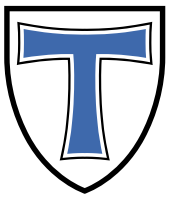
Justus Liebig University Giessen



Dissertation

LEVERAGING UNSTRUCTURED DATA TO ADDRESS SOCIETAL CHALLENGES IN THE DIGITAL AGE

Submitted in fulfilment of the requirements for the degree of

DOCTOR RERUM POLITICARUM (Dr. rer. pol.)

at the

Justus Liebig University Giessen
Faculty of Economics and Business Studies
Chair for Data Science & Digitization

by

Kirill Solovev
October 20, 2024

Druckdatum: 21.10.2024

URL: https://doi.org/10.22029/jlupub-19035

Justus-Liebig-Universität Gießen
Fachbereich Wirtschaftswissenschaften
Professur für Data Science & Digitalisierung
Licher Straße 62
35394 Gießen

Dekanin:	Prof. Dr. Corinna Ewelt-Knauer
Erstgutacher:	Prof. Dr. Nicolas Pröllochs
Zweitgutachterin:	Prof. Dr. Jella Pfeiffer

Datum des Promotionsbeschlusses: 23.09.2024

Acknowledgements

First of all, I would like to sincerely thank my supervisor Prof. Dr. Nicolas Pröllochs, without whose guidance and support this dissertation would not have existed. He supported me in my academic and professional endeavors, and I am incredibly fortunate to have had a chance to collaborate with and work under him. I also thank my second supervisor, Prof. Dr. Jella Pfeiffer, for her support and positivity during the dissertation process and for her valuable feedback on my initial research proposal.

I would also like to thank Markus Rosenfelder, who helped me develop my master's thesis, which became my first published paper, and Prof. Dr. Dirk Neumann, on whose academic chair I gained the knowledge necessary to write it.

Last but not least, I'd like to thank my family, without whose support I couldn't realize my academic passion in Germany.

Special thanks go to my cat Diego, who always provided emotional support during the late-night writing sessions.

Contents

Chapter 1

Introduction

1.1 Motivation

The Digital Age, also commonly referred to as the Information Age or New Media Age, has transformed our increasingly interconnected society (e. g., Barrett et al., 2015; Larson & DeChurch, 2020; Steelman et al., 2014; Wu et al., 2019). This shift is characterized by a move from traditional industrial production methods to economies and societies that rely on the efficient management of vast amounts of information (Webster, 2014). The digital revolution, driven by the widespread use of computing technology, has paved the way for innovative methods of storing and processing information, and has fundamentally altered the way individuals, institutions, and societies interact and function (Baym, 2015).

The rise of the digital age has promoted a steadily quickening shift from traditional data structures and applications to digitized versions. For instance, this shift transforms information dissemination in the media space from a one-to-may communication style typically used by traditional media channels, where information flows from a single source to a widespread audience, to a many-to-many communication style facilitated by social media platforms (Kramer et al., 2014). This alteration empowers increased cross-interaction (Kong et al., 2013), collaborative efforts (Bak-Coleman et al., 2021), and participatory activities (Boulianne, 2015). Rather than being passive recipients of one-way transmissions from authoritative figures or subject matter experts, users now participate in the conversation, allowing for two-way communication and self-generation of content. Similarly, in economic contexts, the digital age has reduced the need for intermediaries and has alleviated information asymmetries. For example, in real estate, property owners can now independently list their living spaces online, securing fair prices while offering prospective buyers and renters a comprehensive market overview (T. Jiang et al., 2019). Such improvements have the potential to directly translate into better living conditions for those previously confined to less desirable options. Overall, these applications in the digital

age can be viewed as sources of data, offering the potential for exhaustive real-time data processing, focused content distribution, and pervasive accessibility. This heralds an unparalleled resource that, when skillfully integrated with data science techniques, has demonstrated immense potential in tackling societal challenges and generating value for the economy and society.

Despite the benefits of increased utilization of the emerging data sources, several challenges must be addressed. First, challenges arise from the unregulated use of digital platforms, the rapid generation and dissemination of digital data, and the growing influence of data-driven decision-making. For instance, on social media, the user-generated nature of information online makes it susceptible to the potential spread of misinformation (e. g., Cinelli et al., 2020; Del Vicario et al., 2016; Vraga & Bode, 2017). This is especially evident during significant global events such as political campaigns and elections (e. g., Bovet & Makse, 2019; Grinberg et al., 2019), or public health crises (Krittanawong et al., 2020). When inaccurate, false, or misleading information is disseminated through various digital channels without rigorous fact-checking or source verification, it can not only sway public opinion but also incite panic, contribute to health hazards, or undermine the public's trust in institutions and the concept of truth itself (Ecker et al., 2022). Another emerging concern in the digital realm is the growing prevalence of hate speech, mobbing, and online harassment (Mathew et al., 2019). Pseudonymity provided by online platforms, combined with the power and reach of digital connectivity, often emboldens individuals to express harmful, violent, or extreme sentiments (Munger, 2016). The subsequent escalated level of hate speech, online bullying, and harassment is becoming a significant issue, particularly in the context of political discourse on social media platforms. In such cases, the misuse of freedom of speech and the platform can exacerbate societal divisions, create discord, and breed hostility (Hopp et al., 2020).

On the other hand, extracting knowledge from user-generated data also poses a challenge due to its unstructured format, necessitating advanced data science methods to extract meaningful insights. Each data format, whether text- or image-based, requires unique approaches and methodologies to extract complex information from unstructured data. While humans can interpret such data relatively quickly, it often presents a convoluted process further exacerbated by the need to analyze large amounts of heterogeneous data. Text-based data, which constitutes a significant portion of unstructured data from social media and web resources, presents semantic (Guan et al., 2016), contextual, and linguistic as well as resource challenges (Ranathunga et al., 2023). This requires refined linguistic and statistical representations to interpret text data in a meaningful way. Techniques like entity extraction (Al-Moslmi et al., 2020), sentiment

analysis (Birjali et al., 2021), and topic modeling (Vayansky & Kumar, 2020) are necessary to extract insights from such data. Image data poses a further set of complications that require the application of feature extraction techniques to prepare them for machine learning algorithms. These can include edge detection, shape recognition, or texture analysis for static images. Various programmatic (e. g., OpenCV) and machine-learning approaches have facilitated this process. Currently, this field is dominated by neural networks (Dhillon & Verma, 2019).

Addressing these challenges is further complicated by the nature of the user-generated information. In particular, the overwhelming magnitude of information continuously generated online may prove too demanding for manual analysis techniques designed for more static, manageable volumes of information. According to X (formerly Twitter) Engineering, their social media platform generates approximately 400 billion events and petabytes of data daily (L. Zhang & Malife, 2023). Analysis of such "big data" necessitates the development and adoption of automated methods capable of processing and analyzing large amounts of data within reasonable time frames and computational constraints (Fan et al., 2014).

This doctoral dissertation aims to develop advanced data science-based methodologies to address the challenges presented by unstructured data and leverage its potential to solve societal problems in the Digital Age. Through applications like hate speech detection, analysis of the spread of misinformation, and price appraisal for real estate, we aim to offer valuable insights, develop practical techniques and tools, and advance knowledge in the respective research domains. The primary objectives of this work are threefold: (1) develop advanced techniques for processing, analyzing and interpreting unstructured data to help bridge the gap between the increasing prevalence of unstructured digital data and the ability to transform that information into insightful knowledge. (2) Apply these methods to solve pressing societal challenges while also broadening our understanding of them, thereby contributing to societal well-being and development. (3) Illustrate the significant amount of untapped information in user-generated data online. My contributions and findings broaden our understanding of these challenges and provide enhanced data science techniques to mitigate them.

1.2 Opportunities and Challenges of Unstructured Data

The advent of the digital age incurs considerable data (Hilbert, 2022) with the majority of data generated online being unstructured in nature (Mughal, 2018). Unstructured data, being relatively unfiltered and spontaneous, has the potential

of providing an authentic representation of societal and economic phenomena (Blazquez & Domenech, 2018). At the same time, the unstructured nature and the fact that large amounts of data are rapidly generated may lead to new challenges when it comes to capturing the underlying information and knowledge (e. g., Balducci & Marinova, 2018; Grossman & Pedahzur, 2020; Patil et al., 2018). As such, unstructured data presents not only a promising resource with several potential applications and insights for research, business, and society in general, but also a formidable challenge for analysis.

In an economic context, unstructured data has emerged as a pivotal tool for businesses to remain sustainable and competitive in highly dynamic marketplaces (Eberendu et al., 2016). By analyzing unstructured data, such as product reviews (Haque et al., 2018), browsing patterns (Su & Chen, 2015), and customer communications (Rese et al., 2020), businesses can tailor their offerings to better meet customer needs and preferences. Further, insights gained from unstructured data analysis can drive marketing strategies and improve overall business performance (Balducci & Marinova, 2018). However, the inherent complexity of unstructured data necessitates significant investment in infrastructure and specialist expertise for effective data management (Venkatraman & Venkatraman, 2019), and careful attention to issues around data ownership and misuse (e. g., Duch-Brown et al., 2017; Livingstone et al., 2019; Martin & Murphy, 2016; Soussan & Trovati, 2021).

Societally, the pervasive reach of online platforms and analysis of user-generated context could catalyze meaningful changes. Social media platforms, in particular, hold great promise in enhancing democratic processes by facilitating broad-based discussions and political mobilizations (Jackson & Lilleker, 2009; Larsson, 2014). For instance, analyzing politicians' social media posts makes it possible to measure public sentiment and forecast electoral prospects. In contrast to traditional media, social media platforms allow for direct feedback and discussion from constituents (Enli & Skogerbø, 2013), providing an unprecedented data source. Responsible use of unstructured data can provide the foundation for evidence-based policies, driving societal progress (e. g., Asensio et al., 2020; Fraisl et al., 2022; Gök et al., 2022; Ilieva & McPhearson, 2018). On the other hand, however, the same platforms can be used by paid agencies (Isaak & Hanna, 2018), foreign adversaries (Eady et al., 2023), or unintentional participants of misinformation campaigns (Bovet & Makse, 2019) to affect democratic processes, thereby compromising public discourse and societal integrity.

In summary, unstructured data accompanying the digital age presents formidable challenges and untapped opportunities. Irrespective of the complexities of leveraging this abundant resource, its potential to drive critical insights across

diverse domains is undeniable. Therefore, the fundamental task is to devise robust methodologies to prioritize and harness this resource effectively while maintaining the highest ethical standards. Integral to this task is the imperative to mitigate associated challenges, ensuring data-driven advances serve the interest of a more fair, inclusive, and prosperous society.

1.3 Unstructured Data Analysis

Unstructured data lack conventional structure or organization, making it inherently more challenging to search, understand, or process using standard algorithms (Eberendu et al., 2016). Any data that cannot be presented using standard data tables can be categorized as unstructured. This can be textual data, such as business reports, video data, such as social media posts, or a combination of both, such as listings on real estate platforms, all of which can contain complex data, providing rich insights that may not be available from traditional structured data sources (e. g., Ngiam & Khor, 2019; Poursaeed et al., 2018; Teinemaa et al., 2016; D. Zhang et al., 2020).

Evidence suggests that unstructured data comprise the majority of all digital data, with an estimated 85 % of all data being unstructured (Eberendu et al., 2016). Furthermore, there is an ever-increasing amount of unstructured data generated by new and emerging online platforms emblematic of Web 2.0 (Bao & Shang, 2021), colloquially referred to as participatory or social web. Unstructured data is innately simple to produce and expedites communication; consider a typical internet user capturing an image and swiftly posting it to a social media platform, accompanied by a succinct caption, making it instantaneously accessible to their followers. Equally advantageous to the data receivers, unstructured data, even those as straightforward as an image, can offer a wealth of information that would conventionally demand more effort to articulate using only textual descriptions. This highlights the flexibility and multi-modality of the unstructured online data. Hence, unstructured data holds considerable explanatory potential in deciphering less explored facets of societal interactions in the digital age. Understanding these interactions is crucial for addressing societal challenges, thus making unstructured data analysis exceptionally relevant in a rapidly evolving informational ecosystem.

However, transforming unstructured data into valuable insights at scale poses complex challenges given its non-standardized and hence variably interpretable format. Thus, it is necessary to employ advanced data science tools (Das & Kumar, 2013). Textual and visual (e. g., image) data are prominent components of

unstructured online data. For example, microblogs, like tweets, present foundational challenges due to their inherently informal and aphoristic nature (Qiang et al., 2022). These challenges span from basic tasks such as language identification (DeLucia et al., 2022), sentiment or opinion detection (Giachanou & Crestani, 2016), and thematic categorization (Garcia & Berton, 2021) to more intricate endeavors like hate-speech detection (Ayo et al., 2020) and establishing broader contextual relevance (Bamman & Smith, 2021). Image data necessitates a complex spectrum of image recognition and object classification algorithms (Sharma & Mir, 2020), deep learning frameworks (Z. Li et al., 2021), and transformers (Touvron et al., 2021). Moreover, unlocking the potential of multimodal information (e. g., Huang et al., 2020; Joshi et al., 2021), which combines text and image data, demands heightened levels of data handling. This presents challenges in uncovering semantic relationships (Cohn, 2016) and synchronicities between diverse data modalities (Belcavello et al., 2020). Such demands highlight the sophistication required to address these challenges effectively. With more advanced and complex data science techniques, we can gain a deeper and more intricate understanding of the hidden complex features contained in unstructured data, unraveling its potential to address pertinent societal challenges in the digital era. Thus, the current climate of rapid data generation provides unprecedented challenges and opportunities for research, requiring progressive adaptation, robust data strategies, and ongoing technical refinements to maximize the informative potential of varying types of unstructured data.

When working with text, the amount of information and data-engineered dimensions is virtually limitless, but the more prominent ones include topic modeling and sentiment analysis. Topic modeling is a prevalent method for text data categorization (Vayansky & Kumar, 2020), involving the categorization of large volumes of documents into various topics and transforming them into interpretable structures. A popular algorithm for topic modeling is the Latent Dirichlet Allocation (LDA), operating under the premise that each document is a confluence of specific topics, each being a distribution of words (Blei et al., 2003). Despite its widespread popularity, its suitability varies depending on the data type. For instance, shorter texts may pose a challenge (Nigam et al., 2000; Quan et al., 2015; Yan et al., 2013). Recently, neural networks and large language model (LLM) approaches have started to gain popularity. For example, LLM RoBerta can be fine-tuned to predict topics with a high degree of accuracy (Guo et al., 2021), supports multiple languages, and can even address issues of low-resource languages (Deng et al., 2020). Regardless of the methodology used, effectively-crafted topic models can unveil the latent semantic structures within vast corpora of textual data, aiding in data exploration and knowledge discovery.

Another prominent methodology in textual research is sentiment analysis or opinion mining. It leverages natural language processing, text analysis, and machine learning to discern subjective information or sentiments in the data. As with the topic modeling, there are multiple methodologies, from dictionary-based analysis with Plutchik's wheel of emotions theory as a backbone (Plutchik, 1984), to neural networks (Adoma et al., 2020) and commercially available LLMs (Wang et al., 2023), though the latter are currently in its infancy. Automatization of sentiment analysis allows researchers to extract the sentiment of people on a variety of topics, including market research (Rambocas & Pacheco, 2018), political science (Ceron et al., 2013), and epidemiology (Daghriri et al., 2022).

In the realm of visual data, the automation of data extraction is mainly dominated by machine learning methodologies and computer vision techniques. Central to this process are Convolutional Neural Networks (CNNs), a type of deep learning model standard in image classification, segmentation, and detection (Gu et al., 2018; Z. Li et al., 2021). CNNs are particularly efficacious due to their distinctly engineered architecture comprising convolutional, pooling, and fully connected layers, which systematically process image inputs and reduce them into high-quality features. These networks exploit spatial correlations by applying filters that convolve along the breadth and height of an input, generating feature maps that retain spatial information. Consequently, CNNs pose remarkable advantages over traditional ML algorithms that treat image pixels independently. State-of-the-art CNN models drastically reduce potential overfitting and have demonstrated superior performance in image analysis tasks (e. g., Yu et al., 2021). In addition to the state-of-the-art neural networks, graphical data analysis has been bolstered by open-source tools like OpenCV (Open Source Computer Vision Library) (Culjak et al., 2012). Predominantly, OpenCV has gained a reputation for providing a comprehensive suite of functions that can aid researchers and practitioners in extracting valuable insights in real-time from static images and video streams. From fundamental image manipulation tasks, such as cropping, blurring, and thresholding, to more advanced operations like contour detection, object detection, and tracking, OpenCV provides capabilities for a variety of image analysis operations by integrating with machine learning libraries, allowing for use alongside the CNNs, GANs and other methodologies for more advanced tasks (Harriat Christa et al., 2021). Consequently, OpenCV's contribution to the field exposition is not to be undermined, and it continues to serve as a crucial toolkit in the grander discourse of unstructured data manipulation and information extraction.

In summary, unstructured data, encompassing a wide variety of digital content such as text and images, is proliferating at an unprecedented scale, mainly

due to user-generated contributions on increasingly participatory online platforms. This presents both a unique challenge and an opportunity within the field of data science. The inherent complexity and variability of unstructured data require meticulous and sophisticated techniques for extracting, analyzing, and presenting the underlying information. Advanced tools and methodologies, notably natural language processing and machine learning for text analysis and convolutional neural networks for image extraction, are instrumental in transforming this raw, unstructured data into meaningful insights capable of addressing sophisticated societal challenges. Harnessing this potential necessitates continuous refinement and innovation with cutting-edge techniques, offering considerable advancements in tackling the challenges of unstructured data. Affirmed by the continuous advancement and refinement of data analysis methodologies, researchers now have a more sophisticated, capable toolbox to tackle the intricacies of unstructured data.

1.4 Research Framework

1.4.1 Hate Speech in Political Discourse

In recent years, social media has emerged as a crucial tool in the political communication and discourse (e. g., Gainous & Wagner, 2014; Lewandowsky et al., 2020). It enables politicians to connect with people personally, share messages on a far larger scale, and gain broader support (Stier et al., 2020). Social media is unique in its ability to facilitate dynamic, two-way communication in real time, which can help to reinforce relationships with the public and allow the real-time adjustment of political strategies (Schöll et al., 2023). The interactive nature of social media also promotes political mobilization, allowing politicians to use social networks for campaigns, volunteer recruitment, and to rally supporters (González-Bailón et al., 2022). Regardless of geographical boundaries, social media empowers politicians to extend their sphere of influence, reach larger audiences, and maximize the impact of their communications (Petrova et al., 2021).

However, as a new means of political communication, social media has its challenges. The openness of social media forums that fosters democratic expression also allows for the propagation of hate speech (Castaño-Pulgarín et al., 2021) and exacerbates political polarization (Waller & Anderson, 2021). As such, social media does not necessarily improve the civility or quality of political discussions, as it often creates digital echo chambers (Barberá et al., 2015) and reinforces a divisive “us versus them” mentality (Mondal et al., 2017). This can result in

cyberbullying, harassment, and, in particular, an increase in hate speech (Cinelli et al., 2021). Hate speech refers to abusive or threatening language demonstrating prejudice against particular groups (Sellars, 2016). The presence of hate speech tends to be particularly pronounced in emotionally charged discussions, making it specifically widespread in political communication, where divisive rhetoric can further trigger extreme responses (Wagner, 2020). One of the reasons is implied pseudonymity provided by the online communities, which is a fertile ground for those whose primary intent is to instigate and fan the flames of discord (Himma & Tavani, 2008; Mondal et al., 2017).

Hate speech in political discourse could co-opt and hijack legitimate discourse and democratic mechanisms. One recent example is the aftermath of the British referendum on the European Union membership, colloquially referred to as "Brexit," which was co-opted by Islamophobic rhetoric on Twitter, replacing what could have been constructive political discussions with a discourse that promotes symbolic violence, essentially removing the possibility of a civil discourse (Evolvi, 2019). Similarly, in Colombia, the rejection of the peace agreement between the government and the rebel group FARC was heavily influenced by misinformation and hate speech propagated through social media, which, among other factors, targeted religious and ethnic groups (Branton et al., 2019). Presidential campaigns are equally susceptible to the rhetoric of hate, as illustrated by the U. S. elections, where conservative clusters were particularly active in disseminating anti-Muslim hashtags accompanying a narrative steeped in racial, anti-immigration, and white nationalist sentiments (Sainudiin et al., 2019). Some politicians have tried to trivialize these harmful narratives or construed them as virtuous acts, which has further contributed to the proliferation of harmful stereotypes involving race, ethnicity, religion, and gender in political discourse (Ben-David & Fernández, 2016).

The growing prevalence of hate speech in political discourse presents a significant threat to society at all of its levels. At the individual level, this phenomenon contributes to reputation damage and the risk of long-term mental health issues (Vidgen et al., 2021). It can prevent candidates from participating in the democratic processes, such as Rodrigo Londoño, who was forced to withdraw from the presidential race in Colombia after facing an onslaught of online and offline hatred (Tabares Higuita et al., 2018). It further precipitates intense political polarization at the societal level, worsening relations between different political and social groups (Piazza, 2020). The toxic environment created by such polarization serves as fertile ground for ideologically driven misinformation, hence distorting reality and compromising the quality of conversations in the digital public sphere (Freelon et al., 2020; Pröllochs, 2022; Solovev & Pröllochs, 2022).

Accordingly, the role of social media in shaping political discourse has become a double-edged sword, presenting both potential harms and advantages (Hong et al., 2019).

Existing research suggests that there is a difference in the way different groups of people communicate online. Political party affiliation in the U. S. influences speech patterns (Sylwester & Purver, 2015), as does gender, with women displaying less expressive and negative emotions (Davis, 1995). Society's gender biases further emphasize these differences, often leading to women being overlooked or rejected online when displaying assertive behavior (Winkler et al., 2017). This stereotyping extends to racial and ethnic identification (Tatum, 2017). Despite existing qualitative insights and summary statistics, there is a clear research gap in quantitative research, empirically modeling the effects of personal characteristics on the likelihood of receiving hate speech.

Overall, the rise of hate speech in political discussions on social media has proven to be a growing concern. While the digital age has made it easier for people to communicate and participate in politics, it has also created a need for tools to monitor and control negative behavior. Although social media has helped spread political awareness and mobilization (Enli & Skogerbø, 2013; Gainous & Wagner, 2014; Graham et al., 2013; Hong & Nadler, 2012; Jackson & Lilleker, 2009; Larsson, 2014; Ross et al., 2014), we need to be cautious and take measures to prevent hate speech from becoming the dominant narrative in political conversations online.

1.4.2 The Role of Moral Emotions in the Propagation of Hate Speech

The proliferation of hate speech on social media has been thoroughly established (Mathew et al., 2019), yet the precise underlying mechanisms facilitating its widespread dissemination remain under-explored. We bridge this research gap by scrutinizing the semantic attributes of the underlying messages. Specifically, supported by previous works (Brady et al., 2017, 2020; Heltzel & Laurin, 2020; Hoover et al., 2021; Pretus et al., 2022; Sternberg, 2003), we postulate that the moral and emotional contents of the discourse might influence the spread of hate speech.

We contribute to the research by exploring a morality-focused perspective, providing a hypothesis regarding the spread of hate speech in the context of social media, along with supporting analysis. Beyond delivering factual information, unstructured messages can additionally convey moralized content (Brady et al., 2017). Content is deemed moralized when it refers to concepts, entities, or events viewed in the context of a system larger than the individual, such as society as a whole (Brady et al., 2020). Considering that interconnected social

media users often share similar ideas and intuitions, moralized content tends to drive information diffusion on these platforms (Brady et al., 2017), but risk polarizing the users (Heltzel & Laurin, 2020), leading to contention, animosity, and malice from groups with opposing ideologies. Previous research suggests that moral issues tend to be the differentiating factor between hate and dislike (Pretus et al., 2022), where hate could reflect perceived moral wrongdoings by the outgroup (Sternberg, 2003), leading to the perceived righteousness of their hateful actions (Hoover et al., 2021). Furthermore, moralized content assists in building group identity (Brady et al., 2020), implying that it might stir up hate within ideologically similar groups against an outgroup. With this understanding, if moralized content incites affective responses in users, then its transmission could be integral to the propagation of hate speech. Hence, we postulate that there is a connection between moralized language in social media posts and an increased propensity to receive hate speech across the fields of politics, news media, and activism.

1.4.3 Drivers of Misinformation During the COVID-19 Pandemic

Building on our analysis of the role of moral emotions in the propagation of hate speech, we now turn to the related issue of misinformation. The virality of misinformation, particularly during crises such as the COVID-19 pandemic, is well-documented (Pröllochs et al., 2021; Vosoughi et al., 2018). Related research has established that the effects of misinformation propagate into the real world, e. g., by obstructing public health efforts (Mosleh & Rand, 2022; Pennycook et al., 2020; Ricard & Medeiros, 2020).

Prior research analyzed the spreading dynamics of rumors vs. non-rumors, but has not focused on diffusion differences across veracities (e. g., Bessi et al., 2015; Del Vicario et al., 2016). Additionally, it did not examine the differences in diffusion context of global events, such as COVID-19 (e. g., Vosoughi et al., 2018), and focused on summary statistics of small sets of hand-labeled rumors and source-based misinformation identification approaches (e. g., Kouzy et al., 2020; Singh et al., 2020). For example, Cinelli et al. (2020) classify sources as reliable and not-reliable, potentially misclassifying false rumors from sources perceived as reputable. In summary, while previous research – at least for non-crisis situations – suggests that false rumors on social media tend to be more viral than the truth (Pröllochs et al., 2021; Vosoughi et al., 2018), the mechanism underlying its viral spread, though critical, remains unresolved.

We address this research gap by collecting a unique large dataset and conducting a comprehensive quantitative analysis into the differences in false vs. true rumor diffusion through the lenses of morality and emotion. Based on prior

studies exploring the idea of false rumors being more viral than true ones (Pröllochs et al., 2021; Vosoughi et al., 2018), we extend the research by analyzing how this spread occurs, particularly in light of moral emotions. Prior research established the impact of moral-emotional signals on the propagation of rumors (Brady et al., 2017; Tangney et al., 2007; Wheatley & Haidt, 2005), which led us to the premise that such emotions might be instrumental in explaining the flow of information within highly polarized environments.

Our central hypothesis is that the disparity in the virality of true and false COVID-19 rumors can be attributed to the moral emotions they contain (Pröllochs et al., 2021; Vosoughi et al., 2018). In particular, we focus on two types of complex emotions: other-condemning and self-conscious emotions. Misinformation tends to proliferate within polarized digital echo chambers, where ideological stances often preempt attempts to verify the truth (Choi et al., 2020; Kim, 2017; Moravec et al., 2019; Weng et al., 2013). We hypothesize that within such discourse environments, other-condemning emotions originating in the initial tweets potentiate the propagation of false rumors, whereas self-conscious emotions dampen the spread.

1.4.4 Leveraging Unstructured Data for Rent Price Appraisal

Expanding from the purely textual unstructured data sources to a multi-modal one, the real estate industry provides a compelling setting where such data sources can be exploited. The advent of online real estate platforms has substantially transformed the industry's landscape, offering a comprehensive display of potential properties based on chosen locations and greatly expediting the search for potential homes (Yuan et al., 2013). In particular, online real estate platforms contain multi-modal unstructured data in form of text, images, and even videos. Pioneering online platforms such as Zillow have demonstrated exponential growth in their user base, from 25 million unique visitors in 2011 to 224 million unique monthly visitors in 2023 (Zillow Group, 2023), hinting at a robust increasing growing demand from home buyers and renters (Kaklauskas et al., 2021; Zumpano et al., 2003). The spikes in usage underscore the imperative for sophisticated tools and algorithms capable of accurately assessing real estate prices (e. g., Y. Jiang et al., 2020; Yuan et al., 2013).

Contemporary research regarding real estate valuation extensively revolves around applying hedonic price models (Monson, 2009; Sopranzetti, 2010; Wallace & Meese, 1997). These models represent the theory that property values or rents are an amalgamation of their varied features (Wallace & Meese, 1997). While the hedonic approach effectively weighs the numerous structural and

locational attributes in property valuation, it often overlooks additional information available on modern online real estate platforms. Real estate has considerably benefited from the infusion of computer vision and visual analytics, data visualization, sentiment analysis, and feature extraction (e. g., Bappy et al., 2017; Glaeser et al., 2018; Kiyota, 2021). Pioneering endeavors leveraging machine learning have simplified real estate data visualization, making it more amenable to interpretation (M. Li et al., 2018; Sun et al., 2013). Moreover, prior works have extracted and analyzed sentiments from images to provide a more comprehensive understanding of property prices (Glaeser et al., 2018; H. Ahmed & Moustafa, 2016; Naumzik & Feuerriegel, 2020; You et al., 2017). For instance, satellite images have been utilized to glean neighborhood data as an integral component of hedonic modeling processes (Bency et al., 2017). Ground-based snapshots of properties and more zoomed-out satellite imagery have become increasingly popular data sources. For example, some works consider the aesthetics of the property exterior and the surrounding neighborhood as well as their effect on the accuracy of hedonic pricing models (Glaeser et al., 2018).

Despite the increased attention to the visual features, one common yet overlooked feature is the floor plan of the property. Attesting to the significance of this feature, a report on rightmove.co.uk highlighted that 90 % of home buyers deemed a floor plan a crucial factor when searching for real estate. Floor plans offer a bird's-eye view of the property's layout, elucidating the spatial relationship between different areas. Much of the information derived from a floor plan is scarcely available in a structured format or comprehensively expressed in text (Goncu et al., 2015). We address this deficiency by devising and implementing a method of extracting the relevant information from the images of floor plans.

1.5 Thesis Structure

In this doctoral dissertation, I present a collection of four research papers (see Table 1.1). Each of the papers covers one of the aspects motivated in the preceding sections and contextualized through Section 1.4. This section provides a brief introduction to each of the papers in this dissertation and outlines the structure of the following chapters.

Chapter 2: Hate Speech in the Political Discourse on Social Media: Disparities Across Parties, Gender, and Ethnicity

In Chapter 2, we quantitatively scrutinize the increasing prevalence of hate speech in political discourse on the social media platform X (called Twitter at

the time of the original publication). This empirical investigation employs X's Historical API to gather an extensive dataset comprising tweets by members of the 117th U. S. Congress over an observation period exceeding half a year. The research has appeared in the following conference proceedings:

> **Solovev, K.**, & Pröllochs, N. (2022). Hate speech in the political discourse on social media: Disparities across parties, gender, and ethnicity. *Proceedings of the WWW*

We quantify the effect of personal characteristics, such as party affiliation, gender, and ethnicity of the U. S. congresspeople, on the likelihood of receiving hate speech in replies to tweets. To this end, we leverage machine learning for hate speech prediction and hierarchical regression models for discerning significant variations.

The results of our analysis demonstrate that replies containing hate speech are statistically more likely when the source tweets are authored by Democrats of color, white Republicans, or female politicians. Subsequent sentiment analysis reveals that the higher the negativity in the source tweet, the higher the associated concentration of hate speech in its replies. Noteworthy, this association is not present uniformly across parties, with Democrats attracting more hate speech in comparison to Republicans when the source tweet sentiment is negative.

In summary, our findings in this paper illuminate stark disparities in politicians' treatment on X, contingent on their party association, gender, and ethnicity. Our results are especially relevant in the context of societal dynamics, as targeted hate speech can discourage participation in politics, jeopardizing diversity in political representation. On a broader scale, the documented presence of hate speech is indicative of heightened social polarization, which poses a threat to the effective functioning of democratic institutions. Hence, the unveiled disparities in the treatment of politicians online prompt pressing issues related to social equity, public health, political participation, and democratic functionality in today's digital society. This lays the ground for the subsequent paper that further explores the subject of hate speech on social media and sets to derive an explanation behind the proliferation of hateful rhetoric.

Chapter 3: Moralized Language Predicts Hate Speech on Social Media

In Chapter 3, we delve deeper into the mechanisms driving hate speech proliferation on digital platforms – a phenomenon with considerable implications for societal safety and individual mental health. Here, we postulate a direct link between moralized language usage and the spread of hate speech. This chapter has appeared in the following journal article:

Solovev, K., & Pröllochs, N. (2023). Moralized language predicts hate speech on social media (J. Van Bavel, Ed.). *PNAS Nexus*, *2*(1), pgac281

We collected three large-scale datasets, amassing approximately 691,234 social media posts and 35.5 million associated replies, spanning the entirety of 2021. In addition to posts made by the politicians, we gather conversations from news people and activists, facilitating a more diverse understanding. A combination of textual analysis and machine learning underpins our investigative approach, aiding in deciphering the association between moralized language in source tweets and the manifestation of hate speech within replies.

