Emotions and News Structure: An Analysis of the Language of Fake News in Spanish

Benedetta Togni[1], Mariona Coll Ardanuy[1], Berta Chulvi[2,3] and Paolo Rosso[1,4]

[1]*PRHLT Research Center, Universitat Politècnica de València, Spain*
[2]*Symanto Research, Spain*
[3]*Social Psychology Department, Universitat de València, Spain*
[4]*Valencian Graduate School and Research Network of Artificial Intelligence (ValgrAI), Spain*

Abstract

Research has repeatedly demonstrated that fake news tend to appeal to the emotions, but it is less common to consider the presence of emotions in relation to the inverted pyramid, a key aspect of journalistic writing. In this work, we adopt an existing neural model for fake news detection, FakeFlow, one of the few approaches to consider the salience of emotions with respect to the structure of news articles, and adapt it for Spanish. We conduct our experiments on the Spanish Fake News Corpus, introduced at the Fake News Detection in Spanish shared task (FakeDeS), with the goal of gaining a better understanding of the characteristics underlying such texts. In our analyses, we show that both the distribution of affective features and the attention mechanism of the model validate the importance of considering the inverted pyramid structure for detecting fake news.

Keywords

Fake news detection, inverted pyramid, emotion analysis

1. Introduction

In an increasingly connected world, research on detecting fake news is more necessary than ever [1, 2, 3]. For years, researchers from different disciplines (such as sociology, NLP and network analysis) have been joining efforts to understand this phenomenon and find effective ways of addressing it. From a language perspective, to date most research has focused on English data. Even for Spanish, one of the most-spoken languages in the world, detection of fake news is a much under-researched topic, despite its becoming an increasingly important social concern.[1] The Fake News Detection in Spanish shared tasks (FakeDeS), held at the IberLEF 2020 and 2021 workshops, are amongst the most notable efforts towards addressing this problem from a NLP perspective. The organisers introduced a new dataset, the Spanish Fake News Corpus (henceforth *SFNC*) [4, 5, 6], to date the only dataset in Spanish for the task of fake news detection. The best-performing approach in the competition achieved an $F_1 score$ of 0.76, leaving plenty of room (and need) for improvement.

The aim of this paper is to gain a better understanding of the challenges of detecting fake news, from a linguistic perspective, and with a focus on Spanish. The goal of our research is to shed some light on this phenomenon and to obtain insights which may help researchers to informatively build better classifiers. In particular, we look at the presence of affective content in the language of fake news. In doing so, we take into consideration a distinctive feature of journalistic writing: the inverted pyramid structure, according to which the most important information is presented at the beginning, with less essential details following in descending order of importance. This style allows readers to grasp the essential elements of a story quickly, even if they only read the first few sentences [7].

In text classification tasks, the information structure of news articles is not often accounted for, even though some research exists in this direction. In fact, the best-performing approach at the FakeDeS 2021 competition [8] used BERT [9] to encode the first and last 512 tokens of the news articles, and concatenated the two embeddings together with an additional memory embedding intended to capture the relationship between samples. The approach was built on the assumption that the middle part of a document is the least informative. This assumption was to some extent demonstrated on two datasets in [10], who showed that removing the middle part of a document was a successful strategy to truncate

SEPLN-2024: 40th Conference of the Spanish Society for Natural Language Processing. Valladolid, Spain. 24-27 September 2024.

btogni1@upvnet.upv.es (B. Togni); mcoll@prhlt.upv.es (M. Coll Ardanuy); berta.chulvi@symanto.com (B. Chulvi); prosso@dsic.upv.es (P. Rosso)
0000-0001-8455-7196 (M. Coll Ardanuy); 0000-0003-1169-0978 (B. Chulvi); 0000-0002-8922-1242 (P. Rosso)

CEUR Workshop Proceedings (CEUR-WS.org)

[1] In May 2024, the Spanish Centre for Sociological Research (CIS) conducted a survey which revealed that 3.3% of the respondents considered the role of the media and social networks (misinformation, information manipulation, dissemination of hoaxes) as one of the three main problems in the country. See: https://www.cis.es/catalogo-estudios/resultados-definidos/barometros, accessed on May 28, 2024.

long articles when fine-tuning a transformer model for text classification. However, only one of the two datasets consisted of news articles, and full details of these experiments were not provided.

Research has repeatedly demonstrated that fake news tend to appeal to the emotions [11, 2], especially negative emotions [12], with the intent of triggering a specific reaction from the reader [13]. Several approaches have taken this information into account in their models [14]. One such approach is FakeFlow [15], a neural model which was developed with the specific goal of learning the flow of semantic and affective information in news articles, and using this information for detecting fake news. In this paper, we adapt FakeFlow for Spanish and use its inherently interpretable capabilities for analysing the distinction between fake and true news, focusing on the SFNC dataset. Unlike in English, the FakeFlow-based classifier for Spanish underperforms when compared with a RoBERTa-based classifier, but nevertheless offers useful insights on the data. We extensively analyse our results both in relation to the affective information and the structure of the news articles, and show why a generalisable solution to fake news detection is hard to achieve.

