Universiteit
Leiden
The Netherlands

Contributions to the computational processing of diachronic linguistic corpora

Cunha, E. Landulfo Teixeira Paradela

Citation

Cunha, E. L. T. P. (2020, March 19). *Contributions to the computational processing of diachronic linguistic corpora*. *LOT dissertation series*. LOT, Amsterdam. Retrieved from https://hdl.handle.net/1887/133504

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/133504

Note: To cite this publication please use the final published version (if applicable).

Cover Page

Universiteit Leiden

Leiden University
Repository

The handle http://hdl.handle.net/1887/133504 holds various files of this Leiden University dissertation.

Author: Cunha, E. Landulfo Teixeira Paradela
Title: Contributions to the computational processing of diachronic linguistic corpora
Issue date: 2020-03-19

CHAPTER 4

Diachronic corpora and quantitative approaches to the lexicon: the case of the term *fake news*[1]

4.1 Introduction

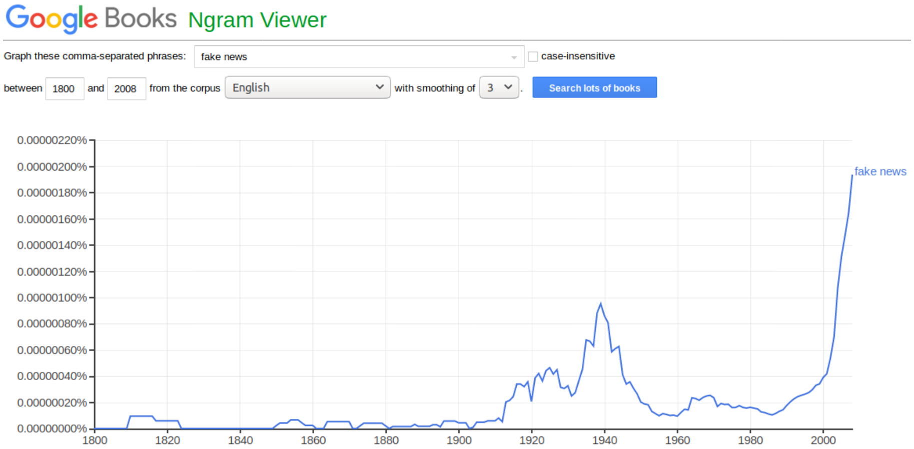
The term *fake news*, defined by Collins Dictionary as "false, often sensational, information disseminated under the guise of news reporting", gained so much attention that it was named the English word of the year 2017 by both Collins Dictionary (2017) and American Dialect Society (2018). Even though the concept of news arti-

[1] This chapter reproduces with minor changes the article "Quantifying the conceptualization of the term 'fake news' in Brazilian and English-speaking media sources" (Cunha et al., under review), submitted for publication. This article is, in its turn, an extended version of the paper "Fake news as we feel it: Perception and conceptualization of the term 'fake news' in the media" (Cunha et al., 2018), published as a chapter of *Social informatics* (eds. Steffen Staab, Olessia Koltsova and Dmitry I. Ignatov) and presented at the *10th International Conference on Social Informatics (SocInfo 2018)*, held in Saint Petersburg, Russia, in September 2018. See Appendix C for more information.

cles aimed to mislead readers is by no means new (Standage, 2017), there seems to exist a relationship between the very expression *fake news* with the 2016 presidential election in the United States of America: Davies (2017a), 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” (Davies, 2017a). Data from the Google Books Ngram Viewer[2], however, shows that the use of the term *fake news* had already peaked in earlier periods. Figure 4.1 reproduces the output of the query for *fake news* in this tool – which, at the time of writing of this text, only includes data until the year 2008. We observe that the relative frequency of this expression in the Google Books corpus experienced an increase in the 1910s/1920s, then peaked around 1940, and then started to rise again in the first decade of the 21st century. Since the data provided by this tool does not reach the 2010s, we cannot compare these frequencies with those from 2016 onwards. In any case, there is evidence that the relative frequencies of this expression are even higher in the years after the 2016 presidential election in the United States of America.

Despite its recent success, the widespread use of the term *fake news* has received much criticism. Members of the British Parliament recommended in a report that the Government rejects this expression, since it “is bandied around with no clear idea of what it means, or agreed definition”, and it “has taken on a variety of meanings, including a description of any statement that is not liked or agreed with by the reader” (Parliament of the United Kingdom,

[2] The Google Books Ngram Viewer is a searchable interface for Google Books, available at `https://books.google.com/ngrams` .



Graph these comma-separated phrases: fake news □ case-insensitive

between 1800 and 2008 from the corpus English with smoothing of 3

Figure 4.1: Output of the query for *fake news* in the Google Books Ngram Viewer. The chart shows relative frequencies from 1800 to 2008.

2018). It has already been suggested that the expression *fake news* is doing great harm (Habgood-Coote, 2018) and that it should be retired (Sullivan, 2017). In addition, the dissemination of the term *fake news* went beyond the English language. The Commission for the Enrichment of the French Language, for example, declared that "[t]he Anglo-Saxon expression 'fake news' (...) has rapidly prospered in French" (BBC News, 2018). The same phenomenon happened in other languages, including in Brazilian Portuguese, in which the borrowed term seems to have spread mainly during 2018 – once again in a context of national elections. Lees (2018) considers that *fake news* "has become an international political catchphrase", since, at the time of the publication of her report, "more than 20 political leaders worldwide[3], from authoritarian regimes to European

[3] "Over the past year, political leaders in Burma, Cambodia, China, Egypt, France, Germany, Hong Kong, Hungary, Kuwait, Libya, Malaysia, the Philippines, Poland, Russia, Singapore, Somalia, Syria, Tanzania, Thailand, Turkey, the USA and Venezuela have publicly accused journalists of reporting, or being,

democracies, have used the term to accuse reporters of spreading lies as a way to discredit journalism they do not like" (p. 88).

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

The specific goal of this chapter is to provide a closer look at how newspapers and magazines across the world shaped 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 the second decade of the 21st century. We investigate the perception and the conceptualization of this expression through the quantitative analysis of two corpora of news published in 21 countries from 2009 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 world. We complement our investigation with data collected from online search queries that help us to measure how the public interest in the expression *fake news* and in the concepts around it changed over time in different places.

fake news" (Lees, 2018, p. 88). Brazil is not on the list only because far-right Jair Bolsonaro took power in January 2019, i.e., after the publication of Lees' article.

Our general goal, however, goes beyond the investigation of an individual expression. By studying the diachronic change in the conceptualization of a term through replicable computational and quantitative methods, we initially propose a framework that can be applied to other cases. In this framework, we opt to establish analogies between concepts and means of expression that exceed the strictly linguistic analysis of the lexicon. To achieve this goal, we employ already established analytical methods which, when put together, are able to delineate the semantic framing of a linguistic item. It is interesting to note that, to the best of our knowledge, this is the first attempt to merge these specific methods into one framework that aims to investigate the diachronic change in the conceptualization of an expression. In addition, we also contribute to the research on diachronic corpus linguistics in Brazilian Portuguese – which, although far from being a low-resource language, is much less studied than English in this domain.

4.1.1 Research question

Our main research question here 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 (Matoré, 1953; Cambraia, 2013), 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 to 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 with this concept and in the overall contextual polarity of the pieces of text around this expression in English-written media articles after 2016 and in Brazilian articles after 2018.

This chapter is structured as follows: first, we mention previous works related to the concept of *fake news* and to the usage of large language datasets to investigate social phenomena; in Section 4.2, we present the process of acquisition and preparation of the data sources used in our investigations; in Section 4.3, we describe our analyses, present the results found and discuss their implications; finally, in Section 4.4, we summarize the outcomes of our study and conclude this chapter by discussing possible future outlooks.

4.1.2 Related work

In the years prior to the publication of this study, the amount of scholarly papers mentioning the term *fake news* has grown dramatically. To illustrate this trend, we show in Figure 4.2 the number of academic publications per year returned by the query for *fake news* on Scopus and Web of Science databases[4]. From respectively ten and eight articles in 2016, the numbers increase to 221 and 160 in 2017, and then to 551 and 367 in 2018. These values were obtained through queries performed on October 8, 2019, which is the reason why there is a small drop in the numbers relative to 2019. Still, the amount of publications shown in the graph in this year are likewise substantial. Of course, the fact that the numbers of articles returned by these queries have greatly increased after 2017 does not mean that the research on disinformation and on "yellow journalism" has started in this period, but only that the use of the term *fake news* has increased in this context.