Across all the studied datasets, we observe that an increase in the usage of moral and moral-emotional words within a tweet portends a heightened likelihood of corresponding replies containing hate speech with increments of 10.66–16.48 % and 9.35–20.63 % respectively. A noteworthy highlight is the potent out-of-sample predictive ability of moralized language, providing clear evidence of its integral role in hate speech proliferation.

Thus, this second paper uncovers novel insights into the drivers of hate speech on social media, potentially guiding efficient intervention strategies. It bridges a gap in social psychological theories by providing empirical evidence for the longstanding theoretical relationship between morality and hate. It substantiates the idea that moralized language can act as a predictor for hate speech on social media platforms, affirming prior psychological theories that associate hate with perceived moral deficiencies in targets. Moreover, the findings can help understand the proliferation mechanisms of hate speech and inform social media literacy, educational programs, counter-strategies, and automatic hate speech detection techniques, laying the groundwork for reducing the spread of hate speech online.

Chapter 4: Moral Emotions Shape the Virality of COVID-19 Misinformation on Social Media

In Chapter 4, we transition from the propagation of hate speech, shifting our attention towards the diffusion of misinformation during the COVID-19 pandemic. In this paper, we explore the relationship between moral emotions and the virality of rumors and rumor cascades. This research has appeared in the following conference proceedings:

Solovev, K., & Pröllochs, N. (2022). Moral emotions shape the virality of COVID-19 misinformation on social media. *Proceedings of the WWW*

We study rumor cascades on X by identifying COVID-19-related rumors from large third-party fact-checking organizations. We reconstruct 10,610 rumor

cascades, accounting for over 24 million retweets between January 2020 and April 2021. To study whether misinformation is more viral than the truth and whether differences in the diffusion of true and false rumors are moderated by the moral emotions they convey, we employ a dictionary-based approach validated by research assistants, focusing on self-conscious and other-condemning emotions.

Our findings suggest that, on average, misinformation about COVID-19 exhibits higher virality than the truth, corroborating previous research. This association, however, is moderated by moral emotions: in situations where source tweets contain a larger share of other-condemning words, false rumors are likely to be more viral. Conversely, an increase in self-conscious words corresponds to a lesser degree of virality. These trends are observable across the entire dataset as well as when separated into health-related and political topics.

This paper broadens our understanding of the dynamics of rumors, emphasizing the pivotal role moral emotions play in their dissemination. This insight underscores the need to consider the hidden and complex features when aiming to understand social media discourse, especially in the context of false information spread during crises. Our research draws connections between the spread of misinformation and its potential exacerbation of health and political crises. It underscores the role of emotional polarization in catalyzing the diffusion of false rumors. The insights afford practical implications, suggesting that considering moral emotions in source tweets could enhance the accuracy of false rumor detection and inform strategies to mitigate rumor spread. Therefore, within the broader landscape of understanding and managing misinformation, this study spotlights the consideration of moral emotions as a crucial factor shaping the dynamics of rumor diffusion in digital communication environments.

Chapter 5: Integrating Floor Plans into Hedonic Models for Rent Price Appraisal

Finally, in Chapter 5, the application of unstructured data analysis is extended beyond social media to include the real estate sector. This research endeavors to aid users and businesses in making informed living and investment decisions by studying how latent information within floor plans can enhance price prediction capabilities on real estate websites. This research has appeared in the following conference proceedings:

> **Solovev, K.**, & Pröllochs, N. (2021). Integrating floor plans into hedonic models for rent price appraisal. *Proceedings of the WWW*

In contrast to the previous papers, which predominantly worked with communication in the digital landscape and looked at false rumors and hate speech diffusions on social media, we now aim to improve user experience on real estate platforms and enhance traditional hedonic pricing models for real estate, which primarily focus on structured data. Using state-of-the-art methods in an automated visual analysis of apartment floor plans, we investigate the possibility to augment these models. To this end, we propose a two-staged deep learning approach to learn price-relevant floor plan designs from historical data. The predictions obtained are then integrated into hedonic rent price models, accounting for an apartment's structural and locational characteristics.

Our empirical research leverages a unique dataset of 9,174 real estate listings to reveal the underutilization of available data of current hedonic models. We find that the visual design of floor plans holds significant explanatory power regarding rent prices over and above the fundamental apartment characteristics. Incorporating floor plans resulted in an up to 10.56 % improvement in out-of-sample prediction performance, offering exceptionally high predictive performance for older and smaller apartments.

This paper contributes to the existing literature by establishing a link between the visual design of floor plans and real estate prices, highlighting the value of harnessing unstructured data – in this case, floor plans – to deepen our understanding of socially relevant phenomena. By demonstrating that the visual design aspects of floor plans contribute to and can improve property valuation, the study highlights implications for optimizing investor portfolios, enhancing user experiences on online platforms, reducing information asymmetry, and improving marketplace transparency. As the last paper in this collection, it extends the unstructured data to a visual domain and encapsulates the overarching theme of this dissertation: Leveraging Unstructured Data to Address Societal Challenges in the Digital Age.

In summary, this dissertation aims to research and remedy several issues that are characteristic of the digital age. Namely, we look deeper into the phenomenon of hate speech in political discourse and general online communication, misinformation on social media, and information inefficiencies on online real estate platforms. In every paper, we leverage advanced data science methods to extract and utilize complex and hidden information in the ever-increasing deluge of user-generated digital data to unlock new insights. Each of the papers included in this dissertation has been peer-reviewed and published in renowned conferences and journals. Table 1.1 presents the list of papers, their publication status, and the authors' contributions. The following sections contain the published papers with slight modifications to fit the overall style of the dissertation.

Table 1.1: Contributions table.

Paper	**Hate Speech in the Political Discourse on Social Media: Disparities Across Parties, Gender, and Ethnicity**
Co-authors	Prof. Dr. Nicolas Pröllochs
Status	Published in the *Proceedings of The Web Conference (WWW'22)*
Key contributions	My contribution is 75 %: • Conceptualized the research focus and methodology • Collected and analyzed the data • Wrote major parts of the manuscript
Paper	**Moralized Language Predicts Hate Speech on Social Media**
Co-authors	Prof. Dr. Nicolas Pröllochs
Status	Published in the *PNAS Nexus*, Volume 2, Issue 1, Article 4
Key contributions	My contribution is 75 %: • Conceptualized the research focus and methodology • Collected and analyzed the data • Wrote major parts of the manuscript
Paper	**Moral Emotions Shape the Virality of COVID-19 Misinformation on Social Media**
Co-authors	Prof. Dr. Nicolas Pröllochs
Status	Published in the *Proceedings of The Web Conference (WWW'22)*
Key contributions	My contribution is 75 %: • Conceptualized the research focus and methodology • Collected and analyzed the data • Conceptualized and implemented rumor cascade reconstruction • Wrote major parts of the manuscript
Paper	**Integrating Floor Plans into Hedonic Models for Rent Price Appraisal**
Co-authors	Prof. Dr. Nicolas Pröllochs
Status	Published in the *Proceedings of the Web Conference (WWW'21)*
Key contributions	My contribution is 85 %: • Conceptualized the research focus and methodology • Collected and analyzed the data • Conceptualized and implemented image data extraction • Wrote major parts of the manuscript

Bibliography

Adoma, A. F., Henry, N.-M., & Chen, W. (2020). Comparative analyses of BERT, RoBERTa, DistilBERT, and XLNet for text-based emotion recognition. *Proceedings of the ICCWAMTIP*.

Al-Moslmi, T., Gallofre Ocana, M., L. Opdahl, A., & Veres, C. (2020). Named entity extraction for knowledge graphs: A literature overview. *IEEE Access*, *8*, 32862–32881.

Asensio, O. I., Alvarez, K., Dror, A., Wenzel, E., Hollauer, C., & Ha, S. (2020). Real-time data from mobile platforms to evaluate sustainable transportation infrastructure. *Nature Sustainability*, *3*(6), 463–471.

Ayo, F. E., Folorunso, O., Ibharalu, F. T., & Osinuga, I. A. (2020). Machine learning techniques for hate speech classification of Twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review*, *38*, 100311.

Bak-Coleman, J. B., Alfano, M., Barfuss, W., Bergstrom, C. T., Centeno, M. A., Couzin, I. D., Donges, J. F., Galesic, M., Gersick, A. S., Jacquet, J., Kao, A. B., Moran, R. E., Romanczuk, P., Rubenstein, D. I., Tombak, K. J., Van Bavel, J. J., & Weber, E. U. (2021). Stewardship of global collective behavior. *PNAS*, *118*(27), e2025764118.

Balducci, B., & Marinova, D. (2018). Unstructured data in marketing. *Journal of the Academy of Marketing Science*, *46*(4), 557–590.

Bamman, D., & Smith, N. (2021). Contextualized sarcasm detection on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*.

Bao, Z., & Shang, B. (2021). Self-efficacy and continuance intention of Web 2.0 platforms: A meta-analysis. *Data Technologies and Applications*, *55*(4), 511–526.

Bappy, J. H., Barr, J. R., Srinivasan, N., & Roy-Chowdhury, A. K. (2017). Real estate image classification. *Proceedings of the WACV*.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531–1542.

Barrett, M., Davidson, E., Prabhu, J., & Vargo, S. L. (2015). Service innovation in the digital age. *MIS Quarterly*, *39*(1), 135–154.

Baym, N. K. (2015). *Personal connections in the digital age* (2nd ed.). Polity.

Belcavello, F., Viridiano, M., da Costa, A. D., da Silva Matos, E. E., & Torrent, T. T. (2020). Frame-based annotation of multimodal corpora: Tracking (a) synchronies in meaning construction. *Proceedings of the IFNW*.

Bency, A. J., Rallapalli, S., Ganti, R. K., Srivatsa, M., & Manjunath, B. S. (2017). Beyond spatial auto-regressive models: Predicting housing prices with satellite imagery. *Proceedings of the WACV*.

Ben-David, A., & Fernández, A. M. (2016). Hate speech and covert discrimination on social media: Monitoring the facebook pages of extreme-right political parties in Spain. *International Journal of Communication*, *10*, 1167–1193.

Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015). Science vs conspiracy: Collective narratives in the age of misinformation. *PLOS ONE*, *10*(2), e0118093.

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*, 107134.

Blazquez, D., & Domenech, J. (2018). Big data sources and methods for social and economic analyses. *Technological Forecasting and Social Change*, *130*, 99–113.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Boulianne, S. (2015). Social media use and participation: A meta-analysis of current research. *Information, Communication & Society*, *18*(5), 524–538.

Bovet, A., & Makse, H. A. (2019). Influence of fake news in Twitter during the 2016 U.S. presidential election. *Nature Communications*, *10*(1), 7.

Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (2020). The mad model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science*, *15*(4), 978–1010.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *PNAS*, *114*(28), 7313–7318.

Branton, Demeritt, Pulido, A., & Meernik. (2019). Violence, voting & peace: Explaining public support for the peace referendum in Colombia. *Electoral Studies*, *61*, 102067.

Castaño-Pulgarín, S. A., Suárez-Betancur, N., Vega, L. M. T., & López, H. M. H. (2021). Internet, social media and online hate speech. Systematic review. *Aggression and Violent Behavior*, *58*, 101608.

Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2013). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society*, *16*(2), 340–358.

Choi, D., Chun, S., Oh, H., Han, J., & Kwon, T. " (2020). Rumor propagation is amplified by echo chambers in social media. *Scientific Reports*, *10*(1), 310.

Cinelli, M., Pelicon, A., Mozetič, I., Quattrociocchi, W., Novak, P. K., & Zollo, F. (2021). Dynamics of online hate and misinformation. *Scientific Reports*, *11*(1), 22083.

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), 1–10.

Cohn, N. (2016). A multimodal parallel architecture: A cognitive framework for multimodal interactions. *Cognition*, *146*, 304–323.

Culjak, I., Abram, D., Pribanic, T., Dzapo, H., & Cifrek, M. (2012). A brief introduction to OpenCV. *Proceedings of the MIPRO*.

Daghriri, T., Proctor, M., & Matthews, S. (2022). Evolution of select epidemiological modeling and the rise of population sentiment analysis: A literature review and COVID-19 sentiment illustration. *International Journal of Environmental Research and Public Health*, *19*(6), 3230.

Das, T. K., & Kumar, P. M. (2013). Big data analytics: A framework for unstructured data analysis. *International Journal of Engineering Science & Technology*, *5*(1), 153.

Davis, T. L. (1995). Gender differences in masking negative emotions: Ability or motivation? *Developmental Psychology*, *31*(4), 660–667.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *PNAS*, *113*(3), 554–559.

DeLucia, A., Wu, S., Mueller, A., Aguirre, C., Resnik, P., & Dredze, M. (2022). Bernice: A multilingual pre-trained encoder for Twitter. *Proceedings of the EMNLP*.

Deng, S., Zhang, N., Sun, Z., Chen, J., & Chen, H. (2020). When low resource NLP meets unsupervised language model: Meta-pretraining then meta-learning for few-shot text classification (student abstract). *Proceedings of the AAAI*.

Dhillon, A., & Verma, G. K. (2019). Convolutional neural network: A review of models, methodologies and applications to object detection. *Progress in Artificial Intelligence*, *9*(2), 85–112.

Duch-Brown, N., Martens, B., & Mueller-Langer, F. (2017). The economics of ownership, access and trade in digital data. *SSRN Electronic Journal*.

Eady, G., Paskhalis, T., Zilinsky, J., Bonneau, R., Nagler, J., & Tucker, J. A. (2023). Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 U.S. election and its relationship to attitudes and voting behavior. *Nature Communications*, *14*(1), 62.

Eberendu, A. C., et al. (2016). Unstructured data: An overview of the data of big data. *International Journal of Computer Trends and Technology*, *38*(1), 46–50.

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13–29.

Enli, G. S., & Skogerbø, E. (2013). Personalized campaigns in party-centred politics: Twitter and Facebook as arenas for political communication. *Information, Communication & Society*, *16*(5), 757–774.

Evolvi, G. (2019). #Islamexit: Inter-group antagonism on Twitter. *Information, Communication & Society*, *22*(3), 386–401.

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National Science Review*, *1*(2), 293–314.

Fraisl, D., Hager, G., Bedessem, B., Gold, M., Hsing, P.-Y., Danielsen, F., Hitchcock, C. B., Hulbert, J. M., Piera, J., Spiers, H., Thiel, M., & Haklay, M. (2022). Citizen science in environmental and ecological sciences. *Nature Reviews Methods Primers*, *2*(1), 64.

Freelon, D., Marwick, A., & Kreiss, D. (2020). False equivalencies: Online activism from left to right. *Science*, *369*(6508), 1197–1201.

Gainous, J., & Wagner, K. M. (2014). *Tweeting to power: The social media revolution in American politics*. Oxford University Press.

Garcia, K., & Berton, L. (2021). Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. *Applied Soft Computing*, *101*, 107057.

Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of Twitter sentiment analysis methods. *ACM Computing Surveys*, *49*(2), 1–41.

Glaeser, E., Kincaid, M. S., & Naik, N. (2018). Computer vision and real estate: Do looks matter and do incentives determine looks. *NBER Working Paper*, *27164*.

Gök, A., Antai, R., Milošević, N., & Al-Nabki, W. (2022). Building the European social innovation database with natural language processing and machine learning. *Scientific Data*, *9*(1), 697.

Goncu, C., Madugalla, A., Marinai, S., & Marriott, K. (2015). Accessible on-line floor plans. *Proceedings of the WWW*.

González-Bailón, S., d'Andrea, V., Freelon, D., & De Domenico, M. (2022). The advantage of the right in social media news sharing. *PNAS Nexus*, *1*(3), pgac137.

Graham, T., Broersma, M., Hazelhoff, K., & van 't Haar, G. (2013). Between broadcasting political messages and interacting with voters: The use of Twitter

during the 2010 UK general election campaign. *Information, Communication & Society*, *16*(5), 692–716.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, *363*(6425), 374–378.

Grossman, J., & Pedahzur, A. (2020). Political science and big data: Structured data, unstructured data, and how to use them. *Political Science Quarterly*, *135*(2), 225–257.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., & Chen, T. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, *77*, 354–377.

Guan, J., Levitan, A. S., & Goyal, S. (2016). Text mining using latent semantic analysis: An illustration through examination of 30 years of research at JIS. *Journal of Information Systems*, *32*(1), 67–86.

Guo, Z., Zhu, L., & Han, L. (2021). Research on short text classification based on RoBERTa-TextRCNN. *Proceedings of the CISAI*.

H. Ahmed, E., & Moustafa, M. (2016). House price estimation from visual and textual features. *Proceedings of the IJCCI*.

Haque, T. U., Saber, N. N., & Shah, F. M. (2018). Sentiment analysis on large scale Amazon product reviews. *Proceedings of the ICIRD*.

Harriat Christa, G., Jesica, J., Anisha, K., & Sagayam, K. M. (2021). CNN-based mask detection system using OpenCV and MobileNetV2. *Proceedings of the ICPSC*.

Heltzel, G., & Laurin, K. (2020). Polarization in America: Two possible futures. *Current Opinion in Behavioral Sciences*, *34*, 179–184.

Hilbert, M. (2022). Information quantity. In L. A. Schintler & C. L. McNeely (Eds.), *Encyclopedia of big data* (pp. 568–571). Springer International Publishing.

Himma, K. E., & Tavani, H. T. (2008). *The handbook of information and computer ethics* (H. T. Tavani & K. E. Himma, Eds.). Wiley.

Hong, S., Choi, H., & Kim, T. K. (2019). Why do politicians tweet? Extremists, underdogs, and opposing parties as political tweeters. *Policy & Internet*, *11*(3), 305–323.

Hong, S., & Nadler, D. (2012). Which candidates do the public discuss online in an election campaign? The use of social media by 2012 presidential candidates and its impact on candidate salience. *Government Information Quarterly*, *29*(4), 455–461.

Hoover, J., Atari, M., Mostafazadeh Davani, A., Kennedy, B., Portillo-Wightman, G., Yeh, L., & Dehghani, M. (2021). Investigating the role of group-based

morality in extreme behavioral expressions of prejudice. *Nature Communications*, *12*(1), 1–13.

Hopp, T., Ferrucci, P., & Vargo, C. J. (2020). Why do people share ideologically extreme, false, and misleading content on social media? A self-report and trace data–based analysis of countermedia content dissemination on Facebook and Twitter. *Human Communication Research*, *46*(4), 357–384.

Huang, S.-C., Pareek, A., Zamanian, R., Banerjee, I., & Lungren, M. P. (2020). Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: A case-study in pulmonary embolism detection. *Scientific Reports*, *10*(1), 22147.

Ilieva, R. T., & McPhearson, T. (2018). Social-media data for urban sustainability. *Nature Sustainability*, *1*(10), 553–565.

Isaak, J., & Hanna, M. J. (2018). User data privacy: Facebook, Cambridge Analytica, and privacy protection. *Computer*, *51*(8), 56–59.

Jackson, N. A., & Lilleker, D. G. (2009). Building an architecture of participation? Political parties and Web 2.0 in Britain. *Journal of Information Technology & Politics*, *6*(3–4), 232–250.

Jiang, T., Guo, Q., Chen, S., & Yang, J. (2019). What prompts users to click on news headlines? Evidence from unobtrusive data analysis. *Aslib Journal of Information Management*, *72*(1), 49–66.

Jiang, Y., Ho, Y.-C. (, Yan, X., & Tan, Y. (2020). When online lending meets real estate: Examining investment decisions in lending-based real estate crowdfunding. *Information Systems Research*, *31*(3), 715–730.

Joshi, G., Walambe, R., & Kotecha, K. (2021). A review on explainability in multimodal deep neural nets. *IEEE Access*, *9*, 59800–59821.

Kaklauskas, A., Zavadskas, E. K., Lepkova, N., Raslanas, S., Dauksys, K., Vetloviene, I., & Ubarte, I. (2021). Sustainable construction investment, real estate development, and COVID-19: A review of literature in the field. *Sustainability*, *13*(13), 7420.

Kim, A. (2017). Says who? The effects of presentation format and source rating on fake news in social media. *MIS Quarterly*, *43*(3), 1025–1039.

Kiyota, Y. (2021). Frontiers of computer vision technologies on real estate property photographs and floorplans. In Y. Asami, Y. Higano, & H. Fukui (Eds.), *Frontiers of real estate science in Japan* (pp. 325–337, Vol. 325). Springer Singapore.

Kong, X., Zhang, J., & Yu, P. S. (2013). Inferring anchor links across multiple heterogeneous social networks. *Proceedings of the CIKM*.

Kouzy, R., Abi Jaoude, J., Kraitem, A., El Alam, M. B., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E., & Baddour, K. (2020). Coronavirus goes viral:

Quantifying the COVID-19 misinformation epidemic on Twitter. *Cureus*, *12*(3).

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *PNAS*, *111*(24), 8788–8790.

Krittanawong, C., Narasimhan, B., Virk, H. U. H., Narasimhan, H., Hahn, J., Wang, Z., & Tang, W. H. W. (2020). Misinformation dissemination in Twitter in the COVID-19 era. *The American Journal of Medicine*, *133*(12), 1367–1369.

Larson, L., & DeChurch, L. A. (2020). Leading teams in the digital age: Four perspectives on technology and what they mean for leading teams. *The Leadership Quarterly*, *31*(1), 101377.

Larsson, A. O. (2014). Pandering, protesting, engaging. Norwegian party leaders on Facebook during the 2013 'Short campaign'. *Information, Communication & Society*, *18*(4), 459–473.

Lewandowsky, S., Jetter, M., & Ecker, U. K. H. (2020). Using the president's tweets to understand political diversion in the age of social media. *Nature Communications*, *11*(1), 5764.

Li, M., Bao, Z., Sellis, T., Yan, S., & Zhang, R. (2018). Homeseeker: A visual analytics system of real estate data. *Journal of Visual Languages & Computing*, *45*, 1–16.

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(12), 6999–7019.

Livingstone, S., Stoilova, M., & Nandagiri, R. (2019). *Children's data and privacy online: Growing up in a digital age: An evidence review*. London School of Economics; Political Science, Department of Media and ...

Martin, K. D., & Murphy, P. E. (2016). The role of data privacy in marketing. *Journal of the Academy of Marketing Science*, *45*(2), 135–155.

Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019). Spread of hate speech in online social media. *Proceedings of the WebSci*.

Mondal, M., Silva, L. A., & Benevenuto, F. (2017). A measurement study of hate speech in social media. *Proceedings of the ACMHT*.

Monson, M. (2009). Valuation using hedonic pricing models. *Cornell Real Estate Review*, *7*(1), 62–73.

Moravec, P., Minas, R., & Dennis, A. R. (2019). Fake news on social media: People believe what they want to believe when it makes no sense at all. *MIS Quarterly*, *43*(4), 1343–1360.

Mosleh, M., & Rand, D. G. (2022). Measuring exposure to misinformation from political elites on Twitter. *Nature Communications*, *13*(1), 7144.

Mughal, M. J. H. (2018). Data mining: Web data mining techniques, tools and algorithms: An overview. *International Journal of Advanced Computer Science and Applications*, *9*(6), 208–215.

Munger, K. (2016). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, *39*(3), 629–649.

Naumzik, C., & Feuerriegel, S. (2020). One picture is worth a thousand words? The pricing power of images in e-commerce. *Proceedings of the WWW*.

Ngiam, K. Y., & Khor, I. W. (2019). Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, *20*(5), e262–e273.

Nigam, K., Mccallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, *39*(2/3), 103–134.

Patil, N. S., Kiran, P., Kiran, N. P., & Naresh Patel, K. M. (2018). A survey on graph database management techniques for huge unstructured data. *International Journal of Electrical and Computer Engineering*, *8*(2), 1140.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, *31*(7), 770–780.

Petrova, M., Sen, A., & Yildirim, P. (2021). Social media and political contributions: The impact of new technology on political competition. *Management Science*, *67*(5), 2997–3021.

Piazza, J. A. (2020). Politician hate speech and domestic terrorism. *International Interactions*, *46*(3), 431–453.

Plutchik, R. (1984). *Emotion: Theory, research, and experience: Theory, research, and experience* (2nd ed., Vol. 1). Academic Press.

Poursaeed, O., Matera, T., & Belongie, S. (2018). Vision-based real estate price estimation. *Machine Vision and Applications*, *29*(4), 667–676.

Pretus, C., Ray, J. L., Granot, Y., Cunningham, W. A., & Van Bavel, J. J. (2022). The psychology of hate: Moral concerns differentiate hate from dislike. *European Journal of Social Psychology*, *53*(2), 336–353.

Pröllochs, N. (2022). Community-based fact-checking on Twitter's Birdwatch platform. *Proceedings of the ICWSM*.

Pröllochs, N., Bär, D., & Feuerriegel, S. (2021). Emotions explain differences in the diffusion of true vs. false social media rumors. *Scientific Reports*, *11*(22721).

Qiang, J., Qian, Z., Li, Y., Yuan, Y., & Wu, X. (2022). Short text topic modeling techniques, applications, and performance: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *34*(3), 1427–1445.

Quan, X., Kit, C., Ge, Y., & Pan, S. J. (2015). Short and sparse text topic modeling via self-aggregation. *Proceedings of the IJCAI*.

Rambocas, M., & Pacheco, B. G. (2018). Online sentiment analysis in marketing research: A review. *Journal of Research in Interactive Marketing*, *12*(2), 146–163.

Ranathunga, S., Lee, E.-S. A., Prifti Skenduli, M., Shekhar, R., Alam, M., & Kaur, R. (2023). Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, *55*(11), 1–37.

Rese, A., Ganster, L., & Baier, D. (2020). Chatbots in retailers' customer communication: How to measure their acceptance? *Journal of Retailing and Consumer Services*, *56*, 102176.

Ricard, J., & Medeiros, J. (2020). Using misinformation as a political weapon: COVID-19 and Bolsonaro in Brazil. *Harvard Kennedy School Misinformation Review*, *1*(3).

Ross, K., Fountaine, S., & Comrie, M. (2014). Facing up to Facebook: Politicians, publics and the social media (ted) turn in New Zealand. *Media, Culture & Society*, *37*(2), 251–269.

Sainudiin, R., Yogeeswaran, K., Nash, K., & Sahioun, R. (2019). Characterizing the Twitter network of prominent politicians and SPLC-defined hate groups in the 2016 U.S. presidential election. *Social Network Analysis and Mining*, *9*(1), 1–15.

Schöll, N., Gallego, A., & Le Mens, G. (2023). How politicians learn from citizens' feedback: The case of gender on Twitter. *American Journal of Political Science*.

Sellars, A. (2016). Defining hate speech. *Berkman Klein Center Research Publication*, *2016*(20), 16–48.

Sharma, V. K., & Mir, R. N. (2020). A comprehensive and systematic look up into deep learning based object detection techniques: A review. *Computer Science Review*, *38*, 100301.

Singh, L., Bansal, S., Bode, L., Budak, C., Chi, G., Kawintiranon, K., Padden, C., Vanarsdall, R., Vraga, E., & Wang, Y. (2020). A first look at COVID-19 information and misinformation sharing on Twitter. *arXiv*, *2003.13907*.

Solovev, K., & Pröllochs, N. (2022). Moral emotions shape the virality of COVID-19 misinformation on social media. *Proceedings of the WWW*.

Sopranzetti, B. J. (2010). Hedonic regression analysis in real estate markets: A primer. In C.-F. Lee, A. C. Lee, & J. Lee (Eds.), *Handbook of quantitative finance and risk management* (pp. 1201–1207). Springer US.

Soussan, T., & Trovati, M. (2021). Social media data misuse. In L. Barolli, H.-C. Chen, & H. Miwa (Eds.), *Proceedings of the incos*. Springer International Publishing.

Steelman, Z. R., Hammer, B. I., & Limayem, M. (2014). Data collection in the digital age. *MIS Quarterly*, *38*(2), 355–378.

Sternberg, R. J. (2003). A duplex theory of hate: Development and application to terrorism, massacres, and genocide. *Review of General Psychology*, *7*(3), 299–328.

Stier, S., Bleier, A., Lietz, H., & Strohmaier, M. (2020). Election campaigning on social media: Politicians, audiences, and the mediation of political communication on Facebook and Twitter. In L. Bode & E. K. Vraga (Eds.), *Studying politics across media* (pp. 50–74). Routledge.

Su, Q., & Chen, L. (2015). A method for discovering clusters of e-commerce interest patterns using click-stream data. *Electronic Commerce Research and Applications*, *14*(1), 1–13.

Sun, G., Liang, R., Wu, F., & Qu, H. (2013). A web-based visual analytics system for real estate data. *Science China Information Sciences*, *56*(5), 1–13.

Sylwester, K., & Purver, M. (2015). Twitter language use reflects psychological differences between democrats and republicans. *PLOS ONE*, *10*(9), e0137422.

Tabares Higuita, L. X., et al. (2018). Análisis del discurso violento y de odio en dos grupos de Facebook contra la candidatura de Rodrigo Londoño 'Timochenko'a la presidencia de Colombia. *Comunicación digital en Iberoamérica*, *8(3)*, 157–184.

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology*, *58*(1), 345–372.

Tatum, B. D. (2017). *Why are all the black kids sitting together in the cafeteria? And other conversations about race: And other conversations about race* (3rd ed.). Basic Books.

Teinemaa, I., Dumas, M., Maggi, F. M., & Di Francescomarino, C. (2016). Predictive business process monitoring with structured and unstructured data. *Proceedings of the BPM*.

Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., & Jégou, H. (2021). Going deeper with image transformers. *Proceedings of the ICCV*.

Vayansky, I., & Kumar, S. A. P. (2020). A review of topic modeling methods. *Information Systems*, *94*, 101582.

Venkatraman, R., & Venkatraman, S. (2019). Big data infrastructure, data visualisation and challenges. *Proceedings of the BDIOT*.

Vidgen, B., Burden, E., & Margetts, H. (2021). *Understanding online hate: VSP regulation and the broader context*. London, UK, Ofcom.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Vraga, E. K., & Bode, L. (2017). Using expert sources to correct health misinformation in social media. *Science Communication*, *39*(5), 621–645.

Wagner, A. (2020). Tolerating the trolls? Gendered perceptions of online harassment of politicians in Canada. *Feminist Media Studies*, *22*(1), 32–47.

Wallace, N. E., & Meese, R. A. (1997). The construction of residential housing price indices: A comparison of repeat-sales, hedonic-regression, and hybrid approaches. *The Journal of Real Estate Finance and Economics*, *14*(1/2), 51–73.

Waller, I., & Anderson, A. (2021). Quantifying social organization and political polarization in online platforms. *Nature*, *600*(7888), 264–268.

Wang, Z., Xie, Q., Ding, Z., Feng, Y., & Xia, R. (2023). Is ChatGPT a good sentiment analyzer? A preliminary study. *arXiv*, *2304.04339*.

Webster, F. (2014). *Theories of the information society* (4th ed.). Taylor & Francis Group.

Weng, L., Menczer, F., & Ahn, Y.-Y. (2013). Virality prediction and community structure in social networks. *Scientific Reports*, *3*(1).

Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science*, *16*(10), 780–784.

Winkler, J., Halfmann, A., & Freudenthaler, R. (2017). Backlash effects in online discussions: Effects of gender and counter-stereotypical communication on persuasiveness and likeability. *Proceedings of the ICA*.

Wu, J., Huang, L., & Zhao, J. L. (2019). Operationalizing regulatory focus in the digital age: Evidence from an e-commerce context. *MIS Quarterly*, *43*(3), 745–764.

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. *Proceedings of the WWW*.

You, Q., Pang, R., Cao, L., & Luo, J. (2017). Image-based appraisal of real estate properties. *IEEE Transactions on Multimedia*, *19*(12), 2751–2759.

Yu, H., Yang, L. T., Zhang, Q., Armstrong, D., & Deen, M. J. (2021). Convolutional neural networks for medical image analysis: State-of-the-art, comparisons, improvement and perspectives. *Neurocomputing*, *444*, 92–110.

Yuan, X., Lee, J.-H., Kim, S.-J., & Kim, Y.-H. (2013). Toward a user-oriented recommendation system for real estate websites. *Information Systems*, *38*(2), 231–243.

