Fake News as We Feel It: Perception and Conceptualization of the Term "Fake News" in the Media

Evandro Cunha[1,2](✉), Gabriel Magno[1], Josemar Caetano[1], Douglas Teixeira[1], and Virgilio Almeida[1,3]

[1] Department of Computer Science, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil
{evandrocunha,magno,josemarcaetano,douglas,virgilio}@dcc.ufmg.br
[2] Leiden University Centre for Linguistics (LUCL), Leiden, The Netherlands
[3] Berkman Klein Center for Internet and Society, Harvard University, Cambridge, USA

Abstract. In this article, we quantitatively analyze how the term "fake news" is being shaped in news media in recent years. We study the perception and the conceptualization of this term in the traditional media using eight years of data collected from news outlets based in 20 countries. Our results not only corroborate previous indications of a high increase in the usage of the expression "fake news", but also show contextual changes around this expression after the United States presidential election of 2016. Among other results, we found changes in the related vocabulary, in the mentioned entities, in the surrounding topics and in the contextual polarity around the term "fake news", suggesting that this expression underwent a change in perception and conceptualization after 2016. These outcomes expand the understandings on the usage of the term "fake news", helping to comprehend and more accurately characterize this relevant social phenomenon linked to misinformation and manipulation.

Keywords: Social computing · Digital humanities
Corpus linguistics · Misinformation · Fake news

1 Introduction

The term "fake news", defined as "false, often sensational, information disseminated under the guise of news reporting" [6], gained so much attention that it was named the Collins Word of the Year 2017 due to its unprecedented usage increase of 365% in the Collins Corpus [6]. Even though the concept of news articles aimed to mislead readers is by no means new [24], it seems to exist a relationship between the very expression "fake news" with the 2016 presidential election in the United States of America: Davies [8], using data from the NOW Corpus, shows that "there is almost no mention of 'fake news' until the first week

of November [2016] (...) and then it explodes in Nov 11–20, and has stayed very high since then". The author adds that the reason "why people all of the sudden started talking about something that had really not been mentioned much at all until that time" was "the US elections, which were held on November 9, 2016".

The sudden popularization of an already existing term (that is, not a neologism) in a language poses interesting questions regarding how concepts around this term are perceived by the speakers of that language. We might ask, for instance: what changed (if anything) in terms of conceptualization of this expression after its boom? Was there any kind of shift in the meaning of this expression when it became widely employed? If so, was this shift uniform across different varieties of the language? These are some of the issues of interest in *lexicology*, the area of linguistics focused in the study of the lexicon, that has been fostered thanks to advances in *corpus linguistics*, concerned with the use of big real-world corpora to the study of language.

The goal of this article is to provide a closer look at how newspapers and magazines across the world shape the term "fake news" – which is a relevant social phenomenon linked to misinformation and manipulation, and that has been facilitated by the rise of the Internet and online social media in recent years. We investigate the perception and the conceptualization of this expression through the quantitative analysis of a large corpus of news published in 20 countries from 2010 to 2018, thus making it possible to examine not only the diachronic development of this term, but also its synchronic usage in different parts of the English-speaking world. We complement our investigation with data collected from online search queries that help to measure how the public interest in the expression "fake news" and in the concepts around it changed over time in different places.

1.1 Related Work

In 2010, Michel et al. [17] coined the term *culturomics* meaning a method to study human behavior, cultural trends and language change through the quantitative diachronic analysis of texts, including of digitized books provided by the project Google Books. Several studies explore this method to investigate topics such as the dynamics of birth and death of words [20], semantic change [12], emotions in literary texts [1] and general characteristics of modern societies [22]. However, many criticisms arose regarding limitations of inferences derived from the analysis of Google Books due to factors that range from optical character recognition errors and overabundance of scientific literature [19] to the lack of metadata in the corpus [13].

Leetaru [15] proposes a somewhat complementary approach that he calls *culturomics 2.0*, which uses historical news data instead of books and can, according to the author, "yield intriguing new understandings of human society". In the same vein, Flaounas et al. analyze the European mediasphere [11] and the writing style, gender bias and the popularity of particular topics [10] in large corpora of news articles. Landsall-Welfare et al. [14], also using a large dataset of media reports, observe a change of framing and sentiment associated with

nuclear power after the Fukushima nuclear disaster. They detected effects on attention, sentiment, conceptual associations and in the network of actors and actions linked to nuclear power following the accident.