[4] Scopus (`https://www.scopus.com/`) and Web of Science (`https://www.webofknowledge.com`) are two citation databases that provide information on published scientific articles, journals and conference proceedings.

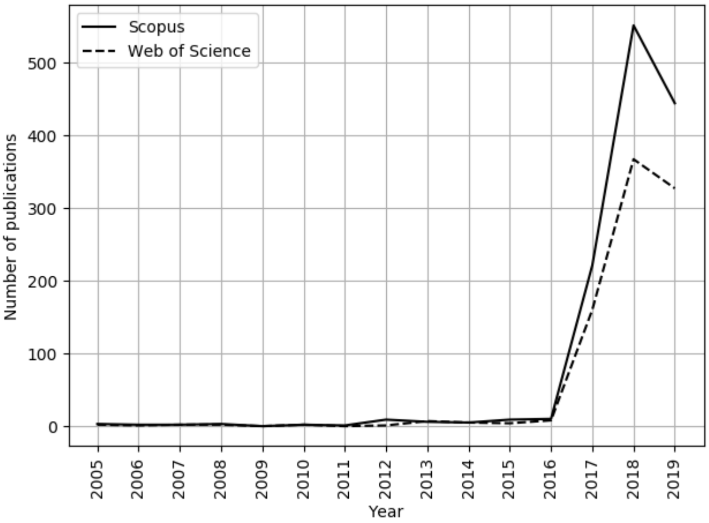


Figure 4.2: Numbers of academic publications per year returned by the query for *fake news* on Scopus and Web of Science databases. These values were obtained through queries performed on October 8, 2019.

On the phenomenon of fake news

The focus of this study is definitely not on the *phenomenon* of fake news, but rather on the use of this very expression in the language. Still, we cite some works on the phenomenon of fake news for the sake of contextualization, since the use of an expression is intimately related to the entity that it represents.

Van Hout and Burger (2015) allude to the boom of satirical fake news sources in the years before the publication of their study. In Sections 4.3.2 and 4.3.4, we show how this is confirmed by our data, since satirical TV shows and hosts often co-occur with the term *fake news* in the period before the 2016 presidential election in the United States of America. Allcott and Gentzkow (2017) study the dissemination of fake news in the particular case of this election and allege that “[f]ollowing the 2016 election, a specific concern has been the effect of false stories – ‘fake news,’ as it has been dubbed – circulated on social media” (p. 212). They analyze, from an economic perspective, the consumption of fake news before and during this election, identifying an important role of social media in this context and confirming “that fake news was both widely shared and heavily tilted in favor of Donald Trump” (Allcott and Gentzkow, 2017, p. 212). As we show in the next sections, these results are also corroborated through our corpus-based methods. This fact gives robustness to our methodology, since it shows that it is able to replicate observations obtained from more fine-grained analyses. Vosoughi et al. (2018), using a big dataset collected from Twitter, empirically demonstrate that fake news diffuse “significantly farther, faster, deeper, and more broadly than the truth in all categories of information” (p. 1146). They also show that the effects of misinformation are “more pronounced for false political news than for false news about terrorism, natural disasters, science, urban legends, or financial information” (Vosoughi et al., 2018, p. 1146).

The task of fighting fake news is raised by Lazer et al. (2018), who call for a multidisciplinary effort to address the problem. The authors identify two categoris of interventions that might be effective to combat fake news and their influence: “(i) those aimed at empowering individuals to evaluate the fake news they encounter, and (ii) structural changes aimed at preventing exposure of individuals to fake news in the first instance” (Lazer et al., 2018, p. 1095). While the former concerns essentially fact checking, the latter is mostly performed through platform-based detection and intervention via algorithms and bots. Burger et al. (2019), in their turn, discuss a specific element in the junk news (i.e. low-quality news) ecosystem: junk news (part of which may be fake) that are commercially motivated – “i.e. money-driven, highly shareable clickbait with low journalistic production standards" (p. 1). By studying the Dutch case, they show that, during the period from 2013 to 2017, the average number of user interactions with junk news significantly exceeded that with mainstream news. They also show that more than half of the Dutch Facebook users interacted with a junk news post at least once in this period.

Some studies focus on the scenarios of developing countries, which bring their own characteristics. For illustration, here we cite Pate and Ibrahim (2019), who analyze the impacts of fake news on the Nigerian democratic system; Glowacki et al. (2018), who target political news and information shared over Twitter and Facebook during the 2018 Mexican presidential election; and Arnaudo (2017), who investigates the use of automated accounts (bots) that spread misinformation in three Brazilian political moments: the 2014 presidential elections, the impeachment of former president Dilma Rousseff, and the 2016 municipal elections in Rio de Janeiro. It is also worth mentioning the project entitled *Eleições Sem Fake*[5] (i.e., *Elections Without Fake*), focused on the development of computational systems to fight against the spread of fake news during Brazilian elections. The cases of other countries, including Romania (Bârgăoanu and Radu, 2018) and South Africa (Wasserman, 2017), have been studied as well.

A significant number of other computationally-driven studies on misinformation on the Web have already been performed, especially on the topics of fake news characterization (e.g. Rashkin et al., 2017; Arif et al., 2018) and its automatic detection (e.g. Conroy et al., 2016; Rubin et al., 2016; Shu et al., 2017; Tschiatschek et al., 2018; Reis et al., 2019). Zannettou et al. (2017), for instance, analyze news published on three online platforms (4chan, Reddit and Twitter) in order to identify and characterize the flow of mainstream and fake news between them, shedding some light

[5] Available at `http://www.eleicoes-sem-fake.dcc.ufmg.br/` .

on the important topic of cross-platform misinformation spread. Vicario et al. (2019) present a framework for identifying polarizing content on social media and, thus, predicting potential targets for hoaxes and fake news. Ruchansky et al. (2017) propose a model for misinformation detection that captures three already observed common characteristics of fake news: (a) low quality of the text, including mismatches between the headline and the body of the article; (b) response to provocation, since "fake news often contains opinionated and inflammatory language, crafted as click bait or to incite confusion" (p. 797), which motivates responses with a high emotional content; and (c) doubtful source, i.e., lack of credibility of the URL, media source and author that published the news story. Despite the good results achieved by the model presented in their study, the classification of fake news still remains, at the writing of this dissertation, a challenging problem with many open questions.

On the expression *fake news*

More related to the object of this study are the investigations on the use of the very expression *fake news*. However, as put by Gelfert (2018), "[w]hile much ink has been spilled, by academics and pundits alike, on [the] disruptive potential and deceptive nature [of the fake news phenomenon], somewhat less attention has been paid to analyzing and defining the term 'fake news' " (p. 85). According to this author, "[i]t is (...) quite natural that a term as recent and controversial as 'fake news' should be used in a variety of (sometimes conflicting) ways, thereby making conceptual analysis more difficult" (Gelfert, 2018, p. 85). Gelfert recognizes that the term *fake news* has evolved rapidly and argues that "it should be reserved for cases of deliberate presentation of (typically) false or misleading claims as news, where these are misleading *by design*" (p. 84, emphasis in original)[6].

Nielsen and Graves (2017), after analyzing data from focus groups and surveys, found that

> [w]hen asked to provide examples of fake news, people identify poor journalism, propaganda (including both lying politicians and hyperpartisan content), and some kinds of advertising more frequently than false information designed to masquerade as news reports. (Nielsen and Graves, 2017, p. 1)

They also observe that people are aware that *fake news* is often employed as "a politicized buzzword used by politicians and others to criticize news media and platform companies" (Nielsen and Graves, 2017, p. 1). This fact reinforces the observation that the term has been carrying an imprecise definition, which is one of the reasons why Habgood-Coote (2019) argues that academics and journalists should stop using it[7]. Tandoc Jr. et al. (2018) contribute to this discussion by reviewing 34 academic articles in order to understand how studies from between 2013 and 2017 have used this expression. The authors classify the found definitions into six categories: news satire, news parody, news fabrication, photo manipulation, advertising and public relations, and propaganda. Even though the use of this term is not analyzed from a diachronic perspective (such as the one proposed in this chapter), Tandoc Jr. et al. (2018) acknowledge that "[e]arlier studies have applied the term to define related but distinct types of content, such as news parodies, political satires, and news propaganda", but that "it is currently used

[6] The author then continues: "[t]he phrase 'by design' here refers to systemic features of the design of the sources and channels by which fake news propagates and, thereby, manipulates the audience's cognitive processes" (Gelfert, 2018, p. 84).

[7] Of course, we assume that Habgood-Coote (2019) is not opposing studies like the one presented here, i.e., that investigate the very use of the term.

to describe false stories spreading on social media" (p. 138). This observation sustains at least one of our results, that suggests a stronger link between the term and parody/satire before the 2016 presidential election in the United States of America, and an association with social networking sites (particularly Facebook, Twitter and WhatsApp) during and after the election campaign.