Zhang, D., Yin, C., Zeng, J., Yuan, X., & Zhang, P. (2020). Combining structured and unstructured data for predictive models: A deep learning approach. *BMC Medical Informatics and Decision Making*, *20*(1), 1–11.

Zhang, L., & Malife, C. (2023). *Processing billions of events in real time at Twitter*. https://blog.twitter.com/engineering/en_us/topics/infrastructure/2021/processing-billions-of-events-in-real-time-at-twitter-

Zillow Group. (2023). *Investor relations stats*. https://investors.zillowgroup.com/investors/overview/default.aspx

Zumpano, L. V., Johnson, K. H., & Anderson, R. I. (2003). Internet use and real estate brokerage market intermediation. *Journal of Housing Economics*, *12*(2), 134–150.

Chapter 2

Hate Speech in the Political Discourse on Social Media: Disparities Across Parties, Gender, and Ethnicity

2.1 Introduction

Social media has become an indispensable communication channel for politicians in the U. S. and around the world. Compared to traditional media, it provides a number of key benefits for politicians: (i) social media provides a tool to spread messages to the public at scale, thereby increasing people's awareness of their (political) agenda (Graham et al., 2013; Hong & Nadler, 2012; Ross et al., 2014). (ii) Social media encourages the dialogue between politicians and users, allowing for direct feedback from constituents and discussions of political ideas (Enli & Skogerbø, 2013). (iii) Due to its interactive nature, social media can be used as a tool for political mobilization (Jackson & Lilleker, 2009; Larsson, 2014). These benefits are further reinforced by the openness of social media as politicians are no longer restricted by geography, scope, or content and can reach significantly wider audiences (Gainous & Wagner, 2014).

However, the shift from traditional channels towards social media does not necessarily improve the quality of the political discourse. Instead, social media is known to foster echo chambers and "us versus them" rhetoric (Mondal et al., 2017). These factors correlate with cyber-bullying, harassment, and, in particular, hate speech (Erjavec & Kovačič, 2012). Broadly speaking, hate speech refers to abusive or threatening speech (or writing) that expresses prejudice against a particular group, often on the basis of ethnicity or sexual orientation (Sellars, 2016). Hate speech often originates from semi-anonymous trolls (Himma & Tavani, 2008; Mondal et al., 2017), and is particularly frequent in discussions

that cause a strong emotional response, such as in political topics (Wagner, 2020). The adoption of social media by politicians is a double-edged sword posing risks both to themselves and society as a whole (Hong et al., 2019). At the individual level, hate speech can threaten reputations and may even lead to long-run mental health issues (Vidgen et al., 2021). At the societal level, it fosters political polarization (Piazza, 2020), which can have severe consequences. Examples include erosion of intergroup political relations and increased opportunities for the spread of ideologically branded misinformation (Freelon et al., 2020; Pröllochs, 2022; Solovev & Pröllochs, 2022).

Research Goal: In this study, we empirically analyze how the user base on Twitter responds to posts from members of the U. S. Congress. We are interested in understanding whether differences in the prevalence of hate speech can be explained by personal characteristics of politicians, such as their party affiliation, gender, and ethnicity. More precisely, we address the following research questions:

- **(RQ1)** *Are members of the U. S. Congress more likely to receive hate speech in the replies to their tweets depending on their party affiliation, gender, and ethnicity?*
- **(RQ2)** *Does hate speech in the replies to tweets depend on the sentiment of the source tweet? Does the strength of the association differ depending on their party, gender, and ethnicity?*

Data & Methods: To address our research questions, we employ the Twitter Historical API to collect all tweets from members of the 117th U. S. Congress between the first session on January 3, 2021 and the end of July 2021. In addition, we collect replies to each source tweet. We then use machine learning to determine the share of replies of each tweet that embeds hate speech. Subsequently, we implement a multilevel binomial regression model with random effects to estimate whether Twitter users are more likely to respond with hate speech depending on the party affiliation, gender, and ethnicity of the politician that has posted the tweet.

Contributions: To the best of our knowledge, this study is the first to empirically model how hate speech in replies to tweets from politicians depends on their personal characteristics (party affiliation, gender, ethnicity). All else being equal, we find that tweets are more likely to receive hate speech in replies if they are authored by (i) persons of color from the Democratic party, (ii) white Republicans, and (iii) women. As an additional contribution, our analysis reveals that more negative sentiment (in the source tweet) is associated with more hate

speech (in replies). However, the association varies across parties: negative sentiment attracts more hate speech for Democrats (vs. Republicans). Altogether, our findings fuel new insights into ongoing discussions on political polarization on social media and highlight disparities in how politicians are treated depending on their party affiliation, gender, and ethnicity

2.2 Background

Political communication on Twitter: The use of social media by U. S. politicians has experienced a rapid surge. At the start of 2009, only 69 individual members of Congress had a Twitter account (Golbeck et al., 2010). Today, every member of the U. S. Congress has a professional Twitter account and oftentimes a second personal account being active at the same time. Existing studies suggest that there are three main reasons *why* politicians adopt social media (Hong et al., 2019). First, social media allows for *unidirectional delivery* of information to the public. Compared to classical media, there is less moderation and real time scrutiny allowing politicians to freely express themselves (Allcott & Gentzkow, 2017). Second, social media enables *dialogue* between politicians and the public. Politicians can use social media as a tool to connect with constituents to discuss political issues and receive feedback (Enli & Skogerbø, 2013). Engaged users may further spread the message with likes and/or reshares. Third, social media can be seen as a tool for *political mobilization*. Specifically, it allows politicians to rally for projects, events, and movements (Theocharis et al., 2014), though it does not guarantee success (Margetts, 2016).

Hate speech: Although there is no all-encompassing definition (Benesch, 2014), hate speech is typically considered to refer to abusive or threatening speech (or writing) that expresses prejudice against a particular group, often on the basis of ethnicity or sexual orientation (Sellars, 2016). While research on hate speech has received increasing attention lately (e. g., Akhtar et al., 2020; Chopra et al., 2020; Davidson et al., 2017; ElSherief et al., 2018; Mossie, 2020; Nagar et al., 2021; Olteanu et al., 2018; Saha et al., 2021; Wich et al., 2021; Zannettou et al., 2018), studies that analyze hate speech in the context of political communication are scant. The few existing works typically focus on qualitative insights or analysis of summary statistics. For instance, previous works have studied hate speech towards female Japanese politicians (Fuchs & Schäfer, 2020), far-right political party discourse in Spain (Ben-David & Fernández, 2016), hateful propaganda towards politicians in Macedonia (Ben-David & Fernández, 2016), hate speech against Members of Parliament in the U.K. (Agarwal et al., 2021), and hate against German politicians (de Smedt & Jaki, 2018). We are aware of only one paper

analyzing hate speech and incivility in the context of tweets from members of the U. S. Congress (Theocharis et al., 2020). However, this study again focuses on summary statistics. In particular, it does not model the effects of personal characteristics of politicians (e. g., ethnicity) on the likelihood of receiving hate speech.

Disparities across parties, gender, and ethnicity: Existing research suggest that political party leanings in the U. S. correlate with different speech patterns: Democrats tend to use more swear words and higher sentiment, while Republicans prefer to communicate more negative sentiment and group identity (Sylwester & Purver, 2015). Besides party differences, a vast strand of studies has shown that there are discrepancies in communication behavior across genders. For instance, women are more likely to hide expressive and negative emotions (Davis, 1995), and are guided by a greater focus on care in moral dilemmas (Nguyen et al., 2007). This is directly applicable to the domain of social media, where women are more likely to report messages targeting racial minorities and women (Downs & Cowan, 2012). Gender differences are further reinforced by widespread stereotypes regarding the role of women in society (Prentice & Carranza, 2003), who are perceived as less persuasive and are often outright dismissed when displaying aggressive and forceful behavior online (Winkler et al., 2017). Furthermore, survey studies suggest that women more often tend to be a target of cyber-bullying and hateful attacks (Beckman et al., 2013), especially if they present an openly active stance, such as feminism (Hardaker & McGlashan, 2016). Ethnicities and racial stereotypes play a similar role in offline and online discourse and differ greatly across countries (Tatum, 2017). For instance, for the U. S., existing studies suggest frequent hate speech against African Americans (Kwok & Wang, 2013).

Research gap: Existing research on hate speech in the political discourse focuses either on qualitative insights or on summary statistics. We are not aware of previous works empirically modeling the effect of personal characteristics on the likelihood of a politician to receive hate speech. This presents our contribution.

2.3 Dataset

Members of the U. S. Congress: We analyze tweets from all 541 members of the 117th U. S. Congress that convened on January 3, 2021. Data on the members of Congress was gathered from the official webpage of the U. S. Congress (Library of Congress, 2021), which provides links to personal and campaign web pages. By following these links, we collected the following information about each politician: (i) party affiliation, (ii) branch of Congress in which the politician serves,

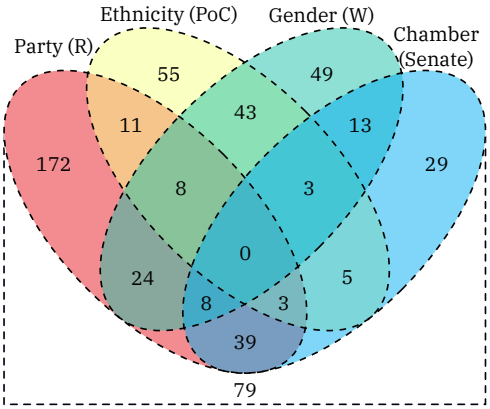


Figure 2.1: Venn Diagram visualizes the composition of the 117th U. S. Congress.

(iii) time served in Congress, (iv) gender, and (v) ethnicity. Figure 2.1 provides an overview of the composition of the 117th U. S. Congress. Most voting seats are held by members of the two major political parties with 269 Democrats (D) and 263 Republicans (R), while 2 seats are occupied by independent senators. Women (W) hold 27 % of all Congress Seats, accounting for 39 % of all Democrats and 15 % of all Republicans, respectively. Notably, the 117th U. S. Congress is the most ethnically diverse so far with 39 % of Democrats and 8 % of Republicans identifying as people of color (PoC).

For the sake of simplicity and interpretability, we focus our later empirical analysis on tweets from Republican and Democratic members; and exclude tweets from the two independent senators.

Collection of tweets: Twitter handles (user names) of every politician in the U. S. Congress are provided by the University of California San Diego library (Smith, 2021). We employed the Twitter Historical API to download the complete timelines of every politician between January 3, 2021 and the end of July 2021. Here we collected the entire tweet history of each person, excluding retweets and replies, resulting in a total number of 199,294 tweets. The average number of tweets per politician is 368.38. We additionally queried Twitter's Historical API to gather the replies to every source tweet in our data set. To ensure feasibility, we restricted the data collection to up to 250 replies for each original tweet, starting with the earliest reply. The crawling process resulted in a total number of 8,362,555 replies.

2.4 Methods

2.4.1 Hate Speech Detection

In this work, we use machine learning to detect hate speech in replies to tweets. Compared to dictionary-based methods that merely count hate-related words (Alrehili, 2019), this approach is generally considered as being more accurate (Badjatiya et al., 2017). Nonetheless, as part of our robustness checks, we validate our results with the frequently-employed Hatebase dictionary (Hatebase, 2021), finding confirmatory results.

We implement machine learning for hate speech detection as follows: we employ the annotated Twitter dataset from (Davidson et al., 2017), containing 25,000 tweets labeled as hateful or not hateful. Each tweet was annotated by at least 3 users who were explicitly instructed to think about the context of the message and not only the words contained within (Davidson et al., 2017). We use the annotated tweets to implement a deep neural network classifier that predicts whether or not a tweet is hateful.[1] The hate speech classifier is then used to predict a binary label of whether or not a tweet is hateful ($= 1$ if true; otherwise $= 0$) for each reply tweet in our dataset. For each source tweet, we calculate the share of replies that are hateful. The resulting variable ranges from 0 to 1, with 0 indicating the lack of hate speech in replies, and 1 indicating that every reply is hateful.

2.4.2 Explanatory Regression Model

We implement a multilevel binomial regression to estimate the effects of party, gender, and ethnicity on the likelihood of a tweet receiving hate speech.

Formally, we model the number of hate speech replies, $HReplies$, as a binomial variable with probability parameter θ. The number of trials is given by the total number of replies a tweet receives ($Replies$). The key explanatory variables are the politicians' party affiliation (*Party*; $= 1$ if Republican, otherwise 0), gender (*Gender*; $= 1$ if Man, otherwise 0), and ethnicity (*Ethnicity*; $= 1$ if Person of Color, otherwise 0). Furthermore, for each source tweet, we calculate a sentiment score (*Source Sentiment*) using `SentiStrength` (Thelwall et al., 2010). We also control for the age of the members of Congress (*Age*), the number of years served (*Years in Office*), whether media was attached to the tweet (*Attached Media*; $= 1$ if true, otherwise 0), and the chamber of Congress at which the politician serves

[1] We use Universal Sentence Encoder (USE) (Cer et al., 2018) as text representation. The machine learning classifier yields a weighted out-of-sample $F1$ score of 0.89, which is similar to previous works (Davidson et al., 2017) and can be seen as reasonably accurate in the context of our study. The model is implemented in Python 3.8.5 using TensorFlow 2.6.0.

(*Chamber*; $= 1$ if Senate, otherwise 0). Based on these variables, we specify the following regression model:

$$\begin{aligned} \text{logit}(\theta) = &\beta_0 + \beta_1 \textit{Party} + \beta_2 \textit{Gender} + \beta_3 \textit{Ethnicity} \\ &+ \beta_4 \textit{Source Sentiment} + \beta_5 \textit{Years in Office} + \beta_6 \textit{Age} \\ &+ \beta_7 \textit{Attached Media} + \beta_8 \textit{Chamber} \\ &+ u_{\text{user}} + \varepsilon, \end{aligned} \tag{2.1}$$

$$HReplies \sim Binomial[Replies, \theta], \tag{2.2}$$

with intercept β_0, error term ε, and user-specific random effects u_{user}. Note that the latter is important as it allows us to control for heterogeneity in users' social influence (e. g., some accounts have many followers and reach different audiences) (Pröllochs et al., 2021a, 2021b).

We estimate Equation (2.1) and Equation (2.2) using MLE and generalized linear models. To facilitate the interpretability of our findings, we z-standardize all variables, so that we can compare the effects of regression coefficients on the dependent variable measured in standard deviations. Our regression analyses are implemented in R 4.0.5 using the `lme4` package (Bates et al., 2021).

2.5 Empirical Analysis

2.5.1 Summary Statistics

We start our analysis by evaluating summary statistics. The average share of hateful replies per tweet in our dataset amounts to 1.99 %. We perform both $t-$tests and Kolmogorov-Smirnov (KS) tests to evaluate whether there are statistically significant differences across parties, genders, and ethnicities. Our findings are as follows: (i) tweets from Democrats (vs. Republicans) receive, on average, a 3.67 % higher share of hate replies. (ii) Tweets from women (vs. men) politicians receive 7.71 % higher share of hate replies. (iii) Tweets from persons of color (vs. whites) receive 37.75 % higher share of hate replies. For each of these comparisons, two-sided $t-$tests confirm that the differences in means are statistically significant ($p < 0.01$). In Section 2.5.1, we visualize the complementary cumulative distribution functions (CCDFs) for the ratio of hate speech in replies. We again find that Democrats, women and persons of color receive more hate speech. KS-tests confirm that all differences in distributions are statistically significant ($p < 0.001$).

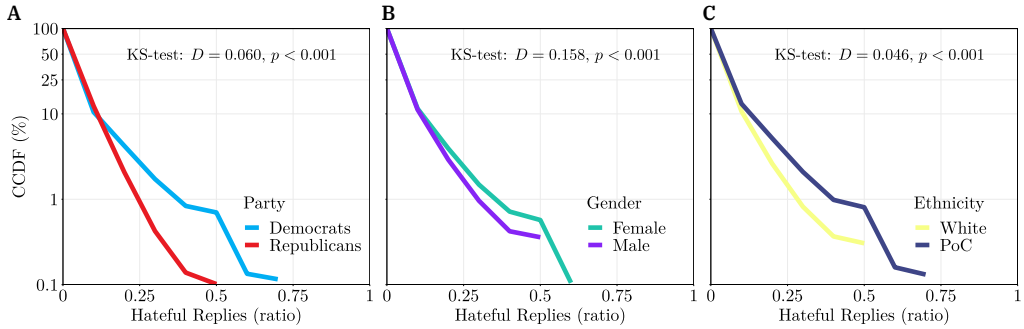


Figure 2.2: CCDFs for the ratio of hate speech in replies separated by (A) party, (B) gender, and (C) ethnicity.

2.5.2 Regression Analysis

We estimate a multilevel binomial regression to understand the effects of party affiliation, gender, and ethnicity on the likelihood of a tweet receiving hate speech (see model w/o interactions in Figure 2.3). In contrast to summary statistics, this allows us to estimate effect sizes *after* controlling for confounding effects. The largest effect size is estimated for *Ethnicity* with a coefficient of 0.346 ($p < 0.01$), which implies that the odds of receiving hate speech for persons of color are $e^{0.346} \approx 1.41$ times the odds for whites. We further observe pronounced party and gender effects. Compared to Democrats, the odds for tweets from Republicans to receive hate speech are 22.02 % higher ($\beta = 0.199$, $p < 0.01$). The odds for men to receive hate speech are 8.33 % ($\beta = -0.087$, $p < 0.05$) lower than for women. We also find that a more negative sentiment in the source tweet is associated with more hate speech in replies. A one standard deviation increase in *Source Sentiment* is associated with a 25.99 % ($\beta = -0.301$, $p < 0.01$) decrease in the odds of receiving hate speech. We find no statistically significant effects from a politician's age, time in office, chambers, and media attachments.

We add interaction terms to test whether users react differently to gender, ethnicity, and sentiment depending on the party affiliation (see model w/ interactions in Figure 2.3). Here we find a statistically significant interaction term between *Party* and *Ethnicity* ($\beta = -0.287$, $p < 0.01$). This implies that persons of color from the Democratic party have higher odds for receiving hate speech than persons of color from the Republican party. Furthermore, the strength of the association between sentiment in the source tweet and hate speech varies across parties ($\beta = 0.235$, $p < 0.01$). Specifically, negative sentiment attracts more hate speech for Democrats. The interaction between party affiliation and gender is not significant at common statistical significance thresholds.

Altogether, our analysis implies that three groups of politicians are particularly likely to receive hate speech in response to their tweets: (i) persons of color

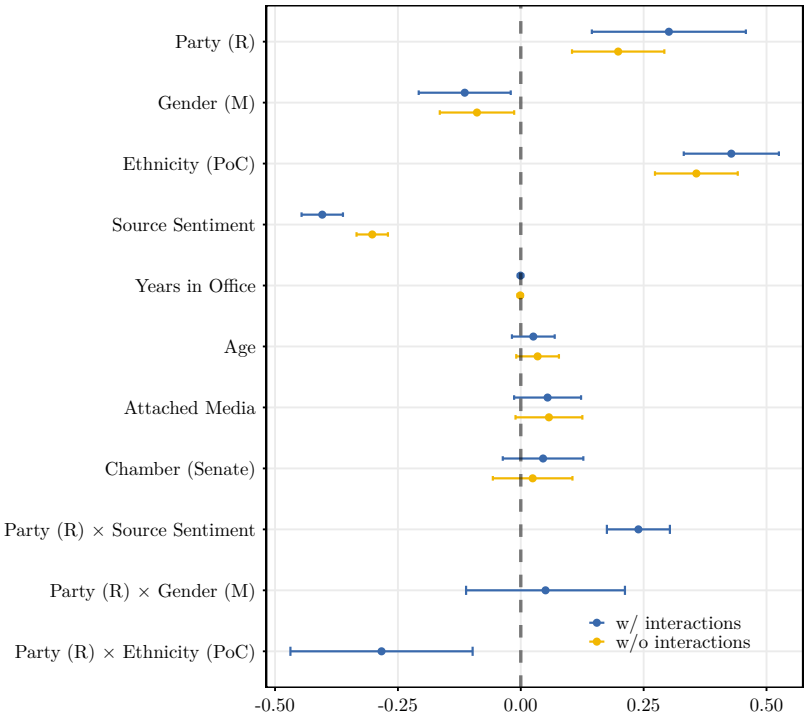


Figure 2.3: Coefficient estimates for binomial regression w/o (mustard yellow) and w/ (navy blue) interaction terms for political party. The horizontal bars represent 95 % confidence intervals. User-specific random effects are included.

from the Democratic party, (ii) white Republicans, and (iii) women.

2.5.3 Robustness Checks

We conducted additional checks to validate the robustness of our analysis: (1) We repeated our analysis with a dictionary-based approach for hate speech detection, specifically the Hatebase dictionary (Hatebase, 2021). (2) We calculated variance inflation factors for all independent variables in our regression model and found that all remain below the critical threshold of four. (3) We repeated our analysis with alternative estimators (e. g., beta regression), controlled for outliers, tested for quadratic effects, and added multiple interaction terms for each explanatory variable. In all cases, our results are robust and consistently support our findings.

2.6 Discussion

Summary of findings: This work empirically models how the amount of hate speech in replies to tweets from politicians depends on their personal characteristics (party affiliation, gender, ethnicity). All else being equal, we find that

Tweets are particularly likely to receive hate speech replies if they are authored by (i) persons of color from the Democratic party, (ii) white Republicans, and (iii) women. Furthermore, our analysis reveals that more negative sentiment (in the source tweet) is associated with more hate speech (in replies). However, the association varies across parties: negative sentiment attracts more hate speech for Democrats (vs. Republicans). Altogether, our empirical findings imply statistically significant differences in how politicians are treated on social media depending on their party affiliation, gender, and ethnicity.

Implications: Our findings are relevant both for politicians and from a societal perspective. Politicians should be aware that social media is a double-edged sword as it comes with the risk of receiving vast numbers of hate comments. This is concerning as hate speech can destroy reputations and may even lead to long-run mental health consequences (Vidgen et al., 2021). Given that hate speech can affect peoples' decision to participate in politics (Scott, 2019), this may also impede diversity in the composition of political institutions. Furthermore, hate speech goes hand in hand with increased polarization, hyper-partisanship, and less common ground between opposing political sides (Finkel et al., 2020), thereby threatening the functioning of democracy itself.

Bibliography

Agarwal, P., Hawkins, O., Amaxopoulou, M., Dempsey, N., Sastry, N., & Wood, E. (2021). Hate speech in political discourse: A case study of UK MPs on Twitter. *Proceedings of the ACMHT*.

Akhtar, S., Basile, V., & Patti, V. (2020). Modeling annotator perspective and polarized opinions to improve hate speech detection. *Proceedings of the HCOMP*.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236.

Alrehili, A. (2019). Automatic hate speech detection on social media: A brief survey. *Proceedings of the AICCSA*.

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. *Proceedings of the WWW Companion*.

Bates, D., Sarkar, D., Bates, M. D., & Matrix, L. (2021). *Lme4* [version 1.1.27]. https://cran.r-project.org/web/packages/lme4/index.html

Beckman, L., Hagquist, C., & Hellström, L. (2013). Discrepant gender patterns for cyberbullying and traditional bullying – An analysis of Swedish adolescent data. *Computers in Human Behavior*, *29*(5), 1896–1903.

Ben-David, A., & Fernández, A. M. (2016). Hate speech and covert discrimination on social media: Monitoring the facebook pages of extreme-right political parties in Spain. *International Journal of Communication*, *10*, 1167–1193.

Benesch, S. (2014). Defining and diminishing hate speech. *State of the World's Minorities and Indigenous Peoples*, *2014*, 18–25.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., & Kurzweil, R. (2018). Universal sentence encoder. *arXiv*, *1803.11175*.

Chopra, S., Sawhney, R., Mathur, P., & Ratn Shah, R. (2020). Hindi-english hate speech detection: Author profiling, debiasing, and practical perspectives. *Proceedings of the AAAI*.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the ICWSM*, *11*(1), 512–515.

Davis, T. L. (1995). Gender differences in masking negative emotions: Ability or motivation? *Developmental Psychology*, *31*(4), 660–667.

de Smedt, T., & Jaki, S. (2018). The Polly corpus: Online political debate in Germany. *Proceedings of the CMC*.

Downs, D. M., & Cowan, G. (2012). Predicting the importance of freedom of speech and the perceived harm of hate speech. *Journal of Applied Social Psychology*, *42*(6), 1353–1375.

ElSherief, M., Kulkarni, V., Nguyen, D., Yang Wang, W., & Belding, E. (2018). Hate lingo: A target-based linguistic analysis of hate speech in social media. *Proceedings of the ICWSM*.

Enli, G. S., & Skogerbø, E. (2013). Personalized campaigns in party-centred politics: Twitter and Facebook as arenas for political communication. *Information, Communication & Society*, *16*(5), 757–774.

Erjavec, K., & Kovačič, M. P. (2012). "you don't understand, this is a new war!" Analysis of hate speech in news web sites' comments. *Mass Communication and Society*, *15*(6), 899–920.

Finkel, E. J., Bail, C. A., Cikara, M., Ditto, P. H., Iyengar, S., Klar, S., Mason, L., McGrath, M. C., Nyhan, B., Rand, D. G., Skitka, L. J., Tucker, J. A., Van Bavel, J. J., Wang, C. S., & Druckman, J. N. (2020). Political sectarianism in America. *Science*, *370*(6516), 533–536.

Freelon, D., Marwick, A., & Kreiss, D. (2020). False equivalencies: Online activism from left to right. *Science*, *369*(6508), 1197–1201.

Fuchs, T., & Schäfer, F. (2020). Normalizing misogyny: Hate speech and verbal abuse of female politicians on Japanese Twitter. *Japan Forum*.

Gainous, J., & Wagner, K. M. (2014). *Tweeting to power: The social media revolution in American politics*. Oxford University Press.

Golbeck, J., Grimes, J. M., & Rogers, A. (2010). Twitter use by the U.S. Congress. *Journal of the American Society for Information Science and Technology*, *61*.

Graham, T., Broersma, M., Hazelhoff, K., & van 't Haar, G. (2013). Between broadcasting political messages and interacting with voters: The use of Twitter during the 2010 UK general election campaign. *Information, Communication & Society*, *16*(5), 692–716.

Hardaker, C., & McGlashan, M. (2016). "real men don't hate women": Twitter rape threats and group identity. *Journal of Pragmatics*, *91*, 80–93.

Hatebase. (2021). *A collaborative, regionalized repository of multilingual hate speech*. https://hatebase.org/

Himma, K. E., & Tavani, H. T. (2008). *The handbook of information and computer ethics* (H. T. Tavani & K. E. Himma, Eds.). Wiley.

Hong, S., Choi, H., & Kim, T. K. (2019). Why do politicians tweet? Extremists, underdogs, and opposing parties as political tweeters. *Policy & Internet*, *11*(3), 305–323.

Hong, S., & Nadler, D. (2012). Which candidates do the public discuss online in an election campaign? The use of social media by 2012 presidential candidates and its impact on candidate salience. *Government Information Quarterly*, *29*(4), 455–461.

Jackson, N. A., & Lilleker, D. G. (2009). Building an architecture of participation? Political parties and Web 2.0 in Britain. *Journal of Information Technology & Politics*, *6*(3–4), 232–250.

Kwok, I., & Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. *Proceedings of the AAAI*.

Larsson, A. O. (2014). Pandering, protesting, engaging. Norwegian party leaders on Facebook during the 2013 'Short campaign'. *Information, Communication & Society*, *18*(4), 459–473.

Library of Congress. (2021). *Members of the U.S. Congress*. https://www.congress.gov/members

Margetts, H. (2016). *Political turbulence: How social media shape collective action: How social media shape collective action* (P. John, S. Hales, & T. Yasseri, Eds.). Princeton University Press.

Mondal, M., Silva, L. A., & Benevenuto, F. (2017). A measurement study of hate speech in social media. *Proceedings of the ACMHT*.

Mossie, Z. (2020). Social media dark side content detection using transfer learning emphasis on hate and conflict. *Proceedings of the WWW Companion*.

Nagar, S., Gupta, S., Bahushruth, C. S., Barbhuiya, F. A., & Dey, K. (2021). Empirical assessment and characterization of homophily in classes of hate speeches. *Proceedings of the AAAI Workshop*.

Nguyen, N. T., Basuray, M. T., Smith, W. P., Kopka, D., & McCulloh, D. (2007). Moral issues and gender differences in ethical judgment using Reidenbach and Robin's (1990) multidimensional ethics scale: Implications in teaching of business ethics. *Journal of Business Ethics*, *77*(4), 417–430.

Olteanu, A., Castillo, C., Boy, J., & Varshney, K. (2018). The effect of extremist violence on hateful speech online. *Proceedings of the ICWSM*.

Piazza, J. A. (2020). Politician hate speech and domestic terrorism. *International Interactions*, *46*(3), 431–453.

Prentice, D. A., & Carranza, E. (2003). Sustaining cultural beliefs in the face of their violation: The case of gender stereotypes. In M. Schaller & C. S. Crandall (Eds.), *The psychological foundations of culture* (pp. 268–289). Psychology Press.

Pröllochs, N. (2022). Community-based fact-checking on Twitter's Birdwatch platform. *Proceedings of the ICWSM*.

Pröllochs, N., Bär, D., & Feuerriegel, S. (2021a). Emotions explain differences in the diffusion of true vs. false social media rumors. *Scientific Reports*, *11*(22721).

Pröllochs, N., Bär, D., & Feuerriegel, S. (2021b). Emotions in online rumor diffusion. *EPJ Data Science*, *10*(1), 51.

Ross, K., Fountaine, S., & Comrie, M. (2014). Facing up to Facebook: Politicians, publics and the social media (ted) turn in New Zealand. *Media, Culture & Society*, *37*(2), 251–269.

Saha, P., Mathew, B., Garimella, K., & Mukherjee, A. (2021). "short is the road that leads from fear to hate": Fear speech in Indian WhatsApp groups. *Proceedings of the WWW*.

Scott, J. (2019). *Women MPs say abuse forcing them from politics*. https://www.bbc.com/news/election-2019-50246969

Sellars, A. (2016). Defining hate speech. *Berkman Klein Center Research Publication*, *2016*(20), 16–48.

Smith, K. L. (2021). *LibGuides: Congressional Twitter accounts*. https://ucsd.libguides.com/congress%5C_twitter

Solovev, K., & Pröllochs, N. (2022). Moral emotions shape the virality of COVID-19 misinformation on social media. *Proceedings of the WWW*.

Sylwester, K., & Purver, M. (2015). Twitter language use reflects psychological differences between democrats and republicans. *PLOS ONE*, *10*(9), e0137422.

Tatum, B. D. (2017). *Why are all the black kids sitting together in the cafeteria? And other conversations about race: And other conversations about race* (3rd ed.). Basic Books.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, *61*(12), 2544–2558.

Theocharis, Y., Barberá, P., Fazekas, Z., & Popa, S. A. (2020). The dynamics of political incivility on Twitter. *SAGE Open*, *10*(2), 215824402091944.

Theocharis, Y., Lowe, W., van Deth, J. W., & García-Albacete, G. (2014). Using twitter to mobilize protest action: Online mobilization patterns and action repertoires in the Occupy Wall Street, Indignados, and Aganaktismenoi movements. *Information, Communication & Society*, *18*(2), 202–220.

Vidgen, B., Burden, E., & Margetts, H. (2021). *Understanding online hate: VSP regulation and the broader context*. London, UK, Ofcom.

Wagner, A. (2020). Tolerating the trolls? Gendered perceptions of online harassment of politicians in Canada. *Feminist Media Studies*, *22*(1), 32–47.

Wich, M., Breitinger, M., Strathern, W., Naimarevic, M., Groh, G., & Pfeffer, J. (2021). Are your friends also haters? Identification of hater networks on social media: Data paper. *Proceedings of the WWW Companion*.

Winkler, J., Halfmann, A., & Freudenthaler, R. (2017). Backlash effects in online discussions: Effects of gender and counter-stereotypical communication on persuasiveness and likeability. *Proceedings of the ICA*.

Zannettou, S., Bradlyn, B., De Cristofaro, E., Kwak, H., Sirivianos, M., Stringini, G., & Blackburn, J. (2018). What is Gab: A bastion of free speech or an alt-right echo chamber. *Proceedings of the WWW Companion*.