The rest of the paper is structured as follows: we describe the FakeFlow approach and how we have adapted it for Spanish in Section 2; we describe the classification experiments in Section 3; we analyse and discuss our findings in Section 4; and provide a conclusion and directions for further research in Section 5.[2]

2. Approach

FakeFlow [15] was developed with the specific goal of learning the 'flow'—i.e. the sequence of information salience—of semantic and affective features in news articles. As shown in Figure 1, FakeFlow consists of two communicating branches: on the one hand, the ***topic information*** branch is built upon word embedding representations of the news segments, and is responsible for capturing the semantics of the texts; on the other hand, the ***affective information*** branch is built upon vectors of lexical features which represent the affective aspect of the texts. In order to model the flow of information, each news article is first split in N segments. The neural architecture of the model is designed to capture the joint interaction between both the topic and affective information in each segment, including a self-attention layer which was added with the purpose of highlighting the importance of a segment in relation to the neighbouring segments in the news article. Finally, the two branches are combined and, for a given text as input, a probability over the two classes (*fake* and *true*) is returned.[3]

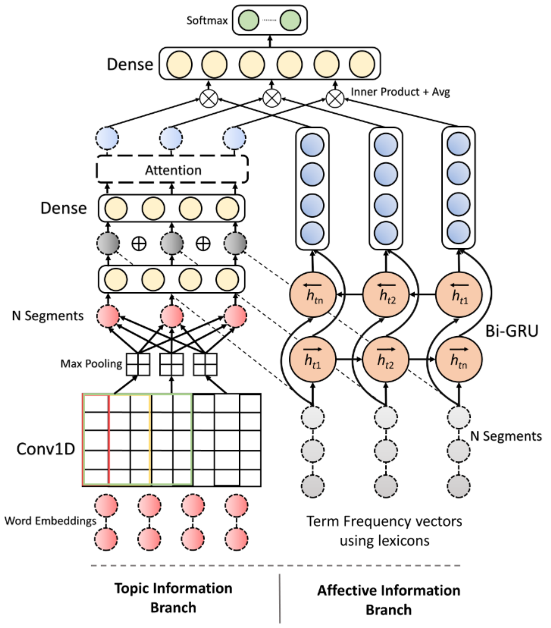


Figure 1: Architecture of the FakeFlow model. Source: Ghanem et al. (2021) [15].

As input for the *topic information* branch, we have used the openly available word2vec [16] embeddings trained on the Spanish Billion Words Corpus[4] [17]. The *affective information* branch requires that each segment is represented as a vector of lexical features. These vectors are built from term frequency representations (based on lexical resources), which are then normalised by the length of the article. In order to adapt FakeFlow for Spanish, we looked for lexical resources in Spanish to capture the affective dimension of the news articles. We have extracted the following features for each news segment:

- **Emotion features:** As in the original paper, we have used the NRC Emotion Lexicon [18, 19],[5] which associates words with eight emotions: anger, fear, anticipation, trust, surprise, sadness, joy, and disgust. After processing the lexicon, the Spanish lexicon consists of 11,326 words.[6] Each

[2] The code for our experiments is available at https://github.com/benedettatogni/FakeFlowSpanish.

[3] Since the model architecture of FakeFlow is not a contribution of this paper, we only provide an overview of its architecture, sufficient for the reader to understand the main idea. For further information, please refer to the original paper [15] or its code repository, openly available at https://github.com/bilalghanem/fake_flow.

[4] https://crscardellino.github.io/SBWCE/.

[5] https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm.

[6] We have followed the same steps to process all the resources: removal of multi-token words, lower-casing, removal of duplicates and stop words, and lemmatisation using the SpaCy library [20].

emotion corresponds to a feature in our model (*8 features*).

- **Sentiment features:** The same NRC Emotion Lexicon also contains associations between words (10,898 after processing) and sentiment polarity: positive and negative. Each is assigned a feature in the model (*2 features*).
- **Hurtful language feature:** The original FakeFlow paper used the Moral Foundations Dictionary [21] to extract a set of morality features. An examination of the content showed that this resource is strongly rooted in American culture. Given that some of the categories in the dictionary are related to the willingness or unwillingness to hurt others, we decided to use *HurtLex* instead, a validated lexicon of offensive, aggressive, and hateful words[7] [22], which is more fitting to the Spanish-speaking context. After processing, the lexicon of hurtful words consists of 2,008 words (*1 feature*).
- **Hyperbolic language feature:** In the original approach, a set of hyperbolic words was manually extracted from clickbait news headlines. We have translated these words into Spanish and have removed uncommon words, obtaining a new list of 292 words that are heavily loaded with a positive or negative sentiment, such as *'abrumador'* and *'sorprendente'* (*1 feature*).
- **Affective-semantic features:** The original model uses the MRC psycholinguistic database [23] to characterise words in terms of their degree of abstractness and imageability. We have replaced it with an equivalent resource in Spanish [24], a dataset that not only provides the semantic information related to the degree of abstraction but also characterises the intensity of the emotion and its valence. After processing, this dictionary consists of 1,400 words, each score normalised between 0 and 1 in terms of their conveyance of valence, arousal, concreteness, and imageability (*4 features*).