The specific use of the term *fake news* by Donald Trump has been the subject of previous works too. Ross and Rivers (2018) investigated a corpus containing 1,416 tweets posted by Trump through comparative keyword analysis and show that this term has been employed by him as a pejorative label used to ridicule the critical mainstream media and "to position himself as the only reliable source of truth" (p. 1). Also, Holan (2017) compares the media's definition of *fake news* with Donald Trump's definition, arguing that

> [w]hen PolitiFact fact-checks fake news, we are calling out fabricated content that intentionally masquerades as news coverage of actual events. When President Donald Trump talks about fake news, he means something else entirely. Instead of referring to fabricated content, Trump uses the term to describe news coverage that is unsympathetic to his administration and his performance, even when the news reports are accurate. (Holan, 2017, p. 121).

Horta Ribeiro et al. (2017) shows that this behavior (that is, labeling as *fake news* any opinions or facts with which one disagrees) is not restricted to Donald Trump and to other populists from around the world. The authors use the suggestive title "Everything I disagree with is #FakeNews" in their article showing that Twitter users also employ the term *fake news* (and other related words and expressions) when designating political content with which they disagree.

Tambini (2017) asks two central questions related to the rise of the expression *fake news*: (a) why have politicians and the media suddenly started talking about fake news? And (b) who benefits from using this concept? According to his view, the three main beneficiaries are the "new populists", who "use the notion of 'fake news' to undermine legitimate opposition, and resist fourth estate accountability"; the "historical losers", who "claim that political changes result from misinformation"; and the "legacy media", that "want to discredit the 'wisdom of crowds' and aim for a return to trusted news brands" (Tambini, 2017, p. 9).

Finally, we cite Brummette et al. (2018), who use social network analysis, content analysis and cluster analysis to explore the use of the term *fake news* on Twitter. Similarly to what we propose here, the authors investigate the prevalent discussions surrounding this expression through the analysis of elements like the most frequently co-occurring words and hashtags. Our approach, however, is innovative for the study of this term not only because it employs different methods and tools, but mainly because it is guided by a diachronic perspective, which allows comparisons between distinct moments in the history of the studied expression.

On the methodology of this study

From a theoretical viewpoint, this chapter is partially inspired by the seminal studies on social lexicology proposed by Matoré (1949, 1953) and further developed as sociohistorical lexicology by Cambraia (2013). These authors argue for the use of models for lexical analysis that take into account social and extralinguistic factors, and address the link between the lexicon and social transformations. Cambraia (2013) considers that important questions in a sociohistorically-based lexicology are, for example: what makes a lexical item earn or lose a seme (or a meaning)? Or what drives a speaker to create a new word for a concept for which another word already existed? According to this approach, the answers to these questions must go beyond the analysis of strictly linguistic factors, but should also contemplate extralinguistic elements. From a methodological perspective, the studies influenced by Cambraia (e.g. Guedes and Mendes, 2016; Dores and Toledo, 2018; Rafael and Simião, 2019) propose an in-depth analysis of specific lexical items, and the textual and social contexts in which these items are inserted are investigated. Here we use these previous works as an inspiration for the proposal of a framework to the analysis of the conceptualization of a given item. Differently from them, however, we employ tools and resources from corpus linguistics and natural language processing that are not considered in the original Cambraia's proposition.

In 2011, Michel et al. (2011) coined the term *culturomics*, meaning a method to study human behavior, cultural trends and language change through the diachronic quantitative analysis of texts, including of digitised 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 (Petersen et al., 2012), semantic change (Gulordava and Baroni, 2011), emotions in literary texts (Acerbi et al., 2013) and general characteristics of modern societies (Roth, 2014), to name a few. Nevertheless, 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 (Pechenick et al., 2015) to the lack of metadata in the corpus (Koplenig, 2017).

Leetaru (2011) proposes a somewhat complementary approach that he calls *culturomics 2.0*, which employs computational analysis of large text archives composed of historical news data (instead of books) and can, according to the author, "yield intriguing new understandings of human society". The author performed sentiment mining and full-text geocoding in order to offer "new insights into how the world views itself and the 'natural civilizations' of the news media" (Leetaru, 2011). In the same vein, Flaounas et al. (2010) analyze the European mediasphere and the writing style, gender bias and the popularity of particular topics (Flaounas et al., 2013) in large corpora of news articles. Lansdall-Welfare et al. (2014), also using a large dataset of media reports, observe a change of framing and sentiment associated with nuclear power after the Fukushima nuclear disaster from 2011. The authors detected effects on attention, sentiment, conceptual associations and in the network of actors and actions linked to nuclear power following the accident. Our work draws a lot of inspiration from this study, as some of the methods (such as the analyses of textual polarity and co-occurring named entities) are similar. Also, Lansdall-Welfare et al.'s investigation contains a temporal element as well, since it compares the media coverage of nuclear power before and after the Fukushima disaster. Nonetheless, the focus of our proposal is the analysis of specific linguistic items, which is the reason why we consider our framework a contribution to the diachronic study of the lexicon from a corpus linguistics perspective.

As mentioned earlier, this chapter extends a previously published work (Cunha et al., 2018) in which we investigate the conceptualization of the term *fake news* in English (i.e., not considering Brazilian news sources as well). After this publication, we also used this incipient framework to analyze news articles that mention the name of the software application WhatsApp in Brazil and in parts of the English-speaking world (Caetano et al., 2018). Among the results obtained, we show that WhatsApp started to be linked to misinformation, politics and criminal scams in 2018. The use of the methodology proposed here for another case study, which also corroborated observations provided by other researchers, shows that our framework is robust enough to be employed in a variety of different contexts. As far as we are concerned, our investigations are the first studies that use a combination of already established methods and tools from corpus linguistics and natural language processing in order to quantitatively examine the history of relevant terms related to technology and online social media, thus helping us to better understand social trends in a fast-changing world.

4.2 Data sources

We use two diachronic corpora of news articles in this study: the first one comes from the Corpus of News on the Web (NOW Corpus), while the second one is a collection of news gathered from Brazilian online newspapers and magazines.

4.2.1 English-written news corpus

The Corpus of News on the Web (NOW Corpus) contains articles written in English and published from 2010 to the present time[8] in online newspapers and magazines based in 20 different countries (Davies, 2013, 2017b). At the time of writing of this study, this corpus is available for download and online exploration[9]. Our analyses are relative to a version of the corpus available in the month of April 2018, containing around six billion words of data.

Using the NOW Corpus exploration tool, we searched for all the occurrences of the term *fake news*. For each occurrence, the online tool provides a concordance line, or *context* – that is, a piece of text of approximately 20-30 words around (before and after) the

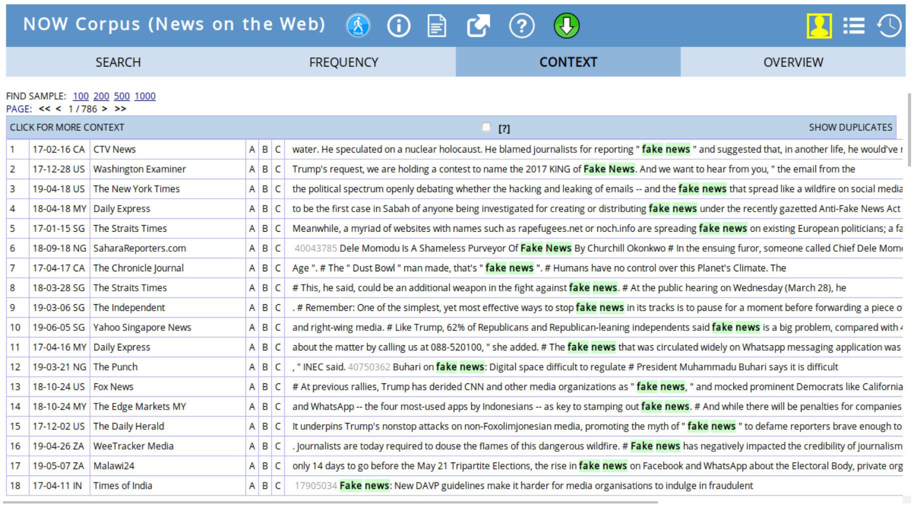
[8] As of the writing of this dissertation, the NOW Corpus is updated daily. It is, thus, a *monitor corpus* (also called a *dynamic corpus*), i.e., a corpus that is "continually growing over time, as opposed to a **static corpus**, which does not change in size once it has been built" (Baker et al., 2006, p. 64, emphasis in original).