Chapter 3

Moralized Language Predicts Hate Speech on Social Media

3.1 Introduction

Social media platforms are a fertile ground for anti-social behavior, including online harassment, cyber-bullying, and, in particular, hate speech (Bilewicz & Soral, 2020). Broadly speaking, hate speech refers to abusive or threatening speech (or writing) that attacks a person or group, typically on the basis of attributes such as ethnicity, religion, sex, or sexual orientation (United Nations, 2020). Hate speech on social media poses severe risks both to the targeted individuals and society as a whole (United Nations, 2020). At the individual level, it threatens the well-being (physically and psychologically) of those affected (Bilewicz & Soral, 2020; Müller & Schwarz, 2020). At the societal level, it fosters political polarization (Piazza, 2020), which can have severe consequences. Examples include increased opportunities for the spread of misinformation about the target group (Freelon & Wells, 2020), erosion of existing antidiscriminatory norms (Bilewicz & Soral, 2020), and even domestic terrorism (Müller & Schwarz, 2020; Piazza, 2020).

While previous research suggests that hate speech on social media is widespread (Mathew et al., 2019), the mechanisms underlying its proliferation, though critical, have remained largely unresolved. In this work, we approach this question through the lens of morality and its triggering role in social media environments. Social media content delivers not only factual information but also carries moralized content (Brady et al., 2017). Broadly defined, content is moralized if it references ideas, objects, or events construed in terms of the good of a unit larger than the individual (e. g., society) (Brady et al., 2020). Since socially connected users often develop similar ideas and intuitions, moralized content is a key driver of information diffusion on social media (Brady et al., 2017). However, moral ideas have also been postulated to be highly polarizing to social media

users (Heltzel & Laurin, 2020) and thus might trigger animosity, hostility, and malice from ideologically opposing groups. Prior research has found that moral concerns differentiate hate from dislike (Pretus et al., 2022) and argued that hate may be a response to perceived moral transgressions or wrongdoing of the outgroup (Sternberg, 2003). In this situation, people may even feel that hurting others is fundamentally right (Hoover et al., 2021). Furthermore, moralized content plays an important role in fulfilling group-identity motives (Brady et al., 2020) and thus may also trigger hate from ideologically concordant groups rallied up against an outgroup. If moralized content on social media triggers such (negative) reactions in users, then its transmission likely plays a significant role in the proliferation of hate speech. Based on this rationale, we hypothesize that moralized language in social media posts is linked to a higher likelihood of receiving hate speech.

In this study, we investigate the link between moralized language and hate speech on social media. Specifically, we empirically analyze whether differences in the prevalence of hate speech in *replies* to social media posts can be explained by moralized language in the *source post*. To address our research question, we perform a large-scale explanatory analysis based on three datasets consisting of $N = 691{,}234$ social media posts from Twitter authored by societal leaders across three domains, namely, politics, news media, and activism (see Methods). We use textual analysis and machine learning to (1) measure moralized language in the source tweets and (2) determine the share of replies to each source tweet that embeds hate speech. Subsequently, we implement multilevel binomial regression models to estimate whether social media users are more likely to receive hate speech if their posts embed moralized language.

3.2 Results

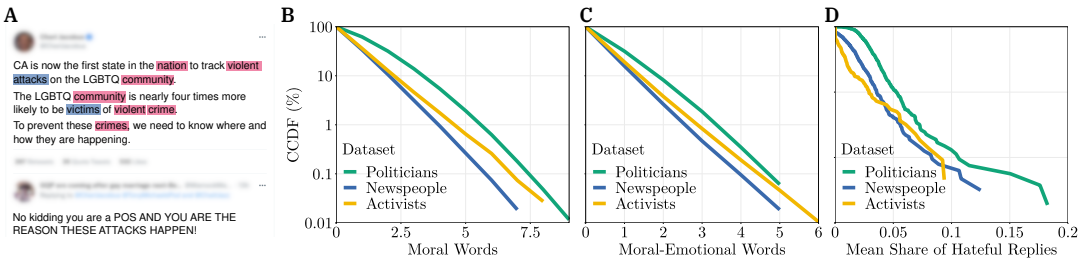


Figure 3.1: (A) Example of moralized language in a source tweet and hate speech in a reply. Here moral words are highlighted in blue and moral-emotional words are highlighted in pink. (B–C) Complementary cumulative distribution functions (CCDFs) for the number of moral and moral-emotional words per source tweet. (D) CCDFs showing the mean share of hateful replies individual users received per source tweet.

We collected three large-scale datasets consisting of 691,234 source tweets and ∼35.5 million corresponding replies in the domains of politics, news media, and activism (see Section 3.A). Specifically, our dataset contained (i) 335,698 tweets that have been authored by the 532 members of the 117th U. S. Congress, (ii) 307,820 tweets from 635 members of five major U. S. TV news networks (CNN, Fox News, NBC News, CBS News, and ABC News), and (iii) 47,716 tweets from 219 influential activists (climate, animal rights, and LGBTQIA+ activists). For each person in the datasets, we collected *all* tweets (excluding retweets and replies) authored during the entire year of 2021, i. e., within an observation period of one year. Politicians were the most active Twitter users, with a monthly average of 52.40 tweets per user. This was followed by newspeople with an average of 40.87 tweets per month and person, and activists with an average of 18.64 tweets per month and person.

We studied whether differences in the prevalence of hate speech in replies to tweets can be explained by moralized language carried in the source tweet (see example in Figure 3.1A). For this purpose, we first used textual analysis to measure moralized language embedded in the source tweets. Specifically, we employed (and validated) a dictionary-based approach (Brady et al., 2017) to count the frequencies of occurrence of moral words and moral-emotional words (see Methods). Politicians tended to use the highest amount of moral and moral-emotional words in their tweets, followed by activists and newspeople (see Figure 3.1B and Figure 3.1C). Second, we employed (and validated) a machine learning model for hate speech detection (Davidson et al., 2017) in order to identify hate speech in replies to tweets (see Methods). The hate speech classifier was used to predict a binary label of whether or not a reply tweet is hateful ($= 1$ if true; otherwise $= 0$) for each reply tweet in our data. On average, the share of hateful replies individual users received per source tweet was highest for politicians (3.26 %), followed by newspeople (2.11 %) and activists (1.61 %). Notably, the distributions were right-skewed, indicating that only a small proportion of users received consistently high shares of hateful replies (see Figure 3.1D).

Subsequently, we fitted explanatory multilevel binomial regression models to estimate the effects of distinctly moral words and moral-emotional words in source tweets on the likelihood of receiving hate speech in the corresponding replies (see Methods). In our binomial regression models, the outcome variable was represented as the proportion of hateful replies relative to all replies. We estimated separate models for each of our three datasets and controlled for previously established content variables that may affect the likelihood of receiving hate speech independent of the main predictors (e. g., number of emotional words, word count, text complexity). The models further included user-specific

random effects to control for heterogeneity at the author level (e. g., differences in users' social influence).

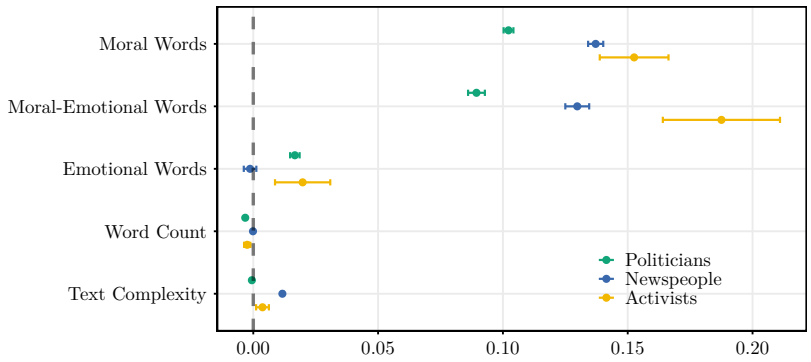


Figure 3.2: Multilevel binomial regression estimating the effects of moral words, moral-emotional words, and further controls on the likelihood of receiving hate speech. Shown are the coefficient estimates with 99 % confidence intervals. User-specific random effects are included.

Figure 3.2 reports the regression results. Across all three datasets, we consistently observed that higher numbers of both moral words and moral-emotional words in source tweets were linked to a higher likelihood of receiving hate speech in replies. For politicians, each additional moral word was associated with 10.66 % higher odds of receiving hate speech (coef = 0.102, 99 % CI = [0.100, 0.105], OR = 1.108, $P < 0.001$). Each moral-emotional word increased the odds of receiving hate speech by 9.35 % (coef = 0.089, 99 % CI = [0.085, 0.094], OR = 1.094, $P < 0.001$). For activists and newspeople, the effects pointed in the same direction. Each additional moral word increased the odds of receiving hate speech of 14.70 % for newspeople (coef = 0.137, 99 % CI = [0.133, 0.141], OR = 1.147, $P < 0.001$) and 16.48 % for activists (coef = 0.153, 99 % CI = [0.134, 0.171], OR = 1.165, $P < 0.001$). Each moral-emotional word was linked to an increase in odds of 20.63 % for activists (coef = 0.188, 99 % CI = [0.157, 0.218], OR = 1.206, $P < 0.001$) and 13.86 % for newspeople (coef = 0.130, 99 % CI = [0.124, 0.136], OR = 1.139, $P < 0.001$). Linear hypothesis tests implied that the estimates of moral and moral-emotional words were significantly different from each other for politicians ($P < 0.001$), newspeople ($P = 0.012$), and activists ($P = 0.018$).

In sum, across all three datasets, we consistently found that higher frequencies of moral and moral-emotional words in source tweets were linked to more hate speech in the corresponding replies. Notably, the effect sizes of moralized language were fairly pronounced. In comparison, purely emotional words only had negligible positive effects on the likelihood of receiving hate for politicians (coef = 0.017, 99 % CI = [0.014, 0.019], OR = 1.017, $P < 0.001$) and activists (coef = 0.020, 99 % CI = [0.005, 0.034], OR = 1.020, $P < 0.001$), and were not significant for newspeople (coef = -0.001, 99 % CI = [-0.005, 0.002], OR = 0.999, $P = 0.315$).

Likewise, the effect sizes of other content characteristics, i. e., the word count (coefs between -0.003 and 0.000; $P < 0.001$ for politicians; $P < 0.001$ for activists; $P = 0.365$ for newspeople) and text complexity (coefs between -0.001 and 0.012; $P = 0.012$ for politicians; $P < 0.001$ for newspeople; $P = 0.005$ for activists) were small. Pairwise comparisons among the coefficient estimates (linear hypothesis tests) confirmed that the estimates of moral and moral-emotional words were significantly greater than for any one of the established content characteristics (all $P < 0.001$).

Multiple exploratory analyses extended our results and confirmed their robustness (see Section 3.E). First we compared our model to an implausible model (Burton et al., 2021) and tested whether the number of X's, Y's and Z's in source tweets (i. e., an absurd factor) would have been an equally adequate predictor of hate speech in the replies. Across all three datasets, implausible models resulted in higher AIC values (i. e., lower model adequacy) and effect sizes close to zero. Second, we implemented 10-fold cross-validation to assess the ability of moralized language to predict hate speech prevalence on out-of-sample data. Compared to a baseline model that only used established author and content features, additionally incorporating word counts for moralized language resulted in an out-of-sample R^2 that was 1.22 times higher for politicians, 1.55 times higher for newspeople, and 1.42 times higher for activists. Third, a wide variety of checks confirmed that our findings held for users across both sides of the political spectrum, across different types of hate speech, and when incorporating additional control variables (e. g., the retweet count of the source tweet). Taken together, our exploratory analyses provided confirmatory evidence that moralized language was a robust and meaningful predictor of hate speech.

3.3 Discussion

This study provides observational evidence that moralized language in social media posts is associated with more hate speech in the corresponding replies. We uncovered this link for posts from a diverse set of societal leaders across three domains (politics, news media, activism). On average, each additional moral word was associated with between 10.66 % and 16.48 % higher odds of receiving hate speech. Likewise, each additional moral-emotional word increased the odds of receiving hate speech by between 9.35 % and 20.63 %. Across the three domains, the effect sizes were most pronounced for activists. A possible reason is that the activists in our data were affiliated with politically left-leaning subjects (climate, animal rights, and LGBTQIA+) that may have been particularly likely to trigger hate speech from right-wing groups. In contrast, our data for politicians

and newspeople were fairly balanced and encompassed users from both sides of the political spectrum. Overall, the comparatively large effect sizes underscore the salient role of moralized language on social media. While earlier research has demonstrated that moralized language is associated with greater virality (Brady et al., 2017; Solovev & Pröllochs, 2022), our work implies that it fosters the proliferation of hate speech.

Notably, a connection between morality and hate has been postulated by social psychology theorists for many years, yet empirical evidence has remained scant. Previous work on the psychology of hate and morality argued that hate is rooted in seeing the hated target as morally deficient (Sternberg, 2003), that morality plays a differentiating role between hate and dislike (Pretus et al., 2022), and that perceptions of outgroup moral wrongdoing may (morally) motivate real-world hate groups (Hoover et al., 2021). Our study adds by demonstrating that moralized language predicts hate speech on social media. Future research may expand upon our work by analyzing users not in a societal leadership role (i. e., regular users), hate speech across ideologically opposing vs. concordant groups, and the role of social status in the proliferation of hate speech.

From a practical perspective, observing and understanding the mechanisms underlying the proliferation of hate speech is the first step toward containing it. While we do not advocate that users *should* avoid moralized language in their social media posts, our work still provides a plausible explanation for *why* certain posts / users receive high levels of hate speech. As such, our findings not only help to foster social media literacy but may also inform educational applications, counterspeech strategies, and automated methods for hate speech detection.

3.4 Methods

Moralized Language (Section 3.B): We applied a dictionary-based approach (Brady et al., 2017) to count the number of moral, moral-emotional, and emotional words in each source tweet. To validate the (previously validated Brady et al., 2017) dictionaries, we recruited four trained research assistants. Words from the distinctly moral and moral-emotional word lists were rated as more "moral" than words from the distinctly emotional word list and non-dictionary words ($P < 0.001$). The annotators yielded a relatively high Kendall's coefficient of concordance of $W = 0.67$ ($P = 0.007$).

Hate Speech Detection (Section 3.C): We used the dataset from (Davidson et al., 2017) to train a classifier that predicted a binary label of whether a reply was hateful. As validation, two trained research assistants annotated 2000 reply

tweets classified as hateful/not hateful (Kendall's $W = 0.69$; $P < 0.001$). The classifier achieved a relatively high balanced accuracy of 0.70.

Model Specification (Section 3.D): We implemented multilevel binomial regressions to estimate the effects of moralized language on the likelihood of receiving hate speech. The number of hate speech replies was modeled as a binomial variable, where the number of trials was given by the total number of replies a tweet received. The key explanatory variables were the absolute counts of moral and moral-emotional words in the source tweets. We controlled for established content characteristics (e. g., emotional words, word count, text complexity) and used random effects to account for author-level heterogeneity.

Bibliography

Bilewicz, M., & Soral, W. (2020). Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, *41*(S1), 3–33.

Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (2020). The mad model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science*, *15*(4), 978–1010.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *PNAS*, *114*(28), 7313–7318.

Burton, J. W., Cruz, N., & Hahn, U. (2021). Reconsidering evidence of moral contagion in online social networks. *Nature Human Behaviour*, *5*(12), 1629–1635.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the ICWSM*, *11*(1), 512–515.

Freelon, D., & Wells, C. (2020). Disinformation as political communication. *Political Communication*, *37*(2), 145–156.

Heltzel, G., & Laurin, K. (2020). Polarization in America: Two possible futures. *Current Opinion in Behavioral Sciences*, *34*, 179–184.

Hoover, J., Atari, M., Mostafazadeh Davani, A., Kennedy, B., Portillo-Wightman, G., Yeh, L., & Dehghani, M. (2021). Investigating the role of group-based morality in extreme behavioral expressions of prejudice. *Nature Communications*, *12*(1), 1–13.

Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019). Spread of hate speech in online social media. *Proceedings of the WebSci*.

Müller, K., & Schwarz, C. (2020). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, *19*(4), 2131–2167.

Piazza, J. A. (2020). Politician hate speech and domestic terrorism. *International Interactions*, *46*(3), 431–453.

Pretus, C., Ray, J. L., Granot, Y., Cunningham, W. A., & Van Bavel, J. J. (2022). The psychology of hate: Moral concerns differentiate hate from dislike. *European Journal of Social Psychology*, *53*(2), 336–353.

Solovev, K., & Pröllochs, N. (2022). Hate speech in the political discourse on social media: Disparities across parties, gender, and ethnicity. *Proceedings of the WWW*.

Sternberg, R. J. (2003). A duplex theory of hate: Development and application to terrorism, massacres, and genocide. *Review of General Psychology*, *7*(3), 299–328.

United Nations. (2020). *United Nations strategy and plan of action on hate speech – Detailed guidance on implementation for United Nations field presences*. https://www.un.org/en/genocideprevention/hate-speech-strategy.shtml

Appendix 3.A Data Collection

We collected three large-scale datasets consisting of tweets from societal leaders across three domains:

- *Dataset I (Members of U. S. Congress):* We collected tweets from the 532 members of the 117th U. S. Congress that convened on January 3, 2021. A curated list of Twitter handles of every politician was downloaded from the University of California San Diego library (Smith, 2021). We employed the Twitter API v2 through the Academic Research track (Twitter, 2022) to download the complete tweet history (excluding retweets and replies) of each politician between January 3, 2021, and the end of 2021, i. e., for an observation period of approximately one year. The resulting dataset contained 335,698 tweets.
- *Dataset II (Newspeople):* We collected tweets from 635 hosts, regular contributors, anchors, reporters, and correspondents of five major U. S. TV news networks, namely, CNN, Fox News, NBC News, CBS News, and ABC News. The list of newspeople and their Twitter handles was gathered from the webpages of the TV news networks, their social media pages, and via manual web search. We used the Twitter API to download the entire tweet history (excluding retweets and replies) for each person in 2021, i. e., for an observation period of one year. The resulting dataset contained 307,820 tweets.
- *Dataset III (Activists):* We collected tweets from 219 climate, animal rights, and LGBTQIA+ activists. Since we are not aware of a single database for different groups of activists, we employed publicly available lists of activists from Wikipedia and retrieved the corresponding Twitter handles via manual web search. For the 219 activists, we retrieved all 47,716 tweets (excluding retweets and replies) that have been posted during the entire year of 2021.

After collecting the source tweets for each of the three datasets, we queried Twitter's API to gather replies to every source tweet. To ensure computational feasibility, we restricted the data collection to up to 500 replies for each source tweet, starting with the earliest reply. As a check, we also experimented with an alternative variant using random samples of replies. Here, we observed consistent results. Note that there is a possibility that Twitter may have removed some particularly egregious hate speech replies, which were, therefore, not available for our current analyses. In total, our three datasets contained 691,234 source tweets and 35,548,076 replies.

Appendix 3.B Measurement of Moralized Language

We used a dictionary-based approach to measure the extent to which moral language is embedded in the source tweets. For this purpose, we first applied standard preprocessing steps from text mining. Specifically, the running text was converted into lower-case and tokenized, and special characters (e. g., hashtags, emoticons) were removed. Subsequently, we employed the dictionary from (Brady et al., 2017), which consists of three word lists: (i) a set of distinctly moral words ($N = 343$), (ii) a set of distinctly emotional words ($N = 848$), and (iii) a set of moral-emotional words ($N = 68$) representing words that are both moral and emotional. We used these word lists to calculate the absolute frequencies of moral words, moral-emotional words, and emotional words in each source tweet.

The discriminant validity of each of the three word lists has previously been validated (Brady et al., 2017). However, as an additional check, we recruited four trained research assistants and repeated the validation procedure. Analogous to the study from (Brady et al., 2017), participants were presented with words that were randomly sampled from each of the word lists ($n = 40$ for strictly moral words, $n = 40$ for strictly emotional words, and $n = 10$ for moral-emotional words). Additionally, we sampled 40 random words from the Linguistic Inquiry and Word Count (Pennebaker et al., 2015) that are not present in the aforementioned word lists (i. e., non-dictionary words). The participants were then asked to rate each word on continuous dimensions of morality and emotions. For this, participants had to answer the question "How related is this word to [morality, emotions]" on a 5-point Likert scale ranging from "Completely Unrelated" to "Completely Related".

In our pilot study, words from the distinctly moral and moral-emotional word lists were rated as more "moral" than words from the distinctly emotional word list and non-dictionary words ($P < 0.001$). Words from the distinctly emotional word list were rated as more "emotional" than words from the distinctly moral word list and non-dictionary words ($P < 0.001$). Words from the moral-emotional word list were rated as more "moral" and "emotional" than non-dictionary words ($P < 0.001$). We observed a relatively high Kendall's coefficient of concordance of $W = 0.67$ ($P = 0.007$) for the moral ratings and $W = 0.68$ ($P = 0.004$) for the emotion ratings. Altogether, our pilot study supported the discriminant validity of the dictionaries.

Appendix 3.C Hate Speech Detection

We used machine learning to detect hate speech in replies to tweets. Compared to dictionary-based methods that merely count hate-related words, this approach is generally considered as being more accurate (Badjatiya et al., 2017).

We implemented machine learning for hate speech detection as follows: we employed the annotated Twitter dataset from (Davidson et al., 2017), containing 25,000 tweets labeled as hateful or not hateful. Each tweet was annotated by at least three users who were explicitly instructed to think about the context of the message and not only the words contained within (Davidson et al., 2017). Purely offensive (non-hateful) language was not considered as hate speech. We used the annotated tweets to implement a deep neural network classifier that predicted whether or not a reply tweet was hateful. Here we used Universal Sentence Encoder (USE) (Cer et al., 2018) as text representation. The hate speech classifier was used to predict a binary hate speech label ($= 1$ if true; otherwise $= 0$) for each reply tweet in our dataset. The machine learning model was implemented in Python 3.8.10 using TensorFlow 2.8.0.

We used a two-pronged approach to evaluate the prediction performance of the machine learning model: (i) we evaluated the out-of-sample prediction performance on the dataset from (Davidson et al., 2017). Here the machine learning classifier yielded an out-of-sample *balanced accuracy* of 0.77 (using 5-fold cross-validation). The predictive performance is similar to previous works (Davidson et al., 2017) and can be seen as reasonably accurate in the context of our study. (ii) We employed two trained research assistants to annotate random subsets of reply tweets that were classified as hateful and not hateful (500 replies per category for each dataset) by the machine learning model. The annotators yielded a Kendall's coefficient of concordance of $W = 0.69$ ($P < 0.001$) and the classifier achieved a *balanced accuracy* of 0.70. This implied that the machine learning classifier was capable of producing relatively reliable hate speech predictions for our data.

Appendix 3.D Regression Analysis

We implemented a multilevel binomial regression to estimate the effects of moralized language of a tweet on receiving hate speech. Formally, we modeled the number of hate speech replies, *HReplies*, as a binomial variable with probability parameter θ. The number of trials was given by the total number of replies a tweet received ($Replies$). The key explanatory variables were the number of moral words (*Moral Words*) and moral-emotional words (*Moral-Emotional Words*).

We controlled for the number of purely emotional words (*Emotional Words*), the word count (*Word Count*), and used the Gunning Fog Index (Gunning, 1968) as a measure of text complexity (*Text Complexity*). In addition, we used binary variables to control for whether media was attached to the tweet (*Media Attached*; $= 1$ if true, otherwise 0) and whether the tweet was a quote tweet (*Quote*; $= 1$ if true, otherwise 0). Based on these variables, we specified the following regression model:

$$\begin{aligned} \text{logit}(\theta) = \beta_0 &+ \beta_1 \textit{ Moral Words} + \beta_2 \textit{ Moral-Emotional Words} \\ &+ \beta_3 \textit{ Emotional Words} + \beta_4 \textit{ Word Count} + \beta_5 \textit{ Text Complexity} \\ &+ \beta_6 \textit{ Media Attached} + \beta_7 \textit{ Quote} + u_{\text{user}} + \varepsilon, \end{aligned} \tag{3.1}$$

$$HReplies \sim Binomial[Replies, \theta], \tag{3.2}$$

with intercept β_0, error term ε, and user-specific random effects u_{user}. Note that the latter was important as it allowed us to control for heterogeneity across users that have authored the source tweets (e. g., varying social influence, different audiences, etc.).

We estimated Equation (3.1) and Equation (3.2) using MLE and generalized linear models. Our regression analyses were implemented in R 4.2.0 using the `lme4` package (Bates et al., 2021).

Appendix 3.E Robustness Checks and Exploratory Analyses

We performed a broad set of checks and exploratory analyses to validate the robustness of our findings. In all cases, our results were robust and consistently supported our findings. In the following, we summarize the main results.

We calculated variance inflation factors for all explanatory variables in our analysis (Table 3.1). The VIFs ranged from 1.046 to 1.531 and were thus substantially below the critical threshold of five (Akinwande et al., 2015). This indicates that multicollinearity was not an issue in our analysis.

We tested alternative model specifications in which we (i) analyzed ratios of word counts (i. e., word frequencies divided by word counts) instead of word counts (Table 3.2); (ii) coded variables for moralized language as dichotomous (i. e., with a binary variable indicating whether the source tweet contained one or more moral / moral-emotional words, or none; see Table 3.3). Furthermore, we repeated our analysis with a zero-one-inflated beta regression that used the

share of hateful replies as the dependent variable (Table 3.4). In all cases, the results were robust and continued to support our findings.

We tested whether the strength of the association between moralized language and hate speech varied depending on the length of the tweet. For this purpose, we extended the regression models from our main analysis with interaction terms between moralized language and the word count (Table 3.5). The coefficient estimate for the interaction *Moral Words* $\times$ *Word Count* was statistically significant and negative for politicians (coef = -0.001, 99 % CI = [-0.001, 0.000], OR = 0.999, $P < 0.001$), newspeople (coef = -0.003, 99 % CI = [-0.003, -0.002], OR = 0.997, $P < 0.001$), and activists (coef = -0.004, 99 % CI = [-0.006, -0.003], OR = 0.996, $P < 0.001$). The coefficient for the interaction *Moral-Emotional Words* $\times$ *Word Count* was negative for newspeople (-0.002, 99 % CI = [-0.003, -0.002], OR = 0.998, $P < 0.001$) and activists (coef = -0.005, 99 % CI = [-0.007, -0.002], OR = 0.995, $P < 0.001$). We observed no statistically significant coefficient for politicians (coef = -0.000, 99 % CI=[-0.001, 0.000], OR = -0.000, $P = 0.254$). Overall, these results suggested that the link between moralized language and hate speech tended to be (slightly) stronger for shorter source tweets.

We explored whether the link between moralized language and hate speech differed across different types of hate speech. Hate speech can be directed at a specific user (directed hate speech) or at a general group of individuals (generalized hate speech) (ElSherief et al., 2018). Previous research (ElSherief et al., 2018) has shown that directed hate speech is correlated with higher use of second-person pronouns (e. g., *you*, *your*), whereas generalized hate speech is correlated with higher use of third-person plural pronouns (e. g., *they*, *themselves*). We thus employed the LIWC dictionary (Pennebaker et al., 2015) to identify the presence of second-person pronouns (as a proxy for directed hate speech) and third-person plural pronouns (as a proxy for generalized hate speech) in each hateful reply in our datasets. Second-person pronouns were present in 68 % of the hateful replies to source tweets from politicians, in 41 % of the hateful replies to source tweets from newspeople, and in 53 % of the hateful replies to source tweets from activists. For third-person plural pronouns, these numbers amounted to 10 % for politicians, 15 % for newspeople, and 13 % for activists. Thus, across all three datasets, directed hate speech was more prevalent than generalized hate speech.

Subsequently, we repeated our analysis with a regression model in which we replaced *HReplies* (i. e., the number of hateful replies) with count variables that measured the number of directed hate speech replies (i. e., the number of replies that were both hateful and contained second-person pronouns) and

generalized hate speech replies (i. e., the number of replies that were both hateful and contained third-person plural pronouns). The regression results are reported in Table 3.6. Across all three datasets, the effect sizes of moralized language were larger for generalized hate speech than for directed hate speech. For politicians, the coefficient of *Moral Words* was 0.162 (99 % CI = [0.155, 0.170], OR = 1.176, $P < 0.001$) for generalized hate speech and 0.092 (99 % CI = [0.089, 0.095], OR = 1.097, $P < 0.001$) for directed hate speech. For newspeople, the coefficient of *Moral Words* was 0.187 (99 % CI = [0.177, 0.196], OR = 1.205, $P < 0.001$) for generalized hate speech and 0.105 (99 % CI = [0.099, 0.111], OR = 1.111, $P < 0.001$) for directed hate speech. For activists, the coefficient of *Moral Words* was 0.202 (99 % CI = [0.159, 0.246], OR = 1.224, $P < 0.001$) for generalized hate speech and 0.145 (99 % CI = [0.121, 0.170], OR = 1.156, $P < 0.001$) for directed hate speech. The patterns were similar for moral-emotional words. For politicians, the coefficient of *Moral-Emotional Words* was 0.136 (99 % CI = [0.123, 0.149], OR = 1.146, $P < 0.001$) for generalized hate speech and 0.088 (99 % CI = [0.083, 0.094], OR = 1.092, $P < 0.001$) for directed hate speech. For newspeople, the coefficient of *Moral-Emotional Words* was 0.158 (99 % CI = [0.143, 0.173], OR = 1.171, $P < 0.001$) for generalized hate speech and 0.107 (99 % CI = [0.098, 0.117], OR = 1.113, $P < 0.001$) for directed hate speech. For activists, the coefficient of *Moral-Emotional Words* was 0.238 (99 % CI = [0.161, 0.314], OR = 1.269, $P < 0.001$) for generalized hate speech and 0.174 (99 % CI = [0.132, 0.215], OR = 1.190, $P < 0.001$) for directed hate speech.

In sum, across all three datasets, moralized language predicted both directed and generalized hate speech (with larger effect sizes for generalized hate speech). These findings add to the validity of our results.

The variable *Emotional Words* in our main analysis measured the total number of distinctly positive and negative emotion words in the source tweets (see Brady et al., 2017). As a check, we used the LIWC dictionary (Pennebaker et al., 2015) to measure the number of distinctly positive and negative words separately; and repeated our regression analysis. The regression results are reported in Table 3.7. For every distinctly negative emotional word, the odds of receiving hate speech were 4.65 % higher for politicians (coef = 0.045, 99 % CI=[0.041, 0.050], OR = 1.047, $P < 0.001$), 4.49 % higher for newspeople (coef = 0.044, 99 % CI=[0.039, 0.049], OR = 1.045, $P < 0.001$), and 12.18 % higher for activists (coef = 0.115, 99 % CI=[0.091, 0.139], OR = 1.122, $P < 0.001$). For every distinctly positive emotional word, the odds of receiving hate speech were 2.56 % lower for newspeople (coef = -0.026, 99 % CI=[-0.031, -0.021], OR = 0.974, $P < 0.001$), 3.34 % lower for activists (coef = -0.034, 99 % CI=[$-0.0.54$, -0.014], OR = 0.967, $P < 0.001$), and 0.26 % higher (coef = 0.003, 99 % CI=[-0.001, 0.006], OR = 1.003, $P = 0.042$) for politicians. Overall,

these findings indicate that a more negative sentiment in the source tweets was linked to more hate speech in the replies. All findings for moralized language remained robust.

We tested whether hate speech in the source tweets was linked to more hate speech in the corresponding replies. For this purpose, we used our machine learning classifier to predict a binary hate speech label ($= 1$ if true; otherwise $= 0$) for each source tweet in our dataset. We then repeated our regression analysis with this additional explanatory variable (see Table 3.8). The coefficient for *Hateful Source* was positive and statistically significant for politicians (coef = 0.734, OR = 2.082, $P < 0.001$, 99 % CI=[0.713, 0.754], newspeople (coef = 1.018, 99 % CI=[1.001, 1.036], OR = 2.769, $P < 0.001$), and activists (coef = 0.905, 99 % CI=[0.813, 0.998], OR = 2.472, $P < 0.001$). This indicated that hate speech in source tweets was linked to more hate speech in the corresponding replies. All findings for moralized language remained robust.

We constructed additional regression to control for the virality / the level of engagement with the source tweet. Specifically, we implemented two model variants that either used the number of retweets or the number of likes as an additional explanatory variable. Note that we used separate regression models as both variables were highly correlated (correlation of 0.895, 0.886, and 0.880 for politicians, newspeople, and activists, respectively). Across all datasets, the coefficient estimates for the number of retweets (Table 3.9) and the number of likes (Table 3.10) were positive and statistically significant. However, the effect sizes for both the number of retweets (standardized coefficients between 0.007 and 0.022) and the number of likes (standardized coefficients between 0.008 and 0.017) were rather small. All findings for moralized language remained robust.