To summarise, each segment in an article is represented as a vector of these *16 features*, which are used as input of the affective information branch of the model. Since Spanish is a significantly more inflected language than English, we lemmatised the article segments to maximise the association between words and features, and only kept nouns, verbs, adjectives, and adverbs.

3. Experiments

We describe the dataset we have used in our investigation in Section 3.1 and describe the classification experiments and results in Section 3.2.

3.1. Dataset

We have conducted our experiments on the Spanish Fake News Corpus (SFNC)[8] [6, 4]. The dataset consists of 1,543 texts, and was compiled in two stages:

- The training and development sets were collected as part of the first edition of the shared task [6, 5], and consist of 971 news articles (676 for training and 295 for development) that were collected from 134 different media websites between January and July 2018, covering the following topics: science, sports, economy, education, entertainment, politics, health, security, and society.
- The test set was collected for the second edition of the shared task [4]. It consists of 572 texts, collected between November 2020 and March 2021, in the midst of the Covid-19 pandemic. The test set consists of articles from four of the topics already covered in the training and development sets (science, sports, politics and society) and introduces three new topics: Covid-19, environment, and international. The texts come from a wide-range of sources and are written in different varieties of Spanish.

Table 1 provides a summary of the dataset. As can be seen, the dataset is heavily imbalanced in its distribution of topics. Furthermore, there is a significant temporal shift between the training and development sets (January–July 2018) and the test set (November 2020–March 2021). These two aspects make this a more challenging task than it may seem *a priori*. In our experiments, we have decided to keep these splits (instead of mixing them back together and re-splitting the dataset into more balanced subsets) because we think this is a more realistic scenario, given that, in real applications, the training set will always be 'historical' and may be topically different.

3.2. Classification Experiments

Since the task is very similar to that for which FakeFlow was originally developed, we eventually used the

[7] https://github.com/valeriobasile/hurtlex. We have used only words categorised as: negative stereotypes and ethnic slurs, physical disabilities and diversity, cognitive disabilities and diversity, moral and behavioural defects, words related to social and economic disadvantage, words related to prostitution, words related to homosexuality, words with potential negative connotations, derogatory words, felonies and words related to crime and immoral behaviour, and words related to the seven deadly sins of the Christian tradition.

[8] https://github.com/jpposadas/FakeNewsCorpusSpanish.

Table 1
Number of instances and distribution of topics and sources in the SFNC dataset.

		Train		Dev		Test	
		True	Fake	True	Fake	True	Fake
# instances		338	338	153	142	286	286
# words	mean	460	294	504	324	637	410
	std	307	155	422	167	438	423
Topics	Education	6	9	4	3	—	—
	Society	41	52	19	22	99	96
	Science	32	30	14	13	6	7
	Security	11	18	6	7	—	—
	Health	16	16	7	7	—	—
	Economy	18	12	6	7	—	—
	Sports	45	41	21	17	1	1
	Politics	121	105	54	43	53	54
	Entertainment	48	55	22	23	—	—
	Covid-19	—	—	—	—	118	119
	International	—	—	—	—	7	7
	Environment	—	—	—	—	2	2
Sources	# unique sources	69	67	46	36	99	118
	instances x source	4.9	5.04	3.33	3.94	2.89	2.42

same hyperparameters than for English:[9] while we experimented with different choices of hyperparameters, we did not consistently obtain significant improvements. We show the results of adapting FakeFlow for Spanish in Table 2. For comparison, we provide the most common class baseline (in this case, '*true*') and a RoBERTa-based baseline.[10] While the original FakeFlow in English surpassed transformer-based approaches [15], in our case we see that the RoBERTa-based classifier is by far a better performing approach, with a difference of 8 points in macro F1-score with respect to the FakeFlow approach. We also compare the FakeFlow model with two simplified versions of the approach, using only a subset of the features: *(1)* ***only emotions*** uses only the eight emotions as features; *(2)* ***only negative*** considers only negative features: 'anger', 'disgust', 'fear', 'sadness', 'negative' and 'hurtful'. Neither of the simplified versions of FakeFlow yield a better performance, in both cases providing a lower F_1score for the true class and a higher score for the fake class. In the following sections, we show how a careful inspection of the dataset, using the features and the outputs of the FakeFlow model, can shed some new light to help us interpret these results.

