UNIVERSIDAD DE MÁLAGA

TESIS DOCTORAL

POR COMPENDIO DE PUBLICACIONES

The Fabric of Digital Discourse:

Analytical Techniques for Social Media Corpora

María García Gámez

Director:

Dr. Antonio Jesús Moreno Ortiz

Facultad de Filosofía y Letras

Programa de Doctorado en Lingüística, Literatura y Traducción

Málaga, 2024

AUTORA: María García Gámez

https://orcid.org/0000-0001-7068-0038

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA PARA OBTENER EL TÍTULO DE DOCTOR

D./Dña MARÍA GARCÍA GÁMEZ
Estudiante del programa de doctorado LINGÜÍSTICA, LITERATURA Y TRADUCCIÓN de la Universidad de Málaga, autor/a de la tesis, presentada para la obtención del título de doctor por la Universidad de Málaga, titulada: THE FABRIC OF DIGITAL DISCOURSE: ANALYTICAL TECHNIQUES FOR SOCIAL MEDIA CORPORA

Realizada bajo la tutorización de ANTONIO JESÚS MORENO ORTIZ y dirección de ANTONIO JESÚS MORENO ORTIZ (si tuviera varios directores deberá hacer constar el nombre de todos)

DECLARO QUE:

La tesis presentada es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, conforme al ordenamiento jurídico vigente (Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), modificado por la Ley 2/2019, de 1 de marzo.

Igualmente asumo, ante a la Universidad de Málaga y ante cualquier otra instancia, la responsabilidad que pudiera derivarse en caso de plagio de contenidos en la tesis presentada, conforme al ordenamiento jurídico vigente.

En Málaga, a 22 de FEBRERO de 2024

Fdo.: MARÍA GARCÍA GÁMEZ Doctorando/a	Fdo.: ANTONIO JESÚS MORENO ORTIZ Tutor/a
Fdo.: ANTONIO JESÚS MORENO ORTIZ Director/es de tesis	

Málaga a 22 de febrero de 2024

Antonio Jesús Moreno Ortiz, profesor de la Facultad de Filosofía y Letras (Universidad de Málaga),

HACE CONSTAR

Que María García Gámez, con DN/NIE/pasaporte , es estudiante de doctorado del Programa de Doctorado "Lingüística, Literatura y Traducción", con matrícula activa, y que ha realizado bajo mi dirección, la Tesis Doctoral titulada

The Fabric of Digital Discourse: Analytical Techniques for Social Media Corpora

Revisado el presente trabajo estimo que reúne los requisitos establecidos según la normativa vigente. Por lo tanto, **AUTORIZO** la admisión a trámite y defensa pública de esta Tesis Doctoral para optar al grado de Doctor en la Universidad de Málaga.

Y para que así conste, lo firmo en Málaga a 22 de febrero de 2024,

Fdo. Antonio Jesús Moreno Ortiz

Antonio Jesús Moreno Ortiz, coautor de los siguientes artículos de investigación:

I. Moreno-Ortiz, A. & García-Gámez, M. (2022). Corpus Annotation and Analysis of Sarcasm on Twitter: #CatsMovie vs. #TheRiseOfSkywalker. *Atlantis. Journal of the Spanish Association of Anglo-American Studies*, *44*(1), 186-207. http://doi.org/10.28914/Atlantis-2022-44.1.11

II. Moreno-Ortiz, A. & García-Gámez, M. (2023). Strategies for the Analysis of Large Social Media Corpora: Sampling and Keyword Extraction Methods. *Corpus Pragmatics*, 1-25. https://doi.org/10.1007/s41701-023-00143-0

III. García-Gámez, M. & Moreno-Ortiz, A. (2024, en prensa). The Politics of Eurovision: A Case Study of the United Kingdom's 2021 and 2022 Participations as Expressed on Social Media. *Revista de Lingüística y Lenguas Aplicadas, 19*.

ACEPTA QUE

María García Gámez, alumna del Programa de Doctorado de Lingüística, Literatura y Traducción, presente dichos artículos como parte de su tesis doctoral por compendio. Asimismo, renuncio a utilizar estas publicaciones en futuras tesis doctorales en la Universidad de Málaga o en cualquier otra universidad.

Y para que conste y surta los efectos oportunos, lo firmo en Málaga, a 22 de febrero de 2024.

Dr. Antonio Jesús Moreno Ortiz

Acknowledgements

To my supervisor, Dr. Antonio Moreno (also known as *The Boss*), for his patience. When we started this academic adventure in 2019, I never thought I would be able to make it to the end with a sane mental health. But that was just a very bad year that was, in fact, followed by some of the best years in my life. Thank you for giving me all the resources you have, for trusting me even when I thought I could not do it, for being so critic with everything I do because you know I can do it even better. You have constantly pushed me to the edge, always taking me out of my comfort zone and watching me from the sidelines, waiting for me to solve the problem by myself. And sometimes I have done so, and sometimes I have not – yet you never felt disappointed (or at least you never showed it to me), you never ran out of patience with me. If I needed help, you would always be there. I am incredibly lucky to work with you.

To Dr. Rosario Arias, for showing me that a different, healthier, and more humane academia exists, and that I deserve a space in that academia. For not allowing me to give up, for backing me up, for telling me what to do at a time when I could not do anything by or for myself. Thank you for including me in everything you do.

To my parents, Margarita and Andrés, for respecting me every time I have said the words "por favor, no me preguntéis cómo va la tesis". These four years have been marked by a lot of anxiety and uncertainty about the future, and no, I still don't know what will happen after the thesis. The only reason why I am here is because you have supported me nonstop ever since I was a child. You bought me all the books I wanted, you read them to me, you were excited about everything I wrote. I was never the top of the class, but you always made me feel like I was. We have walked this path together, always dreaming of the final day, and it brings me so much joy to see you here with me. I love you.

To my sister and my brother-in-law, Laura and Juanje. You know none of this is the result of good luck – this is just a matter of work, work, patience… and more work. Thank you for coming to see me to Edinburgh, to Italy, to the end of the world. For moving heaven and earth every time I've had the smallest problem, for making me laugh every time you've said "so… do you want it to look like an accident?"

To my little nephew, Martín. I basically started this doctoral thesis when you were born, and ever since then I have been a little bit absent. I am very aware of all the times you have asked me to play together and I have had to say "I can't, I have to work, I'm so sorry". Finding a balance between my work and my life, or Málaga and Alhaurín, has not always been easy, and I wish I was better at that. I wish I knew how to be more present in your life without losing the main focus of mine. I hope someday you will understand why I was not there sometimes and feel as proud of me as I am of you.

To my grandparents. Abuela Frasquita, Abuelo Andrés, I hope you feel proud of me, I hope you can see the human being I've grown to be and believe it resembles your principles and values. Abuela Josefita, I see you in everything that is green, I hear you in the sound of music. I am one of the fruits of your will to learn, of your tenacity when it came to words and writing. Abuelo Gámez, life was so cruel it left us with very little time to learn from each other, but I am convinced I am so curious, analytical, and sarcastic because of you. I fully believe if you were here today you would (jokingly) have asked me to trade my thesis for two chickens. And I actually wish you were here to do that. I hope you all are in a safe place and feel happy, proud, and honoured to be my grandparents.

To Miguel Ángel, who blindly believes that I can do anything. Your trust in my capabilities has empowered me, and your patience when things go awfully wrong has provided me with a place where I can find shelter from the academic world. Your analytic

mind has saved me from more than a couple of rabbit holes, and now I finally know what you mean when you say that we must cross bridges when we get to them. Here's one of those bridges, one we get to cross together.

To Juanjo, Magdalena, Manu, María, Rosa, Paula, and Almudena, who are the best friends I could have dreamed of. You make academia a better, safer, happier, and more sensitive place. You are the safety net that allows me to launch myself and gain self-confidence. My support network. Thank you for listening me venting and ranting about every reviewer to have ever existed, for knowing so much about my personal life it's scary, yet keeping it a secret. I am so proud of this strong-as-hell friendship that we have created together.

To Carlos, Paloma, Vanessa, Jose, Bea, Alejandro, and Holly, for being so patient with my absences and deadlines and missed calls and impossible schedules. I have missed so many meetups, so many events, sometimes because of lack of time, sometimes because of distance, and sometimes because I got home so tired that I just wanted to sleep. Thank you for not having stopped writing or calling me, for having understood that this doctoral thesis required a lot of my time and that it was my duty to put it above all else. I am aware of how lucky I am to call you my friends.

“So, if you care to find me,
look to the western sky.
As someone told me lately,
‘everyone deserves the chance to fly.’
And if I’m flying solo,
at least I’m flying free.
To those who ground me,
take a message back from me!
Tell them how I am defying gravity,
I’m flying high defying gravity.”

“Defying Gravity” – Wicked

“I fell from the pedestal,
right down the rabbit hole,
long story short, it was a bad time.
Pushed from the precipice,
climbed right back up the cliff,
long story short, I survived.”

“Long Story Short” – Taylor Swift

Table of Contents

1 Introduction

Social media platforms have become widely used to exchange opinions and ideas in public, as users have gradually turned to them not only to stay in touch with others, but also to share their own thoughts, feelings, and beliefs with the world. Over 65% of the global population employs social media, and 85% of the world's mobile phone users have at least one social media account (Shewale, 2024a). Its immediacy goes beyond geographical boundaries, thus fostering a sense of interconnectedness that allows users to stay informed and participate in conversations shaping society. Whether disseminating breaking news, advocating for social causes, or expressing personal narratives, these platforms amplify voices, democratise communication, and contribute to the formation of a more inclusive and dynamic global discourse. As a conduit for dialogue, social media play a relevant role in shaping the narrative of our interconnected world, and multiple social phenomena would not be understood without it (Vásquez, 2022).