In this investigation, we combine many of the methods employed in the related works mentioned above. However, as far as we are concerned, this is the first paper that uses these methods to examine in details how the relevant term "fake news" is being reported by news media in different parts of the world and in two distinct periods in the history of this expression.

1.2   Research Question

Our main research question is: *was the rise of the public interest in the term "fake news" accompanied by changes in its conceptualization and in the perception about it?* Based on sociolexicological theories, that defend the existence of a considerable relationship between linguistic and extralinguistic factors with regards to the vocabulary of a language [5,16], our hypothesis is that the change of interest in the phenomenon *fake news* might have altered the general usage of the expression referring to it. Indeed, the results obtained in our investigations indicate, in general, a positive answer for our research question. Among other findings, we show modifications in the related vocabulary and in the mentioned entities accompanying the term "fake news", in addition to changes in the topics associated to this concept and in the overall contextual polarity of the pieces of text around this expression in media articles after 2016.

This article is structured as follows: in the next section, we present the process of acquisition and preparation of the main data source used in our investigations; in Sect. 3, we describe our analyses, present the found results and discuss their implications; finally, in Sect. 4, we summarize the outcomes of our study and conclude this paper by discussing possible future outlooks.

2   Data Source

The main dataset used in this study comes from the Corpus of News on the Web (NOW Corpus), which contains articles from online newspapers and magazines written in English and based in 20 different countries from 2010 to the present time [7]. This corpus is available for download and online exploration at https://corpus.byu.edu/now/ and, according to its authors, "is [as of April 2018] by far the largest corpus (of any language) that is available in full-text format". Our analyses are relative to a version of the corpus available in the month of April 2018, containing around six billion words of data.

In this dataset, we searched for all the occurrences of the term "fake news". For each occurrence, the online version of the corpus provides a concordance line, or *context* – that is, a piece of text of approximately 20–30 words around (before and after) the searched term. For example, for a certain news article published in July 25 2017 in the Kenyan newspaper Daily Nation, the context around the term "fake news" is: *(…) of social media and a study that said 90*

per cent of Kenyans had encountered **fake news**. *WhatsApp and Facebook are the two leading sources of misinformation, often (...).* All of our analyses were performed in these contexts, since words immediately surrounding a key term are more relevant to the conceptualization of this term than words further away from it, though in the same text. Wynne [27] adds that the main reason for using keywords in context (KWICs) in corpus linguistics is that "interesting insights into the structure and usage of a language can be obtained by looking at words in real texts and seeing what patterns of lexis, grammar and meaning surround them".

The total number of occurrences of "fake news" extracted from the NOW Corpus in April 30 2018 is 41,124. These occurrences encompass news articles published in all the 20 countries represented in the corpus, that were grouped in six regions based on their geographical locations (Africa, British Isles, Indian subcontinent, Oceania, Southeast Asia and The Americas), since it has been observed that offline and online news outlets tend to give preference to local and national news, to domesticate news about other countries and to reflect imbalanced information flows between the developed and the developing worlds [2].

These occurrences also cover each year in the corpus (from 2010 to 2018). Due to the previously observed increase in the usage of the term "fake news" during and after the 2016 presidential election in the United States of America (mentioned in Sect. 1), we categorized the occurrences in two periods: before and after the 2016 US election. The election was held in November, but we set the delimitation date between these periods in the end of the first semester of 2016 (June 30) in order to include the political campaign in the period *after US election*. Table 1 shows the number of contexts containing the term "fake news" in our dataset according to the geographical origin of the corresponding news media and the year and period of publication of the news article.

3   Analyses and Results

In this section, we display and examine the outcomes of our investigations. Each analysis is introduced by a description of how it is able to contribute answering to our research question, followed by the methodology employed and finally by a presentation and discussion of the results found.

3.1   Web Search Behavior

Before analyzing the data obtained from the NOW Corpus, we investigate whether it is possible to observe a change in Web search behavior regarding the expression "fake news" corresponding to the high increase in its use during and after the 2016 presidential election in the United States of America mentioned in Sect. 1.

Data obtained from Google Trends[1], an online tool that indicates the frequency of particular terms in the total volume of searches in the Google Search

---

[1] https://trends.google.com/trends/.

Table 1. Number of contexts containing the term "fake news" in our dataset according to (a) the geographical origin of the corresponding news media and (b) the year and period (before or after the 2016 presidential election in the United States of America) of publication of the news article.