We estimated a comprehensive regression model with all additional control variables from our exploratory analysis. The results are reported in Table 3.11. All findings for moralized language remained robust.

Our data for politicians (Dataset I) and newspeople (Dataset II) encompassed users from both sides of the political spectrum. This allowed us to test how the strength of association between moralized language and hate speech varied for people from both political parties and across different political leanings of the TV news networks. For this purpose, we implemented two additional regression models: (i) for politicians, we added interaction terms between the predictors for moralized language and the political party of the author of the source tweet (=1 if Democratic; =0 if Republican); (ii) for newspeople, we added interaction terms between the predictors for moralized language and the political leaning of the TV news networks (i.e. *Left*, *Center* or *Right*). To determine the political leanings of the TV news networks, we utilized the website https://mediabiasfactcheck.com,

which provides assessment of political leanings for a large number of media sources. This resulted in the following categorizations of political leanings: *Left* for CNN; *Center* for ABC News, CBS News, and NBC News; *Right* for Fox News.

The regression results are reported in Table 3.12. For politicians, the coefficient estimate for the interactions between party affiliation and the predictors for moralized language were statistically significant and positive. The coefficient estimate for *Moral Words* × *Democratic* was 0.008 (99 % CI=[0.003, 0.014], OR = 1.009, $P < 0.001$) and the coefficient estimate for the interaction *Moral-Emotional Words* × *Democratic* was 0.041 (99 % CI=[0.032, 0.050], OR = 1.042, $P < 0.001$). At the same time, the coefficients of the direct effects of *Moral Words* and *Moral-Emotional Words* remained statistically significant and similar in magnitude as in our main analysis. These results implied that the link between moralized language and hate speech held for both sides of the political spectrum. However, tweets authored by politicians from the Democratic Party were (slightly) more likely to receive hateful replies in response to moralized language than Republicans.

We observed a similar pattern for the dataset with newspeople (Table 3.12). The coefficient estimate for the interaction *Moral Words* × *Right* was 0.016 (99 % CI=[0.005, 0.0027], OR = 1.016, $P < 0.001$) and the coefficient estimate for the interaction *Moral Words* × *Left* was 0.017 (99 % CI=[0.005, 0.029], OR = 1.017, $P < 0.001$). The coefficient estimate for the interaction *Moral-Emotional Words* × *Right* was 0.019 (99 % CI=[0.001, 0.037], OR = 0.019, $P = 0.007$) and the coefficient estimate for *Moral-Emotional Words* × *Left* lacked statistical significance (coef = 0.008, 99 % CI=[-0.012, 0.028], OR = 1.008, $P = 0.284$). At the same time, the direct effects of *Moral Words* and *Moral-Emotional Words* remained statistically significant and similar in magnitude as in our main analysis. These findings implied that the link between moralized language and hate speech held across all political leanings of TV news networks. The differences in the effect sizes between left-leaning vs. right-leaning TV news networks were rather small. However, tweets from newspeople affiliated with left-leaning and right-leaning TV news networks were significantly more likely to receive hate speech in response to moralized language than those from newspeople affiliated with center-oriented TV news networks.

We further note that both models in Table 3.12 showed statistically significant estimates for the direct effect of the political leanings on the likelihood of receiving hate speech. All else being equal, users affiliated with a right-leaning political party (i. e., Republicans) or TV news network (i. e., Fox News) were more likely to receive hate speech on social media.

Taken together, we found that the hypothesized link between moralized language and hate speech was generalizable to both sides of the political spectrum.

Prior research noted that large-scale observational studies can sometimes yield fragile results and even support patently absurd models (Burton et al., 2021). Following earlier work, we thus compared our model to an implausible XYZ model (Burton et al., 2021). Specifically, we counted the number of X's, Y's and Z's in source tweets (i. e., an absurd factor) and tested whether this variable (*XYZ Count*) would have been an equally adequate predictor of hate speech replies. The regression results are reported in Table 3.13. Across all three datasets, XYZ models resulted in higher AIC values (i. e., lower model adequacy) and effect sizes close to zero. The coefficient of *XYZ Count* was 0.009 ($P < 0.001$) for politicians and -0.015 ($P < 0.001$) for activists. For newspeople, the coefficient of *XYZ Count* was not statistically significant ($P = 0.304$). For comparison, the coefficient of *Moral Words* was 0.102 for politicians ($P < 0.001$), 0.137 for newspeople ($P < 0.001$), and 0.153 ($P < 0.001$) for activists. These findings provided strong evidence that moralized language was a meaningful predictor of hate speech.

We used 10-fold cross-validation to assess the ability of moralized language to predict hate speech prevalence on out-of-sample data. For this purpose, we implemented multiple binomial regression models that took different sets of predictors into account. This approach allowed us to compare the out-of-sample prediction performance across different predictor sets. The individual predictor sets were as follows: (i) author features[1] (the number of followers, the number of followees, the account age, and the verified status), (ii) established content features (*Word Count*, *Text Complexity*, and *Emotional Words*, *Quote*, *Media Attached*), and (iii) word counts for moralized language (*Moral Words*, *Moral-Emotional Words*). Note that we used the same linear model as in our in-sample analysis but replaced the author-specific random effect terms with common author-specific variables from previous work. This ensures that out-of-sample predictions can be made for users not present in the training data. All models were implemented in R 4.2.0 using the `tidymodels` package (Kuhn, 2022).

The prediction results for different feature combinations are reported in Table 3.14. Since we used linear regression models, we report the prediction performance in terms of out-of-sample R^2 (calculated using 10-fold cross-validation). Compared to a baseline model that used only established author and content features, additionally incorporating word counts for moralized language resulted in an out-of-sample R^2 that was 1.26 times, 1.57 times, and 1.42 times higher for politicians, newspeople, and activists, respectively.

While our study focuses on predicting the frequency of hateful replies (i. e., a regression problem), we also tested a variant in which we treated the task of identifying source tweets with a disproportionally high share of hateful replies

[1] All author variables were retrieved from the Twitter API v2.

as a binary classification problem. Specifically, we used a logistic regression model and defined a binary response variable *ExtremelyHateful*, which took the value $= 1$ for source tweets that received a disproportionally high share of hateful replies (otherwise $= 0$). In each dataset, source tweets with the 25 % highest share of hateful replies were considered as being *ExtremelyHateful*. To ensure that the analysis was not driven by outliers (e. g., tweets that received only one hateful reply), we excluded source tweets that have received less than 10 total replies. The out-of-sample ROC curves and ROC-AUC (calculated via 10-fold cross-validation) across different sets of predictors are reported in Figure 3.3. We again observed that moralized language was a meaningful predictor of hate speech. Compared to a baseline model that used only established author and content features, additionally incorporating word counts for moralized language resulted in a ROC-AUC that was 6.62 % higher for politicians, 4.48 % higher for newspeople, and 2.77 % higher for activists. Delongs's tests confirmed that the differences in ROC-AUC between the models w/ and w/o features for moralized language were statistically significant (all $P < 0.001$).

As an additional check, we counted the number of X's, Y's and Z's in source tweets (i. e., an absurd predictor) and tested whether this variable (*XYZ Count*) enhances the prediction performance. Consistent with our in-sample analysis, including *XYZ Count* as an additional predictor resulted in practically no changes in out-of-sample R^2 and ROC-AUC. In sum, our out-of-sample analysis confirmed that moralized language was a robust and meaningful predictor of hate speech. Future research may expand on these results by implementing (non-linear) machine learning models for predicting hate speech prevalence.

Appendix 3.F Ethics

This research uses public tweets only, and, thus, no approval from the Institutional Review Board was required by the authors' institutions.

Table 3.1: Variance inflation factors.

	Politicians	**Newspeople**	**Activists**
Moral Words	1.153	1.143	1.155
Moral-Emotional Words	1.057	1.063	1.084
Emotional Words	1.151	1.198	1.225
Word Count	1.474	1.415	1.531
Text Complexity	1.184	1.050	1.120
Media Attached	1.055	1.055	1.094
Quote	1.046	1.072	1.154

Table 3.2: Regression results with proportions of word counts (i. e., word frequencies divided by word counts).

	Politicians	**Newspeople**	**Activists**
Moral Words	2.625***	3.398***	3.251***
	(0.028)	(0.037)	(0.132)
Moral-Emotional Words	2.111***	2.973***	4.665***
	(0.046)	(0.059)	(0.268)
Emotional Words	0.363***	−0.134***	0.689***
	(0.025)	(0.031)	(0.146)
Word Count	0.002***	0.005***	0.004***
	(0.000)	(0.000)	(0.001)
Text Complexity	−0.001***	0.012***	0.005***
	(0.000)	(0.000)	(0.001)
Media Attached (binary)	0.054***	−0.008	−0.118***
	(0.003)	(0.005)	(0.019)
Quote (binary)	0.046***	0.127***	0.165***
	(0.004)	(0.004)	(0.020)
Intercept	−3.568***	−4.471***	−5.083***
	(0.021)	(0.035)	(0.089)
Random effects (user level)	Included	Included	Included
Observations	335 698	307 820	47 716
AIC	806 293	765 918	48 591

Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; standard errors in parentheses

Table 3.3: Regression results with binary variables for moral and moral-emotional language (=1 if the tweet contained one or more moral / moral-emotional words, otherwise 0).

	Politicians	**Newspeople**	**Activists**
Moral Words (binary)	0.168***	0.236***	0.321***
	(0.003)	(0.003)	(0.016)
Moral-Emotional Words (binary)	0.131***	0.189***	0.265***
	(0.003)	(0.004)	(0.018)
Emotional Words	0.020***	0.000	0.020***
	(0.001)	(0.001)	(0.006)
Word Count	−0.002***	0.001***	−0.002***
	(0.000)	(0.000)	(0.001)
Text Complexity	0.000	0.012***	0.004***
	(0.000)	(0.000)	(0.001)
Media Attached (binary)	0.046***	−0.011*	−0.124***
	(0.003)	(0.005)	(0.019)
Quote (binary)	0.039***	0.121***	0.146***
	(0.004)	(0.004)	(0.020)
Intercept	−3.483***	−4.395***	−4.962***
	(0.021)	(0.035)	(0.088)
Random effects (user level)	Included	Included	Included
Observations	335 698	307 820	47 716
AIC	809 386	767 155	48 591

Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; standard errors in parentheses

Table 3.4: Coefficient estimates for zero-one-inflated beta regression. The dependent variable is the share of hateful replies.

	Politicians	**Newspeople**	**Activists**
Moral Words	0.058***	0.070***	0.071***
	(0.002)	(0.003)	(0.011)
Moral-Emotional Words	0.068***	0.072***	0.073***
	(0.003)	(0.005)	(0.017)
Emotional Words	0.038***	−0.006**	−0.003
	(0.002)	(0.002)	(0.008)
Word Count	−0.003***	−0.003***	−0.003**
	(0.000)	(0.000)	(0.001)
Text Complexity	0.000	0.011***	0.008***
	(0.000)	(0.001)	(0.002)
Media Attached (binary)	0.073***	−0.025**	−0.079**
	(0.005)	(0.009)	(0.027)
Quote (binary)	0.096***	0.178***	0.247***
	(0.006)	(0.007)	(0.029)
Intercept	−2.155***	−2.348***	−2.339***
	(0.029)	(0.035)	(0.094)
Random effects (user level)	Included	Included	Included
Observations	335 698	307 820	47 716
WAIC	−44 240	−75 268	−4 890

significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; standard errors in parentheses

Table 3.5: Regression results with interactions between moralized language and word count.

	Politicians	**Newspeople**	**Activists**
Moral Words × Word Count	−0.001***	−0.003***	−0.004***
	(0.000)	(0.000)	(0.001)
Moral-Emotional Words × Word Count	0.000	−0.002***	−0.005***
	(0.000)	(0.000)	(0.001)
Moral Words	0.123***	0.248***	0.332***
	(0.003)	(0.004)	(0.025)
Moral-Emotional Words	0.096***	0.211***	0.385***
	(0.006)	(0.007)	(0.040)
Emotional Words	0.017***	−0.001	0.020***
	(0.001)	(0.001)	(0.006)
Word Count	−0.003***	0.002***	0.000
	(0.000)	(0.000)	(0.001)
Text Complexity	−0.001**	0.011***	0.004**
	(0.000)	(0.000)	(0.001)
Media Attached (binary)	0.053***	−0.003	−0.123***
	(0.003)	(0.005)	(0.019)
Quote (binary)	0.041***	0.128***	0.160***
	(0.004)	(0.004)	(0.020)
Intercept	−3.458***	−4.386***	−5.007***
	(0.021)	(0.034)	(0.088)
Random effects (user level)	Included	Included	Included
Observations	335 698	307 820	47 716
AIC	803 165	764 419	48 387

Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; standard errors in parentheses

Table 3.6: Regression results for predicting the number of hateful replies that include second-person pronouns (2PP) and third-person plural pronouns (3PP). Second person pronouns tend to occur in conjunction with directed hate speech (i. e., hate against a specific person), whereas third person plural pronouns tend to occur in conjunction with generalized hate speech (i. e., hate against a group) (ElSherief et al., 2018).

	Politicians		**Newspeople**		**Activists**	
	2PP	3PP	2PP	3PP	2PP	3PP
Moral Words	0.092***	0.162***	0.105***	0.187***	0.145***	0.202***
	(0.001)	(0.003)	(0.002)	(0.004)	(0.010)	(0.017)
Moral-Emotional Words	0.088***	0.136***	0.107***	0.158***	0.174***	0.238***
	(0.002)	(0.005)	(0.004)	(0.006)	(0.016)	(0.030)
Emotional Words	0.022***	0.025***	0.023***	−0.025***	0.035***	0.029*
	(0.001)	(0.003)	(0.002)	(0.003)	(0.008)	(0.015)
Word Count	−0.002***	−0.004***	0.004***	0.000	0.000	−0.001
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.002)
Text Complexity	−0.002***	0.003***	0.007***	0.017***	−0.003	0.011**
	(0.000)	(0.001)	(0.000)	(0.001)	(0.002)	(0.004)
Media Attached (binary)	0.063***	0.021*	0.060***	−0.082***	−0.113***	−0.282***
	(0.003)	(0.009)	(0.007)	(0.012)	(0.025)	(0.052)
Quote (binary)	−0.007	0.076***	0.110***	0.086***	0.133***	0.183***
	(0.005)	(0.011)	(0.006)	(0.010)	(0.028)	(0.054)
Intercept	−3.889***	−5.758***	−5.394***	−6.255***	−5.682***	−7.130***
	(0.024)	(0.026)	(0.040)	(0.040)	(0.101)	(0.139)
Random effects (user level)	Included	Included	Included	Included	Included	Included
Observations	335 698	335 698	307 820	307 820	47 716	47 716
AIC	642 572	247 116	414 477	260 770	29 111	13 961

Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; standard errors in parentheses

Table 3.7: Regression results with distinctly positive and distinctly negative emotional words.

	Politicians	**Newspeople**	**Activists**
Moral Words	0.102***	0.135***	0.145***
	(0.001)	(0.002)	(0.007)
Moral-Emotional Words	0.089***	0.127***	0.178***
	(0.002)	(0.002)	(0.012)
Positive Emotional Words	0.003*	−0.026***	−0.034***
	(0.001)	(0.002)	(0.008)
Negative Emotional Words	0.045***	0.044***	0.115***
	(0.002)	(0.002)	(0.009)
Word Count	−0.003***	0.000**	−0.002**
	(0.000)	(0.000)	(0.001)
Text Complexity	−0.001***	0.011***	0.003*
	(0.000)	(0.000)	(0.001)
Media Attached (binary)	0.056***	−0.004	−0.133***
	(0.003)	(0.005)	(0.019)
Quote (binary)	0.040***	0.120***	0.163***
	(0.004)	(0.004)	(0.020)
Intercept	−3.443***	−4.341***	−4.886***
	(0.021)	(0.034)	(0.087)
Random effects (user level)	Included	Included	Included
Observations	335 698	307 820	47 716
AIC	802 699	764 622	48 318

Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; standard errors in parentheses

Table 3.8: Regression results controlling for hate speech in the source tweet (*Hateful Source*).

	Politicians	**Newspeople**	**Activists**
Moral Words	0.097***	0.121***	0.139***
	(0.001)	(0.002)	(0.007)
Moral-Emotional Words	0.085***	0.116***	0.174***
	(0.002)	(0.002)	(0.012)
Emotional Words	0.017***	−0.001	0.018**
	(0.001)	(0.001)	(0.006)
Word Count	−0.003***	0.001***	−0.002*
	(0.000)	(0.000)	(0.001)
Text Complexity	−0.001**	0.011***	0.004**
	(0.000)	(0.000)	(0.001)
Media Attached (binary)	0.053***	−0.014**	−0.120***
	(0.003)	(0.005)	(0.019)
Quote (binary)	0.033***	0.108***	0.145***
	(0.004)	(0.004)	(0.020)
Hateful Source (binary)	0.734***	1.018***	0.905***
	(0.008)	(0.007)	(0.036)
Intercept	−3.467***	−4.362***	−4.935***
	(0.021)	(0.034)	(0.086)
Random effects (user level)	Included	Included	Included
Observations	335 698	307 820	47 716
AIC	796 188	747 931	47 953

Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; standard errors in parentheses

Table 3.9: Regression results controlling for the number of retweets of the source tweet. Due to varying scales, all numeric variables have been standardized.

	Politicians	**Newspeople**	**Activists**
Moral Words	0.128***	0.112***	0.142***
	(0.001)	(0.001)	(0.007)
Moral-Emotional Words	0.064***	0.062***	0.100***
	(0.001)	(0.001)	(0.007)
Emotional Words	0.024***	−0.001	0.025**
	(0.001)	(0.002)	(0.008)
Word Count	−0.039***	−0.003	−0.034***
	(0.001)	(0.002)	(0.010)
Text Complexity	−0.003*	0.069***	0.025**
	(0.001)	(0.002)	(0.008)
Media Attached (binary)	0.055***	−0.003	−0.140***
	(0.003)	(0.005)	(0.019)
Quote (binary)	0.046***	0.137***	0.165***
	(0.004)	(0.004)	(0.020)
Retweet Count	0.007***	0.017***	0.022***
	(0.000)	(0.001)	(0.002)
Intercept	−3.400***	−4.125***	−4.777***
	(0.021)	(0.034)	(0.085)
Random effects (user level)	Included	Included	Included
Observations	335 698	307 820	47 716
AIC	802 908	764 330	48 380

Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; standard errors in parentheses

Table 3.10: Regression results controlling for the number of likes of the source tweet. Due to varying scales, all numeric variables have been standardized.

	Politicians	**Newspeople**	**Activists**
Moral Words	0.128***	0.112***	0.143***
	(0.001)	(0.001)	(0.007)
Moral-Emotional Words	0.064***	0.062***	0.103***
	(0.001)	(0.001)	(0.007)
Emotional Words	0.023***	−0.002	0.025**
	(0.001)	(0.002)	(0.008)
Word Count	−0.037***	−0.001	−0.034***
	(0.001)	(0.002)	(0.010)
Text Complexity	−0.003*	0.070***	0.026**
	(0.001)	(0.002)	(0.008)
Media Attached (binary)	0.055***	−0.006	−0.143***
	(0.003)	(0.005)	(0.019)
Quote (binary)	0.047***	0.129***	0.163***
	(0.004)	(0.004)	(0.020)
Like Count	0.008***	0.008***	0.017***
	(0.000)	(0.001)	(0.002)
Intercept	−3.400***	−4.117***	−4.774***
	(0.021)	(0.034)	(0.085)
Random effects (user level)	Included	Included	Included
Observations	335 698	307 820	47 716
AIC	802 824	765 146	48 423

Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; standard errors in parentheses

Table 3.11: Comprehensive regression model with all additional control variables. Due to varying scales, all numeric variables have been standardized.

	Politicians	**Newspeople**	**Activists**
Moral Words × WordCount	−0.004**	−0.028***	−0.056***
	(0.001)	(0.001)	(0.008)
Moral-Emotional Words × WordCount	−0.002	−0.014***	−0.034***
	(0.001)	(0.001)	(0.008)
Moral Words	0.122***	0.116***	0.174***
	(0.001)	(0.002)	(0.010)
Moral-Emotional Words	0.061***	0.065***	0.111***
	(0.001)	(0.001)	(0.009)
Positive Emotional Words	0.005***	−0.020***	−0.030***
	(0.001)	(0.002)	(0.008)
Negative Emotional Words	0.031***	0.026***	0.071***
	(0.001)	(0.001)	(0.007)
Word Count	−0.035***	0.009***	−0.026**
	(0.001)	(0.002)	(0.010)
Text Complexity	−0.005***	0.062***	0.018*
	(0.001)	(0.002)	(0.008)
Media Attached (binary)	0.059***	0.002	−0.094***
	(0.003)	(0.005)	(0.019)
Quote (binary)	0.039***	0.133***	0.172***
	(0.004)	(0.004)	(0.020)
Hateful Source (binary)	0.726***	1.006***	0.847***
	(0.008)	(0.007)	(0.036)
Retweet Count	0.007***	0.017***	0.019***
	(0.000)	(0.001)	(0.002)
Intercept	−3.413***	−4.124***	−4.787***
	(0.020)	(0.034)	(0.083)
Random effects (user level)	Included	Included	Included
AIC	795 521	745 749	47 687
Observations	335 698	307 820	47 716

Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; standard errors in parentheses

Table 3.12: Regression results with interactions between moralized language and political leanings. For politicians, the binary variable *Democratic* indicates whether a politician is affiliated with the Democratic Party (=1 if true, otherwise 0). The reference category refers to politicians affiliated with the Republican Party. For newspeople, the binary variables *Right* and *Left* indicate whether a newsperson is affiliated with a left-leaning or right-leaning TV news network. The reference category refers to newspeople affiliated with center-oriented TV news networks.

	Politicians	**Newspeople**
Moral Words × Democratic	0.008***	
	(0.002)	
Moral-Emotional Words × Democratic	0.041***	
	(0.003)	
Moral Words × Right		0.016***
		(0.004)
Moral Words × Left		0.017***
		(0.005)
Moral-Emotional Words × Right		0.019**
		(0.007)
Moral-Emotional Words × Left		0.008
		(0.008)
Moral Words	0.100***	0.123***
	(0.001)	(0.004)
Moral-Emotional Words	0.074***	0.117***
	(0.002)	(0.006)
Emotional Words	0.017***	−0.001
	(0.001)	(0.001)
Word Count	−0.003***	0.000
	(0.000)	(0.000)
Text Complexity	−0.001*	0.012***
	(0.000)	(0.000)
Media Attached (binary)	0.052***	−0.009
	(0.003)	(0.005)
Quote (binary)	0.040***	0.120***
	(0.004)	(0.004)
Democratic (binary)	−0.276***	
	(0.040)	
Right (binary)		0.367***
		(0.090)
Left (binary)		0.120
		(0.074)
Intercept	−3.318***	−4.459***
	(0.028)	(0.050)
Random effects (user level)	Included	Included
Observations	335 698	307 820
AIC	803 007	765 301

Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; standard errors in parentheses

Table 3.13: Comparison to an implausible XYZ model (Burton et al., 2021). The variable *XYZ Count* measures the number of X's, Y's and Z's in the source tweets (i. e., an absurd factor).

	Politicians		**Newspeople**		**Activists**	
	XYZ	Main	XYZ	Main	XYZ	Main
Moral Words		0.102***		0.137***		0.153***
		(0.001)		(0.002)		(0.007)
Moral-Emotional Words		0.089***		0.130***		0.188***
		(0.002)		(0.002)		(0.012)
Emotional Words	0.024***	0.017***	0.002	−0.001	0.028***	0.020***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.006)	(0.006)
Word Count	0.000	−0.003***	0.004***	0.000	0.004***	−0.002***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)
Text Complexity	0.001***	−0.001*	0.013***	0.012***	0.005***	0.004**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)
Media Attached (binary)	0.032***	0.052***	−0.036***	−0.009	−0.144***	−0.142***
	(0.003)	(0.003)	(0.005)	(0.005)	(0.019)	(0.019)
Quote (binary)	0.029***	0.040***	0.110***	0.120***	0.162***	0.153***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.020)	(0.020)
XYZ Count	0.009***		0.001		−0.015***	
	(0.001)		(0.001)		(0.004)	
Intercept	−3.462***	−3.448***	−4.387***	−4.352***	−4.923***	−4.899***
	(0.021)	(0.021)	(0.035)	(0.035)	(0.089)	(0.087)
Random effects (user level)	Included	Included	Included	Included	Included	Included
Observations	335 698	335 698	307 820	307 820	47 716	47 716
AIC	816 067	803 205	775 894	765 333	49 232	48 474

Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; standard errors in parentheses

Table 3.14: Out-of-sample R^2 (calculated via 10-fold cross-validation) for different sets of predictors. Author features include common author-specific predictors from previous work (i. e., the number of followers, the number of followees, the account age, and the verified status). Content features include established text characteristics (*Word Count*, *Text Complexity*, *Emotional Words*, *Quote*, and *Media Attached*). The predictors for moralized language include the moral and moral-emotional word counts. The variable *XYZ Count* measures the number of X's, Y's and Z's in source tweets (i. e., an implausible predictor).

Predictors	**Politicians**	**Newspeople**	**Activists**
Author Variables	0.029	0.033	0.065
Author Variables + Content Variables	0.047	0.039	0.077
Author Variables + Content Variables + XYZ Count	0.047	0.039	0.076
Author Variables + Content Variables + Moralized Language	0.059	0.060	0.110

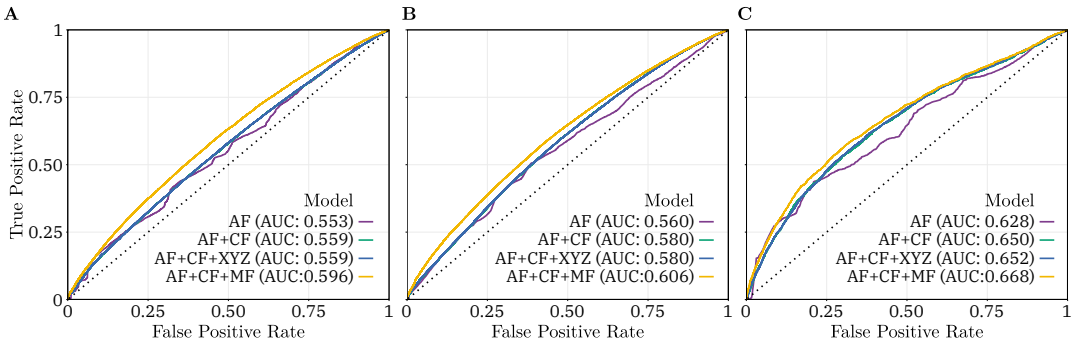


Figure 3.3: Out-of-sample ROC curves and ROC-AUC (calculated via 10-fold cross-validation) across different sets of predictors for (A) Politicians, (B) Newspeople, and (C) Activists. Here, we treated the task of predicting hate speech in replies to source tweets as a binary classification problem. For this, we defined a response variable *ExtremelyHateful*, which took the value $= 1$ for source tweets that received a disproportionally high share of hateful replies (otherwise $= 0$). In each dataset, source tweets with the 25 % highest share of hateful replies were considered as being *ExtremelyHateful*. Source tweets that have received less than 10 replies were excluded. Author features (*AF*) include common author-specific predictors from previous work (i. e., the number of followers, the number of followees, the account age, and the verified status). Content features (*CF*) include established text characteristics (*Word Count*, *Text Complexity*, *Emotional Words*, *Quote*, and *Media Attached*). The predictors for moralized language (*MF*) include the moral and moral-emotional word counts. The predictor *XYZ* measures the number of X's, Y's and Z's in source tweets (i. e., an implausible predictor).

Chapter 4

Moral Emotions Shape the Virality of COVID-19 Misinformation on Social Media

4.1 Introduction

Social media platforms play an ambivalent role during the COVID-19 pandemic. On the one hand, they represent an important source of health information for large parts of society (Limaye et al., 2020). On the other hand, however, this crisis has bred a multitude of rumors (Frenkel et al., 2020; Gallotti et al., 2020; Islam et al., 2020; Kouzy et al., 2020; Pennycook et al., 2020), and verdicts of reputable fact-checking organizations (e. g., politifact.com, snopes.com) suggest that social media is rife with COVID-19 misinformation. COVID-19 misinformation on social media includes, but is not limited to, misinformation about vaccination, "miracle cures," and supposed preventives (Havey, n.d.). False rumors can impact the timely and effective adoption of public health recommendations (Waszak et al., 2018), the effectiveness of the countermeasures deployed by governments (Rapp & Salovich, 2018), and are sometimes even used as a political weapon (Ricard & Medeiros, 2020). Given that exposure to misinformation frequently manifests in offline consequences (Pennycook et al., 2020), there is an urgency to study the spread of rumors on social media in the context of COVID-19. Tedros Adhanom Ghebreyesus, director-general of WHO, and other experts speak of an "infodemic," which must be fought (Zarocostas, 2020).

While previous research – at least for non-crisis situations – suggests that false rumors on social media tend to be more viral than the truth (Pröllochs et al., 2021a; Vosoughi et al., 2018), the mechanism underlying its viral spread, though critical, remains unresolved. In this work, we approach this question through the lenses of morality and emotions and their role in rumor diffusion in polarized social media environments. Social media content delivers not only factual

information but also carries moral ideas and sophisticated emotional signals (Brady et al., 2017). Moral emotions provide the motivational force for humans to do good and to avoid doing bad (Tangney et al., 2007) and can even serve to "moralize" actions that would otherwise be considered non-moral (Wheatley & Haidt, 2005). Since socially connected users often develop similar ideas and intuitions (Brady et al., 2017; Cacioppo et al., 2009; Fowler & Christakis, 2008; Limaye et al., 2020), moral emotions are a key driver of information diffusion in polarized social media environments (Brady et al., 2017). In the context of COVID-19, the overall discussion culture has repeatedly been characterized as highly polarized (Allcott et al., 2020; Cossard et al., 2020; Druckman et al., 2020; Hart et al., 2020; Havey, n.d.; Jing & Ahn, 2021). For instance, people have been observed to be divided in their perceptions of government responses, confidence in scientists, and support for protective actions (Hart et al., 2020; Jing & Ahn, 2021). If COVID-19 rumors are highly polarizing to social media users, then the transmission of moral emotions likely plays a key role in the rumors' diffusion through social networks.

The principal moral emotions can be divided into two families (Haidt, 2003). The families are the "other-condemning" family, comprising the emotions contempt, anger, and disgust, and the "self-conscious" family comprising the emotions shame, pride, and guilt (Tracy & Robins, 2004). Other-condemning emotions, are reactions to the social behavior of others and involve a negative judgment or disapproval of others. In the context of morality, other-condemning emotions are sometimes also referred to as the "hostility triad" (Rozin et al., 1999). Other-condemning emotions are typically associated with perceived moral violations, for example, in the context of individuals' rights and fairness (van Stekelenburg, 2017). While the individual emotions in the other-condemning family (i. e., anger, contempt, and disgust) are often assumed to be not particularly explosive on their own, they can become a dangerous, explosive mix when compressed together (Rozin et al., 1999). Their counterpart is the family of self-conscious emotions, which are evoked by self-reflection and self-evaluation. These emotions motivate individuals to behave in a socially acceptable fashion and are linked to prosocial behaviors such as empathy and altruism (Haidt, 2003; Rozin et al., 1999). As such, self-conscious emotions can enable social healing and avoid triggering the contempt, anger, and disgust of others (Haidt, 2003). While previous research (Brady et al., 2017) broadly distinguished moral vs. non-moral emotions in social media content, we will investigate whether these two clusters of moral emotions (self-conscious vs. other-condemning emotions) have distinct effects on the diffusion of rumors in the context of COVID-19.