As global communication tools, the significance and impact of social media features is undoubtedly amplified by global phenomena. A paradigmatic case in point has been the COVID-19 pandemic, in which social media proved to be crucial tools for navigating the crisis. As global lockdowns were imposed, social media facilitated being connected to others, as well as being informed of what was happening in real time. Their role as a source of emotional support was heightened, and they served as a constructive coping strategy for individuals to reduce feelings of anxiety and loneliness, as they searched for solidarity, empathy, and shared experiences (Cauberghe et al., 2021).

Their importance as global communication tools have turned such platforms into an interesting resource for research in fields where public opinion is relevant, as they

provide access to very large and easily accessible samples of how speakers make their stances in a virtual environment.

This doctoral thesis by compendium aims to cast light on how corpora extracted from social media can be used from the perspective of sentiment and discourse analysis, and how the contents of social media can be approached as a representation of speakers' attitudes. To get a comprehensive understanding of this type of computer-mediated communication, we must begin by describing what social media are. However, the truth is that these communication tools are not so easily defined: Russo et al. (2008) present them as "those that facilitate online communication, networking, and/or collaboration" (p. 22), while Kaplan and Haenlein (2010) describe them as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content" (p. 61). Both these definitions, as noted by Carr and Hayes (2015), entail a problematic factor: they can also be applied to other types of technologies associated with communication (e.g., e-mail) that are unrelated to social media. To avoid such derived problems, they define social media as

> Internet-based channels that allow users to opportunistically interact and selectively self-present, either in real-time or asynchronously, with both broad and narrow audiences who derive value from user-generated content and the perception of interaction with others. (pp. 49-50)

What distinguishes this definition from others is that the authors, to begin with, point out the fact that social media require Internet access but are not necessarily Web-based, which allows us to distinguish them from other platforms that also require an Internet connection and which are Web-based, such as Wikipedia, Gmail, Skype, or Netflix, to name a few. Moreover, it references the possibility of an asynchronous type of communication, thus suggesting that two users need not be committed to communication at the same time, as opposed to face-to-face communication (Walther, 1996).

Carr and Hayes (2015) also point out the relevance of perceived interactivity: it is essential for users to *perceive* an interaction with others to consider a medium to be social. The reason why this interaction should only be perceived (and not necessarily true) is because they often won't receive an answer, yet they will still feel a social connection: for example, when users try to reach celebrities, they may not receive feedback but this will not diminish the feeling of closeness to them that can be achieved thanks to social media.

Masspersonal communication plays an important role in social media too. This concept, as reported by O'Sullivan and Carr (2018), represents instances where (a) mass communication channels are used for interpersonal communication, (b) interpersonal communication channels are used for mass communication, and (c) when individuals engage in in mass and interpersonal communication at the same time. In the case of social media, this type of communication is at its core, as messages are not limited to dyadic interactions, but they can reach wider audiences and be spread more rapidly (Walther et al., 2010).

Another distinguishing feature of the social medium is the fact that the content is user-generated, which means that it is derived "from the contributions from or interactions with other users rather than the content generated by the organisation or individual hosting the medium" (Carr & Hayes, 2015, p. 51). This shows the democratising power of social media, since users are transformed from consumers and passive spectators into creators and active subjects: anybody can produce and share their messages and ideas, thus becoming content generators (Benkler, 2006; Kietzmann et al., 2011).

Within the social media scenario, it is possible to find social networking sites (SNSs henceforth), also known as social media platforms, which are defined by Boyd and Ellison (2007) as services that let individuals

> construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system. (p. 211)

SNSs allow individuals not only to connect with people they already know, but to establish ties with strangers through the articulation of visible profiles. The first social network site was SixDegrees, launched by the company MacroView in 1997, which was advertised as a tool to connect with others (Boyd & Ellison, 2007). This was followed by the creation of more successful sites over the years, such as MySpace, Facebook, or LinkedIn. Ever since then, SNSs have become an increasingly popular way of staying in touch with others, endorsed by the rising number of platforms that have been created in the last 20 years (e.g., Instagram, Twitter/X[1], Tik Tok, Reddit, among others).

Social media, and SNSs more specifically, have allowed us to stay connected with the world using minimal effort. As stated by Foer (2013), each step forward in the development of social media "has made it easier, just a little, to avoid the emotional work of being present, to convey information rather than humanity". In other words, society is evolving into a non-physically present one: a parallel world of social connections has developed within the limits of social media, and it is there where people often share their thoughts or deepen their relations with others. SNSs and online communities generate informational and emotional support, thus providing users with a space where to feel a sense of belonging to a social group (Ballantine & Stephenson, 2011; Hajli, 2014; Liang et al., 2011). This online social bonding has been particularly perspicuous in the context of the past three years, which have been marked by a global health crisis: in the absence of the possibility of being physically close to others, users turned to SNSs as tools to help

[1] In April 2023 Twitter's legal name was changed to X Corp. However, in this doctoral thesis I will refer to the company as Twitter/X to avoid confusion and because all corpora discussed or used were compiled prior to this change of name. For the same reason, these will be referred to as Twitter corpora or datasets.

them feel virtually present. Moreover, SNSs provided access to live information of what was happening anywhere in the world: videos and images were shared indiscriminately, often without even checking the real source. This, obviously, had an impact on the quality of the information that was being shared: on February 2, 2020, the World Health Organisation (WHO hereafter) already warned in their 13th COVID-19 situation report (2020) of the existence of an *infodemic*, i.e., an over-abundance of accurate and inaccurate information that complicates the task of finding trustworthy resources. This information overload may have very serious consequences in all paths of life, but even more at such a particularly difficult time, in which the term conveyed the idea of how misinformation could accelerate the epidemic by fragmenting social response and altering the effectiveness of the measures taken by governments (Cinelli et al., 2020; Kim et al., 2019).

It is also important to take into account that not all SNSs serve the same purposes or have the same audience: while platforms such as Tik Tok, Instagram, and Facebook are more oriented towards sharing visual content, Twitter/X mainly presents short text messages that may also include photos or videos, and which are known as tweets. This microblogging service has become one of the main resources for researchers in fields where public opinion and attitudes are relevant, such as discourse analysis and the social sciences, as Twitter/X data is free, relatively easy to access, and based on user-generated content.

In addition to this, tweets are the perfect expression of evaluative language: these messages are largely based on the verbalisation of emotions, feelings, and ideas, which users often express using extremely polarised language (Kouloumpis et al., 2011). This is specifically relevant from the perspective of sentiment analysis, defined by Liu (2011) as the field of study that analyses "people's opinions, sentiments, evaluations, appraisals,

attitudes, and emotions towards entities such as products, services, organisations, individuals, issues, events, topics, and their attributes" (p. 459). Its main aim is to analyse texts automatically to detect polarity, emotions, and/or intensity. This is generally done through the identification of lexical, iconographic, and structural features, as well as using advanced algorithms to process them, in order to classify a document on a scale that determines its semantic orientation (Lei & Liu, 2021; Moreno-Ortiz, 2019). As to its tools and techniques, these can be machine-learning, lexicon-based, or a combination of both. In the case of the machine-learning approach, it uses a set of features which are learned from annotated corpora. The lexicon-based approach, on the other hand, uses a lexicon to provide the polarity for each word or phrase found in the text (Moreno-Ortiz et al., 2019).

The polarisation of messages on Twitter/X also makes it one of the most interesting social media platforms for the understanding of current events, as this polarisation is precisely what spurs the proliferation of misinformation. In fact, according to Vosoughi et al. (2018), fake news spread faster than fact-based news. In addition, Twitter/X provides direct access to massive amounts of data and it amplifies information of dubious quality very easily. Therefore, Twitter/X is largely responsible for the infodemic previously mentioned, as is Facebook, Instagram, and all other social networks.

In terms of the themes that can be found on tweets, these are as wide and as varied as topics exist. Twitter/X users may talk about anything and everything, ranging from their personal lives to their opinions on something that they have enjoyed, for example. While doing so, they can engage with other users' content through hashtags, which generally function as a way of finding other people's tweets on a particular idea, and which allow researchers to find tweets related to a desired topic or area of study (Zappavigna, 2011). Nevertheless, although hashtags generally summarise a tweet's main

topic, that does not prevent users from also including references to unrelated events that go beyond the main message. In this sense, speakers often relate the contents of their tweets to external circumstances, thus establishing a link between their thoughts and other experiences. This is specifically relevant in terms of practices such as connected co-viewing, i.e., the process of sharing their thoughts on social media while users are watching something, and subsequently engaging with others' comments (Pires & Roig, 2020). Consequently, the language found on Twitter/X cannot be simply typecast according to the hashtags included. Instead, corpus-based discourse analysis techniques are essential to analyse the nature of these messages.

This social media platform also poses a number of challenges for traditional approaches to data mining and Natural Language Processing (NLP henceforth) tasks, as the language used tends to be colloquial, informal, full of slang, and often unstructured and/or ungrammatical. Furthermore, rhetorical devices such as sarcasm are frequent, which often hinders the efforts of accurate sentiment analysis systems. Consequently, automatic sarcasm detection has become the main focus of multiple sentiment analysis shared tasks (Barbieri et al., 2014; Ghosh et al., 2020). Approaches to sarcasm detection can be classified into rule-based, statistical, and deep learning-based approaches. Rule-based approaches aim to identify sarcasm through specific evidence, while statistical approaches use sets of features, such as bag-of-words, to train prediction models. Deep learning-based ones, for their part, have gained a lot of popularity in NLP applications (Joshi et al., 2017).

Nonetheless, the automatic detection of sarcasm is highly dependent on the availability of high-quality annotated datasets, which are scarce. Furthermore, most of the existing annotated datasets have been created in a semi-automatic manner and are, thus, of dubious quality. As reported by Joshi et al. (2017), even the value of those manually

annotated is often questioned, as sarcasm is a subjective phenomenon that might not be perceived equally by everyone. It is for this reason that the annotation schema must be specifically designed, followed by a carefully controlled and guided annotation process, for annotated corpora to be of use for automatic sarcasm detection tasks.