(a) Geographical origin of news media

Region	Country	Occurrences
Southeast Asia	Singapore	3,722
	Malaysia	3,455
	Philippines	3,058
	Hong Kong	171
	Total: 25,3% / 10,406	
The Americas	United States	6,775
	Canada	2,960
	Jamaica	124
	Total: 24,0% / 9,859	
British Isles	Great Britain	4,213
	Ireland	2,035
	Total: 15,2% / 6,248	

Region	Country	Occurrences
Africa	South Africa	2,493
	Nigeria	1,974
	Kenya	1,368
	Ghana	300
	Tanzania	1
	Total: 14,9% / 6,136	
Oceania	Australia	3,052
	New Zealand	1,446
	Total: 10,9% / 4,498	
Indian subcontinent	India	2,961
	Pakistan	772
	Sri Lanka	147
	Bangladesh	97
	Total: 9,7% / 3,977	

(b) Year and period of publication of news article

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018
Occurrences	24	43	57	64	89	95	4,766	25,293	10,693

Period	before US election	after US election
Occurrences	494	40,630

engine, displays that public interest in the term "fake news" was approximately constant from 2010 until 2016, when it greatly and suddenly increased, as indicated by Fig. 1. This data also shows that, in the period before the 2016 US presidential elections, most of the countries with the highest proportions of searches for the term "fake news" were from the Eastern world. However, after the US election, the proportion of searches for this expression in Western countries increased considerably, especially in Europe. The 10 countries with the highest proportion of searches for the term "fake news" in both periods are listed in Table 2.

A closer look at the data from Google Trends also reveals that the great increase in the public interest for the expression "fake news" coincided with a change in the focus of Web searches. Table 3 shows the five most frequent search terms employed by users who also searched for "fake news" in the periods before and after the US election. We observe that, before the US election, searches for "fake news" were generic and regarded topics related to the media industry
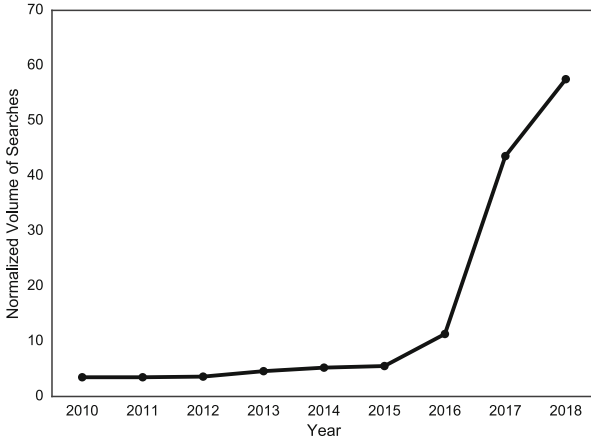
Fig. 1. Normalized volume of searches for the expression "fake news" on Google Search from 2010 to 2018.

Table 2. Countries with the highest proportion of searches for "fake news" on Google Search before and after US election.

Period	Countries
Before US election	India, United Arab Emirates, Singapore, United States, Macedonia, Qatar, New Zealand, Canada, Pakistan, Australia
After US election	Singapore, Philippines, United States, Canada, South Africa, Norway, Denmark, Ireland, United Kingdom, Switzerland

itself, like "article", "stories" and "report"; after the US election, however, these searches started to be more focused on political affairs and in the spread of fake news, mentioning entities like the elected president of the United States of America in 2016 (Donald Trump), the television news channel CNN (that devotes large amounts of its coverage to US politics) and the social media Facebook (considered a major source of fake news on the Internet).

Table 3. Most frequent search terms related to "fake news" on Google Search before and after US election.

Period	Search terms
Before US election	fake news generator, fake news article, fake news stories, make fake news, fake news report
After US election	trump news, the fake news, fake news trump, cnn news, fake news facebook

In this section, we used data obtained from the Google Trends tool. From the next section on, however, all of our analyses use the data described in Sect. 2, obtained from the NOW Corpus.

3.2   Co-occurring Named Entities

The analysis of *named entities* – that is, real-world entities such as persons, organizations and locations that can be denoted with proper names [26] – co-occurring with certain terms is an interesting way to contextualize these concepts. In our case, by identifying which entities are linked to the expression "fake news" in different periods of time and in different parts of the world, we are able to observe relationships of "who and where" in the recent history of our key-term.