[9] At `https://www.english-corpora.org/now/` .

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 (...).* For illustration, Figure 4.3 shows a small selection of contexts[10] including *fake news* in the NOW Corpus online exploration tool. 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 – as put by Baker et al. (2006), "concordances provide information about the 'company that a word keeps' " (p. 43). Wynne (2008) 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" (p. 711).

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 then grouped into the following six regions based on their geographical locations: Africa, British Isles, Indian subcontinent, Oceania, Southeast Asia and the Americas. Some countries with very different histories and cultures were included in the same group – for example, Nigeria and South Africa were grouped together under the label *Africa*, as were India and Pakistan under the label *Indian subcontinent.* This is obviously a result of the simplification needed to carry out the study proposed here. Researchers interested in further analyzing a specific part of the world should of course take into account the differences between

[10] A small selection of random concordance lines is also called a *thinned concordance* (Baker et al., 2006).



1	17-02-16 CA	CTV News	A	B	C	water. He speculated on a nuclear holocaust. He blamed journalists for reporting " **fake news** " and suggested that, in another life, he would've r
2	17-12-28 US	Washington Examiner	A	B	C	Trump's request, we are holding a contest to name the 2017 KING of **Fake News**. And we want to hear from you, " the email from the
3	19-04-18 US	The New York Times	A	B	C	the political spectrum openly debating whether the hacking and leaking of emails -- and the **fake news** that spread like a wildfire on social media
4	18-04-18 MY	Daily Express	A	B	C	to be the first case in Sabah of anyone being investigated for creating or distributing **fake news** under the recently gazetted Anti-Fake News Act
5	17-01-15 SG	The Straits Times	A	B	C	Meanwhile, a myriad of websites with names such as rapefugees.net or noch.info are spreading **fake news** on existing European politicians; a fa
6	18-09-18 NG	SaharaReporters.com	A	B	C	40043785 Dele Momodu Is A Shameless Purveyor Of **Fake News** By Churchill Okonkwo # In the ensuing furor, someone called Chief Dele Mom
7	17-04-17 CA	The Chronicle Journal	A	B	C	Age ". # The " Dust Bowl " man made, that's " **fake news** ". # Humans have no control over this Planet's Climate. The
8	18-03-28 SG	The Straits Times	A	B	C	# This, he said, could be an additional weapon in the fight against **fake news**. # At the public hearing on Wednesday (March 28), he
9	19-03-06 SG	The Independent	A	B	C	. # Remember: One of the simplest, yet most effective ways to stop **fake news** in its tracks is to pause for a moment before forwarding a piece o
10	19-06-05 SG	Yahoo Singapore News	A	B	C	and right-wing media. # Like Trump, 62% of Republicans and Republican-leaning independents said **fake news** is a big problem, compared with
11	17-04-16 MY	Daily Express	A	B	C	about the matter by calling us at 088-520100, " she added. # The **fake news** that was circulated widely on Whatsapp messaging application was
12	19-03-21 NG	The Punch	A	B	C	, " INEC said. 40750362 Buhari on **fake news**: Digital space difficult to regulate # President Muhammadu Buhari says it is difficult
13	18-10-24 US	Fox News	A	B	C	# At previous rallies, Trump has derided CNN and other media organizations as " **fake news**, " and mocked prominent Democrats like California
14	18-10-24 MY	The Edge Markets MY	A	B	C	and WhatsApp -- the four most-used apps by Indonesians -- as key to stamping out **fake news**. # And while there will be penalties for companies
15	17-12-02 US	The Daily Herald	A	B	C	It underpins Trump's nonstop attacks on non-Foxolimjonesian media, promoting the myth of " **fake news** " to defame reporters brave enough to
16	19-04-26 ZA	WeeTracker Media	A	B	C	. Journalists are today required to douse the flames of this dangerous wildfire. # **Fake news** has negatively impacted the credibility of journalism
17	19-05-07 ZA	Malawi24	A	B	C	only 14 days to go before the May 21 Tripartite Elections, the rise in **fake news** on Facebook and WhatsApp about the Electoral Body, private org
18	17-04-11 IN	Times of India	A	B	C	17905034 **Fake news**: New DAVP guidelines make it harder for media organisations to indulge in fraudulent

Figure 4.3: A small selection of contexts including *fake news* in the NOW Corpus online exploration tool.

its specific regions. In any case, this grouping was motivated by the fact 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 (Berger, 2009). To illustrate the outcome of this division, we show, in Figure 4.4, a map highlighting the countries considered in the English-written corpus, grouped by region.

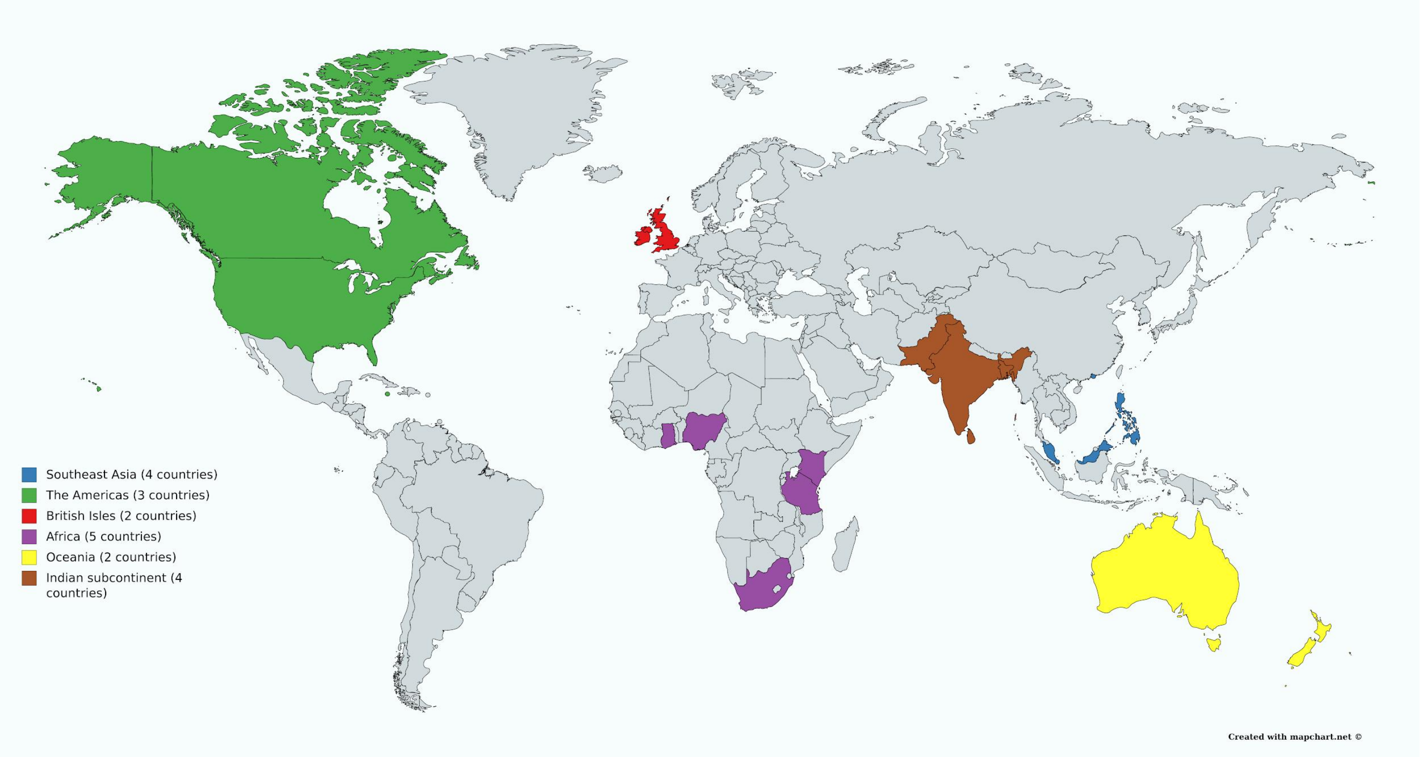

🟦 Southeast Asia (4 countries)
🟩 The Americas (3 countries)
🟥 British Isles (2 countries)
🟪 Africa (5 countries)
🟨 Oceania (2 countries)
🟫 Indian subcontinent (4 countries)
Created with mapchart.net ©

Figure 4.4: Map highlighting the countries considered in the English-written news corpus, grouped by region.

These occurrences also cover each year in the corpus until our data collection (i.e., 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 Section4.1), we categorized the occurrences into 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 4.1 and Table 4.2 show, respectively, the number of contexts containing the term *fake news* in this corpus according to the geographical origin of the corresponding news media and to the year and period of publication of the news article.