4. Analysis and Discussion

In this section, we analyse the differences observed between the fake and true subsets of the SFNC dataset in terms of their distributions of the features used by the FakeFlow classifier (in Section 4.1) and the attention weights produced by the classifier (in Section 4.2), and discuss our findings.

4.1. Feature Analysis

An analysis of the distribution of the features on the full dataset reveals that fake news have on average higher emotional content than true news, as is also reported in previous works [11, 15]. In Figure 2, we show the changing salience of the sixteen features over the article segments. First, we see that, on average (i.e., the dotted lines), fake news have higher values than true news, and that, both in true and fake news, emotion values are higher at the beginning of the article. This pattern is particularly noticeable on some features, such as 'anger', 'anticipation', and 'fear'. We observe that 'negative' is more prominent in fake news, whereas the signal is less clear for 'positive', in the same vein as previous work indicates [11]. The 'hurtful' feature is consistently higher in fake news, and so is 'hyperbolic', even though less prominently. Finally, the four affective-semantic features are higher in true news, except in the first segment. In the case of 'concreteness' and 'imageability', this information could be used to corroborate findings from previous research, according to which fake news tend to require a smaller cognitive effort from the reader in order to process the content [11]. Further research using additional

[9] With the exception of the learning rate, which we set to 0.0001.

[10] We have used an existing model available on the HuggingFace hub: `Narrativaai/fake-news-detection-spanish`, which was fine-tuned for text classification on the training and development sets of the SFNC dataset. It is based on the `PlanTL-GOB-ES/roberta-large-bne` model, pre-trained on the largest existing corpus in Spanish [25]. This RoBERTa-based classifier came out after the shared task, and outperforms all the approaches that participated in the task. We have formatted the input of the classifier according to the indications of the authors, by concatenating the headline to the special token '`[SEP]`' and the text.

Table 2
Results on the FakeDes dataset.

Model	P_{fake}	R_{fake}	$F1_{fake}$	$F1_{true}$	$F1_{macro}$
most-common	0.00	0.00	0.00	0.74	0.37
RoBERTa-classifier	0.81	0.72	0.76	0.79	0.78
FakeFlow	0.76	0.60	0.67	0.73	0.70
FakeFlow: only emotions	0.70	0.66	0.68	0.69	0.69
FakeFlow: only negative	0.73	0.63	0.68	0.72	0.70

lexical resources, such as [26], and on additional datasets is needed to validate these insights and to investigate these phenomena in more detail.

A breakdown of the dataset into topics reveals different distributions of features. In Figure 3, we show the distribution of the eight emotions (plus 'hurtful') on three topics: politics, entertainment, and Covid-19.[11] Here, we distinguish between the training set and the test set, to account both for the topic imbalance and also the different temporal coverage of both subsets. First, focusing on the training set (i.e., the figures on the left, built from news articles collected during the first half of 2018), we observe that 'hurtful' stands out both on true and fake news on entertainment, whereas the other emotions are, with the exception of 'surprise' and 'joy' on the fake subset, less prominent than on politics. In general, the language of political news is more emotional, the main difference between true and fake being higher 'hurtful', 'anger', and 'joy' values in the latter.

On the test set (i.e., the figures on the right, built from news articles collected between November 2020 and March 2021), we see that articles on Covid-19 have a very different distribution of features, especially patent in the case of 'fear' and 'sadness', which stand out both for true and fake news. Fake news on Covid-19 have consistently higher values than true news for all emotions and for hurtful language. Finally, it is interesting to note the evolution of the emotional content in political news from 2018 (i.e., the training set) to 2020–2021 (i.e., the test set): we see true news becoming more similar to fake news, particularly with respect to 'anger' and 'hurtful'. Given that almost three convulsed years span between the articles in both datasets, it remains to be investigated whether this difference is specific of this dataset or it is the product of the larger question of whether, in recent years, the style of mainstream news has become more similar to that of fake news, perhaps as a strategy to compete with the appeal of fake news.

4.2. Attention Analysis

The authors of the original FakeFlow paper [15] observed that the model attended more at the beginning of the articles, concluding that in English most of the information useful for discriminating between fake and true news is presented at the beginning of the news articles. As described in Section 2, the FakeFlow architecture has a self-attention mechanism which highlights the importance of a segment in relation to the other segments in the news article. We therefore extracted the matrices of attention weights from the test set and averaged them.

In Figure 4, we show the matrices of averaged attention weights of the texts that have been correctly classified as true and those correctly classified as fake. We can see that, both in fake and true news, the model attends more at the beginning of the article—as in English, according to [15]—whereas the middle part is less attended, and the last part is the least attended. This suggests that the first part of the articles is on average more useful for discriminating fake from true news, but further investigation on the role of inter-segment attention with respect to the classification decision is needed. Finally, our version of FakeFlow allows the user to inspect the different articles in the dataset, and visualise how much the classifier attended to each of the segments in the article, as illustrated in Figure 5.