The availability of massive amounts of data on Twitter/X, while an asset, can also become an issue for corpus linguistics researchers. Large social media corpora require users to implement their own NLP techniques if they wish to analyse big data, as manual, qualitative analysis is simply unfeasible on this scale. The problem is that such techniques can be quite difficult to learn, which usually becomes a limitation for researchers. Moreover, widely-known corpus tools such as *WordSmith* (Scott, 1996*)*, *AntConc* (Anthony, 2022), or *SketchEngine* (Kilgarriff et al., 2014) simply cannot handle such massive amounts of text.

In summary, working with Twitter/X data entails some challenges that cannot be ignored, such as the presence of rhetorical devices, the management of massive amounts of data, and the presence of opinionated language. Therefore, this doctoral thesis by compendium of publications aims to shed light on these issues by presenting three articles that attempt to provide answers to these concerns regarding Twitter/X, from the perspective of corpus linguistics, sentiment, and discourse analysis.

1.1 Objectives

The main aim of this doctoral thesis is to investigate the use of Twitter/X data as a source to unveil linguistic phenomena specifically related to computer-mediated communication, which is used globally and pervasively in this digital era. Some of these phenomena, such as the presence and expression of sarcasm, are particularly interesting from the perspective of corpus linguistics, while the use of Twitter/X to (strongly) express

opinions and emotions is of great interest for both sentiment and discourse analysis. In addition, with more than 500 million monthly users and over 230 million daily active users (Shewale, 2024b), this doctoral thesis had to tackle the challenge of managing such large-scale corpora, as any corpus extracted from this source is deemed to be extremely big. Thus, this thesis includes three general objectives, each with its own specific objectives, as listed below:

- General objective 1: to create an annotated corpus for automatic sarcasm detection using data extracted from Twitter.
    - Specific objective 1.1: to design an effective annotation schema for sarcasm detection and typology.
    - Specific objective 1.2: to explore sarcasm mechanisms and their formal markers in written text.
    - Specific objective 1.3: to check the real impact that sarcasm has on sentiment analysis systems.
- General objective 2: to provide methodological cues and strategies on how to manage and explore the contents of large-scale Twitter/X corpora.
    - Specific objective 2.1: to compare different sampling sizes to decide which one optimally represents the whole corpus for keyword extraction while keeping data to a manageable size using accessible computing resources.
    - Specific objective 2.2: to compare two different methods of keyword extraction, i.e., the reference-corpus method commonly used in corpus linguistics and the graph-based method commonly used in NLP.
    - Specific objective 2.3: to study which method is more appropriate according to the purposes of each piece of research.

- General objective 3: to analyse the expression of ideological opinions on Twitter/X to study the linguistic patterns generally used in evaluative language.
    - Specific objective 3.1: to identify the semantic orientation of tweets about two contestants in a media event that generates an intense social talk and has a highly active fan community on Twitter/X.
    - Specific objective 3.2: to gauge the actual degree of politicisation expressed by Twitter/X users in relation to such media event.
    - Specific objective 3.3: to gain insights into the political discourse contained in these tweets.

1.2 Thesis structure

This doctoral thesis is presented as a compendium of three articles, listed hereafter, two of which have been published and one that is in press. These three articles can be found in the second, third, and fourth chapters in consecutive order. Relevant information on the characteristics of the journals and the impact and bibliometric indices of each publication are included before each article.

I. Moreno-Ortiz, A. & García-Gámez, M. (2022). Corpus Annotation and Analysis of Sarcasm on Twitter: #CatsMovie vs. #TheRiseOfSkywalker. *Atlantis. Journal of the Spanish Association of Anglo-American Studies*, *44*(1), 186-207. http://doi.org/10.28914/Atlantis-2022-44.1.11

II. Moreno-Ortiz, A. & García-Gámez, M. (2023). Strategies for the Analysis of Large Social Media Corpora: Sampling and Keyword Extraction Methods. *Corpus Pragmatics*, 7, 241-265. https://doi.org/10.1007/s41701-023-00143-0

III. García-Gámez, M. & Moreno-Ortiz, A. (2024, in press). The Politics of Eurovision: A Case Study of the United Kingdom's 2021 and 2022 Participations

as Expressed on Social Media. *Revista de Lingüística y Lenguas Aplicadas, 19*.

Each of the research articles that conform this doctoral thesis by compendium include their own theoretical background, although these are deeply interrelated, in the sense that each of them investigates linguistic mechanisms that Twitter/X users employ to express themselves. Moreover, they interweave three consecutive studies designed specifically to achieve the aforementioned objectives, so that general objective 1 is tackled by the first article, while general objectives 2 and 3 are undertaken by the second and the third article, correspondingly.

2 Corpus Annotation and Analysis of Sarcasm on Twitter: #CatsMovie vs. #TheRiseOfSkywalker

Journal: *Atlantis - Journal of the Spanish Association of Anglo-American Studies*

ISSN: 0210-6124

Research area: English Studies; Linguistics and Language

Review process: Blind peer review

External reviewers: Yes

Open access: Yes

Impact

Atlantis is indexed in the following Thomson Reuters services:

- Journal Citation Reports
- Arts and Humanities Citation Index®
- Current Contents®/Arts & Humanities
- Emerging Source Citation Index
- Social Sciences Citation Index®
- Current Contents®/Social and Behavioral Sciences

Atlantis is also indexed or abstracted, among others, in the following databases or directories:

- SCIMAGO Journal and Country Rank
- SCOPUS
- RESH, Revistas Españolas de Ciencias Sociales y Humanidades
- CARHUS Plus+
- European Society for the Study of English (ESSE)
- DIALNETPlus

- DOAJ Directory of Open Access Journals
- ERIHPlus
- Latindex
- MIAR Matriz de Información para el Análisis de Revistas

This journal holds the Quality Seal for Excellence in Academic Journals awarded by the Spanish Foundation for Science and Technology.

Bibliometric indicators

Database	Impact Factor	Rank/Category
JCR	JIF: 0.405	Ranking by JIF: Q4 (173/194) (Linguistics). Percentile: 11.08
	JCI: 0.80	Ranking by JCI: - Q2 (119/370) (Language & Linguistics). Percentile: 67.9 - Q2 (131/274) (Linguistics). Percentile: 52.37
Scopus	SJR: 0.12	Language & Linguistics: (476/968) 50th percentile Linguistics & Language: (519/1032) 49th percentile
	SNIP: 0.871	
Scimago	SJR: 0.12	Q3 (Linguistics & Language)
	H index: 12	
FECYT	Linguistics: C1 (13/80) 37.76	Quality Seal for Excellence renewed for 2024

Referencia bibliográfica
Moreno-Ortiz, A. & García-Gámez, M. (2022). Corpus Annotation and Analysis of Sarcasm on Twitter: #CatsMovie vs. #TheRiseOfSkywalker. *Atlantis. Journal of the Spanish Association of Anglo-American Studies, 44*(1), 186-207. http://doi.org/10.28914/Atlantis-2022-44.1.11

Resumen en inglés

Sentiment analysis is a natural language processing task that has received increased attention in the last decade due to the vast amount of opinionated data on social media platforms such as Twitter. Although the methodologies employed have grown in number and sophistication, analysing irony and sarcasm still poses a severe problem. From the linguistic perspective, sarcasm has been studied in discourse analysis from several perspectives, but little attention has been given to specific metrics that measure its relevance. In this paper we describe the creation of a manually-annotated dataset where detailed text markers are included. This dataset is a sample from a larger corpus of tweets (n= 76,764) on two highly controversial films: Cats and Star Wars: The Rise of Skywalker. We took two different samples for each film, one before and one after their release, to compare reception and presence of sarcasm. We then used a sentiment analysis tool to measure the impact of sarcasm in polarity detection and then manually classified the mechanisms of sarcasm generation. The resulting corpus will be useful for machine learning approaches to sarcasm detection as well as discourse analysis studies on irony and sarcasm.

Resumen en español

El análisis de sentimiento es una de las aplicaciones del procesamiento del lenguaje natural que más atención ha recibido en la última década, principalmente debido a la cantidad de opiniones vertidas en redes sociales como Twitter. Pese a que las metodologías empleadas son cada vez más sofisticadas, el sarcasmo sigue siendo un gran

problema. Aunque el sarcasmo ha sido estudiado desde varias perspectivas en el análisis del discurso, no se ha prestado mucha atención a su presencia y relevancia real, aportando métricas concretas. En este trabajo se describe la creación de un dataset anotado manualmente en el que se incluyen marcadores textuales. Dicho dataset es la muestra de un corpus de tweets (n= 76.764) sobre dos películas controvertidas: Cats y Star Wars. El Ascenso de Skywalker. Tomamos dos muestras para cada película, antes y después de su estreno, para comparar su acogida. Empleamos una herramienta de análisis de sentimiento para medir el impacto del sarcasmo en la detección de la polaridad, y posteriormente identificamos y clasificamos los mecanismos de generación de sarcasmo. Este corpus puede ser de gran utilidad para la detección del sarcasmo mediante aprendizaje automático, así como para estudios de análisis del discurso sobre la expresión del sarcasmo.