Research hypothesis: In this work, we propose that the virality of true vs.

false COVID-19 rumors can be explained by the moral emotions they carry. Although previous research suggests that false rumors are statistically more often retweeted (Pröllochs et al., 2021a; Vosoughi et al., 2018), not every false rumor is necessarily more viral than a truthful rumor. Rather, misinformation going viral is oftentimes spread through echo chambers with exacerbated ideological polarization (Choi et al., 2020). In these environments, ideological identity is more salient in guiding user behavior (Weng et al., 2013) and users are moved towards more extreme positions (Cinelli et al., 2021). Polarization not only reduces verification behavior (Kim, 2017; Moravec et al., 2019) but also makes users more receptive to hostility against others, e. g., for political attacks (Tucker et al., 2018). Here other-condemning emotions embedded in the *source tweets*, which start the rumor cascade, may function as accelerators and amplifiers (van Stekelenburg & Klandermans, 2017). In polarizing discussions about COVID-19, this would imply that radical ideas and beliefs are strengthened and are more likely to translate into action. Given increased ideological polarization for false rumors (Vicario et al., 2019), the explosive mix of other-condemning emotions should thus accelerate their spread within social networks. The same reasoning suggests that false rumors embedding self-conscious emotions (that avoid triggering other-condemning emotions Haidt, 2003) should be less contagious on social media. In sum, we hypothesize that rumors with a stronger combination of false content and other-condemning emotions in the source tweets reach more people, whereas the combination of false content and self-conscious emotions reaches fewer people.

Data: We collected a *unique* dataset of COVID-19 rumor cascades propagating on Twitter between January 2020 and the end of April 2021. Each rumor cascade was investigated and fact-checked by at least one of three independent fact-checking organizations (snopes.com, politifact.com, truthorfiction.com). Our data include 10 610 rumor cascades that have been retweeted 24.34 million times.

Methodology: We use textual analysis to extract fine-grained moral emotions (self-conscious and other-condemning) embedded in rumor cascades. Specifically, we employ (and validated) a dictionary-based approach to count the frequency of occurrence of self-conscious and other-condemning emotion words in the source tweets that have initiated the rumor cascades. To measure the diffusion of each rumor cascade, we employ the Twitter Historical API to obtain the number of retweets, that is, the number of users interacting with the rumor cascade. We then fit *explanatory* regression models to evaluate how variations in moral emotions are associated with differences in the number of retweets

for true vs. false rumor cascades. In our regression analysis, we follow previous works (Brady et al., 2017; Vosoughi et al., 2018) by controlling for variables known to affect the retweet rate independent of the main predictors, i. e., the number of followers, the account age, etc.

Findings: We observe that, on average, COVID-19 misinformation is more likely to go viral than truthful information. However, the veracity effect is moderated by moral emotions: false rumors are more viral than the truth if the source tweets embed a high number of other-condemning emotion words, whereas a higher number of self-conscious emotion words is linked to a less viral spread. The effects are pronounced both for health misinformation and false political rumors. These findings offer insights into how true vs. false rumors spread and highlight the importance of considering emotions from the moral emotion families in social media content.

4.2 Background

4.2.1 Misinformation on Social Media

Social media has shifted quality control for the content from trained journalists to regular users (Kim, 2017). The lack of oversight from experts makes social media vulnerable to the spread of misinformation (Shao et al., 2016). Social media has indeed repeatedly been observed to be a medium that disseminates vast amounts of misinformation (e. g., Pröllochs, 2022; Vosoughi et al., 2018). The presence of misinformation on social media also has detrimental consequences on how opinions are formed in the offline world (Allcott & Gentzkow, 2017; Bakshy et al., 2015; Del Vicario et al., 2016; Oh et al., 2013). As a result, it not only threatens the reputation of individuals and organizations, but also society at large.

Several works have focused on the question of *why* misinformation is widespread on social media. These studies suggest that it is difficult for users to spot misinformation as it is often intentionally written to mislead others (Wu et al., 2019). Moreover, social media users are often in a hedonic mindset and avoid cognitive reasoning such as verification behavior (Moravec et al., 2019). The vast majority of social media users do not fact-check articles they read (Geeng et al., 2020; Vo & Lee, 2018). A recent study further suggests that the current platform design may discourage users from reflecting on accuracy (Pennycook et al., 2021). Online social networks are also characterized by (political) polarization (Levy, 2021; Pröllochs, 2022; Solovev & Pröllochs, 2022) and echo chambers (Barberá et al., 2015). In these information environments with low content diversity and strong social reinforcement, users tend to selectively consume information that

shares similar views or ideologies while disregarding contradictory arguments (Ecker et al., 2010). These effects can even be exaggerated in the presence of repeated exposure: once misinformation has been absorbed, users are less likely to change their beliefs even when the misinformation is debunked (Pennycook et al., 2018).

4.2.2 Research on Rumor Spreading

Several studies have analyzed the spreading dynamics of rumors vs. non-rumors on social media. This includes analyses of summary statistics with regard to, for instance, the number of retweets (e. g., Bessi et al., 2015; Friggeri et al., 2014) and the rumor lifetime (e. g., Bessi et al., 2015; Castillo et al., 2011; Del Vicario et al., 2016). However, these works discern cascades from rumors vs. non-rumors, and do not focus on differences across veracity. Another stream of literature has analyzed rumors concerning specific events (e. g., the 2013 Boston Marathon bombing) with regard to the overall tweet volume or content (e. g., De Domenico et al., 2013; Starbird, 2017; Starbird et al., 2014). These works analyze how the user base responds to rumors but again do not analyze the diffusion dynamics of true vs. false rumors.

Only a few works have analyzed differences in the spread of true vs. false rumors. Friggeri et al. (2014) classified the veracity of ≈4,000 rumors from Facebook based on fact-checking assessments from snopes.com. The authors find that a majority of resharing of false rumors occurs after fact-checking. This suggests that social media users likely do not notice the fact-checks; or intentionally ignore their verdict. Closest to our work is the study from Vosoughi et al. (2018), which provides a comprehensive analysis of summary statistics of true vs. false rumors on Twitter, finding that false rumors spread significantly farther, faster, and more broadly than the truth. However, this work does not analyze the spread of true vs. false rumors in the context of COVID-19. The same dataset (Vosoughi et al., 2018) has also been used in a recent study (Pröllochs et al., 2021a) that measures emotions embedded in the *replies* to rumor cascades. The authors find that higher frequencies of certain emotions (e. g., anger) are associated with more viral cascades for false rumors.

In the context of COVID-19, research providing large-scale quantitative analyses of the spread of true and false rumors is scant. Existing works have primarily focused on summary statistics of small sets of hand-labeled rumors or source-based approaches to identify COVID-19 misinformation (e. g., Cinelli et al., 2020; Kouzy et al., 2020; Singh et al., 2020). For example, Cinelli et al. (2020) classify news sources into reliable and non-reliable sources in order to analyze the spread of COVID-19-related content. The authors find no significant differences

Table 4.1: Tags used to identify COVID-19-related fact-checks from fact-checking organizations.

Fact-Checking Organization	Tag	#Fact-Checks
politifact.org	Coronavirus	403
snopes.com	COVID-19	265
truthorfiction.com	covid-19	44

regarding the spreading dynamics. Notably, however, categorizations of reliable vs. non-reliable sources do not necessarily correspond to true vs. false rumors. In addition, source-based approaches ignore false rumors from influential individuals, emerging websites, and misclassify false rumors from websites that are generally considered as being reliable. Note that there are other recent papers reporting that COVID-19 misinformation is widespread on social media, characterizing COVID-19 misinformation, and expressing concerns about consequences for public health (e. g., Gallotti et al., 2020; Griffith et al., 2021; Islam et al., 2020; Kouzy et al., 2020; Pennycook et al., 2020). However, these works do not focus on modeling differences in the diffusion of true vs. false COVID-19 rumor cascades.

Our contributions: This work makes two key contributions. (1) We collected a unique dataset of COVID-19 rumor cascades and demonstrate that misinformation is, on average, more viral than the truth. Here, our study connects to previous works (Pröllochs et al., 2021a; Vosoughi et al., 2018), which yielded similar conclusions, yet not in the context of COVID-19. (2) The mechanisms underlying the viral spread of false rumors, though critical, have remained largely unresolved in previous research. Our work is the first to approach the question through the lenses of morality and emotions – finding that moral emotions embedded in *source tweets* shape the diffusion of false rumors on social media.

4.3 Methods

4.3.1 Data Collection

Fact-checks: We identified three fact-checking organizations that thoroughly investigate rumors related to COVID-19. The names of the fact-checking organizations are: politifact.com, truthorfiction.com, and snopes.com. These fact-checking organizations list COVID-19 rumors in separate categories or tag them with a topic label (e. g., "COVID-19", "Coronavirus") which allows us to distinguish COVID-19-related rumors from other rumors (see Table 4.1). We scraped all COVID-19-related fact-checks from these platforms.

The fact-checking organizations have different ways of labeling the veracity of a rumor. For example, politifact.com articles are given a "Pants on Fire" rating for false rumors, whereas snopes.com assigns a "false" label. Consistent with Vosoughi et al. (2018), we normalized the veracity labels across the different sites by mapping them to a score of 1 to 5. All rumors with a score of 1 or 2 were categorized as "false," whereas rumors with a score of 4 or 5 were categorized as "true." Rumors with a score of 3 were categorized as "mixed." In some cases, the same rumors have been investigated by multiple fact-checking organizations. Previous research has shown that fact-checking websites show high pairwise agreement (Vosoughi et al., 2018), ranging between 95 % and 98 %. Rumors classified as "true" or "false" even showed a perfect pairwise agreement of 100 % (Vosoughi et al., 2018). The resulting collection of fact-checks contained the following information: (i) the veracity label ("true", "false", "mixed"), (ii) links to the articles of the fact-checking organizations, and (iii) the headline of the article that is being verified.

Rumor cascades on Twitter: We followed the approach from Vosoughi et al. (2018) to identify rumor cascades on Twitter: A rumor cascade on Twitter starts with a user making an assertion about a topic such as tweeting a text message or a link to an article. Social media users then propagate the rumor by retweeting it. Oftentimes, people also reply to the original tweet. These replies sometimes contain links to fact-checking organizations that either confirm or debunk the rumor in the original tweet. We used such cascades to identify rumor cascades that are propagating on Twitter.

We employed the Twitter Historical API to map the rumors to retweet cascades on Twitter as follows. First, we collected all tweets that contain a link to any of the websites from the fact-checking organizations. Second, for each reply tweet, we extracted the original tweet and the number of retweets of the original tweet. Here, special care is needed to ensure that the replies containing a link to any of the trusted websites address the original tweet. We followed the approach from Vosoughi et al. (2018) to address this important issue: (i) we considered only replies to the original tweet and exclude replies to replies. (ii) To ensure that we study how unverified and contested information diffuses on Twitter, we removed all original tweets that are directly linking to one of the fact-checking websites. Note that tweets linking to one of the fact-checking websites do not qualify as they are no longer unverified. (iii) We compared the headline of the linked article to that of the original tweet. For this purpose, we used Universal Sentence Encoder (Cer et al., 2018) to convert the headline of the fact-check and the original tweet to vector representations that capture their semantic content. We then used cosine similarity to measure the distance between the vectors. If

Table 4.2: Summary statistics for tweets of rumor starters. Mean values are highlighted in bold, standard deviations are shown in parentheses. All Twitter variables were obtained from the Twitter Historical API.

Variable	All cascades	Politics	Health	Other
Dates collected	01/02/20 – 05/13/21	01/02/20 – 05/13/21	01/27/20 – 05/13/21	01/27/20 – 05/12/21
Number of cascades	10,610	8,157	4,116	1,297
Number of retweets	24,339,625	20,374,097	10,231,382	1,416,474
Retweet count range	0 – 260,637	0 – 260,637	0 – 207,155	0 – 76,092
Proportion *True*	35.3 %	39.0 %	34.7 %	19.7 %
Proportion *False*	46.9 %	42.7 %	48.3 %	64.8 %
Proportion *Mixed*	17.7 %	18.3 %	17.0 %	15.6 %
Followers	**2,256,095** *(7,700,566)*	**2,545,874** *(8,260,175)*	**2,526,538** *(9,166,450)*	**816,527.4** *(3,859,619)*
Followees	**9,193.9** *(34,750.39)*	**10,124.4** *(37,507.33)*	**9,320.23** *(40,249.53)*	**5,952.10** *(25,541.03)*
Account age	**3,333.35** *(1,383.38)*	**3,374.93** *(1,376.82)*	**3,321.95** *(1,391.67)*	**3,098.33** *(1,386.78)*
Verified users	55.1 %	60.2 %	56 %	31.9 %
Includes media	28.6 %	27 %	26.4 %	38.2 %
Other-condemning emotions	**0.167** *(0.217)*	**0.164** *(0.201)*	**0.153** *(0.190)*	**0.189** *(0.291)*
Self-conscious emotions	**0.300** *(0.209)*	**0.294** *(0.198)*	**0.317** *(0.196)*	**0.321** *(0.256)*

the cosine similarity was lower than 0.4, the tweet was discarded.

The retweet cascades remaining after these filtering steps then represent rumors propagating on Twitter – for which a veracity label is known based on the assessment from the fact-checking organization. In our data, the frequencies of fact-checking labels at cascade level are: 3,748 (=true), 4,979 (=false), and 1,883 (=mixed). These 10,610 rumor cascades have received more than 24.33 million retweets by Twitter users.

Following previous works (Brady et al., 2017; Vosoughi et al., 2018), we employed the Twitter API to collect a set of additional user variables for each source tweet, i. e., the number of followers, the account age, etc. These variables are known to affect the retweet rate and are later used as control variables in our regression model. Summary statistics of our dataset are reported in Table 4.2.

4.3.2 Calculation of Emotion Scores

The "other-condemning" family of moral emotions comprises the emotions *anger*, *disgust*, and *contempt*, whereas the "self-conscious" family comprises the emotions *shame*, *pride*, and *guilt* (Haidt, 2003; Tracy & Robins, 2004). We employed text mining methods to measure the extent to which these emotions are embedded in the source tweets. For this purpose, we first applied standard preprocessing steps from text mining. Specifically, the running text was converted into lower-case and tokenized, and special characters (e. g., hashtags, emoticons)

were removed. Subsequently, we applied (and validated) a dictionary-based approach analogous to earlier research (Brady et al., 2017; Pröllochs et al., 2021a; Vosoughi et al., 2018).

We measured other-condemning and self-conscious emotions embedded in the source tweets based on the NRC emotion lexicon (Mohammad & Turney, 2013). This lexicon comprises 181,820 English words that are classified according to the emotions of Plutchik's emotion model (Plutchik, 1984). Plutchik's emotion model defines 8 basic emotions and 24 emotional dyads. The emotional dyads represent complex emotions, which are derived as a combination of two basic emotions (Pröllochs et al., 2021b). We used the NRC dictionary to count the frequency of words in the tweets that belong to each of the emotions. Afterwards, we divided the word counts by the total number of dictionary words in the text, so that the vector is normalized to sum to one across the emotions (Pröllochs et al., 2021a; Vosoughi et al., 2018). In our data, 78.15 % of all source tweets contained at least one emotion word from the NRC lexicon. We filtered out tweets that do not contain any emotional words since, otherwise, the denominator is not defined (Pröllochs et al., 2021a; Vosoughi et al., 2018). However, our later analysis yields qualitatively identical results when including these observations (i. e., assigning zero values). Based on the scores for the 8 basic emotions and the 24 derived emotions, and the definitions of the two moral emotion families (Tracy & Robins, 2004), we calculated other-condemning emotions by taking the sum of *anger*, *disgust*, and *contempt*. Self-conscious emotions were calculated by taking the sum of *shame*, *pride*, and *guilt*.

User study: In order to test the construct validity of our dictionary-based approach, we employed two trained research assistants to annotate a random subset of 200 tweets that were categorized as being more other-condemning than self-conscious based on the dictionaries; and a random subset of 200 tweets that were categorized as being more self-conscious than other-condemning. For each of the 400 tweets, the annotators were asked to what extent the tweet relates to other-condemning and self-conscious emotions on two 5-point Likert scales, ranging from 1 ("not related to [other-condemning, self-conscious] emotions at all") to 5 ("very related to [other-condemning, self-conscious] emotions"). The annotators viewed the tweets in randomized order and were explained the difference between other-condemning and self-conscious emotions. The annotators exhibited a statistically significant inter-rater agreement according to Kendall's W ($p < 0.01$). Furthermore, the annotators rated the random subset of other-condemning tweets as more "other-condemning" than "self-conscious" $[t = 6.53, p < 0.001]$; and the random subset of self-conscious tweets as more "self-conscious" than "other-condemning" $[t = 4.50, p < 0.001]$.

Table 4.3: Exemplary tweets of rumor starters for each topic.

Topic	**Veracity**	**Twitter Message**
Politics	True	Trump fired the Pandemic response team in 2018... He did not replace them... #TrumpYoureKilling
Politics	False	Sick: Nancy Pelosi tried to insert abortion funding measures into the Chinese Coronavirus response stimulus package I never want to hear that Donald Trump is politicizing this pandemic again while Democrats try this stunt This is a disgrace—Speaker Pelosi should be ashamed
Health	True	More police officers have died from Covid-19 this year than have been killed on patrol. Gunfire is the second-highest cause of death.
Health	False	80 % of People Taking Maderna Vaccine Had Significant Side-Effects. While the killer Bill Gates laughs all the way to the bank. Stop this insanity now!
Other	True	This is the first day of school in Paulding County, Georgia.
Other	False	I thought this was supposed to be a conspiracy theory. But here it is, straight from Trudeau's mouth. The pandemic is the excuse for a "Great Reset" of the world, led by the UN.

4.3.3 Rumor Topics

We employed a weakly supervised machine learning framework (Yao et al., 2020) to infer the topics in the source tweets that have initiated the rumor cascades. The benefit of this state-of-the-art approach is that (i) it is regarded as superior to conventional topic modeling (i. e., Latent Dirichlet Allocation) for short texts (Yao et al., 2020), and (ii) its weakly supervised nature allows for an ex-ante selection of topics that we perceive as being particularly relevant in the context of COVID-19. We categorized the rumor cascades into three (not mutually exclusive) topics: Health (e. g., rumors about the safety of vaccines), Politics (e. g., allegations of political opponents), and Other (i. e., rumors that do not fall into one of the other categories). Example tweets for each topic are provided in Table 4.3.

Our weakly supervised machine learning framework proceeded in three steps (see Yao et al., 2020 for methodological details): (1) We started to identify topic-related tweets based on a set of manually selected keywords for each topic. For instance, for the topic Health, we searched for all tweets containing words such as "vaccine," "flu," "mask," etc. (see list of keywords in the Appendix). (2) We conducted clustering-assisted manual word sense disambiguation on the

keyword-identified tweets (Yao et al., 2020). Here we used the k-means clustering algorithm with Silhouette criterion to cluster the keyword-identified tweets for each topic. We then manually inspect random tweets sampled from each cluster and assessed whether the tweets in the cluster refer to the topic. We excluded each tweet cluster that does not show the pertinent meaning of the topic keyword. This allowed us to significantly clean and improve the quality of the keyword-identified tweets. (3) We used the created labeled data to train a deep neural network classifier and learn to predict whether or not individual Twitter messages belong to a certain topic. The input data for the training machine learning classifier was a vector representation of the (cleaned) keyword-identified tweets and the topic label. To create vector representations of tweets, we used neural language models in the form of the Universal Sentence Encoder (Cer et al., 2018). In our deep neural network classifier, we treated the task of predicting topic labels for (vector representations of) tweets as a multi-label problem considering that one tweet may belong to multiple topics (i. e., Health and Politics). In training, we used an equal number of 1 000 keyword-identified Tweets for each topic as positive training instances. In addition, we used the excluded tweets from step (2) and randomly sampled unlabeled tweets equal to the sum of labeled tweets as negative training instances, i. e., with a topic label Other.

User study: To ensure that the topic predictions are accurate, we tested for the presence of errant tweets with the help of two trained research assistants. We randomly sampled 200 tweets for each topic, and instructed the research assistants to annotate the tweets. Each annotator was asked to judge the validity of the topic label on a 5-point Likert scale, ranging from 1 ("not related to [topic] at all") to 5 ("very related to [topic]"). When comparing the human annotations to the predicted topic labels, we found very few misclassified instances. On average, the share of tweets that were not classified as at least "somewhat related to [topic]" was lower than 8.5 % (see Appendix).

4.3.4 Model Specification

We specified regression models with interaction terms that explain the number of retweets based on rumor veracity and other-condemning emotions and self-conscious emotions. Let *Retweet Count*$_i$ denote the number of retweets for rumor cascade i. Furthermore, let *Other Condemning*$_i$ denote the proportion of other-condemning emotions, *Self Conscious*$_i$ the proportion of self-conscious emotions, and *Falsehood*$_i$ the veracity. Here we define a true rumor as $Falsehood_i = 0$ and a false rumor as *Falsehood*$_i$ *= 1*. We adjusted for variables known to affect retweet rate (Brady et al., 2017; Pröllochs, 2022; Pröllochs et al., 2021a, 2021b; Stieglitz &

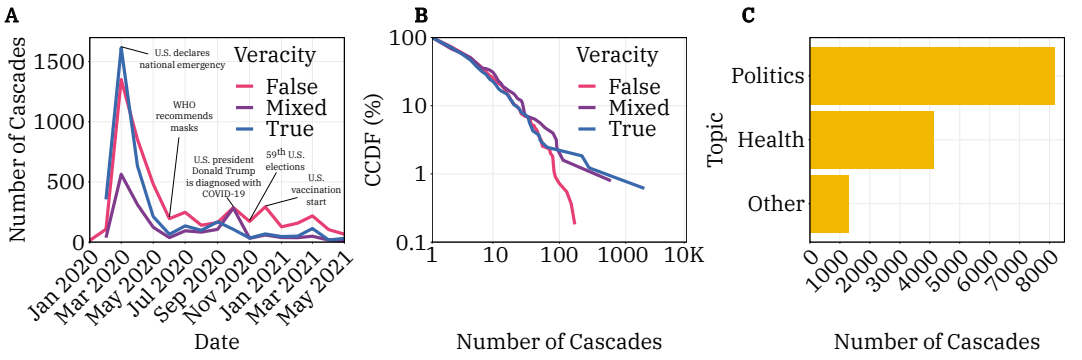


Figure 4.1: COVID-19 rumor cascades propagating on Twitter between January 2020 and the end of April 2021. (A) Monthly counts of true, false, and mixed rumor cascades. (B) Complementary cumulative distribution functions (CCDFs) of true, false, and mixed rumor cascades. (C) Number of rumor cascades across different topics.

Dang-Xuan, 2013; Vosoughi et al., 2018), which included the number of followers (*Followers*$_i$) and followees (*Followees*$_i$) of the author of the tweet, the account age (*Account Age*$_i$), whether the author was verified by Twitter (*Verified*$_i$), and whether media was attached to the tweet (*Has Media*$_i$). Each of these factors was extracted from the Twitter API. We $z-$standardized all continuous predictors in order to facilitate interpretability.

Based on the above variables, we specified the following generalized linear model for our analysis:

$$\begin{aligned}\log(\mathrm{E}(RetweetCount_i \mid {}^*)) = {} & \beta_0 + \beta_1\, \textit{Falsehood}_i \\ & + \beta_2\, \textit{Falsehood}_i \times \textit{Other Condemning}_i \\ & + \beta_3\, \textit{Falsehood}_i \times \textit{Self Conscious}_i \\ & + \beta_4 \textit{Other Condemning}_i + \beta_5\, \textit{Self Conscious}_i \\ & + \beta_6\, \textit{Followers}_i + \beta_7\, \textit{Followees}_i + \beta_8\, \textit{Account Age}_i \\ & + \beta_9\, \textit{Has Media}_i + \beta_{10}\, \textit{Verified}_i\end{aligned} \tag{4.1}$$

with intercept β_0.

RetweetCount is a non-negative count variable, and its variance is larger than the mean. To adjust for overdispersion, we drew upon a negative binomial regression (Brady et al., 2017; Stieglitz & Dang-Xuan, 2013). Note that because we estimate a negative binomial regression model with interaction terms, the coefficients cannot be interpreted as the change in the mean of the dependent variable for a one unit (i. e., standard deviation) increase in the respective predictor variable, with all other predictors remaining constant. The reason is that in nonlinear regression models with interaction terms, marginal effects are nonlinear functions of the coefficients and the levels of the explanatory variables

(Buis, 2010). Instead, the coefficients can be interpreted on a multiplicative scale by calculating the incidence rate ratio (IRR), which is equal to the exponent of the coefficient of the respective variable (Buis, 2010). Here the coefficients can be interpreted as the natural logarithm of a multiplying factor by which the predicted number of retweets changes, given a one unit increase in the predictor variable, holding all other predictor variables constant (Buis, 2010).

4.4 Results

Our data include 10 610 rumor cascades that have been retweeted 24.34 million times. The total number of COVID-19 rumor cascades peaked in March 2020 when the U. S. government declared a national emergency concerning the coronavirus disease and again in October 2020, the month prior to the U. S. presidential elections (Figure 4.1 (A)). The three fact-checking organizations have categorized 46.9 % of all rumors as false, 35.3 % as true, and 17.7 % as being of mixed veracity. While the absolute number of rumor cascades has decreased over the course of the pandemic, the relative share of false vs. true rumors has increased (Figure 4.1 (A)). Compared to false rumors, a greater fraction of true rumors experienced more than 100 rumor cascades (Figure 4.1 (B)). COVID-19 rumors are not constrained exclusively to health topics (e. g., rumors about the safety of vaccines). Rather, a sizable number of COVID-19 rumors concern political topics (e. g., true or false allegations of political opponents) (Cossard et al., 2020). We thus applied topic modeling to categorize the rumor cascades in our dataset into three (not mutually exclusive) topics: Politics, Health, and Other. Figure 4.1 (C) shows that a large proportion of COVID-19 rumors were thematically related to Politics (76.9 %), Health (38.8 %), while only 12.2 % concerned Other topics (e. g., conspiracy theories). A total share of 34.1 % of rumor cascades were thematically related to both Politics and Health.

Regression analysis: We fitted explanatory regression models to evaluate how variations in moral emotions are associated with differences in the number of retweets for true vs. false rumor cascades. In our regression analysis, we followed previous works (Brady et al., 2017; Vosoughi et al., 2018) by controlling for variables known to affect the retweet rate independent of the main predictors, i. e., the number of followers, the account age, etc.

As a baseline, we started our regression analysis with a negative binomial regression explaining the number of retweets solely based on the veracity label and control variables (see SI, Table 4.6). Here false rumors (Falsehood $= 1$) were estimated to receive 15.66 % more retweets than true rumors (IRR 1.16;

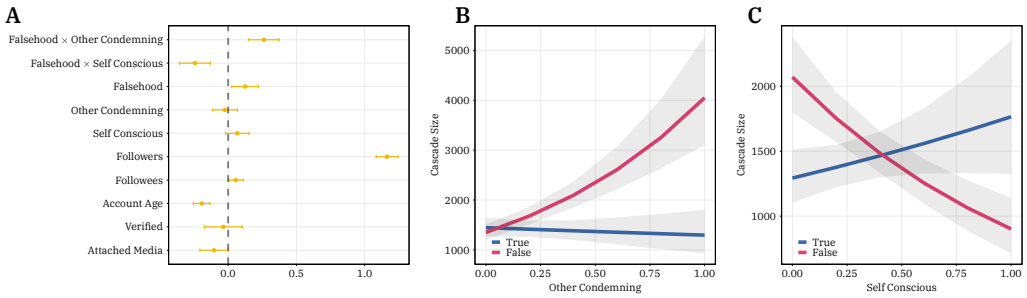


Figure 4.2: Increases in other-condemning emotions predict higher retweet counts for false rumors, whereas increases in self-conscious emotions predict less retweets. (A) Coefficient estimates for negative binomial regression with 95 % confidence intervals. The dependent variable is the number of retweets. (B–C) Predicted marginal means of the number of retweets for other-condemning emotions and self-conscious emotions. The 95 % confidence intervals are highlighted in gray.

$p < 0.01$). Subsequently, we extended the negative binomial regression by including interaction terms between rumor veracity and other-condemning emotions, and between rumor veracity and self-conscious emotions (Figure 4.2 (A)). The coefficient estimates for these two interaction terms were statistically significant, which implies that false rumors' virality depended on the moral emotions embedded in the source tweet. Specifically, a one standard deviation increase in other-condemning emotions for false rumors was linked to a 26.99 % increase in the number of retweets (IRR 1.27; $p < 0.01$). In contrast, a one standard deviation increase in self-conscious emotions for false rumors was linked to a 23.43 % decrease in the number of retweets (IRR 1.23; $p < 0.01$). We found no statistically significant effect of other-condemning and self-conscious emotion words for true rumors. In sum, we observed that false rumors were more viral than the truth if the source tweet embedded a high proportion of other-condemning emotion words, whereas a high proportion of self-conscious emotion words was linked to a less viral spread (see Figure 4.2 (B), Figure 4.2 (C)).

Analysis across topics: We also examined the effect of moral emotions across different topics. For each topic from Figure 4.1 C, we generated observation subsets and re-estimated our regression model (Figure 4.3). We observed differences in the effects of moral emotions on the number of retweets. The effect of other-condemning emotions on the number of retweets was pronounced both for false rumors from the Health category (IRR 1.34; $p < 0.01$) and for false rumors from the Politics category (IRR 1.61; $p < 0.01$). For the Other category (with comparatively smaller sample size), the coefficients pointed in the same directions but were not statistically significant at common significance thresholds.

Analysis of deleted tweets: Major social media platforms such as Twitter have intensified their efforts to combat the spread of misinformation on their platforms by deleting misinformation (BBC, 2017). While the Twitter API does

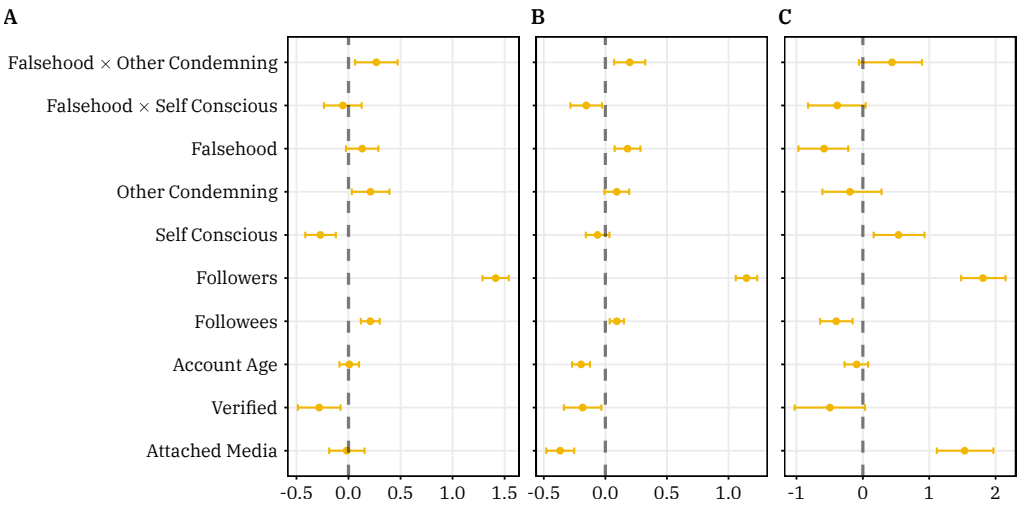


Figure 4.3: Coefficient estimates for negative binomial regressions with 95 % confidence intervals for rumor cascades filtered by topic (*A:* Health, *B:* Politics, and *C:* Other). The dependent variable is the number of retweets.

not provide access to the content of source tweets that have been deleted, we were still able to analyze some of their characteristics. As part of an exploratory analysis, we found that 3 663 potential rumor starter tweets have been deleted (either by Twitter or by the users themselves). An overwhelming majority of those (68.35 %) are potentially false rumors. Hence, even though these numbers suggest that a relevant proportion of false rumors on Twitter has been deleted, the vast majority of false rumors continue to circulate. For those rumors, our results demonstrate that falsehood can be more viral than the truth.

Additional checks[1]**:** Numerous exploratory analyses and checks validated our results and confirmed their robustness: (i) Since self-conscious emotions can be regarded as the counterpart of other-condemning emotions, we tested an alternative model specification in which we included the *difference* between the emotion scores for other-condemning and self-conscious emotions instead of two individual variables. Consistent with our main analysis, we found that false rumors are more viral than the truth if the source tweet embeds a high proportion of other-condemning emotion words, whereas a high proportion of self-conscious emotion words is linked to a less viral spread. (ii) In our main analysis, we focused on rumors that are clearly true or false. However, 17.7 %, of all rumors have been categorized as being of mixed veracity by the fact-checking organizations. We tested whether counting rumors of mixed veracity as either true or false affects the validity of our results. We find that our results are robust and that the combination of other-condemning emotion and mixed veracity

[1] Detailed results are reported in the Appendix.

is similarly viral as the combination of other-condemning emotions and false veracity. (iii) In our data, 9.48 % of rumor starters have started more than one retweet cascade. To ensure that our models are not biased due to this source of non-independence, we dropped all users with clustering and reestimated the models. The results are robust and support our findings. We also repeated our analysis with monthly fixed effects to control for differences in the virality of rumor cascades due to different start dates. Also here, the results confirmed the findings from our main analysis. (iv) We repeated our analysis for subsets of rumor cascades that have been started by users that are either verified or not verified by Twitter. We find that our main findings hold for both user groups.