4.2.2 Brazilian news corpus

Our second data source includes articles collected from Brazilian online newspapers and magazines, all written in Portuguese, also containing the term *fake news*. To collect this material, we used the tool *Selenium*[11] to automate exact match searches for the term *fake news* in the following ten major Brazilian news websites: *Exame*, *Folha de S. Paulo*, *Gazeta do Povo*, *G1*, *O Estado de S. Paulo*, *R7*, *Terra*, *Universo Online* (*UOL*), *Valor Econômico* and *Veja*. The total number of occurrences of *fake news* extracted from these websites on December 31, 2018 is 4,936. These occurrences appeared in 2,464 unique news articles. Then, we used the Python library *newspaper*[12] to collect the full texts of the news articles containing these occurrences. Finally, we gathered the fifteen words before and after the key term *fake news* to create the contexts/concordance lines that will be analyzed in the next sections.

[11] Available at `https://www.seleniumhq.org/` .

[12] Available at `https://pypi.org/project/newspaper/` .

Table 4.1: Number of contexts containing the term *fake news* in the English-written news corpus according to the geographical origin of the corresponding 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
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

Table 4.2: Number of contexts containing the term *fake news* in both English-written and Brazilian news corpora according to the year and period (before or after the 2016 and 2018 presidential elections in the United States of America and Brazil, respectively) of publication of the news article.

English-written news corpus										
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: 494; after US election: 40,630								
Brazilian news corpus										
Year	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Occurrences	1	4	0	0	10	0	3	4	158	4,756
Period	before BR election: 923; after BR election: 4,013									

Similarly to what was observed in the 2016 US election, the increase in the usage of the term *fake news* in Brazil seems to correspond to the 2018 Brazilian general election (see Section 4.3.1). For this reason, we also categorized the occurrences in this dataset into two periods: before and after the 2018 BR election. The election was held in October, but, again, we set the delimitation date between these periods in the end of the first semester of the year (June 30, 2018) to include the political campaign in the period *after BR election*. Table 4.2 also shows the number of contexts containing the term *fake news* in the Brazilian news corpus according to the year and period of publication of the news article.

It is noteworthy to mention the increase in the frequency of the expression *fake news* in the English-written corpus from 2015 (95 occurrences) to 2016 (4,766 occurrences), and in the Brazilian corpus from 2016 (4 occurrences) to 2017 (158 occurrences) and then to 2018 (4,756 occurrences).

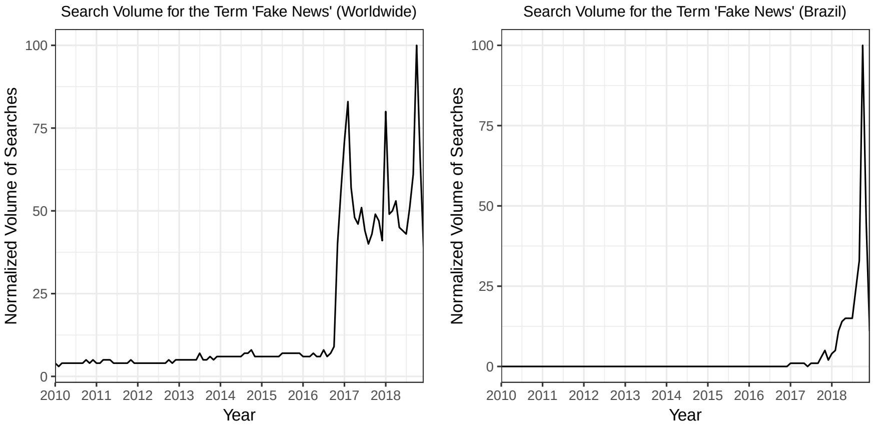
4.3 Analyses and results

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.

4.3.1 Web search behavior

Before analyzing the data obtained from our English-written and Brazilian news corpora, we investigate whether it is possible to observe changes in Web search behavior regarding the expression *fake news* corresponding to the high increase in its use during and after the 2016 and 2018 elections in the United States of America and in Brazil, as mentioned in Section 4.1 and observed in Table 4.2.

Data obtained from Google Trends[13], an online tool that indicates the frequency of particular terms in the total volume of searches in the Google Search engine, informs that, according to this metric, the worldwide public interest in the term *fake news* was approximately constant from 2010 until mid 2016, when it greatly and suddenly increased, as indicated by Figure 4.5a. This corresponds to the period of the campaign for the 2016 US presidential election. When we examine Web searches for the term *fake news* in Brazil over time (Figure 4.5b), we also observe a significant increase. However, in this case, the most noteworthy growth happened in mid 2018, which also corresponds to the period of a major political event: the campaign for the 2018 Brazilian general election.



(a) Search volume worldwide

(b) Search volume in Brazil

Figure 4.5: Normalized volumes of searches for the expression *fake news* on Google Search from 2010 to 2018. Values represent search volumes relative to the highest point on the chart. A value of 100 is the peak popularity for the term.

[13] `https://trends.google.com/trends/` .

Google Trends data also show a spatial change regarding queries for the term *fake news*. The ten countries with the highest proportions of searches for *fake news* in each period (before and after the 2016 US presidential election) are listed in Table 4.3. In the period before the 2016 US election, a significant part of the countries with the highest proportions of searches are from the Eastern world (India, United Arab Emirates, Singapore, Qatar, Pakistan). However, after the US election, the proportion of searches for this expression in Western countries increased considerably, especially in Europe (Norway, Denmark, Ireland, United Kingdom, Switzerland). We are not able to provide an explanation for such observation.

Table 4.3: Countries with the highest proportions of searches for *fake news* on Google Search before and after 2016 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

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 4.4 shows the five most frequent search terms employed by users who also searched for *fake news* in the periods before and after the 2016 US election. We observe that, before the election, searches for *fake news* were generic and regarded terms related to the media industry itself, like *article*, *stories* and *report*; after the election, however, these searches started to be more focused on political affairs and in the spread of false information, 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 online social media Facebook (sometimes considered a major source of fake news on the Internet).

Table 4.4: Most frequent search terms related to *fake news* on Google Search before and after US and BR elections.

Period	**Search terms**
Worldwide	
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
Brazil	
before BR election	fake news redação, o que fake news, fake news brasil, fake news o que é, fake news significado
after BR election	tse fake news, fake news eleições, fake news kit gay, redação enem 2018, kit gay

Table 4.4 also displays the most frequent terms queried by users who also searched for *fake news* in Brazil, but now in the periods before and after the 2018 Brazilian general election. Before the election, most of the top searches concerned the meaning of the expression *fake news* itself, such as *o que* and *o que é* (i.e., *what* and *what is it*), and *significado* (i.e., *meaning*). This is understand-

able, since a significant part of the Brazilian population does not speak English. During and after the campaign, however, the focus changed, again, mostly to queries related to politically related terms, such as *tse* (acronym for the Brazilian Superior Electoral Court) and *eleições* (i.e., *elections*), and to campaign controversies (*kit gay*[14]).

In this section, we used data obtained from the Google Trends tool. From now on, all of our analyses use the data described in Section 4.2, obtained from the NOW Corpus and from our collection of Brazilian news articles.

4.3.2 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 (Tjong Kim Sang and De Meulder, 2003) – co-occurring with certain terms is an interesting way to contextualize these terms. 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 corpora 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 con-

[14] The "gay kit controversy" was one of the most contentious topics during the Brazilian presidential campaign of 2018. In short, far-right candidate Jair Bolsonaro accused center-left candidate Fernando Haddad of planning to distribute "gay kits" in schools – a reference to sexual education materials that, according to him, were aimed to "pervert" youngsters and encourage homosexuality. At some point, the Superior Electoral Court considered this information a piece of fake news and ordered Bolsonaro to remove it from his campaign. For more information on fake news in Brazilian politics, see Harden (2019).

texts; then, we manually analyzed the most frequent capitalized words in each subdivision of the corpora (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*). This "semi-manual method" proved to be more effective than the use of automatic named entity recognition tools probably because of the lack of completeness of the contexts analyzed – which ignore sentence boundaries and punctuation, and may start and finish in indiscriminate positions of the texts (cf. examples of context in Section 4.2.1).

Table 4.5 shows the five most mentioned named entities in the English-written news corpus 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 (The Daily Show, Onion News Network) and hosts (Jon Stewart, Stephen Colbert) 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 4.4) shift of interest towards political affairs and the spread of fake news on the Internet observed in Web searches.