DOI: http://doi.org/10.28914/Atlantis-2022-44.1.11

3 Strategies for the Analysis of Large Social Media Corpora: Sampling and Keyword Extraction Methods

Journal: *Corpus Pragmatics*

ISSN: 2509-9515

Research area: Linguistics and Language; Corpus Linguistics

Review process: Blind peer review

External reviewers: Yes

Open Access: Yes

Impact

Corpus Pragmatics is indexed in the following Thomson Reuters services:

- Emerging Sources Citation Index

Corpus Pragmatics is also indexed or abstracted, among others, in the following databases or directories:

- SCIMAGO Journal and Country Rank
- SCOPUS
- ProQuest-ExLibris Primo
- ProQuest-ExLibris Summon
- EBSCO Discovery Service

Bibliometric indicators

Database	Impact Factor	Rank/Category
Scopus	SJR: 0.247 SNIP: 0.774	Language & Linguistics: (181/1001) 81st percentile Linguistics & Language: (211/1078) 80th percentile
Scimago	SJR: 0.247 H index: 8	Q2 (Linguistics & Language)

Referencia bibliográfica

Moreno-Ortiz, A. & García-Gámez, M. (2023). Strategies for the Analysis of Large Social Media Corpora: Sampling and Keyword Extraction Methods. *Corpus Pragmatics, 7*, 241-265. https://doi.org/10.1007/s41701-023-00143-0

Resumen en inglés

In the context of the COVID-19 pandemic, social media platforms such as Twitter have been of great importance for users to exchange news, ideas, and perceptions. Researchers from fields such as discourse analysis and the social sciences have resorted to this content to explore public opinion and stance on this topic, and they have tried to gather information through the compilation of large-scale corpora. However, the size of such corpora is both an advantage and a drawback, as simple text retrieval techniques and tools may prove to be impractical or altogether incapable of handling such masses of data. This study provides methodological and practical cues on how to manage the contents of a large-scale social media corpus such as Chen et al. (JMIR Public Health Surveill 6(2):e19273, 2020) COVID-19 corpus. We compare and evaluate, in terms of efficiency and efficacy, available methods to handle such a large corpus. First, we compare different sample sizes to assess whether it is possible to achieve similar results despite the size difference and evaluate sampling methods following a specific data management approach to storing the original corpus. Second, we examine two keyword extraction methodologies commonly used to obtain a compact representation of the main subject and topics of a text: the traditional method used in corpus linguistics, which compares word frequencies using a reference corpus, and graph-based techniques as developed in Natural Language Processing tasks. The methods and strategies discussed in this study enable valuable quantitative and qualitative analyses of an otherwise intractable mass of social media data.

Resumen en español

En el contexto de la pandemia de COVID-19, las plataformas de medios sociales como Twitter han sido de gran importancia para que los usuarios intercambien noticias, ideas y percepciones. Investigadores de campos como el análisis del discurso y las ciencias sociales han recurrido a estos contenidos para explorar la opinión y la postura públicas sobre este tema, y han tratado de recabar información mediante la compilación de corpus a gran escala. Sin embargo, el tamaño de estos corpus es a la vez una ventaja y un inconveniente, ya que las técnicas y herramientas sencillas de recuperación de textos pueden resultar poco prácticas o del todo incapaces de manejar tales masas de datos. Este estudio proporciona pistas metodológicas y prácticas sobre cómo gestionar los contenidos de un corpus de medios sociales a gran escala como el corpus COVID-19 de Chen et al. (2020). Comparamos y evaluamos, en términos de eficiencia y eficacia, los métodos disponibles para manejar un corpus tan grande. En primer lugar, comparamos diferentes tamaños de muestra para evaluar si es posible obtener resultados similares a pesar de la diferencia de tamaño y evaluamos métodos de muestreo que siguen un enfoque específico de gestión de datos para almacenar el corpus original. En segundo lugar, examinamos dos metodologías de extracción de palabras clave utilizadas habitualmente para obtener una representación compacta del tema principal y los tópicos de un texto: el método tradicional empleado en lingüística de corpus, que compara las frecuencias de palabras utilizando un corpus de referencia, y las técnicas basadas en grafos, tal y como se han desarrollado en tareas de Procesamiento del Lenguaje Natural.

DOI: https://doi.org/10.1007/s41701-023-00143-0

4 The Politics of Eurovision: A Case Study of the United Kingdom's 2021 and 2022 Participations as Expressed on Social Media

Journal: *Revista de Lingüística y Lenguas Aplicadas*

ISSN: 1886-6298

Research area: Linguistics; Discourse Analysis

Review process: Blind peer review

External reviewers: Yes

Open Access: Yes

Impact

Revista de Lingüística y Lenguas Aplicadas is indexed or abstracted, among others, in the following databases or directories:

- SCOPUS
- ProQuest
- EBSCO Discovery Service
- DOAJ
- ERIHPlus
- MIAR

This journal holds the Quality Seal for Excellence in Academic Journals awarded by the Spanish Foundation for Science and Technology.

Bibliometic indicators

Database	Impact Factor	Rank/Category
Scopus	SJR: 0.174	Language & Linguistics: (333/1001) 66th percentile Linguistics & Language: (381/1078) 64th percentile
	SNIP: 0.874	
Scimago	SJR: 0.174	Q2 (Linguistics & Language)
	H index: 7	
FECYT	Linguistics: C1 (12/80) 38.13	Quality Seal for Excellence renewed for 2024

El equipo editorial de la *Revista de Lingüística y Lenguas Aplicadas* hace constar que el artículo «THE POLITICS OF EUROVISION: A CASE STUDY OF THE UNITED KINGDOM'S 2021 AND 2022 PARTICIPATIONS AS EXPRESSED ON SOCIAL MEDIA», presentado por MARÍA GARCÍA-GÁMEZ y ANTONIO MORENO-ORTIZ, ha sido evaluado positivamente por los revisores de nuestra publicación y, por consiguiente, aparecerá en el número 19 de la *Revista de Lingüística y Lenguas Aplicadas* en 2024.

Y para que conste a los efectos oportunos, se expide el siguiente documento a petición de los interesados.

Valencia, 31 de enero de 2024

Equipo editorial
Revista de Lingüística y Lenguas Aplicadas
https://polipapers.upv.es/index.php/rdlyla
Departamento de Lingüística Aplicada
Universitat Politècnica de València

Referencia bibliográfica

García-Gámez, M. & Moreno-Ortiz, A. (2024, en prensa). The Politics of Eurovision: A Case Study of the United Kingdom's 2021 and 2022 Participations as Expressed on Social Media. *Revista de Lingüística y Lenguas Aplicadas, 19*.

Resumen en inglés

In recent years, the opinion that the Eurovision Song Contest has become highly politicised is prevalent in the media and the popular voice, although not much research exists that can attest to this claim. In this work we conduct a case study that applies sentiment and discourse analysis methodologies to the assessment of political opinions in social media regarding this artistic and social event. The main objective is to explore to what extent and in what form this supposed politicisation has an expression on Twitter, as illustrated by the cases of artists Sam Ryder and James Newman, the United Kingdom's representatives in the 2022 and 2021 editions of the contest, respectively. We examine references to two historicalpolitical contexts that have had a severe impact on the European society over the last few years, and which have determined, among many other social aspects, the reception of Eurovision results since they took place: Brexit and the Russian invasion of Ukraine.

Resumen en español

En los últimos años, prevalece en los medios de comunicación y en la voz popular la opinión de que el festival de Eurovisión se ha politizado enormemente, aunque no existen muchas investigaciones que puedan dar fe de esta afirmación. En este trabajo realizamos un estudio de caso que aplica metodologías de análisis del sentimiento y del discurso a la evaluación de las opiniones políticas en los medios sociales sobre este acontecimiento artístico y social. El objetivo principal es explorar hasta qué punto y de qué forma esta supuesta politización tiene expresión en Twitter, como ilustran los casos de los artistas

Sam Ryder y James Newman, representantes del Reino Unido en las ediciones de 2022 y 2021 del certamen, respectivamente. Examinamos las referencias a dos contextos histórico-políticos que han tenido un severo impacto en la sociedad europea durante los últimos años, y que han determinado, entre otros muchos aspectos sociales, la recepción de los resultados de Eurovisión desde su celebración: El Brexit y la invasión rusa de Ucrania.

5 Conclusions

Each of the articles included in this body of work detail and discuss their own specific conclusions, but in this final chapter I wish to provide a final recount of the most important results obtained and lessons learned throughout the process of writing this doctoral thesis, in which I have worked intensively with data extracted from Twitter/X, using corpus, sentiment, and discourse theoretical and methodological foundations.

There are several characteristics that give homogeneity, continuity and coherence to the works presented in this thesis. The first, which has involved a long learning process, refers to the research design they present, as they all make use of a series of state-of-the-art analytical tools and resources. The second, also of a methodological nature, refers to the use of mixed quantitative and qualitative research methods for linguistic analysis and the extraction of information from a corpus.

A global learning that can be extracted from the process of writing this doctoral thesis is that Twitter/X serves as a valuable source of unveiling linguistic phenomena, especially in the realm of computer-mediated communication, which has become ubiquitous in our digital era. With its global reach and real-time interactions, social media provide a unique lens into evolving linguistic trends, language innovations, and the dynamic interplay between users across different communities and cultures. Studying user-generated discourse on Twitter/X shows the rapid evolution of language in the digital landscape, and it offers linguists a rich tapestry to unravel the intricacies of contemporary communication.

The path that has led me to achieve this overall learning process can be broken down into several, more specific, lessons learned. The first one is that sarcasm is an inherent part of Twitter/X, but it is present in many ways that often go beyond the limits

of what has been traditionally considered sarcasm, which makes me question *what* sarcasm is exactly. On Twitter/X, sarcasm is not just a shift in the polarity of a word or a phrase, but it can also be dark humour or straight criticism, and it won't always appear in a fixed form. Instead, users are likely to combine semantic and pattern-based mechanisms to be funny, humorous, witty, and original. Moreover, the correct encoding and decoding of sarcasm is necessarily based on the knowledge shared by a given community; therefore, one must be part of that community in order to be able to detect and understand a tweet as sarcastic.