4.5 Discussion

Here we provide evidence that moral emotions play a crucial role in the spread of COVID-19 misinformation on social media. Using a comprehensive dataset of COVID-19 rumors that have been fact-checked by three independent fact-checking organizations (snopes.com, politifact.com, truthorfiction.com), we establish that other-condemning emotions – also known as the hostility triad – are linked to a more viral spread of false rumors.

While false rumors pose a threat to the successful overcoming of this pandemic, an understanding of how rumors diffuse in online social networks is – even for non-crisis situations – still in its infancy. Analyzing the spreading dynamics of fact-checked rumors is to a great extent generalizable to the spread of other (non-fact-checked) rumors on social media (Vosoughi et al., 2018). Our finding that COVID-19 misinformation is, on average, more viral than the truth directly connects to the study from Vosoughi et al. (Vosoughi et al., 2018), which yielded similar findings, yet outside the context of COVID-19. Previous research has also shown that misinformation on social media can have negative offline consequences. Among other instances, this has previously been confirmed to be the case during humanitarian crises (Starbird et al., 2014) and elections (Allcott & Gentzkow, 2017; Aral & Eckles, 2019; Bakshy et al., 2015; Grinberg et al., 2019). Our observation that COVID-19 misinformation is both widespread and viral on social media is at least equally concerning. COVID-19 misinformation not only poses severe health risks to individuals but also undermines the integrity of the political discourse (Rapp & Salovich, 2018).

The results of this study highlight the role of moral emotions in rumor diffusion. Previous research (Brady et al., 2017) broadly distinguished moral vs. non-moral emotional expressions in social media content, while this work

demonstrates that the two clusters of moral emotions (self-conscious vs. other-condemning emotions) have distinct effects on social transmission in the context of true and false rumors. We observe that false rumors receive more retweets than true rumors if the source tweets embed a high share of other-condemning emotions, whereas we find the opposite pattern, yet of smaller magnitude, for self-conscious emotions. Another relevant finding is that the expression of other-condemning emotion on virality is pronounced both for health misinformation and political misinformation. These findings may be partially explained by the high level of polarization of social media users in the context of COVID-19. In polarizing debates, radical ideas and beliefs are strengthened and more likely to translate into action. It thus seems plausible that the explosive mix of other-condemning emotions accelerates the spread of false rumors about those topics within social networks.

From a practical perspective, policy initiatives around the world urge social media platforms to limit the spread of false rumors (Lazer et al., 2018). While previous research has studied emotions in replies to rumor cascades (Pröllochs et al., 2021a, 2021b), our work highlights the importance of considering (moral) emotions in the source tweets that have initiated the rumor cascades. These findings could eventually be leveraged in machine learning models in order to detect false rumors more accurately. Emotion scores for source tweets are available immediately upon the beginning of the diffusion process – a time point at which features from propagation dynamics are scarce (Conti et al., 2017). Our findings may also be relevant with regard to other downstream tasks such as educational applications. Altogether, considering moral emotions in social media posts might help future works to develop more effective strategies against false rumors.

This work is subject to the typical limitations of observational studies. We report associations and refrain from making causal claims. Future work should seek to corroborate our conclusions in controlled laboratory experiments and, in particular, test the causal influence of exposure to moral-emotional language on attitudes and behavior. Our inferences are also limited by the accuracy and availability of our data, specifically those from the three different fact-checking websites. For those, however, our data comprises all COVID-19 rumor cascades on Twitter until the end of April 2021. Despite these limitations, we believe that observing and understanding how misinformation spreads is the first step toward containing it. We hope that our work inspires more research into the causes, consequences, and potential countermeasures for the spread of misinformation – both in crisis and non-crisis situations.

4.6 Conclusion

While false rumors pose a threat to the successful overcoming of this pandemic, an understanding of "what makes false rumors viral" is – even for non-crisis situations – still in its infancy. In this work, we approach this question through the lenses of morality and emotions and their role in rumor diffusion in polarized social media environments. For this purpose, we collected a unique dataset of COVID-19-related rumor cascades from Twitter and empirically analyze their spreading dynamics. We find that COVID-19 misinformation is, on average, more viral than the truth. However, the veracity effect is moderated by moral emotions: false rumors are more viral than the truth if the source tweets embed a high number of other-condemning emotion words, whereas a higher number of self-conscious emotion words is linked to a less viral spread. These findings offer insights into how true vs. false rumors spread and highlight the importance of considering moral emotions in social media content.

Bibliography

Allcott, H., Boxell, L., Conway, J., Gentzkow, M., Thaler, M., & Yang, D. (2020). Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. *Journal of Public Economics*, *191*, 104254.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236.

Aral, S., & Eckles, D. (2019). Protecting elections from social media manipulation. *Science*, *365*(6456), 858–861.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531–1542.

BBC. (2017). *Coronavirus: World leaders' posts deleted over fake news*. https://www.bbc.com/news/technology-52106321

Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015). Science vs conspiracy: Collective narratives in the age of misinformation. *PLOS ONE*, *10*(2), e0118093.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *PNAS*, *114*(28), 7313–7318.

Buis, M. L. (2010). Stata tip 87: Interpretation of interactions in nonlinear models. *The Stata Journal: Promoting communications on statistics and Stata*, *10*(2), 305–308.

Cacioppo, J. T., Fowler, J. H., & Christakis, N. A. (2009). Alone in the crowd: The structure and spread of loneliness in a large social network. *Journal of Personality and Social Psychology*, *97*(6), 977.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on Twitter. *Proceedings of the WWW*.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., & Kurzweil, R. (2018). Universal sentence encoder. *arXiv*, *1803.11175*.

Choi, D., Chun, S., Oh, H., Han, J., & Kwon, T. " (2020). Rumor propagation is amplified by echo chambers in social media. *Scientific Reports*, *10*(1), 310.

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *PNAS*, *118*(9).

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), 1–10.

Conti, M., Lain, D., Lazzeretti, R., Lovisotto, G., & Quattrociocchi, W. (2017). It's always April fools' day! On the difficulty of social network misinformation classification via propagation features. *Proceedings of the WIFS*.

Cossard, A., De Francisci Morales, G., Kalimeri, K., Mejova, Y., Paolotti, D., & Starnini, M. (2020). Falling into the echo chamber: The Italian vaccination debate on Twitter. *Proceedings of the ICWSM*.

De Domenico, M., Lima, A., Mougel, P., & Musolesi, M. (2013). The anatomy of a scientific rumor. *Scientific Reports*, *3*(1).

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *PNAS*, *113*(3), 554–559.

Druckman, J. N., Klar, S., Krupnikov, Y., Levendusky, M., & Ryan, J. B. (2020). How affective polarization shapes Americans' political beliefs: A study of response to the COVID-19 pandemic. *Journal of Experimental Political Science*, *8*(3), 223–234.

Ecker, U. K. H., Lewandowsky, S., & Tang, D. T. W. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition*, *38*(8), 1087–1100.

Fowler, J. H., & Christakis, N. A. (2008). Estimating peer effects on health in social networks: A response to Cohen-Cole and Fletcher; Trogdon, Nonnemaker, Pais. *Journal of Health Economics*, *27*(5), 1400–1405.

Frenkel, S., Alba, D., & Zhong, R. (2020). Surge of virus misinformation stumps Facebook and Twitter. *The New York Times*, *8*.

Friggeri, A., Adamic, L., Eckles, D., & Cheng, J. (2014). Rumor cascades. *Proceedings of the ICWSM*.

Gallotti, R., Valle, F., Castaldo, N., Sacco, P., & De Domenico, M. (2020). Assessing the risks of 'infodemics' in response to COVID-19 epidemics. *Nature Human Behaviour*, *4*(12), 1285–1293.

Geeng, C., Yee, S., & Roesner, F. (2020). Fake news on Facebook and Twitter: Investigating how people (don't) investigate. *Proceedings of the CHI*.

Griffith, J., Marani, H., & Monkman, H. (2021). COVID-19 vaccine hesitancy in Canada: Content analysis of tweets using the theoretical domains framework. *Journal of Medical Internet Research*, *23*(4), e26874.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, *363*(6425), 374–378.

Haidt, J. (2003). The moral emotions. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 852–870, Vol. 11). Oxford University Press.

Hart, P. S., Chinn, S., & Soroka, S. (2020). Politicization and polarization in COVID-19 news coverage. *Science Communication*, *42*(5), 679–697.

Havey, N. F. (n.d.). Partisan public health: How does political ideology influence support for COVID-19 related misinformation? *Journal of Computational Social Science*, *3*(2), 319–342.

Islam, M. S., Sarkar, T., Khan, S. H., Kamal, A.-H. M., Hasan, S. M. M., Kabir, A., Yeasmin, D., Islam, M. A., Chowdhury, K. I. A., Anwar, K. S., et al. (2020). COVID-19-related infodemic and its impact on public health: A global social media analysis. *The American Journal of Tropical Medicine and Hygiene*, *103*(4), 1621.

Jing, E., & Ahn, Y.-Y. (2021). Characterizing partisan political narrative frameworks about COVID-19 on Twitter. *EPJ Data Science*, *10*(1), 53.

Kim, A. (2017). Says who? The effects of presentation format and source rating on fake news in social media. *MIS Quarterly*, *43*(3), 1025–1039.

Kouzy, R., Abi Jaoude, J., Kraitem, A., El Alam, M. B., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E., & Baddour, K. (2020). Coronavirus goes viral: Quantifying the COVID-19 misinformation epidemic on Twitter. *Cureus*, *12*(3).

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M.,

Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096.

Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, *111*(3), 831–870.

Limaye, R. J., Sauer, M., Ali, J., Bernstein, J., Wahl, B., Barnhill, A., & Labrique, A. (2020). Building trust while influencing online COVID-19 content in the social media world. *The Lancet Digital Health*, *2*(6), e277–e278.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, *29*(3), 436–465.

Moravec, P., Minas, R., & Dennis, A. R. (2019). Fake news on social media: People believe what they want to believe when it makes no sense at all. *MIS Quarterly*, *43*(4), 1343–1360.

Oh, O., Agrawal, M., & Rao, H. R. (2013). Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. *MIS Quarterly*, *37*(2), 407–426.

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, *147*(12), 1865–1880.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590–595.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, *31*(7), 770–780.

Plutchik, R. (1984). *Emotion: Theory, research, and experience: Theory, research, and experience* (2nd ed., Vol. 1). Academic Press.

Pröllochs, N. (2022). Community-based fact-checking on Twitter's Birdwatch platform. *Proceedings of the ICWSM*.

Pröllochs, N., Bär, D., & Feuerriegel, S. (2021a). Emotions explain differences in the diffusion of true vs. false social media rumors. *Scientific Reports*, *11*(22721).

Pröllochs, N., Bär, D., & Feuerriegel, S. (2021b). Emotions in online rumor diffusion. *EPJ Data Science*, *10*(1), 51.

Rapp, D. N., & Salovich, N. A. (2018). Can't we just disregard fake news? The consequences of exposure to inaccurate information. *Policy Insights from the Behavioral and Brain Sciences*, *5*(2), 232–239.

Ricard, J., & Medeiros, J. (2020). Using misinformation as a political weapon: COVID-19 and Bolsonaro in Brazil. *Harvard Kennedy School Misinformation Review*, *1*(3).

Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology*, *76*(4), 574–586.

Shao, C., Ciampaglia, G. L., Flammini, A., & Menczer, F. (2016). Hoaxy: A platform for tracking online misinformation. *Proceedings of the WWW Companion*.

Singh, L., Bansal, S., Bode, L., Budak, C., Chi, G., Kawintiranon, K., Padden, C., Vanarsdall, R., Vraga, E., & Wang, Y. (2020). A first look at COVID-19 information and misinformation sharing on Twitter. *arXiv*, *2003.13907*.

Solovev, K., & Pröllochs, N. (2022). Hate speech in the political discourse on social media: Disparities across parties, gender, and ethnicity. *Proceedings of the WWW*.

Starbird, K. (2017). Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter. *Proceedings of the ICWSM*.

Starbird, K., Maddock, J., Orand, M., Achterman, P., & Mason, R. R. T. M. (2014). Rumors, false flags, and digital vigilantes: Misinformation on Twitter after the 2013 Boston marathon bombing. *Proceedings of the iConference*.

Stieglitz, S., & Dang-Xuan, L. (2013). Emotions and information diffusion in social media: Sentiment of microblogs and sharing behavior. *Journal of Management Information Systems*, *29*(4), 217–248.

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology*, *58*(1), 345–372.

Tracy, J. L., & Robins, R. W. (2004). Putting the self into self-conscious emotions: A theoretical model. *Psychological Inquiry*, *15*(2), 103–125.

Tucker, J., Guess, A., Barbera, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., & Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. *SSRN Electronic Journal*, *3144139*.

van Stekelenburg, J. (2017). Radicalization and violent emotions. *PS: Political Science & Politics*, *50*(04), 936–939.

van Stekelenburg, J., & Klandermans, B. (2017). Individuals in movements: A social psychology of contention. In B. Klandermans & C. Roggeband (Eds.), *Handbook of social movements across disciplines* (pp. 103–139). Springer.

Vicario, M. D., Quattrociocchi, W., Scala, A., & Zollo, F. (2019). Polarization and fake news: Early warning of potential misinformation targets. *ACM Transactions on the Web*, *13*(2), 1–22.

Vo, N., & Lee, K. (2018). The rise of guardians: Fact-checking url recommendation to combat fake news. *Proceedings of the SIGIR*.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Waszak, P. M., Kasprzycka-Waszak, W., & Kubanek, A. (2018). The spread of medical fake news in social media – The pilot quantitative study. *Health Policy and Technology*, *7*(2), 115–118.

Weng, L., Menczer, F., & Ahn, Y.-Y. (2013). Virality prediction and community structure in social networks. *Scientific Reports*, *3*(1).

Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science*, *16*(10), 780–784.

Wu, J., Huang, L., & Zhao, J. L. (2019). Operationalizing regulatory focus in the digital age: Evidence from an e-commerce context. *MIS Quarterly*, *43*(3), 745–764.

Yao, W., Zhang, C., Saravanan, S., Huang, R., & Mostafavi, A. (2020). Weakly-supervised fine-grained event recognition on social media texts for disaster management. *Proceedings of the AAAI*.

Zarocostas, J. (2020). How to fight an infodemic. *The Lancet*, *395*(10225), 676.

Appendix 4.A Topic modeling

Table 4.4 shows the manually selected seed words that were used to identify topic-related tweets in weakly supervised learning.

Table 4.4 reports the results for our user study testing for the presence of errant tweets. On average, the share of tweets that were not classified as at least "somewhat related to [topic]" was lower than 8.5 %.

Table 4.4: Seed words used to identify topic-related tweets in weakly supervised learning. Various word forms of the keywords are also considered, e. g., "masks" and "masking" are also considered for the keyword "mask".

Topic	**Seed keywords**
Politics	Bill, Trump, Biden, Obama, Democrats, GOP, Republicans, Tax, Administration, Red, Blue, Pelosi, Economy, Chinavirus
Health	Vaccine, Flu, Mask, Fever, Ebola, SARS, Ibuprofen, Garlic, Health, Infection

Table 4.5: Frequency of errors in topic labeling.

Topic	**Percent Error**
Politics	6.0 %
Health	2.2 %
Other	17.5 %
Mean	8.5 %

Appendix 4.B Analysis of control variables

We tested a model specification in which we only incorporated control variables from previous works. Table 4.6 shows that rumors receive a particularly high number of retweets if they are false and if they have been started by users with a larger number of followers.

Table 4.6: Regression results for control variables only. The dependent variable is the number of retweets.

Dependent Variable: *RetweetCount*	
Falsehood	0.145**
	(0.050)
Followers	1.134***
	(0.036)
Followees	0.046
	(0.025)
AccountAge	−0.217***
	(0.026)
HasMedia	−0.132*
	(0.052)
Verified	0.007
	(0.073)
Intercept	7.250***
	(0.058)
Observations (rumor cascades)	8 727

Sign. levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; standard errors in parentheses

Appendix 4.C Verified vs. unverified users

We repeated our analysis for subsets of rumor cascades that have been started by users that are verified or not-verified by Twitter. Table 4.7 shows that our main findings hold for both user groups.

Table 4.7: Regression results for rumor cascades initiated from Verified (column 1) or Non-verified (column 2) users only.

Dependent Variable: *RetweetCount*		
	Subset: Verified	**Subset: Non-verified**
Falsehood × OtherCondemning	0.238***	0.302**
	(0.058)	(0.100)
Falsehood × SelfConscious	−0.256***	−0.262*
	(0.055)	(0.111)
Falsehood	0.182***	0.062
	(0.049)	(0.104)
OtherCondemning	−0.016	−0.025
	(0.042)	(0.085)
SelfConscious	0.100**	0.073
	(0.037)	(0.096)
Followers	0.871***	2.138***
	(0.043)	(0.075)
Followees	0.101***	−0.512***
	(0.027)	(0.055)
AccountAge	−0.090*	−0.265***
	(0.036)	(0.041)
HasMedia	−0.344***	0.142
	(0.053)	(0.104)
Intercept	7.412***	7.969***
	(0.048)	(0.106)
Observations (rumor cascades)	4 836	3 891

Sign. levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; standard errors in parentheses

Appendix 4.D Rumors with mixed veracity

Table 4.8: Regression results with mixed rumors categorized as false rumors (Model 1) and mixed rumors categorized as true rumors (Model 2).

Dependent Variable: *RetweetCount*		
	Model (1)	**Model (2)**
Falsehood × OtherCondemning	0.298***	0.165***
	(0.046)	(0.050)
Falsehood × SelfConscious	−0.341***	−0.100*
	(0.046)	(0.049)
Falsehood	0.085	0.156***
	(0.044)	(0.046)
OtherCondemning	−0.054	−0.026
	(0.033)	(0.042)
SelfConscious	0.163***	0.076
	(0.033)	(0.040)
Followers	1.175***	1.173***
	(0.033)	(0.033)
Followees	0.039	0.031
	(0.022)	(0.022)
AccountAge	−0.125***	−0.132***
	(0.024)	(0.024)
HasMedia	−0.122*	−0.131**
	(0.048)	(0.048)
Verified	−0.085	−0.077
	(0.066)	(0.066)
Intercept	7.344***	7.278***
	(0.049)	(0.054)
Observations (rumor cascades)	10 610	10 610

Sign. levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; standard errors in parentheses

In our main analysis, we focused on rumors that are clearly true or false. However, 17.7 %, of all rumors have been categorized as being of mixed veracity by the fact-checking organizations. We tested whether counting rumors of mixed veracity as either true or false affects the validity of our results. Table 4.8 shows that our results are robust and that the combination of other-condemning emotion and mixed veracity is similarly viral as the combination of other-condemning emotions and false veracity.

Appendix 4.E Sensitivity to non-independence

In our data, 9.48 % of rumor starters have started more than one retweet cascade. To ensure that our models are not biased due to this source of non-independence, we dropped all users with clustering and reestimated the models. Table 4.9 show that the results are robust and support our findings.

We also repeated our analysis with monthly fixed effects to control for differences in the virality of rumor cascades due to different start dates (Table 4.9). All results confirm the findings from our main analysis.

Table 4.9: Regression results without rumor cascades from users that have started more than one retweet cascade (Model 1) and with monthly fixed effects (Model 2).

Dependent Variable: *RetweetCount*		
	Model (1)	**Model (2)**
Falsehood × OtherCondemning	0.428***	0.225***
	(0.100)	(0.052)
Falsehood × SelfConscious	−0.386***	−0.245***
	(0.104)	(0.052)
Falsehood	0.118	0.226***
	(0.101)	(0.051)
OtherCondemning	−0.101	0.005
	(0.085)	(0.041)
SelfConscious	0.128	0.109**
	(0.089)	(0.040)
Followers	1.937***	1.289***
	(0.081)	(0.036)
Followees	−0.235***	0.061*
	(0.058)	(0.024)
AccountAge	−0.246***	−0.273***
	(0.043)	(0.026)
HasMedia	0.254*	−0.162**
	(0.099)	(0.052)
Verified	−0.326*	0.034
	(0.137)	(0.072)
Intercept	7.857***	7.445***
	(0.122)	(0.586)
Monthly fixed effects	✗	✓
Observations (rumor cascades)	4 139	8 727

Sign. levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; standard errors in parentheses

Appendix 4.F Alternative emotion measure

Since self-conscious emotions can be regarded as the counterpart of other-condemning emotions, we tested an alternative model specification in which we included the *difference* between the emotion scores for other-condemning and self-conscious emotions instead of two individual variables (Table 4.10). Consistent with our main analysis, we found that false rumors are more viral than the truth if the source tweet embeds a high proportion of other-condemning emotion words, whereas a high proportion of self-conscious emotion words is linked to a less viral spread.

Table 4.10: Retweet count as a function of the difference of other-condemning and self-conscious emotions (OtherCondemning–SelfConscious).

Dependent Variable: *RetweetCount*	
Falsehood ×	0.289***
OtherCondemning–SelfConscious	(0.048)
Falsehood	0.127*
	(0.050)
OtherCondemning–SelfConscious	−0.051
	(0.038)
Followers	1.161***
	(0.036)
Followees	0.057*
	(0.025)
AccountAge	−0.196***
	(0.026)
HasMedia	−0.107*
	(0.052)
Verified	−0.035
	(0.073)
Intercept	7.260***
	(0.058)
Observations (rumor cascades)	8 727

Sign. levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; standard errors in parentheses

Chapter 5

Integrating Floor Plans into Hedonic Models for Rent Price Appraisal

5.1 Introduction

Online real estate platforms have become significant marketplaces for the real estate industry. These online platforms provide an overview of available properties in a specific location and facilitate users' search for an apartment or a house (Yuan et al., 2013). In 2011, one of the biggest online real estate providers, Zillow, reported 24 million unique visitors. Today, this number has risen to 196 million and stays on the upwards trend, indicating continual interest from prospective home buyers and tenants (Zumpano et al., 2003). While the number of people referring to online real-estate platforms continues to grow, there is a rising need for tools and recommendation algorithms that allow for an accurate appraisal of real estate prices (e. g., Yuan et al., 2013).

Prior works have primarily studied real estate valuation based on *hedonic price models* (Monson, 2009; Sopranzetti, 2010; Wallace & Meese, 1997). Hedonic pricing theory suggests that the valuation of housing values or rents can be viewed as a weighted sum of its features (Wallace & Meese, 1997). Specifically, the price and rent are determined by the attributes and characteristics of the dwelling unit and the surrounding neighborhood. Hedonic models provide a high degree of flexibility when selecting the attributes, which can roughly be divided into structural and locational attributes (Natividade-Jesus et al., 2007). The structural factors describe an apartment itself (e. g., number of rooms, amenities, parking, pool), while the locational attributes are composed of external features affecting the price. For instance, previous studies (e. g., Peterson & Flanagan, 2009; Sirmans et al., 2005) have examined the impact of the distance to transport hubs and shopping centers on real estate prices. Altogether, the hedonic approach

provides scholars and practitioners with a framework of assessing the value of real estate properties, where the final price entails a precise valuation of a diverse feature package.

While structural and locational attributes are of great importance for real estate price models, modern online real estate marketplaces provide additional information that has been neglected in previous works. A particularly relevant feature for housebuyers and prospective tenants may be real estate *floor plans*. According to rightmove.co.uk, 90 % of home-buyers think that a floor plan is an essential part of the decision making process when finding an apartment or house. A real estate floor plan is a schematic illustration that provides a top-down view of the property. While information about floor plans may be partially provided in structured form (e. g., in floor plan filings), the actual *floor plan images* may contain price-relevant hidden information. For example, the relationship between rooms and spaces, as well as the overall layout are typically only available from the floor plan image itself. Information on floor plans is typically also not available in textual form as the vast majority of platforms either do not include an alternative text description of the floor plan or the text is very general and not helpful (Goncu et al., 2015).

Altogether, current hedonic models for rent price appraisal fail at incorporating the complex data in floor plans, as it is not typically a part of the structured information. Thus, in this study, we investigate to what extent an automated visual analysis of *floor plan images* on online real estate marketplaces can help to enhance real estate appraisal.
Research Question: *Can apartment floor plans on online real estate platforms enhance hedonic rent price appraisal?*

Methodology: We propose a tailored two-staged deep learning approach to determine the hedonic price of apartment floor plans on online real estate marketplaces. Our method does not require any kind of manual labeling by human raters, as it learns price-relevant designs of floor plans based solely on price data. In the first step, we compute adjusted rent prices by regressing the monthly rent price on the structural and locational variables. This allows us to extract the variance in rent that is unexplained by these variables and prevents our model from learning information that is already available from known features. Second, we use convolutional neural networks (CNN) and the adjusted rent prices to predict the sentiment of the floor plans, i. e., the apartment rent price after controlling for locational and structural characteristics of an apartment. We then evaluate the benefits of integrating floor plans into hedonic rent price models as follows: (i) we perform hedonic regression analysis to evaluate the *explanatory power* of floor plans, and (ii) we evaluate the out-of-sample

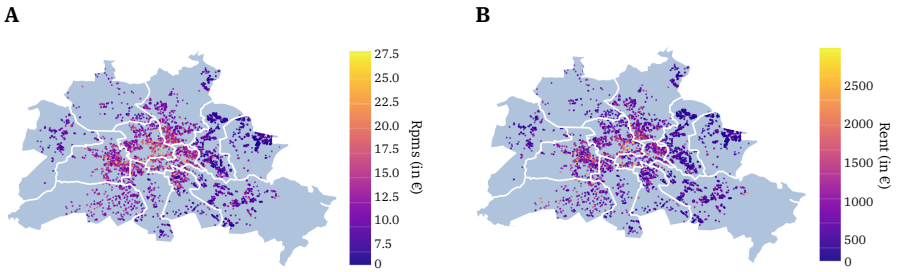


Figure 5.1: Heatmap of monthly apartment costs across districts in Berlin.

prediction performance in rent price prediction.

Main Findings: We collected a unique dataset of 9,174 real estate listings from a leading online real estate marketplace in Germany. For each apartment, we collected a raw image showing the floor plan of the apartment, the monthly rent price, the monthly rent price per m^2, and 36 fields with structural and locational attributes about the apartment. Based on a thorough hedonic analysis of rent prices, we yield the following main findings:

- The visual design of floor plans has significant explanatory power in hedonic regression models for rent price appraisal – even after controlling for structural and locational apartment characteristics.
- Harnessing floor plans results in an up to 10.56 % lower out-of-sample prediction error of rent prices.
- The predictive power of floor plans varies by the nature of the floor plans. Floor plans yield a particularly high gain in prediction performance for older and smaller apartments.

Altogether, our empirical findings contribute to the existing research body by quantifying the hedonic value of floor plans and establishing the link between visual designs of floor plans and rent prices. We show that there is an underutilization of the available data in current hedonic models. To the best of our knowledge, our paper is the first study that demonstrates how harnessing floor plans can enhance real estate appraisal on online real estate platforms.

Implications: Our findings have direct implications for online real estate platforms, providing avenues to improve user experience in their real estate listings. Our results also allow practitioners in the real estate industry to enhance the accuracy of their price estimates without the need for costly and subjective human annotations. The latter equips real estate investors with the clear benefit of improved risk assessment, allowing them to enhance their portfolios and reduce the possibility of misvaluation.

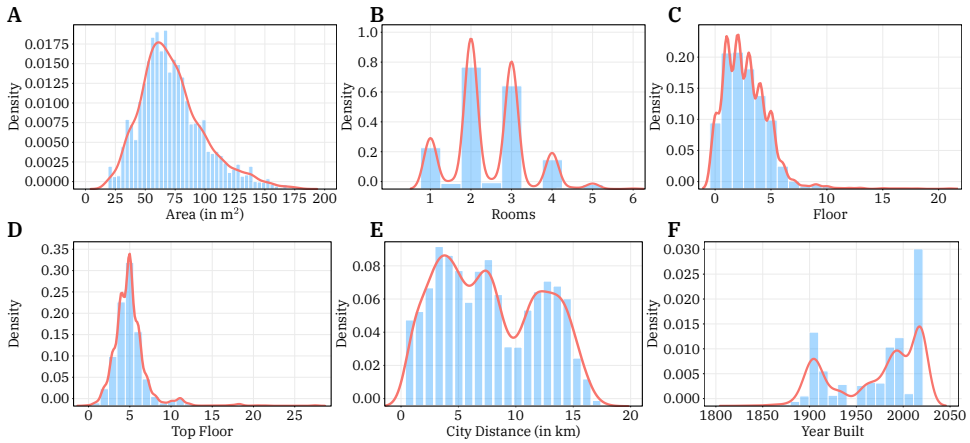


Figure 5.2: Distribution of explanatory variables.

5.2 Related Work

5.2.1 Hedonic Appraisal of Real Estate Prices

Traditional real estate valuation research is based on financial estate theory and constructs measures of fundamental values and price indices (Krainer & Wei, 2004) using a wide variety of methods, including comparison (Krainer & Wei, 2004), assessed value (Clapp & Giaccotto, 1992), and time series analysis (Chiang et al., 2005). Yet these approaches are severely limited by their scope; if a parameter has no clear market value, it is usually ignored. For example, proximity to the transport hubs of the city may be important for the potential renter but is not taken into account by classical models.

Hedonic models deviate from standard assumptions and take into account parameters that are not easily quantifiable in terms of the market price (Pagourtzi et al., 2003; Sirmans et al., 2005). In general, hedonic analysis is concerned with the marginal changes in the target variable (e. g., rent price) by the variable of interest. While still taking the basic structural parameters into account, hedonic analysis is not constrained by a comparison between similar properties and can be extended to incorporate almost any kind of additional information. They place greater weight on the heterogeneity of the market and try to estimate the effects of individual characteristics of the property (Malpezzi, 2002). Hedonic real estate models typically show relatively high explanatory power. For example, in the German market, the hedonic model developed by (Kolbe & Wüstemann, 2014) shows an R^2 of above 0.7. While hedonic models are frequently implemented via linear regression, previous research has demonstrated that the hedonic approach can also be used in combination with neural networks and other machine learning methods e. g., Peterson and Flanagan, 2009.

Hedonic models typically account for structural and locational characteristics of a property (e. g., Natividade-Jesus et al., 2007). They may vary in accordance with the available information and research questions, which is both a benefit and a possible deficiency of the method. The authors in (Sirmans et al., 2005) provide an overview of the parameters used in the previous works and their effects. Popular structural categories include the size and the number of floors of a property, which mostly have a positive effect on the price. Locational characteristics tend to contain distance measures that can have an effect in either direction. For example, school districts mostly have a negative effect (Sirmans et al., 2005). More unconventional specifications of hedonic models may further include the forced nature of the sale (Andersen & Nielsen, 2017), non-euclidean distance metrics (Lu et al., 2014), and news sentiment (D. Sun et al., 2014).

5.2.2 Image Analysis in Real Estate

In the area of real estate, computer vision and visual analytics have found varying degrees of success as tools for visualization, sentiment analysis and feature extraction. In particular, machine learning can be used to visualize real estate data for easier understanding (Li et al., 2018; G. Sun et al., 2013). Furthermore, previous works have used machine learning to extract information and sentiment contained within images to better explain or to predict the property price (Glaeser et al., 2018; H. Ahmed & Moustafa, 2016; Naumzik & Feuerriegel, 2020; You et al., 2017). For instance, photos of surrounding areas, such as satellite images, have been used to extract neighborhood data as a part of a hedonic process (Bency et al., 2017). Other popular data sources are interior and exterior images of the property or neighboring objects made from the ground or with the help of satellites. Glaeser et al. (2018) show that exterior and neighborhood aesthetics can enhance the explanatory power of hedonic models.