In our dataset of news articles, we employed a simple method to identify named entities: we made use of the fact that newspapers and magazines consistently capitalize nouns representing named entities and counted all the words that appear capitalized in the contexts; then, we manually analyzed the most frequent capitalized words in each subdivision of the corpus (i.e. representing each region and period) to remove words not relative to named entities (such as "I", "SMS", "March" and words capitalized for other reasons) and to merge duplicated entities represented more than once (e.g. "Donald" and "Trump").

Table 4 shows the five most mentioned named entities in the periods before and after the 2016 US presidential election, regardless of geographical origin of the corresponding news media. Before the US election, it is possible to observe a strong connection between humor and fake news: with exception of Facebook, all the other most mentioned named entities are related to satirical TV shows and hosts based in the United States of America. On the other side, in the period after the US election, there is a movement towards politically related entities (Donald Trump), traditional media sources (CNN) and social networking services (Facebook and Twitter). It is interesting to notice that this shift matches the already mentioned (in Table 3) shift of interest towards political affairs and the spread of fake news observed in Web searches.

Table 4. Most mentioned entities in the periods before and after US election.

Period	Entities
Before US election	The Daily Show, Jon Stewart, Onion News Network, Facebook, Stephen Colbert
After US election	Donald Trump, Facebook, US, CNN, Twitter

When we make this same diachronic comparison, but now considering the geographical origin of the corresponding news media, we observe a noteworthy phenomenon: the global standardization of the named entities related to *fake news*. Table 5 indicates that local entities are more relevant in the period before

the US election, when names of geographical regions (Ekiti), countries (Nigeria, China), local political parties (PDP – People's Democratic Party of Nigeria, BJP – Bharatiya Janata Party of India) and local personalities (Shahid Afridi, King Salman, Korina Sanchez) appear frequently among the most mentioned entities. In the contexts after the US election, however, Donald Trump, Facebook and US are the three most mentioned entities for nearly all the regions – with the sole exception of The Americas, where CNN replaces US.

Table 5. Most mentioned entities in the periods before and after US election, considering the geographical origin of the corresponding news media.

Region	Period	Entities
Africa	Before	PDP, Ekiti, Nigeria
	After	Donald Trump, Facebook, US
British Isles	Before	Facebook, The Daily Show, Stephen Colbert
	After	Donald Trump, Facebook, US
Indian subcontinent	Before	Shahid Afridi, King Salman of Saudi Arabia, BJP
	After	Facebook, Donald Trump, US
Oceania	Before	Twitter, The Daily Show, NBC
	After	Donald Trump, Facebook, US
Southeast Asia	Before	Korina Sanchez, US, China
	After	Facebook, Donald Trump, US
The Americas	Before	The Daily Show, Jon Stewart, Onion News Network
	After	Donald Trump, Facebook, CNN

3.3   Semantic Fields of the Surrounding Vocabulary

Besides the investigation of the named entities that accompany a given key-term, the analysis of the general vocabulary co-occurring with it is also valuable. In our case, one of the possible methods of performing such analysis is by observing the semantic fields (i.e. groups to which semantically related items belong) of the words co-occurring with the expression "fake news" in our contexts.

For performing this task, we first lemmatized all the words in the contexts by employing the WordNet Lemmatizer function provided by the Natural Language Toolkit [3] and using *verb* as the part-of-speech argument for the lemmatization method. By applying this lemmatization, we grouped together the inflected forms of the words so that they could be analyzed as single items based on their dictionary forms (*lemmas*).

Then, we used Empath [9], "a tool for analyzing text across lexical categories"[2], to classify the lemmatized words according to categories that represent different semantic fields, such as diverse topics and emotions. For every context,

---

[2] https://github.com/Ejhfast/empath-client.

we calculated the percentage of words belonging to each semantic field represented by an Empath category. Due to the high number of categories predefined by Empath (194 in total), we selected eight that showed interesting results and are relevant for our discussion: *government*, *internet*, *journalism*, *leader*, *negative emotion*, *politics*, *social media* and *technology*. By way of example, the category *internet* includes 79 words such as *homepage*, *download* and *hacker*, while the category *journalism* contains 69 words, including *report*, *article* and *newspaper*.