In order to validate whether these findings are specific of the FakeFlow approach or more generic, we have done a text-based ablation study using the RoBERTa classifier. In Table 3, we observe how the performance of the classifier changes depending on the input that is provided: 'first part' contains only segments 1 to 3 (i.e., the first part of the articles), 'middle part' includes segments 4 to 7, and 'last part' includes segments 8 to 10. For comparison, we provide the full text (excluding the headline, hence the difference with respect to Table 2) and the full text removing either the middle or the last part. It is important to note the drop of $F_1 score$ of the true class for individual parts, whereas the $F_1 score$ of the fake class is less (or not) affected. The results show that the middle and last parts are clearly less informative, and we show that removing the last part does not affect the classification (and even improves the $F_1 score$ of the fake class). This is an interesting finding, not only from a scientific

[11] Topics were selected based on their frequency in the dataset: 'politics' is the most popular topic overall, 'entertainment' is the most popular topic in the training and development sets (besides 'politics'), and 'Covid-19' is the most popular topic in the test set.

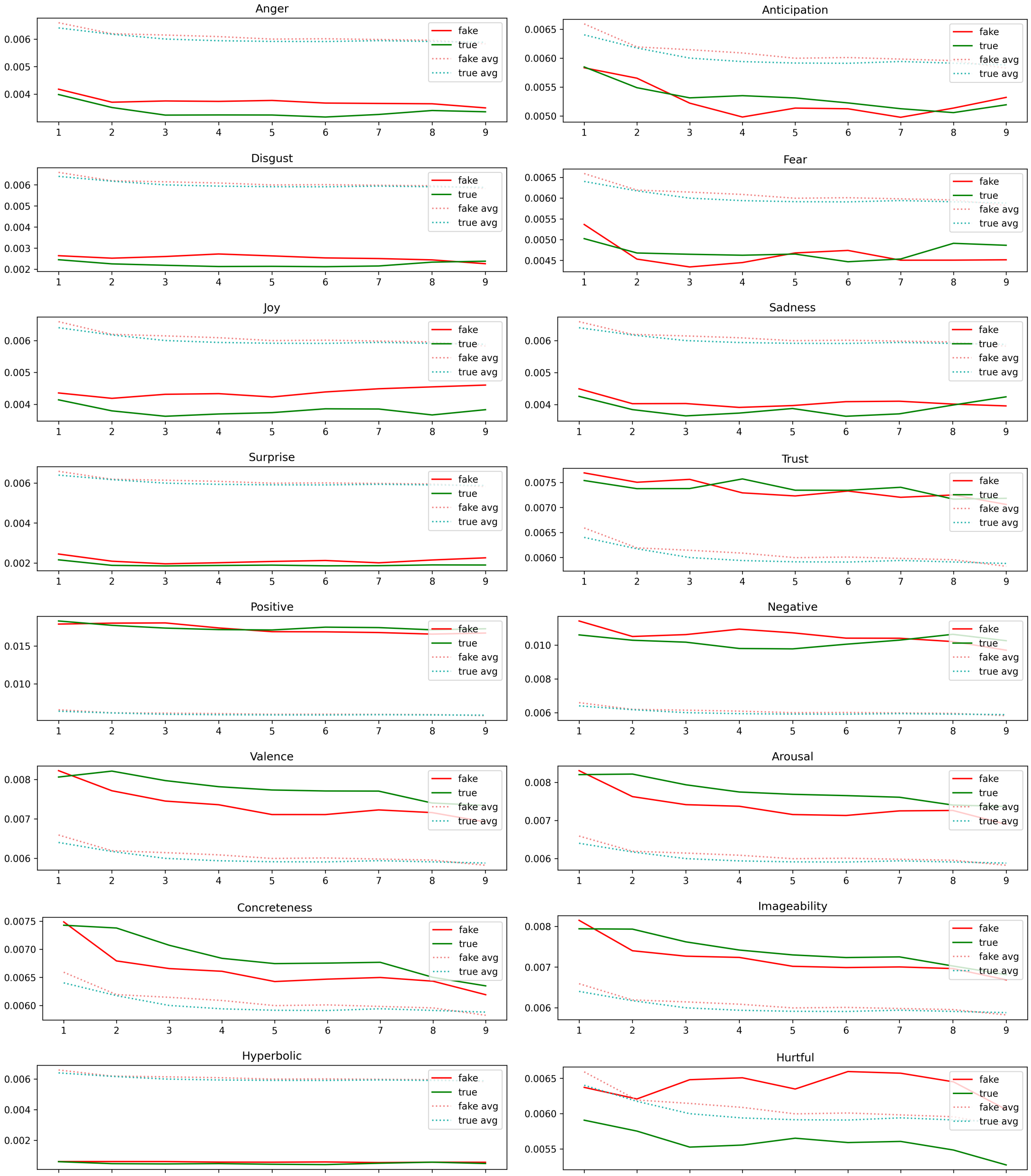
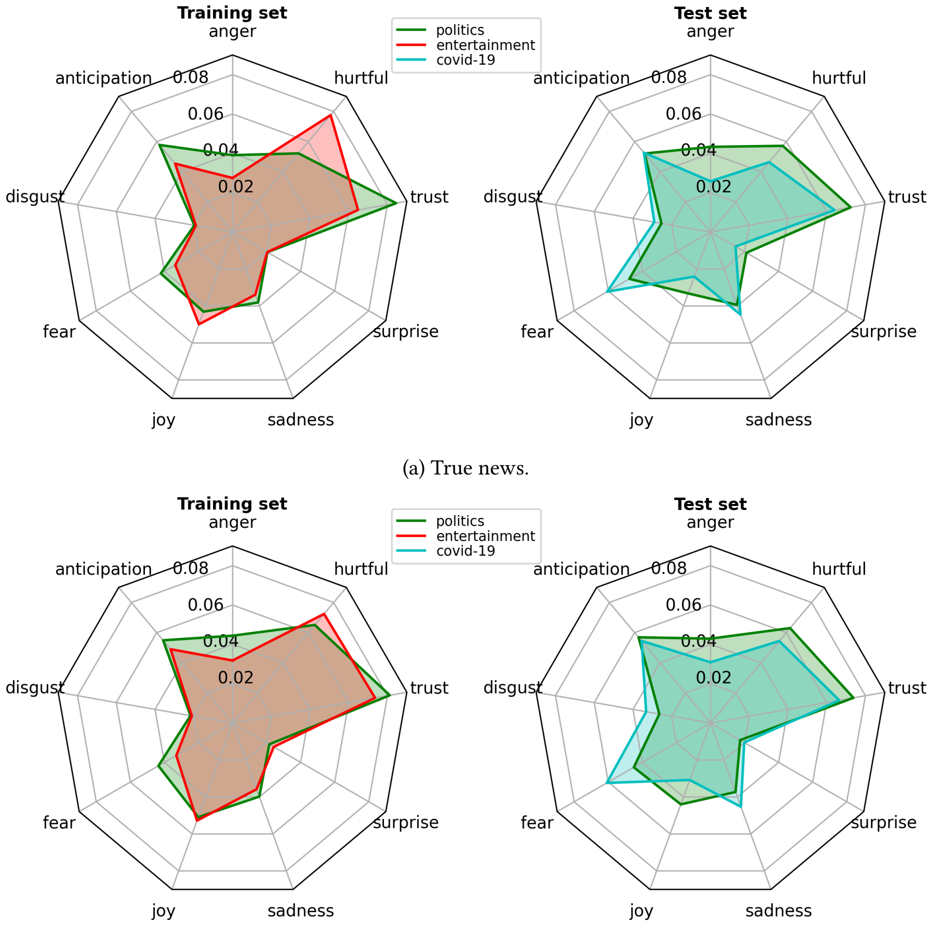


