. The model splits each document into sentences and learns sentence representations from words.
- **BERT**: is a text representation model that showed superior performance on multiple natural language processing (NLP) benchmarks (Devlin et al., 2019). We use the pre-trained *bert-base-uncased* version which has 12-layers and yields output embeddings with a dimension of size 768. We feed the hidden representation of the special [CLS] token, that BERT uses to summarize the full input sentence, to a softmax layer. Experimentally, we found that fine-tuning BERT

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

layers gives a higher performance. It is worth mentioning that BERT input length is limited to 512 word pieces (sub-words level) (Devlin et al., 2019), thus, we discard the rest of the text in long news articles.

- **Fake News Detection Models**: We compare our model to several fake news detection models. We use Horne and Adali (2017) model, FakeNewsDetector (Pérez-Rosas et al., 2018), Rashkin et al. (2017) model, and EIN (Ghanem et al., 2020).[10]

- **Longformer**: Giving that Transformer-based models (i.e. BERT) are unable to process long sequences, we use Longformer (Beltagy et al., 2020), which is a SOTA model for long document tasks. In our experiments, we set the max sequence length to 1500 to handle documents that have more than 512 tokens in the MultiSourceFake dataset (see Figure 2). Also, we found that fine-tuning the Longformer model gives better results and a much faster convergence.

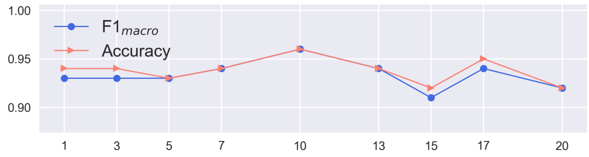


Figure 3: The accuracy and F1 results of the FakeFlow model using different N (number of segments).

Model	Acc.	Prec.	Rec.	F1$_{macro}$
Majority Class	0.59	0.35	0.59	0.37
Horne and Adali (2017)	0.80	0.75	0.78	0.80
FakeNewsDetector	0.86	0.86	0.86	0.86
LSTM	0.91	0.86	0.91	0.90
CNN	0.91	0.89	0.89	0.91
Rashkin et al. (2017)	0.92	0.92	0.92	0.92
BERT	0.93	0.93	0.94	0.93‡
EIN	0.93	0.94	0.93	0.93‡
HAN	0.94	0.94	0.94	0.93‡
Longformer	0.97	0.97	0.97	**0.97**†
FakeFlow	0.96	0.93	0.97	0.96
FakeFlow – Topic only	0.91	0.89	0.90	0.90
FakeFlow – Affective only	0.61	0.38	0.60	0.40

Table 2: Results on the MultiSourceFake dataset. (‡) indicates a statistically significant improvement of *FakeFlow* over the referred model using McNemar test; (†) indicates no statistically significant improvement over *FakeFlow*.

6 Results and Analysis

Table 2 presents the results of our proposed model and the baselines on the MultiSourceFake dataset. Our best result was achieved by using 10 as the number of segments (N, as found on the validation data). In Figure 3 we show the model's performance for segments of different length.[11] In general, the results show that models that are based on either word ngrams or word embeddings are performing better than other models that use handcrafted features, e.g. Horne and Adali (2017). Also, despite the huge amount of data used to train the BERT model, the results show that BERT performs worse than *FakeFlow* and also fails to outperform some of the other models. We speculate that this is due to the fact that the input length in BERT is limited to 512 words, as we mentioned previously, and a large portion of the news articles in the MultiSourceFake dataset has a length greater than 512 words. The results of the Longformer model confirm our claim regarding the documents' length and show a significantly higher F1 score than the BERT model. This emphasizes that despite the strong performance of BERT on multiple NLP benchmarks, it is unable to handle long text documents, in contrast, e.g., to vanilla text categorization (Adhikari et al., 2019). In addition, Longformer's results show a higher F1 score than the FakeFlow model, yet, the difference is statically insignificant.

To isolate the contribution of topical vs. affective information we run two simplified versions of our architecture, each consisting of the networks to capture topical and affective information only. The results show that the flow of the affect information has a weak performance when used alone; this emphasizes that affective information of a news article is a meaningful, yet complementary source of information.