Figure 2 displays the average percentage of words in these categories for all the six regions considered here, both before and after the 2016 US election. By analyzing the graphs presented, we observe interesting differences and trends regarding the quantitative utilization of words from the semantic fields considered. We highlight the high increase in the use of words from the related categories *government*, *leader* and *politics* (and also from the supposedly unrelated category *negative emotion*) and the high decrease in the use of words from the categories *internet*, *journalism* and *technology* (but not *social media*) in almost all regions after the US election.
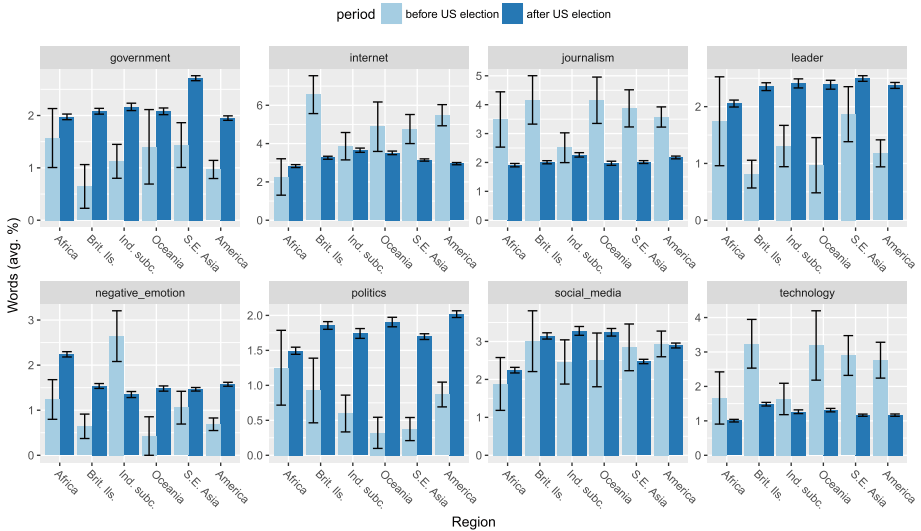

Fig. 2. Percentage of words in each semantic field represented by an Empath category. Error bars indicate standard errors.

We hypothesize that these results indicate a change in the focus of the news considered here: before the 2016 US election, the term "fake news" was probably more mentioned in contexts in which the focus was the *environment* where they occur (Internet, newspapers etc.), sometimes even meta-discussions on the very topic of fake news and its dissemination; during and after the US election, however, the discussion seems to have migrated to themes more close to the *content* of the fake news themselves (politics, elections etc.).

3.4   Co-occurrence Networks

Another possible method of investigating the vocabulary accompanying a key-term in a corpus is through the observation of co-occurrence networks. In our case, this method enables us to visually analyze the words that co-occur with the expression "fake news" in the contexts that we are considering. Here, we compare co-occurrence networks between the periods before and after the 2016 US election, regardless of the geographical origin of the media outlets. These networks are represented here by graphs, in which each node corresponds to a word and each edge corresponds to an association between two given words.
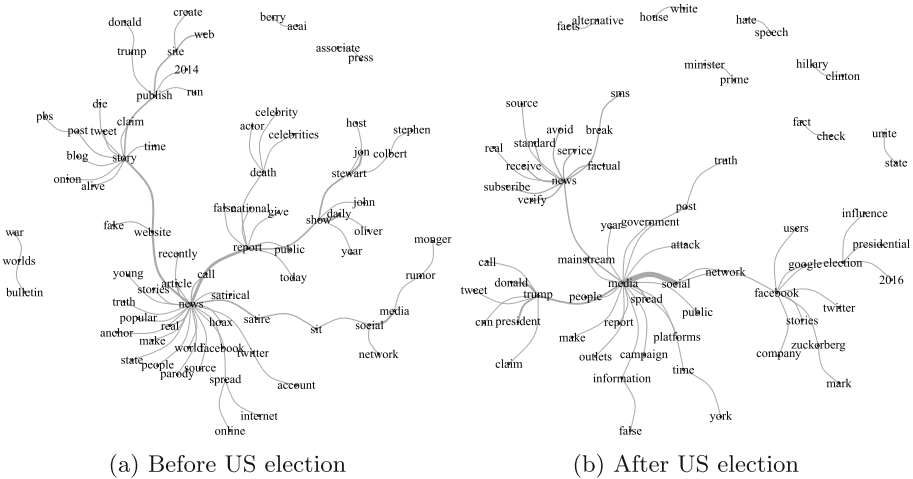

(a) Before US election   (b) After US election

Fig. 3. Co-occurrence networks of words before and after the 2016 US election.