Figure 2: Information flow of the sixteen features. The Y-axis is the normalised frequency of the feature per segment, the X-axis is the sequence of segments into which the article has been split. The solid lines show feature values averaged across the dataset, the dotted lines indicate the average of all features per segment. We applied a moving average of two.

perspective (as it aligns with the well-known inverted pyramid approach to writing news articles), but also for practical application.[12]

[12] Language models usually truncate long input sequences to a maximum length. A better understanding of which parts of the article are more likely to be informative may be a safe strategy to reduce time and computational costs, especially when processing large amounts of text, since the quadratic complexity of transformers means that, the longer the text, the longer the time to process it.



(a) True news.

(b) Fake news.

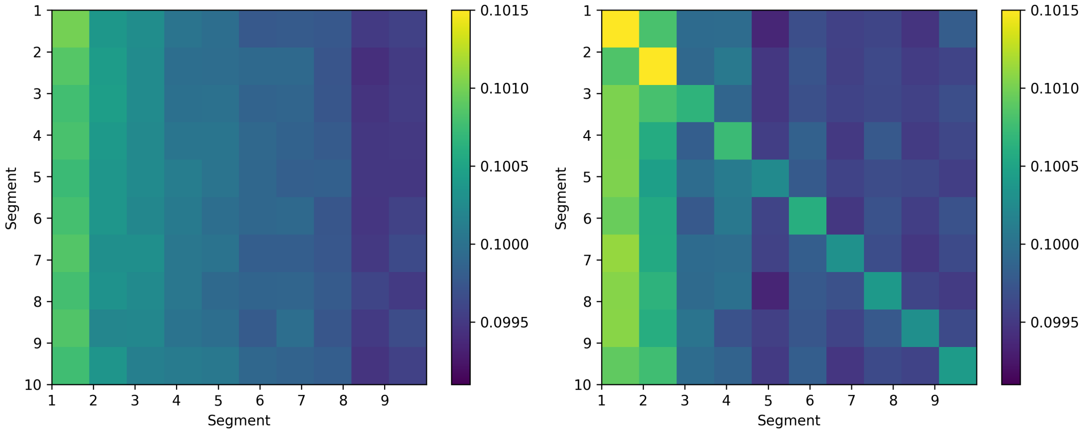
Figure 3: Radar plots of the training and test sets for *true* news and *fake* news.