Performance on Multiple Datasets. In Table 3 we compare the performance of the FakeFlow model to SOTA results on the other datasets we introduced in Section 4. The TruthShades dataset has two test sets, in-domain and out-of-domain. In the in-domain configuration, training and test articles come from the same sources, and from different sources in out-of-domain configuration. The results demonstrate that FakeFlow achieves a better

[10] We only compare TopicAgnostic on the dataset the authors proposed (PoliticalNews).

[11] In the case of N=1 in Figure 3, we set the maximum segment length to 1500 words instead of 800 to not lose parts of the longer articles.

TruthShades	Acc.	Prec.	Rec.	$F1_{macro}$
Out-of-domain				
Rashkin et al. (2017)	0.67	0.70	0.67	0.65
FakeFlow	0.68	0.69	0.68	**0.68**
In-domain				
Rashkin et al. (2017)	0.91	0.91	0.91	0.91
FakeFlow	0.96	0.96	0.96	**0.96**

PoliticalNews	Acc.	Prec.	Rec.	$F1_{weighted}$
TopicAgnostic	0.87	0.87	0.87	0.87
FakeFlow	0.88	0.88	0.88	**0.88**

FakeNewsNet	Acc.	Prec.	Rec.	$F1_{weighted}$
FakeNewsTracker	0.80	0.82	0.75	0.79
One-Hot LR	0.82	0.90	0.72	0.80
FakeFlow	0.86	0.86	0.86	**0.85**

Table 3: Results on multiple datasets. We compare the FakeFlow model to SOTA models on each dataset.

F1 on both test sets. In a similar way, the results on the PoliticalNews dataset show that FakeFlow also outperforms the TopicAgnostic model, although the gap in results is not very large. Finally, regarding the FakeNewsNet dataset, it looks that the deep learning-based model (FakeNewsTracker) does not achieve a good performance comparing to the other proposed baseline by the authors, which is a Logistic Regression (LR) classifier with one-hot vectors of the news articles' text. Furthermore, it seems that a simple word-based model works better than a more sophisticated model that incorporates social media and context information. The FakeFlow model, on the other hand, achieves a better result, outperforming both the FakeNewsTracker and the LR baseline.

Topic-Aware Model. Constantly, new events are covered by news agencies. These events are different from the old ones in terms of discourse and topic. Therefore, a fake news detector trained on news articles from years back is unable to detect recent news. In this experiment, we are evaluating our approach on the PoliticalNews dataset that is constructed from news distributed across different years (2013 to 2018). Following the experimental setup in (Castelo et al., 2019), we train the FakeFlow model on news from one year and test on the other years, one year at a time for testing. For example, we train the model on news from 2013 and we test on news from 2015. Note that each test set is associated with 5 results, one for each year. Figure 4 shows the average accuracy for each test set. We compare FakeFlow to the TopicAgnostic model that proved to be effective at detecting fake news from different years. It is worth mentioning that the features of the TopicAgnostic model have been extracted from both headlines and text of the news articles. However, the results show that both models have a similar performance, except for the 2013 test set where FakeFlow achieves a higher accuracy with a difference of 7%. The experiment shows that FakeFlow is capable of detecting fake news from different years, with a flat performance across the years.

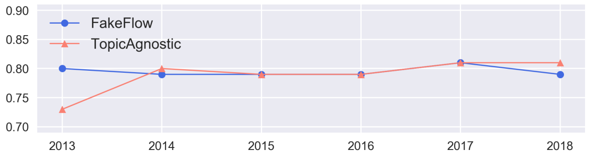


Figure 4: Topic aware experiment's results.

Attention Weights. The proposed FakeFlow model shows that taking into account the flow of affective information in fake news is an important perspective for fake news detection. We argue that being able to better understand the behaviour of the model can make it more transparent to the end-users. Figure 5 illustrates this by showing the attention weights of a fake news article across the 10 segments (left bar).[12] The figure shows that FakeFlow attends more to the beginning of the article. For better understanding, we match the affective information with the attention weights. Regarding the news text in the figure, the *emotions* features[13] show a clear example of how fake news articles try to manipulate the reader. It looks as if the existence of *fear, sadness*, and *surprise emotions* at the beginning of the article have triggered the attention on this part. Towards the end of the article, on the other hand, we can notice that such negative emotions do not exist, while *emotions* like *joy* and *anticipation* appear. This exemplifies how fake news try to attract the readers' attention in the first part of the text. Regarding the *morality* features, we only match the word "kill" with the *harm* category. Also, for the *hyperbolic* feature, we match the words "terrifying" and "powerful". In the same manner, both *morality* and *hyperbolic* features match words that occur at the beginning of the article. Lastly, for both *sentiment* and *im-*

[12] We averaged the attention weight matrix along the timesteps (number of segments) representations.