To build our graphs of co-occurring words, we followed the steps below. First, we counted the number of contexts in which two given words co-occur, so that we have the volume of co-occurrences for each pair of words. Then, we normalized these values in order to work with percentages instead of working with the absolute number of co-occurrences. To improve the visualization of the graphs, we filtered out minor relationships and highlighted the strongest ones by removing nodes and edges when the co-occurrence percentage was lower than 0.8% (for the period before the US election) and 0.5% (for the period after the US election), and by plotting the width of the edges proportionally to the strength of association between two given words. Finally, we obtained the *maximum spanning trees* of both graphs, which are presented in Fig. 3.

This method of investigation enables us to make several qualitative observations. Comparing the two graphs, we notice clear changes in the relationships between the words co-occurring with "fake news" in our contexts. For instance, before the US election, the main cluster contains words related to the news

industry itself ("article", "stories", "hoax") and to Internet ("website", "twitter", "facebook", "account"). Corroborating previous findings (Sect. 3.2), there is also another cluster containing words referring to satirical TV shows and hosts ("daily", "show", "colbert", "oliver", "stewart"). In the graph representing the period after the 2016 US election, we start to observe terms linked to specific events, mainly the US election itself. Interestingly, some terms that surround meta-discussions about fake news also appear, highlighting relevant related concepts: *fact check*, *hate speech*, *post truth* and *alternative facts*.

3.5   Topics Addressed in the Contexts

In addition to studying the vocabulary around a key-term, it is also possible to find the main topics addressed in the pieces of text surrounding the occurrences of the expression "fake news" in our corpus.

For this task, we used latent Dirichlet allocation (LDA) [4], a way of automatically discovering topics discussed in texts. First, we lowercased and tokenized all the words in the dataset. Then, we removed stop words using the list provided by the Natural Language Toolkit – after having added the words "fake" and "news" to this list, since they appear in all contexts. Finally, we ran the LDA algorithm using `gensim` [21], a Python library for topic modeling. We used topic coherence score [18] to choose the optimum number of topics k to be returned by the algorithm. Thus, for each region, we ran the LDA algorithm starting with $k = 2$ and ending with $k = 20$, and chose the best LDA model, that is, the LDA model with highest topic coherence score. All regions had, respectively, $k = 2$ and $k = 14$ for the periods before and after the US election, except The Americas, that had $k = 8$ and $k = 14$. For each region, the LDA returned these k topics containing words ordered by importance in the corresponding context, filtered both by region and topic. We then selected the most important topic as the representative of each region and period. Table 6 shows the main topic for each region in both periods (before and after the US election) and the top ranked ten words produced by our LDA model.

Unlike in the period before the US election, the words related to each topic inferred by LDA are cohesive among each other in the period after the US election. We observe, for all regions, a relevant frequency of words related to politics and social networks in the period after the US election. More specifically, the words "russian", "russia", "election" and "facebook" rank high in this period. In the period before the US election, most of the top words are related to Internet and to the spread of misinformation, in all regions.

3.6   Polarity

Our final analysis explores a different feature of the contexts in which the expression "fake news" appear in our dataset: their *polarities*, that is, whether the expressed opinion in the texts is mostly positive, negative or neutral. Here, we performed sentiment analysis [23] in each one of the contexts in our dataset

Table 6. Main topic for each region. Inside each topic, ten words are presented in order of importance according to the LDA output.

Region	Period	Main topic words
Africa	Before	become, world, party, south, leave, week, online, state, give, member
	After	trump, people, spread, president, truth, propaganda, thing, look, show, nigerian
British Isles	Before	Story, account, real, daily, website,new, show, use, death, state
	After	propaganda, source, russian, american, russia, lie, mean, popular, politics, allegation
Indian subcontinent	Before	create, spread, report, death, also, lot, say, not, social, do
	After	facebook, also, problem, user, issue,company, russian, state, work, zuckerberg
Oceania	Before	people, story, site, report, website,mortgage, would, fool, year, day
	After	election, influence, media, create, russian,question, policy, discuss, presidential, word
Southeast Asia	Before	article, website, story, report, site, celebrity, death, publish, go, viral
	After	public, government, fact, twitter, proliferation,day, official, however, phenomenon, concern
The Americas	Before	release, chip, firm, flurry, blue, target, date, breadcrumb, irresponsible, last
	After	facebook, problem, network, company, also,believe, publish, work, policy, russian

using `SentiStrength`[3] [25], a tool able to estimate the strength of positive and negative sentiment in short texts. Given a piece of text, this tool returns a score that varies from -4 (negative sentiment) to $+4$ (positive sentiment).