The second relevant lesson derived from this study is that sarcasm does not always have a big impact on lexicon-based sentiment analysis systems, precisely because of the aforementioned fact that sarcasm may not necessarily imply a change in the polarity of a word. Tweets tend to be polarised and opinionated, but their sarcastic component does not always mean that sentiment analysis cannot escape it.

The third important lesson learned that can be extracted from this work is that the analysis of sarcasm on Twitter/X relies heavily on the corpus being used. Sarcastic instances (and hence the type of sarcasm mechanisms employed) will vary significantly depending on the contents and the topics included in the dataset being used. In addition, it is essential to work with at least two datasets to be able to put the presence of sarcasm in perspective; otherwise, the results will be biased.

The fourth general conclusion to be drawn from the work presented in this thesis is that keywords and graph-based models are one of the keys for the 2024 linguist to analyse large corpora. Datasets extracted from Twitter/X can be massive, and if the corpus linguistics researcher wishes to enter this otherwise intractable mass of data, one must become familiar with distributional models and learn how to use them. This does not mean, however, that frequency-based keyword extraction tools such as *AntConc* or

SketchEngine should be consigned to eternal oblivion. Instead, they must be viewed as complementary to one another, as the combination of these two approaches to keyword extraction will provide the qualitative researcher with the key to the contents of the corpus.

The second key to the analysis of large corpora constitutes the fifth lesson, which is that although *there is no data like more data*, big samples are not needed for the qualitative researcher to be able to obtain good and generalisable results. Smaller samples are easier to handle from a computational perspective (which is something to keep in mind if we do not have a computer with a powerful GPU) and they also produce comparable results. Big data, then, should not be a limitation for the qualitative researcher – instead, it can be the open door to realistic and accurate results, provided the right keys are used to open it (i.e., keyword extraction and sampling size).

From the point of view of the content expressed by Twitter/X users and its linguistic expression, the sixth lesson learned is that we must distinguish polarisation from politicisation, because not everything on Twitter/X is related to politics – not even a TV show that has been deemed to be strongly influenced by political relations between countries. This reinforces the fact that we, as corpus linguists, must look closely and more than twice at the data and avoid being guided by our own preconceived ideas. This is why qualitative analysis remains important even when we are dealing with large amounts of text, where quantitative analysis is the best entry point, but not the only resource.

The seventh and final lesson learned is that data extracted from Twitter/X must be researched from two complementary perspectives: sentiment and discourse. The presence of opinionated language on Twitter/X cannot be neither obviated nor reduced to a particular hashtag or theme. The nature and content of these messages should not be regarded as static, since they are changeable from speaker to speaker, and vary as events

unfold. This brings us back to the very first chapter of this doctoral thesis, where it was stated that Twitter/X users talk about anything and everything: in other words, the expression of opinions on Twitter/X is subject to various parameters, and consequently, their analysis must also account for such variables. If the researcher's aim is to study the sentiment conveyed in these opinions, such a task must be additionally accompanied by a discourse-based analysis. Otherwise, one runs the risk of wrongly overgeneralising the results obtained: for example, words with a negative polarity do not always entail a negative message. Instead, they might function as emphasisers. Therefore, if we just focus on the sentiment analysis without using discourse analysis techniques that consider the social and cultural context, we might end up with the impression that our corpus is extremely positive or extremely negative, which may not even be the case.

In essence, these lessons collectively serve as a guide for researchers navigating the intricate landscape of Twitter. By acknowledging the multifaceted nature of language expression, the limitations of computational tools, and the contextual dependencies within datasets, researchers can unlock a more nuanced and accurate understanding of sentiment and discourse on this ever-evolving digital platform. The journey through these lessons is not only a reflection of Twitter's intricacies, but also a testament to the evolving methodologies required to capture the essence of online communication from the point of view of linguistic research.

6 References

Anthony, L. (2022). *AntConc (Version 4.0.10)* [Computer software]. Waseda University. https://www.laurenceanthony.net/software

Ballantine, P. W., & Stephenson, R. J. (2011). Help me, I'm fat! Social support in online weight loss networks: Social support in online weight loss networks. *Journal of Consumer Behaviour*, *10*(6), 332–337. https://doi.org/10.1002/cb.374

Barbieri, F., Saggion, H., & Ronzano, F. (2014). Modelling sarcasm in Twitter, a novel approach. *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 50–58. https://doi.org/10.3115/v1/W14-2609

Benkler, Y. (2006). *The wealth of networks*. Yale University Press.

Boyd, D. M., & Ellison, N. B. (2007). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210–230. https://doi.org/10.1111/j.1083-6101.2007.00393.x

Carr, C. T., & Hayes, R. A. (2015). Social Media: Defining, Developing, and Divining. *Atlantic Journal of Communication*, *23*(1), 46–65. https://doi.org/10.1080/15456870.2015.972282

Cauberghe, V., Van Wesenbeeck, I., De Jans, S., Hudders, L., & Ponnet, K. (2021). How Adolescents Use Social Media to Cope with Feelings of Loneliness and Anxiety During COVID-19 Lockdown. *Cyberpsychology, Behavior, and Social Networking*, *24*(4), 250–257. https://doi.org/10.1089/cyber.2020.0478

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., & Scala, A. (2020). The COVID-19 social media

infodemic. *Scientific Reports*, *10*(1), 16598. https://doi.org/10.1038/s41598-020-73510-5

Foer, J. S. (2013, June 8). How not to be alone. *The New York Times*. https://www.nytimes.com/2013/06/09/opinion/sunday/how-not-to-be-alone.html

Ghosh, D., Vajpayee, A., & Muresan, S. (2020). A report on the 2020 sarcasm detection shared task. *arXiv:2005.05814 [Cs]*. http://arxiv.org/abs/2005.05814

Hajli, M. N. (2014). A study of the impact of social media on consumers. *International Journal of Market Research*, *56*(3), 387–404. https://doi.org/10.2501/IJMR-2014-025

Joshi, A., Bhattacharyya, P., & Carman, M. J. (2017). Automatic sarcasm detection: A survey. *ACM Computing Surveys*, *50*(5), 1–22. https://doi.org/10.1145/3124420

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, *54*(3), 241–251. https://doi.org/10.1016/j.bushor.2011.01.005

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, 7–36.

Kim, L., Fast, S. M., & Markuzon, N. (2019). Incorporating media data into a model of infectious disease transmission. *PLOS ONE*, *14*(2), e0197646. https://doi.org/10.1371/journal.pone.0197646

Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter Sentiment Analysis: The good, the bad, and the OMG! *Proceedings of the 5th International Conference on Weblogs and Social Media (ICWSM-11)*, 538–541.

Lei, L., & Liu, D. (2021). *Conducting Sentiment Analysis*. Cambridge University Press. https://www.cambridge.org/core/elements/conducting-sentiment-analysis/B00BACADE638BF1AD5F61972FEE4183D

Liang, T.-P., Ho, Y.-T., Li, Y.-W., & Turban, E. (2011). What Drives Social Commerce: The Role of Social Support and Relationship Quality. *International Journal of Electronic Commerce*, *16*(2), 69–90. https://doi.org/10.2753/JEC1086-4415160204

Liu, B. (2011). *Web data mining: Exploring hyperlinks, contents, and usage data*. Springer.

Moreno-Ortiz, A. (2019). Mi opinión cuenta: La expresión del sentimiento en la Red. In S. Robles Avila & A. Moreno-Ortiz (Eds.), *Comunicación mediada por ordenador: La lengua, el discurso y la imagen* (1a edición, pp. 38–74). Cátedra.

Moreno-Ortiz, A., Salles-Bernal, S., & Orrequia-Barea, A. (2019). Design and validation of annotation schemas for aspect-based sentiment analysis in the tourism sector. *Information Technology & Tourism*, *21*(4), 535–557. https://doi.org/10.1007/s40558-019-00155-0

O'Sullivan, P. B., & Carr, C. T. (2018). Masspersonal communication: A model bridging the mass-interpersonal divide. *New Media & Society*, *20*(3), 1161–1180. https://doi.org/10.1177/1461444816686104

Pires, F., & Roig, A. (2020). All aboard?! Co-viewing with and within connected platforms in the Eurovision Song Contest. *Observatorio (OBS*)*, *14*(4). https://doi.org/10.15847/obsOBS14420201673

Russo, A., Watkins, J., Kelly, L., & Chan, S. (2008). Participatory Communication with Social Media. *Curator: The Museum Journal*, *51*(1), 21–31. https://doi.org/10.1111/j.2151-6952.2008.tb00292.x

Scott, M. (1996). *WordSmith Tools manual*. Oxford University Press.

Shewale, R. (2024a, January 9). *Social Media Users And Statistics For 2024 (Latest Data)*. https://www.demandsage.com/social-media-users/

Shewale, R. (2024b, January 10). *Twitter Statistics For 2024—(Facts After 'X' Rebranding)*. https://www.demandsage.com/twitter-statistics/

Vásquez, C. (2022). Introduction. In C. Vásquez (Ed.), *Research Methods for Digital Discourse Analysis* (pp. 1–18). Bloomsbury. https://www.bloomsbury.com/us/research-methods-for-digital-discourse-analysis-9781350166837/

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Walther, J. B. (1996). Computer-Mediated Communication: Impersonal, Interpersonal, and Hyperpersonal Interaction. *Communication Research*, *23*(1), 3–43. https://doi.org/10.1177/009365096023001001

Walther, J. B., Carr, C. T., Choi, S., DeAndrea, D., Kim, J., Tong, S., & Van Der Heide, B. (2010). Interaction of Interpersonal, Peer, and Media Influence Sources Online: A Research Agenda for Technology Convergence. In Z. Papacharissi (Ed.), *The networked self* (pp. 17–38). Routledge.

World Health Organisation. (2020). *Novel Coronavirus (2019-nCoV) Situation Report—13* (pp. 1–7).