[13] Words with multiple colors mean that they have been annotated with multiple emotion types in the NRC lexicon.

ageability features, we are not able to find a clear interpretation in this example where many words across the segments match.

Real vs. Fake Analysis. In Table 4 we present an analysis on both real and fake news articles. The analysis gives an intuition to the reader on the distribution of the used features across the articles' segments. It shows that an emotion like *fear* has on average a higher difference between the first and the last segment in fake news than in real ones (see Figure 6 for a visualized distribution). Also, a feature like *hyperbolic* has a higher average value and lower standard deviation across all segments for fake news than real news, thus indicating that fake news have a higher amount of hyperbolic words with similarly high values.

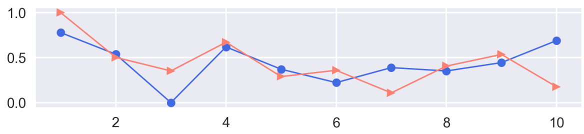


Figure 6: The flow of the *Fear* emotion in **fake** (►) and **real** (●) news articles in the MultiSourceFake dataset. Y-axis presents the average number of *Fear* emotion words in 0-1 scale; the X-axis presents the document text, divided into 10 segments.

7 Conclusion

In this paper we presented FakeFlow, a model that takes into account the flow of affective information (*emotions, sentiment, hyperbolic words*, etc.) in texts to better detect fake news articles. The model receives as input a text, segmented into smaller units, instead of processing one long fragment. This enables it to learn the flow of affective information by modeling the interaction between the topic and affective terms in the news article. We evaluated our model on four different datasets and compared it to several strong baselines. The extensive experiments show the effectiveness of FakeFlow over state-of-the-art models. Although FakeFlow was trained using a limited amount of text, the results demonstrated that it achieves results on-par with resource-hungry models (e.g. BERT and Longformer). In future work, we plan to extend our dataset and study more fine-grained news types, e.g. propaganda, from an emotional perspective. Moreover, we plan to investigate how we can replace the lexicon-based information with language-independent approaches in an attempt to make our model multilingual.

Acknowledgment

The first author would like to thank Ines Rehbein and Ana Uban for their valuable comments and suggestions. The work of the third author was partially funded by the Spanish MICINN under the research project MISMIS-FAKEnHATE on MISinformation and MIScommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31) and by the Generalitat Valenciana under the research project DeepPattern (PROMETEO/2019/121).

References

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. Docbert: Bert for document classification. *arXiv preprint arXiv:1904.08398*.

Ahmet Aker, Leon Derczynski, and Kalina Bontcheva. 2017. Simple open stance classification for rumour analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 31–39.

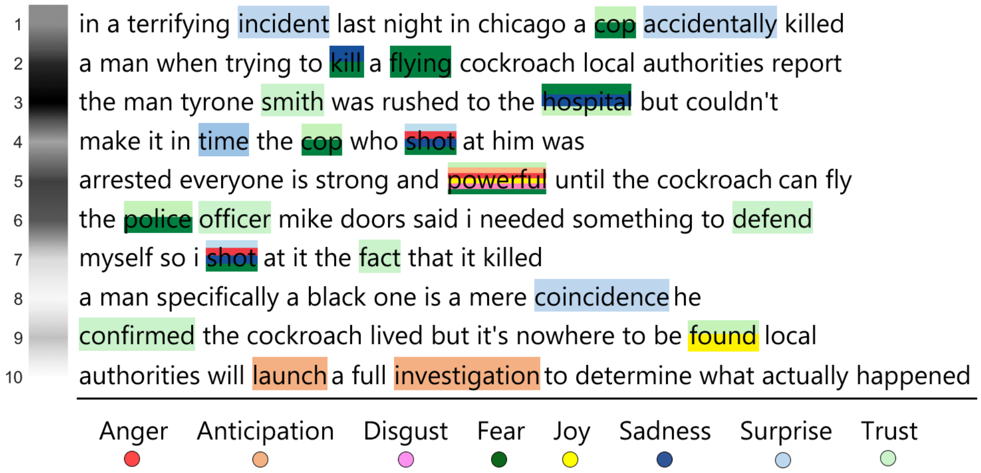
Alberto Barrón-Cedeno, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. 2019. Proppy: A system to unmask propaganda in online news. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9847–9848.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The Long-document Transformer. *arXiv preprint arXiv:2004.05150*.