Table 4.5 also displays the five most mentioned named entities in the Brazilian news corpus in the periods before and after the 2018 Brazilian general election. Here, before the election, we observe the presence of entities linked to the upcoming electoral process, such as TSE (the Brazilian Superior Electoral Court) and its former president Luiz Fux. These two entities are highly mentioned

Table 4.5: Most mentioned named entities in the periods before and after US and BR elections.

Period	**Entities**
English-written news corpus	
before US election	The Daily Show, Jon Stewart, Onion News Network, Facebook, Stephen Colbert
after US election	Donald Trump, Facebook, US, CNN, Twitter
Brazilian news corpus	
before BR election	TSE, Luiz Fux, Facebook, Donald Trump, Brasil
after BR election	TSE, Jair Bolsonaro, Brasil, WhatsApp, Folha de São Paulo

due to Fux's declaration (in June 2018) concerning the possibility of annulment of the election in case of massive fake news influence (Ramalho, 2018). Donald Trump is also mentioned, probably due to influences of the international scenario. Interestingly, after the start of the campaign period, candidate Jair Bolsonaro takes the place of Donald Trump and the online social service WhatsApp replaces Facebook, as a clear reflection of the Brazilian scenario in 2018, more affected by political fake news disseminated through WhatsApp than through Facebook – as alleged by the fifth most mentioned entity, the major newspaper Folha de S. Paulo (Phillips, 2018).

It is particularly interesting to compare the most mentioned named entities in the English-written news corpus after the US election with the most mentioned named entities in the Brazilian news corpus after the BR election. In both cases, we observe the presence of: (a) the name of the country (*US* and *Brasil*); (b) the conservative (and winning) candidate (*Donald Trump* and *Jair Bolsonaro*); (c) social networking services (*Facebook* and *Twitter*, and *WhatsApp*); and (d) a traditional media source (*CNN* and *Folha de S. Paulo*). This fact shows that, although in different scenarios and times, many similarities still emerge.

Table 4.6: 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

When we make this same diachronic comparison (*before vs. after* the elections), but now considering the geographical origin of the corresponding news media in the English-written corpus, we observe a noteworthy phenomenon: the global standardization of the named entities related to *fake news*. Table 4.6 shows the three most mentioned entities in the periods before and after US election in each region, and 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.

4.3.3 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. According to Cunha et al. (2014b), "vocabulary is a system of mapping the world, so this kind of investigation reveals how groups perceive reality" (p. 215). 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 (Bird et al., 2009) 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* (Fast et al., 2016), "a tool for analyzing text across lexical categories"[15], to classify the lemmatized words according to categories that represent different semantic fields, such as diverse topics and emotions. For every context, 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*. The complete lists of words that comprise each one of these categories are displayed in Appendix B. Since Empath is (at the moment of the writing of this dissertation) available only in English, the corpus containing Brazilian news articles was not included in this analysis. In future work, it might be possible to use alternative tools, such as the multilingual Linguistic Inquiry and Word Count – LIWC (Pennebaker et al., 2015), to also consider the Brazilian news corpus in this analysis.

Figure 4.6 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 unre-

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

lated 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.

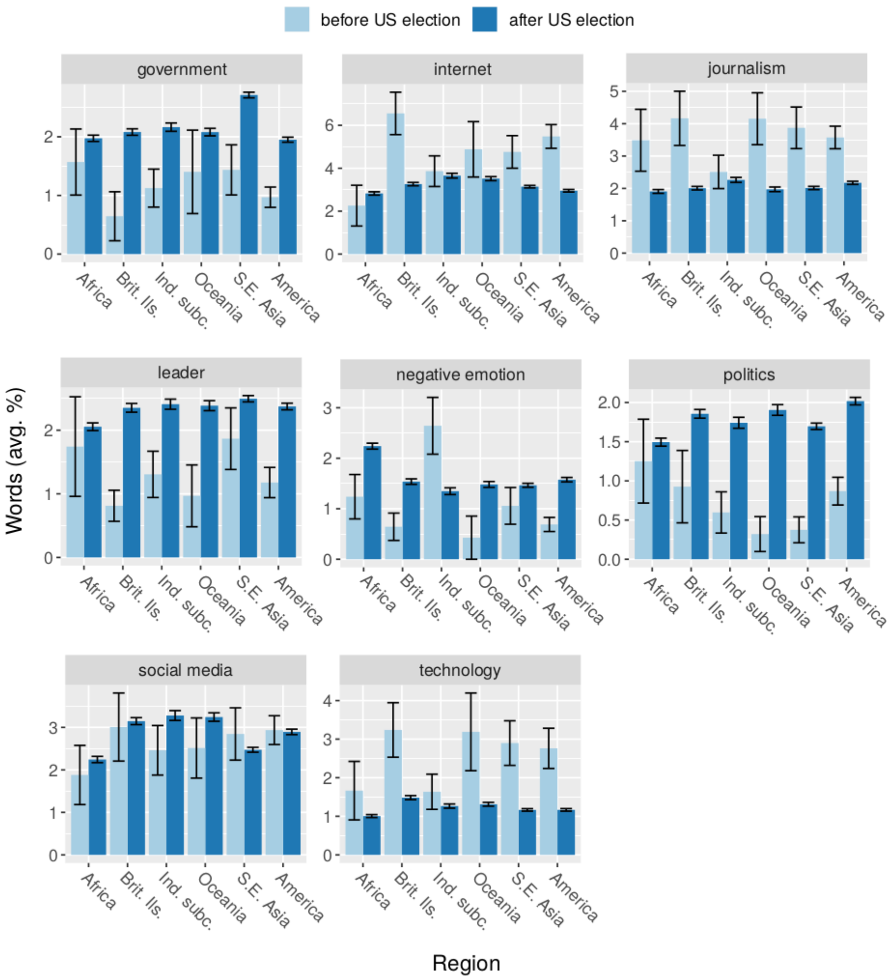


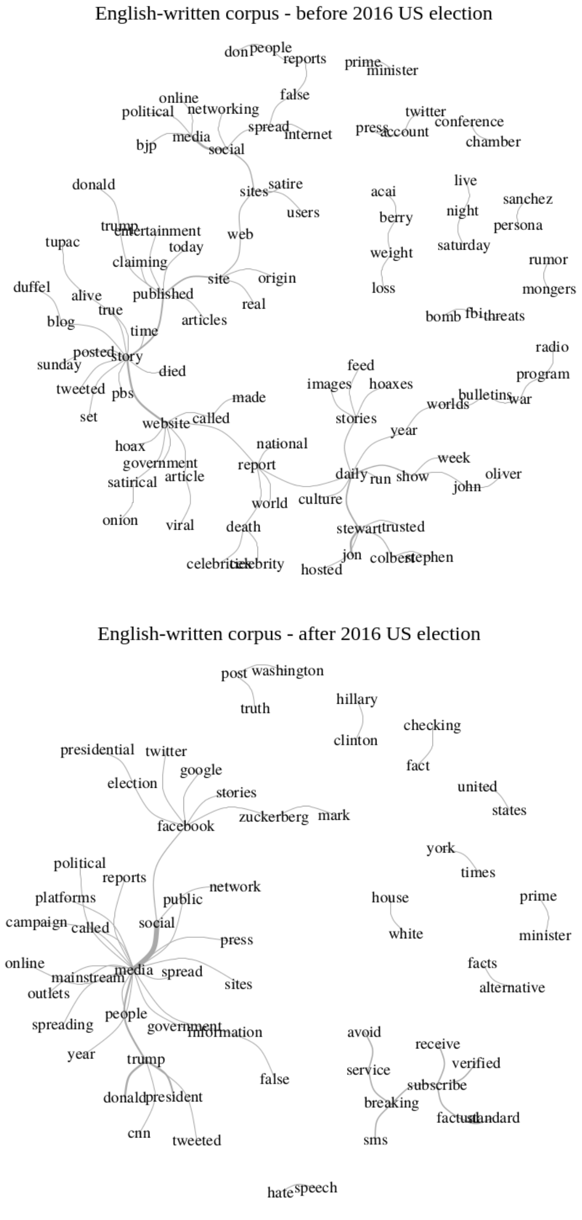
Figure 4.6: Percentage of words in each semantic field represented by an Empath category. Error bars indicate standard errors.

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.).

4.3.4 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 considered. Here we compare co-occurrence networks between the periods before and after the elections. These networks are represented by graphs, in which each node corresponds to a word and each (weighted) edge corresponds to an association between two given words.