Zappavigna, M. (2011). Ambient affiliation: A linguistic perspective on Twitter. *New Media Society*, *13*(5), 788–806. https://doi.org/10.1177/1461444810385097

Resumen

Las redes sociales se han convertido en plataformas ampliamente empleadas para el intercambio dinámico de opiniones, percepciones, e ideas en el ámbito en línea. Esto se debe a que los usuarios han recurrido gradualmente a ellas no solo para mantenerse en contacto con los demás, sino también para compartir con el mundo sus propios pensamientos y creencias. Más del 65% de la población mundial hace uso de las redes sociales, y el 85% de los usuarios de teléfonos móviles del mundo tiene al menos una cuenta en alguna de las redes sociales existentes. Su inmediatez va más allá de las fronteras geográficas, fomentando así una sensación de interconexión que permite a los usuarios mantenerse informados y participar en las conversaciones que dan forma a la sociedad. Ya sea difundiendo noticias de última hora, defendiendo causas sociales o expresando relatos personales, estas plataformas amplifican las voces, democratizan la comunicación y contribuyen a la formación de un discurso global más inclusivo y dinámico. Como conducto para el diálogo, las redes sociales desempeñan un papel relevante en la configuración de la narrativa de nuestro mundo interconectado, y múltiples fenómenos sociales no se entenderían sin ellas.

Como herramientas de comunicación global, el impacto de las redes sociales ha sido, sin duda, amplificado por los fenómenos globales. Un caso paradigmático ha sido la pandemia de COVID-19, en la que estas plataformas resultaron ser herramientas cruciales para navegar la crisis. Mientras se imponían confinamientos a nivel mundial, las redes sociales facilitaron estar conectados, así como estar informados de lo que ocurría en tiempo real. Su papel como fuente de apoyo emocional se acentuó, y sirvieron como estrategia de afrontamiento constructiva para que las personas redujeran sus sentimientos de ansiedad y soledad, mientras buscaban solidaridad, empatía y experiencias

compartidas. De esta manera, ofrecían una línea de vida en tiempo real con el mundo exterior y servían como fuentes vitales de información.

En esta era sin precedentes, las redes sociales se han convertido en recursos inestimables para los investigadores, sobre todo en campos en los que la opinión pública desempeña un papel fundamental. La facilidad de acceso a muestras amplias y fácilmente disponibles del discurso virtual ofrece una vía intrigante para investigar cómo se expresan los individuos en un entorno en línea. La tesis doctoral que aquí se presenta pretende contribuir a este campo de estudio emergente, centrándose en la utilización de corpus extraídos de plataformas de medios sociales como Twitter/X. El objetivo principal es mostrar el potencial de estos conjuntos de datos para la lingüística de corpus, así como el análisis del discurso y el sentimiento, desvelando cómo el contenido de las redes sociales puede interpretarse como un reflejo de las actitudes de los hablantes. Al ahondar en las complejidades del uso del lenguaje en estas plataformas, la investigación trata de desentrañar la naturaleza polifacética de la expresión en línea, ofreciendo una comprensión matizada del sentimiento y la dinámica del discurso. Dado que las redes sociales siguen configurando las interacciones sociales e influyendo en ellas, esta investigación ofrece una exploración oportuna del panorama en evolución de la comunicación digital y sus implicaciones más amplias para la comprensión del sentimiento público en el mundo contemporáneo.

En el capítulo inicial, titulado "Introduction", se establece el marco teórico que sostiene la presente tesis doctoral por compendio. Este capítulo no solo sirve como un preámbulo a la investigación, sino que también proporciona definiciones clave y contextualiza conceptos fundamentales que son la base de este trabajo. Entre otros conceptos, se destaca el análisis de sentimiento, un campo de estudio dedicado a la evaluación automatizada del texto con el fin de extraer la polaridad de los mensajes. Este

análisis se lleva a cabo mediante la identificación de características léxicas, iconográficas y estructurales, junto con la aplicación de algoritmos avanzados para procesarlas, con el objetivo de clasificar un documento en una escala de orientación semántica.

Además, se aborda teóricamente el tema de las plataformas de redes sociales, definiendo y delimitando el concepto de red social para diferenciarlo de otros servicios en línea que también implican interacciones sociales a través de internet, como es el caso del correo electrónico. El énfasis se coloca en la distinción precisa de lo que constituye una red social. Asimismo, se ofrece una descripción detallada de las características más relevantes de la red social en la que se centra esta tesis doctoral: Twitter/X. Se exploran tanto sus elementos distintivos como los desafíos inherentes a su uso en investigaciones académicas.

Este capítulo introductorio no solo establece las bases teóricas esenciales de esta tesis doctoral, sino que también plantea interrogantes y desafíos que serán abordados a lo largo de la investigación. La meticulosa delineación de conceptos y la comprensión teórica de las herramientas y plataformas utilizadas sirven como cimientos sólidos para la exploración más profunda que se llevará a cabo en los siguientes capítulos de la tesis doctoral.

Uno de los desafíos inherentes a Twitter/X radica en la diversidad de temas y tópicos presentes en cada tuit. Los usuarios de esta plataforma pueden abordar desde aspectos íntimos de su vida hasta expresar opiniones sobre algo que les haya gustado. Aunque los hashtags se utilizan para conectar contenido relacionado y facilitar la búsqueda temática, la inclusión de referencias a eventos no relacionados complica la clasificación simple basada en hashtags. Los hablantes tienden a entrelazar sus pensamientos con experiencias personales, generando una riqueza de matices que

requiere técnicas avanzadas de análisis del discurso basadas en corpus para comprender la complejidad y la naturaleza de los mensajes.

Otro desafío crucial es la naturaleza coloquial, informal, y a menudo no estructurada del lenguaje en Twitter/X, repleto de jerga y recursos retóricos como el sarcasmo. Este último, en particular, representa un obstáculo para los sistemas de análisis de sentimiento, ya que la detección automática del sarcasmo depende de conjuntos de datos anotados de alta calidad, que son escasos. Además, la creación de corpus anotados precisa un diseño específico y un proceso de anotación cuidadosamente controlado para que sean efectivos en la tarea de detección automática del sarcasmo.

La abundancia de datos en Twitter/X, aunque es una ventaja, también se convierte en un desafío para los investigadores en lingüística de corpus. El análisis cualitativo manual de grandes corpus es prácticamente inviable, lo que obliga a los investigadores a emplear técnicas computacionales de procesamiento del lenguaje natural. Sin embargo, la curva de aprendizaje de estas técnicas y la necesidad de conocimientos informáticos avanzados pueden representar limitaciones significativas para algunos investigadores. Además, las herramientas convencionales de manejo de corpus, como *Wordsmith*, *AntConc*, o *SketchEngine* se revelan insuficientes para lidiar con la magnitud de datos en Twitter/X, requiriendo el desarrollo y la implementación de enfoques más especializados.

En resumidas palabras, el estudio de Twitter/X como fuente de investigación implica la confrontación de desafíos ineludibles y complejos que van desde la diversidad temática hasta la complejidad del lenguaje y la gestión eficiente de grandes cantidades de datos. No obstante, cada uno de estos desafíos también ofrece oportunidades para desarrollar y mejorar metodologías de investigación, impulsando la comprensión más profunda de la dinámica social en entornos digitales. Por ello, esta tesis doctoral por compendio de publicaciones pretende arrojar luz sobre estas cuestiones mediante la

presentación de tres artículos que intentan dar respuesta a estas cuestiones en torno a Twitter/X desde la perspectiva de la lingüística de corpus, el análisis de sentimiento, y el análisis del discurso. Cada artículo se erige como un pilar conceptual que busca aportar soluciones y perspicacia a los retos que plantea la investigación en esta plataforma, contribuyendo así al avance del conocimiento en estas disciplinas.

Este capítulo introductorio incluye una subsección titulada "Objectives", destinada a delinear los propósitos de la tesis. En consonancia con lo mencionado previamente, el objetivo primordial radica en investigar el uso de datos de Twitter/X como fuente para desvelar fenómenos lingüísticos específicamente relacionados con la comunicación mediada por ordenador, utilizada globalmente en esta era digital. Algunos de estos fenómenos, como la presencia y expresión del sarcasmo, son especialmente interesantes desde la perspectiva de la lingüística de corpus, mientras que el uso de Twitter/X para expresar (fuertemente) opiniones y emociones es de gran interés tanto para el análisis del sentimiento como del discurso. Además, con más de 500 millones de usuarios mensuales y más de 230 millones de usuarios activos diarios, esta tesis doctoral se propone abordar el reto de gestionar corpus de Twitter/X a gran escala, ya que cualquier corpus extraído de esta fuente se considera extremadamente grande. Así, esta tesis incluye tres objetivos generales, cada uno con sus propios objetivos específicos, que se enumeran a continuación:

- Objetivo general 1: crear un corpus anotado para la detección automática del sarcasmo a partir de datos extraídos de Twitter/X.
    - Objetivo específico 1.1: diseñar un esquema de anotación eficaz para la detección y tipología del sarcasmo.
    - Objetivo específico 1.2: explorar los mecanismos del sarcasmo y sus marcadores formales en el texto escrito.

  - Objetivo específico 1.3: comprobar el impacto real que tiene el sarcasmo en los sistemas de análisis de sentimiento.