Sonia Castelo, Thais Almeida, Anas Elghafari, Aécio Santos, Kien Pham, Eduardo Nakamura, and Juliana Freire. 2019. A Topic-Agnostic Approach for Identifying Fake News Pages. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 975–980.

Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. 2016. Stop Clickbait: Detecting and Preventing Clickbaits in Online News Media. In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*, pages 9–16.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186.



in a terrifying incident last night in chicago a cop accidentally killed
a man when trying to kill a flying cockroach local authorities report
the man tyrone smith was rushed to the hospital but couldn't
make it in time the cop who shot at him was
arrested everyone is strong and powerful until the cockroach can fly
the police officer mike doors said i needed something to defend
myself so i shot at it the fact that it killed
a man specifically a black one is a mere coincidence he
confirmed the cockroach lived but it's nowhere to be found local
authorities will launch a full investigation to determine what actually happened

Figure 5: Emotional interpretation of a *fake* news article by showing the attention weights (the bar on the left) and highlighting the emotions in the text.

	Features	Real News				Fake News			
		$\mu\ first_{seg.}$	$\mu\ last_{seg.}$	$\mu\ all_{seg.}$	$\sigma\ all_{seg.}$	$\mu\ first_{seg.}$	$\mu\ last_{seg.}$	$\mu\ all_{seg.}$	$\sigma\ all_{seg.}$
Emotions	Anger	0.175	0.167	0.170	0.003	0.183	0.170	0.171	0.008
	Anticipation	0.301	0.315	0.264	0.025	0.293	0.305	0.260	0.022
	Disgust	0.095	0.101	0.095	0.004	0.096	0.091	0.091	0.007
	Fear	0.254	0.250	0.238	0.010	0.265	0.226	0.238	0.011
	Joy	0.217	0.226	0.183	0.021	0.207	0.203	0.175	0.020
	Sadness	0.161	0.158	0.160	0.006	0.155	0.155	0.158	0.007
	Surprise	0.140	0.144	0.123	0.012	0.142	0.123	0.120	0.008
	Trust	0.446	0.466	0.400	0.031	0.461	0.421	0.401	0.029
Senti.	Positive	0.599	0.623	0.558	0.030	0.608	0.591	0.554	0.032
	Negative	0.369	0.337	0.347	0.011	0.367	0.336	0.350	0.013
Morality	Harm	0.007	0.011	0.007	0.002	0.008	0.013	0.007	0.002
	Care	0.026	0.023	0.019	0.004	0.021	0.022	0.019	0.003
	Fairness	0.003	0.013	0.007	0.002	0.005	0.020	0.009	0.004
	Unfairness	0.000	0.000	0.001	0.000	0.001	0.000	0.001	0.001
	Loyalty	0.016	0.017	0.019	0.002	0.014	0.016	0.019	0.003
	Betrayal	0.004	0.003	0.005	0.001	0.002	0.003	0.004	0.001
	Authority	0.025	0.032	0.026	0.003	0.024	0.028	0.026	0.002
	Subversion	0.005	0.004	0.004	0.001	0.006	0.007	0.005	0.002
	Sanctity	0.005	0.005	0.004	0.001	0.005	0.006	0.005	0.002
	Degradation	0.003	0.004	0.003	0.001	0.006	0.004	0.003	0.001
Img	Imageability	0.845	1.203	1.144	0.122	0.877	1.184	1.145	0.124
	Abstraction	0.424	0.331	0.352	0.028	0.382	0.304	0.342	0.037
	Hyperbolic	0.042	0.05	0.045	0.005	0.046	0.044	0.047	0.003

Table 4: A quantitative analysis of the features existence across articles' segments. We present the average value in the first segment ($\mu\ first_{seg.}$), the average value in the last segment ($\mu\ last_{seg.}$), the average value in the all 10 segments ($\mu\ all_{seg.}$), and the standard deviation ($\sigma\ all_{seg.}$) of a feature across the 10 segments, both in real and fake news.