Table 3
Results of the RoBERTa classifier on the FakeDes dataset. Ablation study of text parts based on insights from the FakeFlow attention matrices. In all cases, we use only the content of the news articles, without the headline.

Model	$F1_{fake}$	$F1_{true}$	$F1_{macro}$
Full text	0.75	0.77	0.76
First part	0.75	0.67	0.71
Middle part	0.71	0.60	0.66
Last part	0.74	0.59	0.67
First and middle part	0.77	0.76	0.76
First and last part	0.75	0.72	0.73

5. Conclusions

While transformer-based models are the state-of-the-art approaches for text classification, it is notorious that they suffer from interpretability issues [27], making it difficult to gain scientifically-interesting new insights on why texts are classified in one way or another. In this paper, we have used an adapted version of an inherently interpretable model, FakeFlow, to inspect the SFNC dataset with the aim of gaining a better understanding of the relation between fake news and affective language, with



(a) Fake news

(b) True news

Figure 4: Matrix of attention weights averaged across correctly classified articles. Both rows and columns represent articles segments, and the intersections are coloured according to how much, at each time-step (i.e., segment), the classifier attends to the other segments or to itself. Yellow indicates a higher attention.

a special focus on Spanish. We have shown that the beginning of the articles is more discriminative between fake and true news text than the middle and last parts. This is consistent with the well-known inverted pyramid approach to writing news articles, in which the most noteworthy information is provided at the beginning. The higher salience of some features (i.e., anger, anticipation, fear, and sadness) at the beginning of the articles could point at the way such information is presented in order to attract the attention of the readers, by appealing to specific emotions.

In NLP, convention dictates that the test set is similar to the training set, but the SFNC dataset is heavily imbalanced in its distribution of topics and displays a significant temporal shift between the training and test set (even more so considering the global impact of the Covid-19 pandemic), representing a more realistic scenario than a better-balanced dataset. An in-depth analysis of the data revealed that emotions show different distributions according to the topics of the news, but also that this distribution may change over time, hence the difficulty of finding a generalisable solution to the issue. Our paper opens new avenues for further investigating the relation between affective language and the information structure of news articles in fake news detection.

Acknowledgments

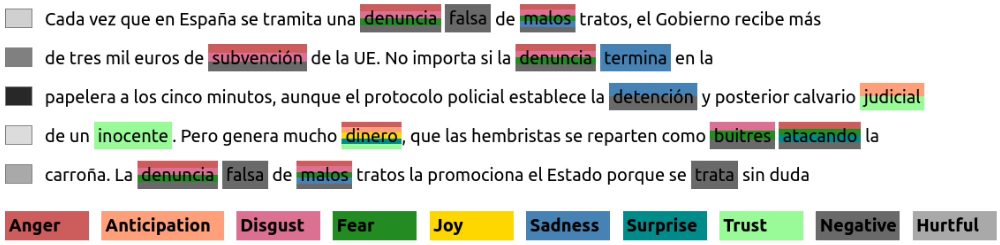
We are grateful to the reviewers for their careful and constructive reviews. We would also like to thank Damir Korenčić and Ivan Grubišić for their feedback.

This work was funded by the research project FAKEnHATE-PdC: FAKE news and HATE speech–Proof of Concept, funded by MCIN/AEI/10.13039/501100011033 and by European Union NextGenerationEU/PRTR. The work of Paolo Rosso was in the framework of the Iberifier Plus project, co-funded by the European Commission under the Call DIGITAL-2023-DEPLOY-04, European Digital Media observatory (EDMO)-National and multinational hubs (reference 101158511), and the MARTINI project (Malicious Actors Profiling and Detection in Online Social Networks Through Arificial Intelligence), grant PCI2022-135008-2, funded by MCIN/AEI/10.13039/501100011033 and by European Union NextGenerationEU/PRTR. The work of Mariona Coll was in the framework of the research project FairTransNLP on Fairness and Transparency for equitable NLP applications in social media, grant PID2021-124361OB-C31, funded by MCIN/AEI/10.13039/501100011033 and by ERDF, EU A way of making Europe. The work of Berta Chulvi was conducted while at Universitat Politècnica de València.

References

[1] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, et al., The science of fake news, Science 359 (2018) 1094–1096.

[2] G. Ruffo, A. Semeraro, A. Giachanou, P. Rosso, Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language, Computer science review 47 (2023) 100531.