We are interested in analyzing how polarity changes over time and in different regions when it comes to "fake news" and how this can be perceived in our dataset. Figure 4 depicts the average polarity of the contexts in each region before and after the 2016 US presidential election. We first observe a clear dominance of negative polarities in all periods and regions, indicating that the term "fake news" is often related to negative words [28] and sentiments – which is not surprising, since the concept of fake news seems to be strongly associated with negative concepts, like misinformation, manipulation and the spread of untrue facts.

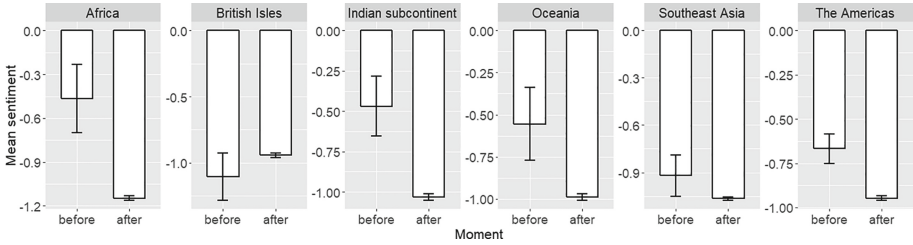---

[3] http://sentistrength.wlv.ac.uk/.

Fig. 4. Average polarity of the contexts in each region before and after US election (bars indicate the standard error of the mean).

In this figure, we also observe that, in general, the polarity expressed in the contexts in the period after the US election is more negative than before. The only exception is in the British Isles, where the difference of polarity between the periods is not statistically relevant. This result seems to corroborate findings presented in previous sections of this article, which demonstrated that, before the 2016 US election, the term "fake news" was often linked to satirical TV shows and more general topics, while in the period during and after the election the topics became more related to the spread of false information in the context of political activity.

3.7   Summary of Results

The most relevant outcomes of the analyses presented in this section can be summarized and integrated as follows:

– the interest for the term "fake news" suddenly increased after the 2016 US election, as indicated by the rise of news about it and of Google Search queries for this expression (Sect. 3.1);
– this growth was accompanied by a change of framing around the term "fake news" – from, for instance, topics regarding the media industry itself to those related to political affairs (Sects. 3.1, 3.3, 3.4 and 3.5);
– the named entities linked to the expression "fake news" not only changed towards political topics, but also suffered from global standardization after the US election (Sect. 3.2);
– the negativity of the news containing the term "fake news" increased after the US election (Sect. 3.6).

All these results suggest that, as hypothesized in Sect. 1, the rise of public interest in the term "fake news" came with changes in its conceptualization and in the perception about it.

4   Concluding Remarks

Due to the increased role of the Internet in modern societies, topics regarding misinformation and manipulation in online environments seem to be subject to

progressively more public debate and interest, including from the traditional media. Understanding how these topics are viewed through the eyes of opinion leaders is crucial to comprehend how public opinion about them is being shaped in present day.

In this article, we present a quantitative analysis on the perception and conceptualization of the term "fake news" in a corpus of news articles published from 2010 to 2018 in 20 countries. We investigate how media sources have been reporting topics related to fake news and whether the rise of the public interest in this very expression during and after the 2016 presidential election in the United States of America was accompanied by changes of perception and shifts in sentiment about it. We observed changes in the vocabulary and in the mentioned entities around the term "fake news" in our corpus, in the topics related to this concept and in the polarity of the texts around it after 2016, as well as in Web search behavior of Google Search users interested in this concept.

We are also interested in understanding whether the term "fake news" is framed differently across the globe – and, if so, which are these differences. The existence of such variations may result in different shifts in the meanings and in the sentiments around these concepts in various regions of the world, which justifies this study as a way to more clearly understand how the public opinion is being steered in the current context in different countries of the English-speaking world.

In this paper, we analyzed the usage of the term "fake news" in a diachronic perspective, but only considered two historical moments: before and after a key event in the history of this expression (the 2016 US presidential election). In the future, we plan to consider a larger spectrum of periods, particularly to understand whether (and, if it is the case, when) the conceptualization of "fake news" changed once again. We also intend to add analyses using data from other relevant sources, including Twitter posts and Wikipedia edits.