Bilal Ghanem, Alessandra Teresa Cignarella, Cristina Bosco, Paolo Rosso, and Francisco Manuel Rangel Pardo. 2019. UPV-28-UNITO at SemEval-2019 Task 7: Exploiting Post's Nesting and Syntax Information for Rumor Stance Classification. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1125–1131.

Bilal Ghanem, Paolo Rosso, and Francisco Rangel. 2020. An emotional analysis of false information in social media and news articles. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–18.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and Conservatives Rely on Different Sets of Moral Foundations. *Journal of personality and social psychology*, 96(5):1029–1046.

Benjamin D Horne and Sibel Adali. 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 *Eleventh International AAAI Conference on Web and Social Media*, pages 759–766.

Sudipta Kar, Suraj Maharjan, and Thamar Solorio. 2018. Folksonomication: Predicting tags for movies from plot synopses using emotion flow encoded neural network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2879–2891.

Georgi Karadzhov, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. Fully automated fact checking using external sources. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 344–353.

Suraj Maharjan, Sudipta Kar, Manuel Montes-y Gómez, Fabio A Gonzalez, and Thamar Solorio. 2018. Letting emotions flow: Success prediction by modeling the flow of emotions in books. *arXiv preprint arXiv:1805.09746*.

Saif M Mohammad and Peter D Turney. 2010. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic Detection of Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 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*, pages 2931–2937.

Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1):1–12.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019a. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 395–405.

Kai Shu, Deepak Mahudeswaran, and Huan Liu. 2019b. FakeNewsTracker: a Tool for Fake News Collection, Detection, and Visualization. *Computational and Mathematical Organization Theory*, 25(1):60–71.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media. *arXiv preprint arXiv:1809.01286*.

Yla R Tausczik and James W Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of language and social psychology*, 29(1):24–54.

Michael Wilson. 1988. MRC Psycholinguistic Database: Machine-usable dictionary, version 2.00. *Behavior research methods, instruments, & computers*, 20(1):6–10.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, and Feifei Li. 2018. Opentag: Open Attribute Value Extraction from Product Profiles. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1049–1058.

Xinyi Zhou and Reza Zafarani. 2018. Fake news: A survey of research, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315*.

Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. 2019. Fact-checking meets fauxtography: Verifying claims about images. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2099–2108.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Kalina Bontcheva, and Peter Tolmie. 2015. Towards Detecting Rumours in Social Media. In *AAAI Workshop: AI for Cities*, pages 35–41.

A Appendices

A.1 Hyper-parameters

For FakeFlow hyper-parameters, we tune the following parameters with the their correspondent search space:

- Dropout: random selection in the range [0.1, 0.6],
- Dense layers: [8, 16, 32, 64, 128],
- Activation functions: [selu, relu, tanh, elu],
- CNN filters' sizes: [(2, 3, 4), (3, 4, 5), (4, 5, 6), (3, 5), (2, 4), (4,), (5,), (3, 5, 7), (3, 6)],
- Numbers of CNN filters: [4, 8, 16, 32, 64, 128],
- Pooling size: [2, 3],
- GRU units: [8, 16, 32, 64, 128],
- Optimization function: [adam, adadelta, rmsprop, sgd],
- For the early stopping, we set the 'patience' parameter to 4 and we set the epochs number to 50.

For the parameters selection, we use hyperopt[14] library that receives the above search space to randomly select different N combination of parameters (trials). We use a small value of N in all of our experiments to avoid overdrawn finetuning; we set N to 35.

A.2 Topic Aware experiments

In Figure 4, we present the average accuracy of our model when we train on different years and test a specific one. In the following we show the results before we averaged them.

Test Train	2013	2014	2015	2016	2017	2018
2013	0.00	0.82	0.74	0.76	0.78	0.74
2014	0.84	0.00	0.79	0.76	0.81	0.74
2015	0.79	0.81	0.00	0.82	0.80	0.82
2016	0.80	0.76	0.87	0.00	0.85	0.79
2017	0.79	0.82	0.76	0.80	0.00	0.85
2018	0.79	0.75	0.81	0.83	0.83	0.00
Average	0.80	0.79	0.79	0.79	0.81	0.79

Table 5: FakeFlow results for each train-test run for the Topic-Aware experiment.

[14] https://github.com/hyperopt/hyperopt