- Objetivo general 2: proporcionar estrategias metodológicas sobre cómo gestionar y explorar los contenidos de corpus de Twitter/X a gran escala.
  - Objetivo específico 2.1: comparar diferentes tamaños de muestreo para decidir cuál representa de forma óptima todo el corpus para la extracción de palabras clave, manteniendo al mismo tiempo los datos en un tamaño fácilmente manejable y utilizando recursos informáticos accesibles.
  - Objetivo específico 2.2: comparar dos métodos diferentes de extracción de palabras clave, es decir, el método del corpus de referencia utilizado habitualmente en lingüística de corpus, y el método basado en grados utilizado habitualmente en el procesamiento del lenguaje natural.
- Objetivo general 3: analizar la expresión de opiniones ideológicas en Twitter/X para estudiar los patrones lingüísticos generalmente utilizados en la expresión del lenguaje evaluativo.
  - Objetivo específico 3.1: identificar la orientación semántica de los tuits sobre dos concursantes de un evento mediático que genera una intensa conversación social y cuenta con una comunidad de fans muy activa en Twitter/X.
  - Objetivo específico 3.2: medir el grado real de politización expresado por los usuarios de Twitter/X en relación con dicho acontecimiento mediático.
  - Objetivo específico 3.3: obtener información sobre el discurso político contenido en estos tuits.

Finalmente, el primer capítulo incluye una segunda subsección, titulada "Thesis structure", donde se especifican las referencias de los tres artículos que componen esta tesis doctoral por compendio, y que son los siguientes:

I. Moreno-Ortiz, A. & García-Gámez, M. (2022). Corpus Annotation and Analysis of Sarcasm on Twitter: #CatsMovie vs. #TheRiseOfSkywalker. *Atlantis. Journal of the Spanish Association of Anglo-American Studies*, *44*(1), 186-207. http://doi.org/10.28914/Atlantis-2022-44.1.11

II. Moreno-Ortiz, A. & García-Gámez, M. (2023). Strategies for the Analysis of Large Social Media Corpora: Sampling and Keyword Extraction Methods. *Corpus Pragmatics*, *7,* 241-265. https://doi.org/10.1007/s41701-023-00143-0

III. García-Gámez, M. & Moreno-Ortiz, A. (2024, en prensa). The Politics of Eurovision: A Case Study of the United Kingdom's 2021 and 2022 Participations as Expressed on Social Media. *Revista de Lingüística y Lenguas Aplicadas, 19*.

Cada uno de estos artículos pretende completar uno de los objetivos anteriormente mencionados, de modo que el objetivo general 1 lo aborda el primer artículo, mientras que los objetivos generales 2 y 3 los cumplen el segundo y el tercer artículo, respectivamente.

El segundo capítulo de esta tesis doctoral se titula "Corpus Annotation and Analysis of Sarcasm on Twitter: #CatsMovie vs. #TheRiseOfSkywalker", y presenta el primero de los tres artículos que constituyen esta tesis doctoral. En dicho artículo se describe la elaboración de un esquema de anotación válido para la detección automática del sarcasmo. Para ello, se crea un corpus de tweets (n=76.764) sobre dos películas controvertidas: *Cats* y *Star Wars: El Ascenso de Skywalker*. Tomamos dos muestras de

500 tuits para cada película (antes y después de su estreno) para así poder comparar su acogida, de tal forma que obtuvimos 1000 tuits por película en total. Dichas muestras fueron anotadas teniendo en cuenta dos aspectos: (i) si los tuits contenían sarcasmo, y en caso de contener sarcasmo, (ii) qué tipo de sarcasmo. Los mecanismos de generación de sarcasmo se dividían en dos categorías: patrones semánticos y contextuales, y patrones formales. Dentro de la primera categoría, encontramos los siguientes recursos: hipérbole, juegos de palabras, sarcasmo intertextual, ironía fraseológica, choques semánticos, e incongruencia implícita. Dentro de la segunda categoría, encontramos los siguientes elementos: comparaciones, preguntas retóricas, efectos tipográficos, locuciones sarcásticas, sarcasmo situacional, y la categoría "otro" (donde incluir aquellos recursos que no pertenecían a ninguna de las opciones indicadas anteriormente). Los resultados obtenidos sugieren que el impacto general del sarcasmo parece depender en gran medida del tema (o, en este caso, la película) y de la comunidad de usuarios. En cuanto a los mecanismos empleados para la expresión del sarcasmo, en el caso de *Cats* se basaban sobre todo en patrones semánticos, aunque también se identificaron otros recursos, como las preguntas retóricas, la hipérbole, o la ironía frasal, a menudo en combinación con la incongruencia semántica. En cambio, en *Star Wars*, los recursos se basaban generalmente en patrones formales, como las locuciones sarcásticas, las preguntas retóricas, y los efectos tipográficos. Estos temas suelen surgir espontáneamente en las redes sociales, y a menudo se plasman y condensas en memes que son posteriormente alimentados por la comunidad de usuarios. Por otro lado, los resultados también muestran que más del 10% de los tuits del corpus contenían algún tipo de sarcasmo, pero solo el 6% planteaba realmente un problema para los sistemas de análisis de sentimiento basados en el léxico, ya que el resto se clasificaba de forma correcta.

El tercer capítulo, llamado “Strategies for the Analysis of Large Social Media Corpora: Sampling and Keyword Extraction Methods”, presenta el segundo de los tres artículos que componen esta tesis doctoral. En este artículo se proporcionan estrategias metodológicas y prácticas sobre cómo gestionar los contenidos de un corpus a gran escala extraído de Twitter/X que se centra en la pandemia de COVID-19. Se comparan y se evalúan, en términos de eficiencia y eficacia, los métodos disponibles para el manejo de un corpus de este tipo. En primer lugar, se comparan diferentes tamaños de muestra (1% y 0.1%) para evaluar si es posible obtener resultados similares a pesar de la diferencia de tamaño, y evaluamos distintos métodos de muestreo que siguen un enfoque específico de gestión de datos para almacenar el corpus original. En segundo lugar, se examinan dos metodologías de extracción de palabras clave utilizadas habitualmente para obtener una representación compacta de los temas principales que aparecen en el texto: (i) el método tradicionalmente empleado en la lingüística de corpus, basado en la comparación de las frecuencias de las palabras utilizando un corpus de referencia, representado por *SketchEngine*; y (ii) las técnicas basadas en grafos desarrolladas en tareas de procesamiento del lenguaje natural, representadas por *TextRank*.

Los resultados de este estudio demuestran que la extracción de palabras clave es un recurso valioso para la exploración de grandes corpus extraídos de redes sociales como Twitter/X, ya que proporciona una vía clara y un punto de entrada para el investigador cualitativo a una masa de datos que, de otro modo, resultaría imposible de acceder y tratar. El empleo de varios métodos y sistemas de extracción de palabras clave, como *TextRank* y *SketchEngine*, ofrece resultados interesantes, ya que la naturaleza de las palabras clave extraídas varía. Los métodos tradicionales basados en el empleo de un corpus de referencia (como es el caso de *SketchEngine*) tienen el problema derivado de la necesidad de seleccionar un corpus concreto, ya que las palabras del corpus de interés que tienen

una frecuencia baja en el corpus de referencia ocuparán un lugar destacado en la lista de palabras clave. Por otro lado, las palabras clave extraídas mediante métodos basados en grafos (como *TextRank*) parecen captar con mayor precisión el contenido general del corpus. Así, el análisis de las palabras clave por año muestra que los resultados obtenidos por *TextRank* parecen captar mejor la naturaleza del corpus, mientras que los proporcionados por *SketchEngine* son más específicos del contenido y dependen de los parámetros de búsqueda empleados. En definitiva, estos dos métodos de extracción de palabras clave pueden ser apropiados para objetivos diferentes: los métodos basados en grafos parecen más adecuados para obtener los temas, eventos, y entidades más destacados de un corpus, mientras que los métodos basados en estadísticas son mejores para extraer términos especializados. A pesar de ello, los resultados proporcionados por cada método no deben entenderse como opuestos, sino más bien como complementarios, ya que podría considerarse que las palabras clave de la intersección entre ambos sistemas representan exhaustivamente no solo los principales acontecimientos y conceptos de la pandemia en un sentido general, sino también aspectos relevantes de la política que también están relacionados. En cuanto al tamaño de las muestras, nuestro estudio sugiere que, con un corpus masivo de datos extraídos de Twitter/X, las muestras más pequeñas producen resultados comparables a las más grandes. Por lo tanto, cuando se trata de corpus masivos de redes sociales, es innecesario extraer muestras más grandes en aras de la representatividad estadística, ya que las muestras más pequeñas también pueden ser representativas y relevantes para el investigador cualitativo, así como más fáciles de procesar desde una perspectiva computacional.

El cuarto capítulo, titulado “The Politics of Eurovision: A Case Study of the United Kingdom’s 2021 and 2022 Participations as Expressed on Social Media”, es el tercer y último artículo que conforma esta tesis doctoral. En dicho artículo se lleva a cabo

un estudio de caso que aplica metodologías de análisis de sentimiento y del discurso a la evaluación de opiniones políticas vertidas en Twitter/X sobre un acontecimiento artístico y social como es Eurovisión. El objetivo principal es explorar hasta qué punto y de qué forma esta supuesta politización toma forma en Twitter. Para ello, se examinan los casos de los artistas Sam Ryder y James Newman, representantes del Reino Unido en las ediciones de 2022 y 2021 del certamen, respectivamente. Además, se tienen en cuenta las referencias a dos contextos histórico-políticos que han tenido un severo impacto en la sociedad europea durante los últimos años, y que han determinado, entre otros muchos aspectos sociales, la recepción de los resultados de Eurovisión desde su celebración: el Brexit y la invasión rusa de Ucrania.