Cada vez que en España se tramita una denuncia falsa de malos tratos, el Gobierno recibe más
de tres mil euros de subvención de la UE. No importa si la denuncia termina en la
papelera a los cinco minutos, aunque el protocolo policial establece la detención y posterior calvario judicial
de un inocente . Pero genera mucho dinero , que las hembristas se reparten como buitres atacando la
carroña. La denuncia falsa de malos tratos la promociona el Estado porque se trata sin duda

Figure 5: Initial segments of a news article, shown as an example, in which words are tagged according to the emotion conveyed (including 'hurtful' and 'negative'). The box on the left is coloured according to the normalised averaged attention weight of the segment, where darker means a higher attention.

[3] X. Zhou, R. Zafarani, A survey of fake news: Fundamental theories, detection methods, and opportunities, ACM Computing Surveys (CSUR) 53 (2020) 1–40.

[4] H. Gómez-Adorno, J. P. Posadas-Durán, G. B. Enguix, C. P. Capetillo, Overview of FakeDes at Iberlef 2021: Fake news detection in Spanish shared task, Procesamiento del lenguaje natural 67 (2021) 223–231.

[5] M. E. Aragón, H. J. Jarquín-Vásquez, M. Montes-y Gómez, H. J. Escalante, L. V. Pineda, H. Gómez-Adorno, J. P. Posadas-Durán, G. Bel-Enguix, Overview of MEX-A3T at IberLEF 2020: Fake news and aggressiveness analysis in Mexican Spanish., in: IberLEF@ SEPLN, 2020, pp. 222–235.

[6] J.-P. Posadas-Durán, H. Gómez-Adorno, G. Sidorov, J. J. M. Escobar, Detection of fake news in a new corpus for the Spanish language, Journal of Intelligent & Fuzzy Systems 36 (2019) 4869–4876.

[7] H. Pöttker, News and its communicative quality: the inverted pyramid–when and why did it appear?, Journalism Studies 4 (2003) 501–511.

[8] X. Huang, J. Xiong, S. Jiang, GDUF-DM at FakeDeS 2021: Spanish fake news detection with BERT and sample memory, in: CEUR Workshop proceedings: Iberian Languages Evaluation Forum, 2021, pp. 621–629.

[9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[10] C. Sun, X. Qiu, Y. Xu, X. Huang, How to fine-tune bert for text classification?, in: Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2019, p. 194–206. doi:10.1007/978-3-030-32381-3_16.

[11] C. Carrasco-Farré, The fingerprints of misinformation: how deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions, Humanities and Social Sciences Communications 9 (2022) 1–18.

[12] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, science 359 (2018) 1146–1151.

[13] V. Bakir, A. McStay, Fake news and the economy of emotions: Problems, causes, solutions, Digital journalism 6 (2018) 154–175.

[14] Z. Liu, T. Zhang, K. Yang, P. Thompson, Z. Yu, S. Ananiadou, Emotion detection for misinformation: A review, Information Fusion (2024) 102300.

[15] B. Ghanem, S. P. Ponzetto, P. Rosso, F. Rangel, Fake-Flow: Fake news detection by modeling the flow of affective information, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021, pp. 679–689.

[16] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Weinberger (Eds.), Advances in Neural Information Processing Systems, volume 26, Curran Associates, Inc., 2013.

[17] C. Cardellino, Spanish Billion Words Corpus and Embeddings, 2019. URL: https://crscardellino.github.io/SBWCE/.

[18] S. Mohammad, P. Turney, 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, Association for Computational Linguistics, Los Angeles, CA, 2010, pp. 26–34. URL: https://aclanthology.org/W10-0204.

[19] S. M. Mohammad, P. D. Turney, Crowdsourcing a word-emotion association lexicon, Computational

Intelligence 29 (2013) 436–465.

[20] M. Honnibal, I. Montani, S. V. Landeghem, A. Boyd, spacy: Industrial-strength natural language processing in python, 2020. doi:10.5281/zenodo.1212303.

[21] J. Graham, J. Haidt, B. A. Nosek, Liberals and conservatives rely on different sets of moral foundations., Journal of personality and social psychology 96 (2009) 1029.

[22] E. Bassignana, V. Basile, V. Patti, Hurtlex: A multilingual lexicon of words to hurt, in: CEUR Workshop proceedings: Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-It), volume 2253, 2018.

[23] M. Wilson, MRC psycholinguistic database: Machine-usable dictionary, version 2.00, Behavior research methods, instruments, & computers 20 (1988) 6–10.

[24] M. Guasch, P. Ferré, I. Fraga, Spanish norms for affective and lexico-semantic variables for 1,400 words, Behavior Research Methods 48 (2016) 1358–1369.

[25] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022).

[26] J. A. García-Díaz, P. J. Vivancos-Vicente, A. Almela, R. Valencia-García, Umutextstats: A linguistic feature extraction tool for spanish, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 6035–6044.

[27] A. Madsen, S. Reddy, S. Chandar, Post-hoc interpretability for neural nlp: A survey, ACM Computing Surveys 55 (2022) 1–42.