References

1. Acerbi, A., Lampos, V., Garnett, P., Bentley, R.A.: The expression of emotions in 20th century books. PLOS ONE **8**(3), e59030 (2013)
2. Berger, G.: How the Internet impacts on international news: exploring paradoxes of the most global medium in a time of 'hyperlocalism'. Int. Commun. Gaz. **71**(5), 355–371 (2009)
3. Bird, S., Loper, E., Klein, E.: Natural Language Processing with Python. O'Reilly Media Inc., Sebastopol (2009)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
5. Cambraia, C.N.: Da lexicologia social a uma lexicologia sócio-histórica: caminhos possíveis. Revista de Estudos da Linguagem **21**(1), 157–188 (2013)
6. Collins Dictionary: Word of the year 2017 (2017). https://www.collinsdictionary.com/woty. Accessed 4 May 2018
7. Davies, M.: Corpus of News on the Web (NOW): 3+ billion words from 20 countries, updated every day (2013). https://corpus.byu.edu/now/

8. Davies, M.: Fake news (2017). https://corpus.byu.edu/now/help/fake-news.asp. Accessed 4 May 2018

9. Fast, E., Chen, B., Bernstein, M.S.: Empath: understanding topic signals in large-scale text. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 4647–4657. ACM (2016)

10. Flaounas, I., et al.: Research methods in the age of digital journalism: massive-scale automated analysis of news-content - topics, style and gender. Digit. Journal. **1**(1), 102–116 (2013). https://doi.org/10.1080/21670811.2012.714928

11. Flaounas, I., Turchi, M., Ali, O., Fyson, N., De Bie, T., Mosdell, N., Lewis, J., Cristianini, N.: The structure of the EU mediasphere. PLOS ONE **5**(12), e14243 (2010)

12. Gulordava, K., Baroni, M.: A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In: Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics, pp. 67–71. Association for Computational Linguistics (2011)

13. Koplenig, A.: The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets - reconstructing the composition of the German corpus in times of WWII. Digit. Sch. Humanit. **32**(1), 169–188 (2017). https://doi.org/10.1093/llc/fqv037

14. Lansdall-Welfare, T., Sudhahar, S., Veltri, G.A., Cristianini, N.: On the coverage of science in the media: a big data study on the impact of the Fukushima disaster. In: 2014 IEEE International Conference on Big Data, pp. 60–66. IEEE (2014)

15. Leetaru, K.: Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. First Monday **16**(9) (2011). http://firstmonday.org/article/view/3663/3040

16. Matoré, G.: La méthode en lexicologie: domaine français. Didier, Paris (1953)

17. Michel, J.B., et al.: Quantitative analysis of culture using millions of digitized books. Science **331**(6014), 176–182 (2011)

18. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT 2010, pp. 100–108. Association for Computational Linguistics, Stroudsburg, PA, USA (2010). http://dl.acm.org/citation.cfm?id=1857999.1858011

19. Pechenick, E.A., Danforth, C.M., Dodds, P.S.: Characterizing the Google Books corpus: strong limits to inferences of socio-cultural and linguistic evolution. PLOS ONE **10**(10), e0137041 (2015)

20. Petersen, A.M., Tenenbaum, J., Havlin, S., Stanley, H.E.: Statistical laws governing fluctuations in word use from word birth to word death. Sci. Rep. **2**, Article number 313 (2012)

21. Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50. ELRA, Valletta, Malta, May 2010. http://is.muni.cz/publication/884893/en

22. Roth, S.: Fashionable functions: a google ngram view of trends in functional differentiation (1800–2000). Int. J. Technol. Hum. Interact. **10**(2), 34–58 (2014)

23. Silva, N.F.F.D., Coletta, L.F.S., Hruschka, E.R.: A survey and comparative study of tweet sentiment analysis via semi-supervised learning. ACM Comput. Surv. **49**(1), 15:1–15:26 (2016). https://doi.org/10.1145/2932708

24. Standage, T.: The true history of fake news. 1843 Magazine (2017). https://bit.ly/2sh9OYQ. Accessed 4 May 2018

25. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment in short strength detection informal text. J. Am. Soc. Inf. Sci. Technol. **61**(12), 2544–2558 (2010). https://doi.org/10.1002/asi.v61:12
26. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In: Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003. Association for Computational Linguistics (2003)
27. Wynne, M.: Searching and concordancing. In: Corpus Linguistics. An International Handbook, vol. 1, pp. 706–737 (2008)
28. Zollo, F.: Emotional dynamics in the age of misinformation. PLOS ONE **10**(9), 1–22 (2015). https://doi.org/10.1371/journal.pone.0138740