To build our graphs of co-occurring words, we followed the steps below. First, we removed stop words using the lists provided by the Natural Language Toolkit (Bird et al., 2009) for English and Portuguese (the words *fake* and *news* were included in these lists as well, since they are present in all contexts). Then, we extracted the most relevant words from each period by using the *term frequency-inverse document frequency* (*tf-idf*) technique, that reflects how important a word is to a document in a corpus (Rajaraman and Ullman, 2011). We calculated the tf-idf for each pair (period, word) and extracted from each period the top 50 words with the highest tf-idf scores. In the following step, we counted the number of co-occurrences of each pair of words. For each period, we obtained the list with its 50 most relevant words (according to tf-idf) and incremented by one the counter relative to each pair of words in



English-written corpus - before 2016 US election

English-written corpus - after 2016 US election

Figure 4.7: Co-occurrence networks before and after the 2016 US election in the English-written corpus.

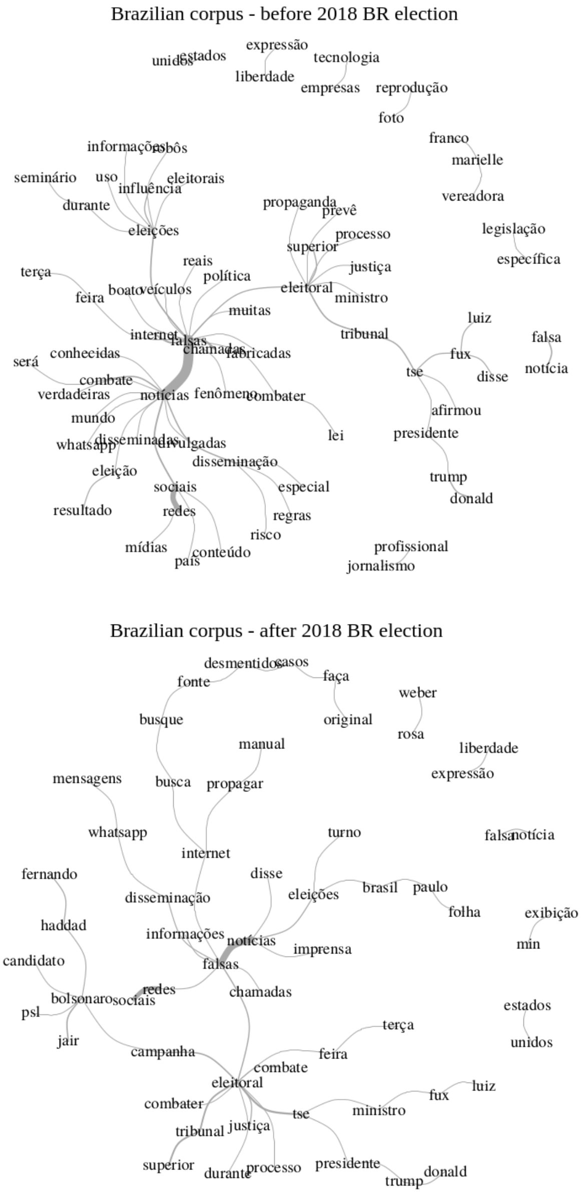


Figure 4.8: Co-occurrence networks before and after the 2018 BR election in the Brazilian corpus.

this list (combination two by two). Instead of using the absolute count of contexts in which two words co-occur, we normalized this value by dividing it by the total number of contexts. At the end of this process, we obtained a graph in which vertices represent words and weighted edges indicate their degree of co-occurrence in the same contexts. Finally, we selected the top 100 edges with the highest weights, calculated the maximum spanning tree out of the remaining graph, and generated trees that depict the most relevant relationships, which are presented in Figures 4.7 (for the English-written corpus) and 4.8 (for the Brazilian corpus).

This method of investigation enables us to make several qualitative observations, which can be further elaborated in specific studies for this purpose. Here we will just draw attention to some clusters that seemed interesting to us. Comparing the two graphs representing the English-written news corpus, we notice, for instance, that before the US election one of the main clusters contains words related to the news industry itself (*articles*, *published*, *story*) and to the Internet (*website*, *tweeted*, *blog*, *posted*). Corroborating previous findings (Section 4.3.2), there is also a cluster containing words referring to satirical TV shows and hosts (*daily*, *show*, *colbert*, *oliver*, *stewart*). There are also a few words related to the political world, including *donald*, *trump* and *government*, and to minor topics, such as a cluster on the use of açaí berry to lose weight. In the graph representing the period after the 2016 US election, we start to observe more terms specifically linked to the US election itself. In the main cluster, *donald*, *trump* and *president* have an important role; *presidential*, *election*, *political* and *campaign* also appear. The thickest edge, however, is between the words *social* and *media*, which appear close to *spread* – suggesting, then, the relationship between social media and the spread of fake news. Some terms that surround meta-discussions about fake news are also present, highlighting relevant related concepts such *alternative facts*, *fact checking*, *hate speech* and *post truth*.

In the Brazilian news corpus, before the 2018 election, most of the relevant co-occurring words in the main cluster are already related to the battle (*combate*) against the dissemination (*disseminação*) of fake news (*notícias*, *falsas*, *fabricadas*, *boatos*), including the battle officialy promoted by the justice due to the upcoming elections (*tse*, *eleitoral*, *ministro*, *fux*). The use of bots (*robôs*) in the elections is also mentioned. The scenario in the United States (*estados*, *unidos*) is also relevant, as seen by the cluster that mentions *presidente*, *donald* and *trump*. During and after the election, however, the co-occurring words in the main cluster regard mostly the campaign itself. We highlight the names of the main presidential candidates (*jair*, *bolsonaro*, *fernando*, *haddad*) and the focus on social networks (*redes*, *sociais*), especially on WhatsApp messages (*whatsapp*, *mensagens*), since this tool was considered the main platform for disinformation during the 2018 elections in Brazil (Bradshaw and Howard, 2018).

4.3.5 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 corpora. For this task, we used *latent Dirichlet allocation* (LDA) (Blei et al., 2003), a way of automatically discovering topics in texts. LDA generates summaries of topics in terms of the keywords relevant for each topic, i.e., it returns a set of keywords that illustrates each topic alluded to in the text.

To perform this analysis, we first lowercased and tokenized all the words in both corpora. 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* (Řehůřek and Sojka, 2010), a Python library for topic modeling. We used topic coherence score (Newman et al., 2010) 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, and Brazil, that had k=16 and k=19. For each region, the LDA returned these k topics containing keywords ordered by importance in the corresponding context, filtered both by region and topic. We then selected the main topic as the representative of each region and period.

Table 4.7 shows the top ranked ten keywords produced by our LDA model that represent the main topic in each region in both English-written and Brazilian media sources, before and after the elections. In this case, the analysis of the LDA output is performed subjectively, by observing and comparing keywords that are representative of each topic. Sometimes, however, the definition of a topic is not very clear. Nevertheless, we can find a few elements that seem to corroborate previous findings of this chapter. Regarding the English-written corpus, for example, we observe, for all regions, a relevant frequency of keywords related to journalism, media and the publishing industry in the period before the US election, like *story*, *website*, *site*, *report* and *article*. In the period after the US election, none of these keywords appears anymore, and we can find examples of keywords linked to politics (like *trump*, *president*, *politics*, *election*, *presidential*, *public* and *government*) and, to a lesser extent, to online social media (like *facebook*, *zuckerberg* and *twitter*). The region that more clearly displays this shift is probably Southeast Asia, whose top keywords in the most relevant topic before 2016 US election are literally *article*, *website*, *story*, *report* and *site*, and the

Table 4.7: Main topic for each region and period. For each topic, ten keywords are presented ordered according to the LDA output.

Region	**Period**	**Main topic keywords**
English-written news corpus		
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
Brazilian news corpus		
before BR election		eleição, americana, redes sociais, propaganda, lado, ganhar, evitar, importante, reporter, seriedade
after BR election		imprensa, redes sociais, influenciar, verdade, proliferação, democracia, candidato, procuradora geral, votar, congresso

top keywords in the most relevant topic in the period after the election are *public*, *government*, *fact*, *twitter* and *proliferation*. These observations corroborate our findings from previous sections, where we found, for instance, a change in the focus of the news mentioning *fake news*: the focus changed from journalism and media to politics and online social networks.

In the Brazilian corpus, we observe that keywords related to the US election (*eleição* (i.e., *election*), *americana* (i.e., *american*)) rank high before the BR election, while after the election we note keywords related to information diffusion (*influenciar* (i.e., *to influence*), *proliferação* (i.e., *proliferation*)) and to the elections (*votar* (i.e., *to vote*), *candidato* (i.e., *candidate*)). It turns out, therefore, that the scenario in the United States was an important topic in the news that mentioned *fake news* in Brazil until mid-2016, but then the most important topic became the proliferation of fake news in the Brazilian electoral scenario itself.