Los resultados obtenidos en este estudio evidencian que los mensajes de felicitación prevalecen en Twitter/X para ambos representantes independientemente de la posición final conseguida por cada uno de ellos, lo que indica que la obtención de la segunda o la última posición en Eurovisión no afecta a las percepciones del público hacia ellos. En ambos casos, ni siquiera los términos negativos se utilizan generalmente con una intención negativa hacia la entidad, sino que suelen emplearse con el objeto de enfatizar mensajes positivos. En cuanto al impacto real de la politización, es muy bajo en términos absolutos, de tal forma que las referencias políticas no prevalecen en las opiniones de los seguidores de los artistas sobre el concurso, ya que la mayoría de los usuarios no justifican los resultados con un razonamiento político. A pesar de ello, la posición obtenida en el concurso sí parece afectar a su asociación con la política: terminar el concurso con un mal resultado puede dar lugar a un mayor porcentaje de referencias políticas, ya que los usuarios culpan al contexto político. No obstante, el nivel real de politización debería medirse en términos relativos, es decir, la pregunta correcta sería "¿es la proporción de referencias políticas mayor en estos dos años que la media de los últimos

10-20 años?". Obviamente, para responder a esta pregunta serían necesarios datos de todos estos años para establecer una línea de base válida. Sin embargo, es evidente que el tema político varía según el representante. El lenguaje político encontrado en el corpus de Ryder gira en torno a la victoria de Ucrania sobre el Reino Unido a causa de la guerra. Además, está implícito: los usuarios se quejan de que dicho país haya ganado el concurso aludiendo al voto por simpatía o dando a entender que en cualquier otro año el Reino Unido habría ganado el concurso, pero en general evitan las referencias directas a la guerra en sí. Por otro lado, el lenguaje político encontrado en el corpus de Newman se centra en el Brexit y su impacto en la visión europea del Reino Unido. En general, los resultados evidencian la necesidad del empleo de técnicas avanzadas de análisis del discurso para poder investigar los diferentes temas que pueden surgir en los tuits que conforman un corpus extraído de Twitter.

El quinto capítulo, titulado "Conclusions", resume de manera global los resultados conseguidos en cada uno de los estudios llevados a cabo como parte de esta tesis doctoral. Además, se describen las conclusiones generales que se han obtenido, así como las lecciones aprendidas gracias a los tres trabajos que conforman esta tesis doctoral. Dicho capítulo puede leerse traducido al español en la siguiente sección, titulada "Conclusiones".

Conclusiones

Cada uno de los artículos incluidos en esta tesis doctoral por compendio detalla y discute sus propias conclusiones específicas, pero en esta última sección quisiera hacer un recuento final de los resultados más importantes y los aprendizajes que se han obtenido durante el proceso de elaboración de esta tesis doctoral, en la que he trabajado intensamente con datos extraídos de Twitter/X, utilizando fundamentos teóricos y metodológicos propios del análisis de corpus, sentimiento, y discurso.

Hay varias características que dotan de homogeneidad, continuidad y coherencia a los trabajos presentados en esta tesis. La primera, que ha implicado un largo proceso de aprendizaje, se refiere al diseño de investigación que presentan, ya que todos ellos hacen uso de una serie de herramientas y recursos analíticos de última generación. La segunda, también de índole metodológica, se refiere al empleo de métodos de investigación mixtos, cuantitativos y cualitativos, para el análisis lingüístico y la extracción de información de un corpus.

Un aprendizaje global que puede extraerse del proceso de redacción de esta tesis doctoral es que Twitter/X sirve como una valiosa fuente para desvelar fenómenos lingüísticos, especialmente en el ámbito de la comunicación mediada por ordenador, que se ha vuelto omnipresente en nuestra era digital. Con su alcance global y sus interacciones en tiempo real, las redes sociales ofrecen una perspectiva única de la evolución de las tendencias e innovaciones lingüísticas y la interacción dinámica entre usuarios de distintas comunidades y culturas. El estudio del discurso generado por los usuarios en Twittter/X muestra la rápida evolución del lenguaje en el panorama digital y ofrece a los lingüistas un rico tapiz a través del cual poder desentrañar los entresijos de la comunicación contemporánea.

El camino que me ha llevado a alcanzar este aprendizaje general se puede desglosar en varias lecciones aprendidas más específicas. La primera es que el sarcasmo es una parte inherente de Twitter/X, pero está presente de muchas maneras que a menudo van más allá de los límites de lo que tradicionalmente se ha considerado sarcasmo, lo que me hace cuestionarme qué es exactamente el sarcasmo. En Twitter/X, el sarcasmo no es sólo un cambio en la polaridad de una palabra o una frase, sino que también puede ser humor negro o crítica directa, y no siempre aparecerá de una forma fija. En cambio, es probable que los usuarios combinen mecanismos semánticos y basados en patrones para mostrarse divertidos, cómicos, ocurrentes y originales. Además, la correcta codificación y descodificación del sarcasmo se basa necesariamente en el conocimiento que comparte una comunidad determinada; por tanto, hay que formar parte de esa comunidad para poder detectar y entender un tuit como sarcástico.

La segunda lección importante que se deriva de este estudio es que el sarcasmo no siempre tiene un gran impacto en los sistemas de análisis de sentimiento basados en el léxico, precisamente por el hecho antes mencionado de que puede no implicar necesariamente un cambio en la polaridad de una palabra. Los tuits tienden a polarizarse y a generar opiniones, pero su componente sarcástico no siempre significa que el análisis de sentimientos no pueda escapar a él.

La tercera lección aprendida es que el análisis del sarcasmo en Twitter/X depende en gran medida del corpus que se utilice. Las instancias sarcásticas (y, por tanto, los tipos de mecanismos de sarcasmo empleados) variarán significativamente en función de los contenidos y los temas incluidos en el conjunto de datos que se utilice. Además, es esencial trabajar con al menos dos conjuntos de datos para poder poner en perspectiva la presencia de sarcasmo; de lo contrario, los resultados estarán sesgados.

La cuarta conclusión que se desprende de los trabajos presentados en esta tesis es que las palabras clave y los modelos basados en grafos son una de las claves para que el lingüista de 2024 pueda analizar grandes corpus. Los conjuntos de datos extraídos de Twitter/X pueden ser masivos, y si el investigador en lingüística de corpus desea adentrarse en esta masa de datos, de otro modo intratable, debe familiarizarse con los modelos distribucionales y aprender a utilizarlos. Esto no significa, sin embargo, que las herramientas de extracción de palabras clave basadas en la frecuencia, como *AntConc* o *SketchEngine*, deban ser relegadas al olvido eterno. Por el contrario, deben considerarse complementarias entre sí, ya que la combinación de estos dos enfoques de la extracción de palabras clave proporcionará al investigador cualitativo la clave del contenido del corpus.

La segunda clave para el análisis de grandes corpus constituye la quinta lección, y es que, aunque no hay nada como tener muchos datos, no se necesitan grandes muestras para que el investigador cualitativo pueda obtener resultados buenos y generalizables. Las muestras más pequeñas son más fáciles de manejar desde el punto de vista computacional (algo que debemos tener en cuenta si no disponemos de un ordenador con una potente GPU) y además producen resultados comparables. Los macrodatos, por tanto, no deben ser una limitación para el investigador cualitativo, sino que pueden ser la puerta abierta a resultados realistas y precisos, siempre que se utilicen las llaves adecuadas para abrirla (es decir, la extracción de palabras clave y el tamaño de la muestra).

Desde el punto de vista del contenido expresado por los usuarios de Twitter/X y su expresión lingüística, la sexta lección aprendida es que debemos distinguir la polarización de la politización, porque no todo en Twitter/X está relacionado con la política, ni siquiera un programa de televisión que se ha considerado fuertemente influido por las relaciones políticas entre países como es el caso de Eurovisión. Esto refuerza el

hecho de que nosotros, como lingüistas de corpus, debemos mirar de cerca y más de dos veces los datos y evitar guiarnos por nuestras propias ideas preconcebidas. Por ello, el análisis cualitativo sigue siendo importante incluso cuando estamos tratando con grandes cantidades de texto, cuyo análisis cuantitativo es la mejor puerta de entrada, pero no el único recurso.

La séptima y última lección aprendida es que los datos extraídos de Twitter/X deben investigarse desde dos perspectivas complementarias: el sentimiento y el discurso. La presencia de lenguaje de opinión en Twitter/X no puede obviarse ni reducirse a un hashtag o tema concreto. La naturaleza y el contenido de estos mensajes no deben considerarse como estáticos, sino que son cambiantes de un hablante a otro, y varían a medida que se desarrollan los acontecimientos. Esto nos remite al primer capítulo de esta tesis doctoral, donde se afirmaba que los usuarios de Twitter/X hablan de cualquier tema: en otras palabras, la expresión de opiniones en Twitter/X está sujeta a diversos parámetros y, en consecuencia, su análisis también debe tener en cuenta dichas variables. Si el objetivo del investigador es estudiar el sentimiento transmitido en estas opiniones, dicha tarea debe ir acompañada adicionalmente de un análisis basado en el discurso. De lo contrario, se corre el riesgo de sobregeneralizar erróneamente los resultados obtenidos: por ejemplo, las palabras con polaridad negativa no siempre implican un mensaje negativo. Por el contrario, podrían funcionar como enfatizadores. Por lo tanto, si sólo nos centramos en el análisis del sentimiento sin utilizar técnicas de análisis del discurso que tengan en cuenta el contexto cultural y social, podríamos acabar con la impresión de que nuestro corpus es extremadamente positivo o extremadamente negativo, lo que puede no ser ni siquiera el caso.

En esencia, estas lecciones sirven colectivamente como guía para los investigadores que navegan por el intrincado panorama de Twitter/X. Al reconocer la

naturaleza polifacética de la expresión lingüística, las limitaciones de las herramientas informáticas, y las dependencias contextuales dentro de los conjuntos de datos, los investigadores pueden obtener una comprensión más matizada y precisa del sentimiento y el discurso en esta plataforma digital en constante evolución. El viaje a través de estas lecciones no es sólo un reflejo de las complejidades de Twitter/X, sino también un testimonio de la evolución de las metodologías necesarias para captar la esencia de la comunicación en línea desde el punto de vista de la investigación lingüística.