To get a better sense of the results that LDA can generate, in future work it might be interesting to explore alternative possibilities for visualizing topics and keywords, and also to consider more than one topic per region and period.

4.3.6 Polarity

Our final analysis explores a different feature of the contexts in which the expression *fake news* appear in our corpora: their *polarities*, that is, whether the expressed opinion in the texts is mostly positive, negative or neutral. Here we performed sentiment analysis[16] (Silva et al., 2016) in each one of the contexts using *SentiStrength*[17] (Thelwall et al., 2010), a tool able to estimate the

[16] "Sentiment analysis is a type of natural language processing for tracking the mood of the public about a particular product or topic" (Vinodhini and Chandrasekaran, 2012, p. 282).

[17] Available at `http://sentistrength.wlv.ac.uk/` .

strength of positive and negative sentiment in short texts based on their vocabulary. Given a piece of text, this tool returns a score that varies from -4 (negative sentiment) to +4 (positive sentiment). Among the several possible tools for analyzing sentiment, we chose SentiStrenght mainly due to two reasons: first, it was developed especially for short texts, like the contexts analyzed here; second, it is available for both English and Portuguese, so we could use the same tool to calculate polarities in our two corpora[18].

The first six graphs in Figure 4.9 depict the average polarity of the contexts in each region of the English-written corpus before and after the 2016 US presidential election, while the rightmost graph shows the average polarity of the contexts in the Brazilian corpus before and after the 2018 Brazilian general 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 (Zollo et al., 2015) and sentiments – which is not surprising, since the idea of fake news seems to be strongly associated with negative concepts like misinformation, manipulation and hostility.

In the charts concerning the English-written corpus, 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 relevant. The main message that we can draw from these results is that media texts on fake news are normally negative, but that the post-election ones are even more. In fact, texts mentioning fake news in politics often involve deception, fraud and accusations. In the graph regarding the Brazilian corpus, the difference of polarity between before and after the 2018 Brazilian general election is almost unnoticeable. Besides that, the

[18] Even though it is not possible to compare polarity values between different languages.

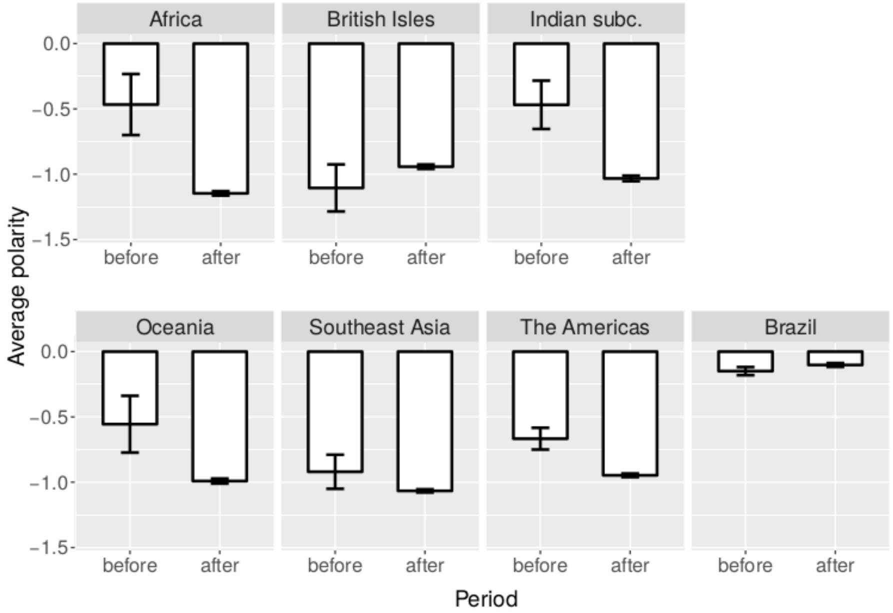


Figure 4.9: Average polarity of the contexts in each region before and after 2016 US election and 2018 BR election. Bars indicate the standard error of the mean.

average polarity in both periods is, albeit also negative, much less negative than in all regions considered in the English-written corpus. The difference between polarities in different languages, however, is not relevant, since sentiment analyses in distinct languages are not completely comparable between themselves.

4.3.7 Summary of results

In this section, we present a series of independent analyses on the texts that accompany the term *fake news* in English-written and Brazilian news. The most relevant outcomes of these analyses can be summarized and integrated as follows:

- globally, the interest for the term *fake news* suddenly increased after the 2016 US election, and in Brazil this interest increased after the 2018 Brazilian election, as indicated by the rise of news mentioning it and of Google Search queries for this expression (Section 4.3.1);

- 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, both in English and Portuguese (Sections 4.3.1, 4.3.3, 4.3.4, 4.3.5);
- 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 (Section 4.3.2);
- in English, the negativity of the news containing the term *fake news* increased after the US election (Section 4.3.6).

All these results suggest that, as hypothesized in Section 4.1, the rise of public interest in the term *fake news* brought changes to its conceptualization and to the perception about it.

4.4 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.

Here we present a quantitative analysis on the perception and conceptualization of the term *fake news* in two corpora of news articles published from 2009 to 2018 in 21 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 and 2018 presidential elections in the United States of America and Brazil, respectively, 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 the topics related to this concept and in the polarity of the texts around it, 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. We understand that, in this way, our study joins the scholarship that "contributes to the nascent debate on the concept of fake news" (Gelfert, 2018, p. 85) both as a linguistic term and, to a lesser extent, as a social phenomenon.

From a journalism studies perspective, our findings come as no surprise: the semantic shift of *fake news* from news satire to political propaganda has already been identified by political scientists and journalism scholars using qualitative methods (see Section 4.1.2). Their findings however, serve to validate the diachronic method employed here. Our study contributes to the literature on the term *fake news* as it is used in news media by adding quantitative data and geographical scope – most of the previous studies focus on the United States of America.

More than just analyzing an isolated case, though, our intention in this chapter is to present an analytical framework that can be applied and replicated in other situations. The basis of our proposal is the diachronic investigation of vocabulary from a "holistic" view, that is, combining and mixing different approaches in order to understand the phenomenon of semantic change from different perspectives. The same investigations implemented here can be performed in the most diverse contexts, using different items as key terms. As an example, we mention the study carried out by Caetano et al. (2018), which was based on the investigations presented here and used as key term the word *whatsapp*. What is important is that the corpora for analysis have a temporal component, so that different periods can be compared and contrasted with each other, thus suggesting changes (or maintenance) in the conceptualization of that key term over time.

Each of the methods employed here has its own features, advantages and disadvantages. The analysis of the Web search behavior (volume of searches, related queries etc.) regarding a key term (Section 4.3.1) allows the measure of the interest in a particular topic employing users' behavior on search engines as a proxy. The main advantage of this method is its simplicity. However, it is important to remember that search engine users are unlikely to be a statistically adequate sample of a population. Concerning our different approaches to investigate the text co-occuring with a key term, it is interesting to note how each approach complements the others. The analysis of co-occurring named entities (Section 4.3.2) makes it possible to identify people, places and institutions related to the key term. This is part of the issue, but it leaves out non-entities related to the term, which can be covered by the more generic analysis of the semantic fields of the surrounding vocabulary (Section 4.3.3). In addition, the investigation on co-occurrence networks (Section 4.3.4), although less clear from an interpretative point of view, allows the relationships between words to become more evident, and supports different and complementary interpretations that add to the simple analysis of co-occurring word lists. None of these approaches, however, deals with the *topic* of the contexts in which the key terms are inserted. For this reason, we include topic analysis (Section 4.3.5) – which is, on the one hand, more complex to be performed; but, on the other hand, is able to capture an aspect present at a higher level of analysis than lexical co-occurrence examinations. Finally, the polarity analysis (Section 4.3.6) has the characteristic of being generic and orthogonal to the semantic content of the text, since it can, independently of the topic, evaluate the level of positivity/negativity in the contexts in which the key term is present. It is, thus, the only sentiment-oriented method proposed in our framework, which complements those previously described. The idea of working with all these methods at the same time is precisely to be able to merge their qualities and thus get a broader view of the studied phenomenon.

In this chapter, we analyzed the usage of the term *fake news* in a diachronic perspective, but, for each corpus, only considered two historical moments: before and after a key event in the history of this expression (in English, the 2016 US presidential election; in Brazilian Portuguese, the 2018 Brazilian general election). In the future, we plan to consider a larger spectrum of periods, in order 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, so to observe the use of this term by different actors of the society.