Research on real estate floor plans: Previous research on real estate floor plans has primarily focused on summarizing and extracting structured information from floor plans, yet with clear differences from our study. For example, Goncu et al. (2015) have created an application that leverages text and geometry recognition software to convert floor plan images into a vector file with contrasting colors. The goal is to provide a simplified, more accessible version of floor plans to sight-impaired users. Other works have focused on facilitating automatic search of homes based on floor plans. Sharma et al. (2019) developed a machine learning approach that maps floor plan images or user sketches of floor plans to floor plan images in a database. An enhanced version of this approach ("FloorNet") was developed by Kato et al. (2020) with the goal of making more accurate real estate recommendations for users on online real estate platforms.

The authors' findings suggest that floor plans contain additional information that is not available in the structured data, such as the way rooms are connected. Another approach is to analyze floor plans using adjacency graphs with nodes labeled as rooms and/or corridors (Hanazato et al., 2005). However, this method requires extensive manual labeling and can only provide partial data contained in the floor plans (Kato et al., 2020).

Research gap: None of the above references has utilized floor plans to enhance rent appraisal in hedonic pricing models. In this paper, we close this research gap by incorporating floor plans into hedonic models for rent price appraisal on online real estate platforms. Unlike exterior, interior, or satellite images, which present only a part of the apartment and are external to the property itself, we expect floor plans to be an integral part of property valuation and purchasing decision-making. Floor plans contain information about relative dimensions, positioning of rooms and utilities, and thus may carry price-relevant information in the real estate market. To the best of our knowledge, our research is the first to demonstrate that harnessing floor plans can enhance real estate appraisal in hedonic pricing models – even after controlling for structured apartment characteristics such as size and location.

5.3 Data

5.3.1 Apartment Listings

The data for this study was extracted from ImmobilienScout24.de, the largest German online real estate aggregator. At the end of 2019, ImmobilienScout24 contained 91,415 active rent listings across Germany, with approximately 44 million visitors per month.

We implemented a Python-based web crawler to download and store information about apartments in the city of Berlin in a database format. The crawler was active for three weeks at the end of 2019, during which it gathered a total number of 15,604 observations. Each observation represents an apartment in Berlin from an active or archived listing on ImmobilienScout24. Our dataset includes apartment listings between mid-2017 and end of 2019 (i. e., spans two and a half years). For each apartment, the crawler collected a raw image showing the floor plan of the apartment, the monthly rent price, the monthly rent price per m^2, the city district in which the apartment is located, and 36 fields with categorical and numerical information about the apartment (e. g., apartment size, etc.)

5.3.2 Variable Definitions

Dependent variables: The target variables in our study are two measures of the monthly costs of an apartment:

- *Rpms*: The monthly rent price (in €) per m^2
- *Rent*: The total monthly rent price (in €) of an apartment

Explanatory variables: Our hedonic rent price model accounts for the following apartment characteristics from previous works:

- *Area*: The total size of the apartment in m^2
- *Rooms*: The total number of rooms in the apartment
- *Floor*: The floor at which the apartment is located
- *Top Floor*: The number of floors of the building in which the apartment is located
- *City Distance*: The distance (in km) from the center of the city to the apartment
- *Year Built*: The construction year of the building in which the apartment is located

Control variables: Our dataset further includes 12 locational dummies that provide information about the city district, and a set of 30 additional control variables in the form of categorical dummies that provide fine-grained information about contractual and structural characteristics of an apartment (e. g., whether the apartment offers access to a parking lot).

5.3.3 Summary Statistics

Dependent variables: Table 5.1 shows summary statistics, whereas Figure 5.3 plots the complementary cumulative distributions (CCDFs) for the dependent variables. The total rent prices (*Rent*) in our data range from €230 per month to almost €3000 with a standard deviation of 476.80. The rent per m^2 (*Rpms*) ranges from 5.38 to 27.80 with a standard deviation of 4.32. Figure 5.1 provides an overview of the monthly apartment costs across different districts in Berlin. Evidently, there is a higher concentration of apartments with high monthly costs near the city center of Berlin.

Explanatory variables: Figure 5.2 plots the distributions of the explanatory variables in our dataset. The size of the apartments (*Area*) ranges from 16.00 to 178.08 square meters with a standard deviation of 26.35. The number of rooms

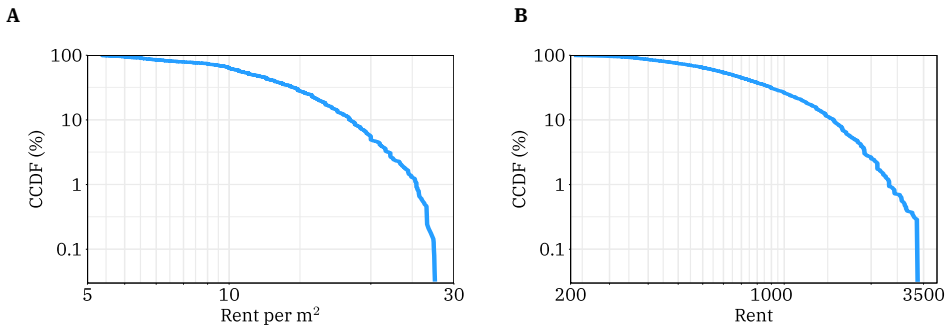


Figure 5.3: Complementary cumulative distribution functions (CCDFs) for *Rpms* and *Rent*.

Table 5.1: Summary statistics of key variables.

Variable	Mean	Median	Min	Max	Std. dev
Dependent variables					
Rpms (in) €	11.73	11.00	5.38	27.80	4.32
Rent (in) €	847.12	720.15	230.19	2 988.00	476.84
Explanatory variables					
Area (in m^2)	71.35	67.41	16.00	178.08	26.35
Rooms	2.44	2.00	1.00	6.00	0.89
Floor	2.89	3.00	0.00	26.00	2.29
Top Floor	5.09	5.00	0.00	27.00	2.21
City Distance (in km)	7.99	7.51	0.37	17.54	4.26
Year Built	1 970	1 985	1 830	2 020	43.76

ranges from 1 to 6, with most of the apartments having either 2 or 3 rooms. The average distance between an apartment and the city center (*City Distance*) is 7.99 km. The variables *Top Floor* and *Floor* have the same minimum of 0 and a maximum of 27 and 26, respectively. The houses in our sample have been built between the years 1830 and 2020. Most buildings are relatively new with a median construction year of 1985.

5.3.4 Cross-correlations

Figure 5.4 provides an overview of cross-correlations between the independent variables. Unsurprisingly, we see a positive correlation between the size of an apartment and the number of rooms (correlation of 0.81). Likewise, we find a positive correlation of 0.35 between *Floor* and *Top Floor*. We observe a negative correlation of -0.14 between *City Distance* and *Top Floor*, indicating that in Berlin, there is a higher concentration of tall buildings in areas close to the city center. The correlation between *Year Built* and *Top Floor* (0.16) suggests that newer buildings tend to have more floors, while the correlation between the construction year and the distance from the city center (0.18) highlights that there has been

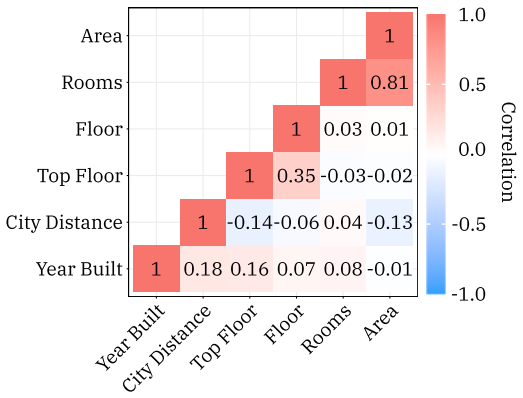


Figure 5.4: Cross-correlations.

more construction activity near the edges of Berlin in recent years. All remaining correlations are fairly small. Importantly, the variance inflation factors of all independent variables in our later analysis are below the critical threshold of 4. Hence there is no evidence that multicollinearity impedes the validity of our findings.

5.3.5 Extraction of Floor Plans from Images

We perform three main preprocessing steps to transform the raw floor plans into a format that allows for further calculations. First, we manually remove erroneous entries containing images that are not floor plans. These include interior and exterior photos, promotional materials, and presentation slides. Second, we exclude technically correct but unrelated floor plans. For example, if there is a floor plan for the parking space, it is removed, so the neural network does not learn unrelated information. These filtering steps reduce our dataset to 9,174 observations.

Once the data set is reduced to contain only the relevant entries, we modify them for the use in neural networks. We crop every picture until only the plan remains. Subsequently, we scale them with proportions preserved, filling empty space with pure black to prevent unnecessary calculations. All floor plans are then adjusted to follow the Xception guidelines (Chollet, 2017) of min-max normalized color images with 299 by 299 pixels in size. For this purpose, we use an ImageMagick script that scales images without shifts in proportions to avoid potential loss of information and to lighten the computational burden. The prepared images are then vectorized, min-max normalized, and concatenated to the data frame. The process is illustrated in Figure 5.5.

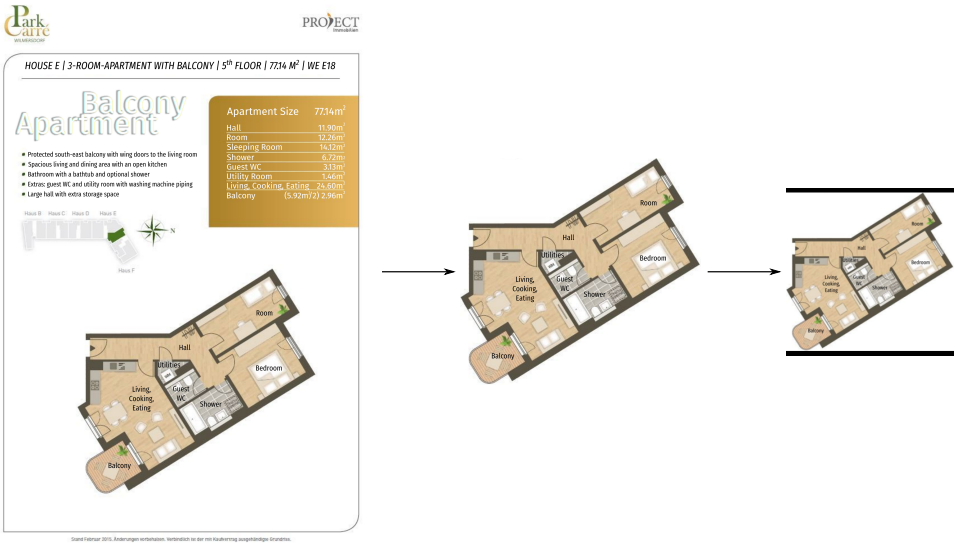


Figure 5.5: Example of cropping and scaling of floor plans.

5.4 Methodology

We propose a two-stage approach to determine the hedonic price of apartment floor plans. As a first step, we compute adjusted rent prices by regressing the monthly apartment costs on the structural and locational variables (Stage 1). This allows us to extract the variance in rent that is unexplained by these variables and prevents our model from learning information that is already available from known features. Second, we use convolutional neural networks (CNN) and the adjusted rent prices to predict the sentiment of the floor plans, i. e., the apartment rent price after controlling for locational and structural characteristics of an apartment (Stage 2).

5.4.1 Computing Adjusted Rent Prices (Stage 1)

Stage (1) performs a hedonic regression to calculate the variance in the apartment costs that is unexplained by the structural and locational attributes of an apartment. This approach follows previous works (Ball et al., 2012; Dehaan et al., 2013; Farrell et al., 2014; Naumzik & Feuerriegel, 2020), where regression residuals are used as a proxy for the proportion of variance of the dependent variable that is unexplained by known determinants. Let y_i denote the logarithmized rent/rent per m^2 belonging to listing $i = 1, \ldots, n$. We then estimate a linear model via ordinary least squares (OLS) regressing y_i on the structural and locational

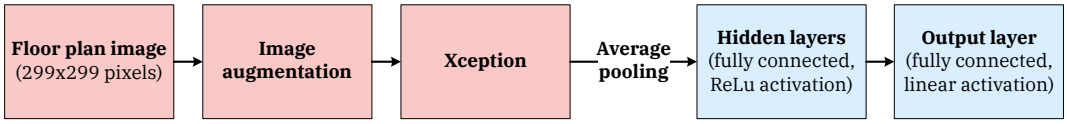


Figure 5.6: Illustration of model architecture.

attributes $c_{i1}, \ldots, c_{iJ}$. Formally, the model is described by

$$y_i = \beta_0 \sum_{j=1}^{J} \beta_j c_{ij} \varepsilon_i, \tag{5.1}$$

with intercept β_0 and error term ε. The model residuals $\tilde{y_i} = y_i - \hat{y_i}$ then represent adjusted rent prices that are used in a neural network model in Stage (2). Importantly, the calculation of the adjusted prices ensures that the neural network does not merely learn to predict easily observable structural apartment characteristics.

5.4.2 Floor Plan Sentiment (Stage 2)

Stage (2) uses convolutional neural networks (CNN) to predict the sentiment of the floor plans. The input to the model is a vector representation of the pre-processed floor plan images (299×299 pixels). The variable to predict is the adjusted rent price from Stage (1), i. e., $\tilde{y_i}$. We implement our learning task using Xception (Chollet, 2017) developed by Google, which is the state-of-the-art in image classification (Li et al., 2018). Specifically, we employ a pretrained Xception model trained on a large dataset of images from ImageNet (Chollet, 2017) and employ a transfer learning strategy (i. e., learn a new task through the transfer of knowledge from a related task that has already been learned) to transfer the pre-trained Xception model to our floor plan prediction task.

Architecture customization: The pretrained Xception model uses a softmax layer as the final layer to make categorical predictions. As our goal is to use floor plan images to predict a continuous variable ($\tilde{y_i}$), we modify the architecture by removing the softmax layer from the pre-trained model. Further, we apply the global average pooling operation on the output layer. We then append a set of fully connected layers with ReLu activation function, followed by an output layer consisting of a single neuron (with linear activation function). The latter ensures that our model produces continuous predictions of $\tilde{y_i}$. The number of hidden layers and nodes are treated as hyperparameters and the optimal numbers are selected by applying a grid search using 5-fold cross-validation. Our model architecture is illustrated in Figure 5.6.

Image augmentation: To increase the number of available floor plans, we implement online image augmentation. In particular, each floor plan has a probability of being mirrored or rotated by an arbitrary number of radians. While this does not change any information in the image itself, it expands the number of images available to the neural network, which may help it to better understand different features and combat overfitting.

Training: We implement the CNN in Keras with GPU-accelerated TensorFlow as backend. Formally, we train[1] the function f_θ by mapping the image vector representations of the (augmented) floor plans x_i to the training labels $\tilde{y_i}$, resulting in an estimated parametrization $\hat{\theta}$. The CNN is trained by minimizing the loss between the trained function f_θ and the price residual $\tilde{y_i}$. In the following, we refer to the resulting predictions as the sentiment of the floor plans (*Floor Plan Sentiment*). These predictions are later used in (i) a hedonic regression model to evaluate the explanatory power of floor plans, and (ii) as an additional feature for out-of-sample prediction of rent prices.

5.5 Empirical Analysis

We now empirically investigate to what extent an automated visual analysis of floor plans can enhance real estate appraisal on online real estate marketplaces. First, we link to earlier research by studying the hedonic pricing value of floor plans in a hedonic regression model. Second, we evaluate the gains in prediction performance when incorporating floor plans.

5.5.1 Hedonic Regression Analysis

Model specification: We apply a log-linear regression model to analyze the role of floor plans for hedonic rent price appraisal. Regression models are generally regarded as an explanatory approach with the ability to document statistical relationships and, in particular, estimating effect sizes (Breiman, 2001). Furthermore, log-linear regression models are widely used to estimate the hedonic rent price models in the real estate sector (cf. Section 5.2). This modeling approach allows us to interpret the model coefficients and statistically test the association between floor plans and rent prices.

The key explanatory variable in our hedonic regression model is the sentiment of the floor plan images (*Floor Plan Sentiment*), which we calculate based on

[1]We use relu activations for the hidden layers, linear activation for the output layer, and Nadam optimizer. We use a validation split of 20 % to determine the optimal stopping point. We optimize the number of layers, the number of neurons per layer for each layer, learning rate, and betas via grid search.

the methodology described in the previous section. Importantly, we use 5-fold cross-validation to predict the *Floor Plan Sentiment*. Hence, *Floor Plan Sentiment* refers to out-of-sample predictions.

$$y_i = \beta_0 \beta_1 FloorPlanSentiment_i \sum_{j=1}^{J} \gamma_j c_{ij} \varepsilon_i, \tag{5.2}$$

where the dependent variable y_i is the log of either *Rpms* or *Rent*, β_0 is the intercept, and the coefficients γ gauge the effects of the explanatory and control variables. The coefficient β_1 captures the marginal effect of the *Floor Plan Sentiment*. This is our parameter of interest as it measures the contribution of the visual design of floor plans on the monthly costs of an apartment. Note that, by combining floor plans and structural/locational in a joint regression model, we can isolate the **marginal** effect of floor plans on the monthly costs of an apartment.

Estimation: We follow previous works and estimate Equation (5.2) via ordinary least squares (OLS). Since rent prices tend to be log-normally distributed, we log-transform the dependent variables. For the sake of interpretability, we also z-standardize all variables, so that the regression coefficients measure the relationship with the dependent variable measured in standard deviations.

Coefficient estimates for rent per m^2 (*Rpms*): The parameter estimates in Figure 5.7 show that the visual design of floor plans has significant explanatory power regarding rent prices. The coefficient for *Floor Plan Sentiment* has the largest positive effect size, and is statistically significant (coef: 0.128; $p < 0.001$). A one standard deviation change in *Floor Plan Sentiment* is estimated to increase the rent price by 13.65 %. The largest negative effect on *Rpms* is estimated for *City Distance* (coef: -0.108; $p < 0.001$). An increase in standard deviation in *City Distance* reduces the price per m^2 by 10.23 %. We further find statistically significant positive effects for *Year BUilt* and *Area*. In contrast, higher values for *Rooms*, *Floor*, and *Top Floor*, have a negative effect on the rent price per m^2.

For comparison, Figure 5.7 also reports the results for a baseline model without *Floor Plan Sentiment*. All coefficients remain relatively stable which confirms the validity of our results. We again find that apartments that are smaller, located at higher floors, and farther away from the city center are associated with a lower rent per m^2. Apartments with fewer rooms are associated with higher rent per m^2.

Control variables: Our regression model controls for 30 additional structural and locational apartment characteristics (see Figure 5.1). The corresponding estimates are omitted from Figure 5.7 for the sake of brevity. In short, more preferable contractual clauses (e. g., non-coal heating system, the allowance of pets or recent renovation) result in higher rent prices. We further observe

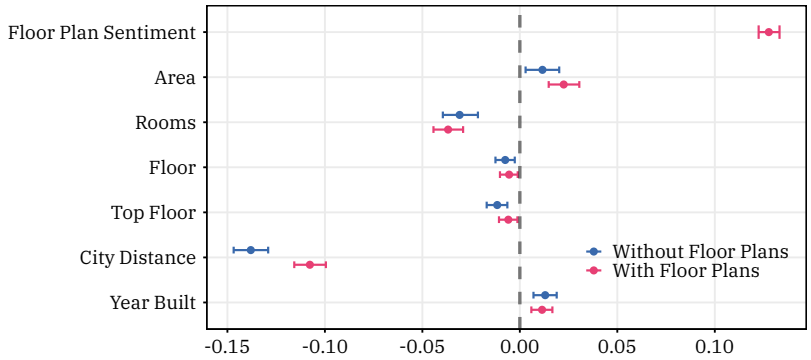


Figure 5.7: Standardized parameter estimates and 95 % confidence intervals.

location effects, i. e., city district dummies that are statistically significant.

Goodness-of-fit: We calculated the adjusted-R^2 for each model, resulting in relatively high values of 0.622 for the baseline hedonic model without floor plans. If we additionally include floor plans, the adjusted-R^2 increases to 0.697. Evidently, our model specification accounts for a large proportion of the variations in the dependent variable and *Floor Plan Sentiment* has significant explanatory power. This observation is supported by the difference in the AIC scores for the model with and without *Floor Plan Sentiment*. For both dependent variables, the difference is greater than 10, indicating strong support for the corresponding candidate models (Burnham & Anderson, 2004). Therefore, the model that incorporates the floor plans is to be preferred.

Analysis of total rent (*Rent*): The estimates for the regression with the total rent (*Rent*) as dependent variable are omitted for brevity. In short, the coefficients are consistent with those for the rent price per m^2 (*Rpms*). Also here, the coefficient for the *Floor Plan Sentiment* variable is statistically significant at the 1 % statistical significance level. The regression results again suggest that the sentiment of the floor plans explains a considerable amount of the variations in rent prices after controlling for structural and locational characteristics of an apartment. The coefficient for *Floor Plan Sentiment* is again among the variables with the largest effect sizes. In other words, floor plans contain price relevant information in online real estate listings.

5.5.2 Prediction Performance

Next, we examine whether floor plans are useful regarding out-of-sample prediction of rent prices. As mentioned in Section 5.2.1, hedonic analysis is not limited to linear regression and can be used in combination with machine learning methods. In the following, we compare the predictive value of incorporating floor plans as an additional predictor for three models: log-linear OLS regression

Method	Floor Plans	MSE	MAE
Benchmark: Log-Linear Regression	×	0.04984	0.1565
Hedonic models			
Log-Linear Regression (with floor plans)	✓	0.04400	0.1347
CatBoost	×	0.00140	0.0262
CatBoost (with floor plans)	✓	0.00134	0.0258
Deep Neural Networks	×	0.00142	0.0214
Deep Neural Networks (with floor plans)	✓	**0.00127**	**0.0205**

Table 5.2: Out-of-sample prediction performance for rent per m^2 (*Rpms*). The lowest values of mean squared error (MSE) and mean absolute error (MAE) are highlighted in bold.

(acting as a benchmark), boosted decision tree ensembles (CatBoost), and deep neural networks. To evaluate the predictive value of floor plans, we split our dataset into training and test sets and compare the effect of incorporating floor plans on the out-of-sample prediction performance of the models. Specifically, we use 5-fold cross-validation to calculate mean squared error (MSE) and mean absolute error (MAE).

Prediction of rent per m^2: Table 5.2 reports the out-of-sample prediction performance for the rent per m^2 ($Rpms$). Incorporating floor plans yields improvements for all considered out-of-sample performance metrics. A log-linear hedonic regression model yields an MSE of 0.0484 and an MAE of 0.1565. Incorporating *Floor Plan Sentiment* reduces the MSE by 11.71 % and the MAE by 13.93 %. Hedonic models trained with machine learning methods yield further improvements. A hedonic model trained with neural networks yields the lowest out-of-sample prediction error (MSE of 0.0014 and MAE of 0.0214). Also here, incorporating floor plans yields a substantially lower error. The MSE is reduced by 10.56 %, whereas the MAE is reduced by 4.21 %. For the decision tree ensembles (CatBoost), we find a reduction of 4.29 % in MSE and a reduction in MAE of 1.53 %. We also conducted paired t-tests to assess the statistical significance of the reductions in prediction error between the machine learning models with vs. without floor plans. We find that incorporating floor plan yields statistically significantly reduced prediction errors for all considered machine learning methods, i. e., log-linear regression ($p < 0.001$), CatBoost ($p < 0.05$), and deep neural networks ($p < 0.001$).

Prediction of total rent: Table 5.3 reports the performance for the prediction of the total rent price (*Rent*). All considered models show a significant reduction of the prediction error when including floor plans. The MSE of log-linear regression, decision tree ensemble, and deep neural networks reduce by 6.05 %, 4.27 %, and 5.08 %. We also observe a lower MAE for log-linear regression (−3.92 %), decision tree ensemble (−2.03 %), and deep neural networks

Method	**Floor Plans**	**MSE**	**MAE**
Benchmark: Log-Linear Regression	×	0.05078	0.1529
Hedonic models			
Log-Linear Regression (with floor plans)	✓	0.04771	0.1459
CatBoost	×	0.00117	0.0246
CatBoost (with floor plans)	✓	0.00112	0.0241
Deep Neural Networks	×	0.00118	0.0203
Deep Neural Networks (with floor plans)	✓	**0.00112**	**0.0199**

Table 5.3: Out-of-sample prediction performance for rent (*Rent*). The lowest values of mean squared error (MSE) and mean absolute error (MAE) are highlighted in bold.

(-1.97 %). Analogous to the analysis of the rent per m^2, paired t-tests show that the reductions in prediction error between the models with vs. without floor plans are statistically significant for each considered machine learning method. Altogether, these results demonstrate that floor plans not only feature significant explanatory power but also serve as highly relevant features for predictive purposes.

5.5.3 Sensitivity Analysis & Robustness Checks

Analysis on data subsets: We repeated our analysis on multiple subsets of the data. While the increase in explanatory power stemming from the floor plans remained significant for every tested subset, we observed a statistically significant more pronounced effect for smaller apartments, as well as for older houses.

Model specification: We repeated our analysis using a number of alternative model specifications, including, but not limited to CNN models (Xception, VGG16, ResNet101V2, DenseNet, EfficientNet), alternative input image formats (color, gray scale, preserved ratios, fill methods), alternative parameter configurations of the neural network (model concatenation, input concatenation, output concatenation, depths), and alternative variable specifications (normalization methods, log-specifications). Additionally, we tested the inclusion of the images directly alongside the structured data and as a standalone feature. In total, we have tested more than 50 various permutations yielding robust findings, i. e., significant improvements in explanatory power and prediction performance when incorporating floor plans into hedonic price models for rent price appraisal.

Additional robustness checks: We conducted additional checks to validate the robustness of our hedonic regression model: (1) We calculated variance inflation factors for all independent variables in our hedonic regression models and found that all remain below the critical threshold of four. (2) We tested

alternative model specifications in which naturally correlated variables such as size and number of rooms of an apartment are iteratively added one by one and tested for statistical significance after each iteration (i. e., stepwise regression). (3) We controlled for outliers in the dependent variables. (4) We added quadratic terms for each explanatory variable to our hedonic regression models. In all cases, our results are robust and consistently support our findings.

5.5.4 Exemplary Apartment Listings

We now explore apartment listings for which floor plans are particularly informative for rent price appraisal on online real estate platforms. Figure 5.8 shows four exemplary floor plans with particularly pronounced differences between the predicted rent prices of the hedonic model with vs. without floor plans. The floor plans (a) and (b) yielded upward price adjustments, while the floor plans (c) and (d) yielded downward price adjustments.

Figure 5.8 suggests that incorporating floor plans yields higher rent price predictions in cases in which floor plans convey relatively complex information. A possible reason is that these floor plans are particularly informative because they convey information that is not clear from the structured data alone. For instance, floor plan (a) in Figure 5.8 has an uncluttered entrance, opening up to the entire building and a comparatively high number of windows, which may facilitate higher valuation. Floor plan (b) in Figure 5.8 shows a spacious apartment with a large and open-spaced living room, as well as two separated and self-contained private sections, each outfitted with personal bathrooms. This may be highly price-relevant information that is only available from the apartment floor plan. On the contrary, floor plan (c) seems to lack features that warrant a higher rent price. The apartment it is separated into relatively small rooms connected by a very long corridor. The only point of entrance to a single balcony is through the room farthest away from the entrance. Likewise, floor plan (d) has only a tiny bathroom and four rooms that are very elongated. The apartment thus exhibits potentially unfavorable characteristics that are not directly observable from from the structured data. Altogether, our exploratory analysis suggests that floor plans contain hidden information that is highly price-relevant – even after accounting for structural and locational characteristics of an apartment.

5.6 Discussion

Research implications: Our empirical findings contribute to the existing research body by quantifying the hedonic value of floor plans and establishing the

A Positive price adjustment

B Positive price adjustment

C Negative price adjustment

D Negative price adjustment

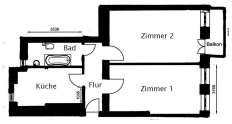


Figure 5.8: Examples of floor plans with positive and negative effects on rent prices.

link between the visual design of floor plans and real estate prices. Prior research (Hill et al., 1997; Nowak & Smith, 2016) suggests that real estate markets tend to contain a variety of non-observable or hidden characteristics that are not taken into account by conventional valuation methods. Our findings show that there is indeed an underutilization of the available data in current hedonic models. Specifically, we show that floor plans contain price relevant information in online real estate listings. This suggests that there are hidden features in floor plans, such as the relative size and positioning between the rooms, that are relevant even after controlling for structural and locational characteristics. To the best of our knowledge, our paper is the first study to demonstrate that harnessing floor plans can enhance real estate appraisal on online real estate platforms.

Practical implications: From a practical perspective, our findings are particularly relevant for online real estate platforms. Currently, the decision of whether or not to upload a floor plan for an apartment is typically left to the discretion of the user. Based on our finding that floor plans contain price-relevant information that helps users to make an informed decision, real estate platforms should consider making them a mandatory part of the listings and show them more prominently. This would allow for greater market transparency for both renters and landlords, as well as home buyers and sellers. In the backend of online real estate platforms, floor plans could also be used as an additional price predictor. For example, online real estate platforms could implement a model similar to ours after the author has filled all fields to recommend an appropriate rent price. Alternatively, the proposed prediction could be listed alongside the price set by the author, such that the renter or home buyer has a clearer picture of the value of the property. Our model could also suggest links to similar lots based not only on the price but on the information from the floor plans, providing a better market overview and diminishing the problem of information asymmetry. While some valuation projects, such as Zillow, have started taking images into account, floor plans remain severely underutilized. Ultimately, our study equips real estate investors with the clear benefit of improved risk assessment, allowing them

to enhance their portfolios and reduce the possibility of misvaluation, which indirectly improves the quality of financial markets (Hoesli & Reka, 2015).

Limitations and future research directions: Our work provides several avenues for future research that could further enhance the accuracy of recommendation systems on online real estate platforms. First, it would be interesting to extend our hedonic model by performing textual analysis (e. g., sentiment analysis Lutz et al., 2019, 2023; Pröllochs et al., 2018) of apartment descriptions in real estate listings. Second, while our study was conducted on the rental market of Berlin, the underlying deep learning approach can easily be applied to listings from other marketplaces and locations. Third, future research could extend our study by studying the hedonic value of different floor plan layouts or other interpretable floor plan features (e. g., windows, doors, relative room sizes). It is also a promising research direction to study the suitability of floor plans with different characteristics for certain groups of potential users.

5.7 Conclusion

Current hedonic price models in the online real estate market fail to incorporate information about floor plans, as they are typically not part of the structured information of online listings, but rather provided in the form of accompanying images. For this purpose, this paper investigates to what extent an automated visual analysis of floor plans on online real estate marketplaces can help to enhance real estate appraisal. We find that the visual design of floor plans has significant explanatory power regarding rent prices – even after controlling for structured apartment characteristics such as size and location. Moreover, we demonstrate that harnessing floor plan sentiment results in 10.56 % more accurate out-of-sample predictions compared to models that only use structural and locational data. From a practical perspective, our findings provide decision support for real estate investors and have direct implications for online real estate platforms to improve user experience in their real estate listings.

Bibliography

Andersen, S., & Nielsen, K. M. (2017). Fire sales and house prices: Evidence from estate sales due to sudden death. *Management Science*, *63*(1), 201–212.

Ball, R., Jayaraman, S., & Shivakumar, L. (2012). Audited financial reporting and voluntary disclosure as complements: A test of the confirmation hypothesis. *Journal of Accounting and Economics*, *53*(1–2), 136–166.

Bency, A. J., Rallapalli, S., Ganti, R. K., Srivatsa, M., & Manjunath, B. S. (2017). Beyond spatial auto-regressive models: Predicting housing prices with satellite imagery. *Proceedings of the WACV*.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–231.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*(2), 261–304.

Chiang, K. C. H., Lee, M.-L., & Wisen, C. H. (2005). On the time-series properties of real estate investment trust betas. *Real Estate Economics*, *33*(2), 381–396.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. *Proceedings of the CVPR*.

Clapp, J. M., & Giaccotto, C. (1992). Estimating price indices for residential property: A comparison of repeat sales and assessed value methods. *Journal of the American Statistical Association*, *87*(418), 300–306.

Dehaan, E., Hodge, F., & Shevlin, T. (2013). Does voluntary adoption of a clawback provision improve financial reporting quality? *Contemporary Accounting Research*, *30*(3), 1027–1062.

Farrell, A. M., Goh, J. O., & White, B. J. (2014). The effect of performance-based incentive contracts on system 1 and system 2 processing in affective decision contexts: FMRI and behavioral evidence. *The Accounting Review*, *89*(6), 1979–2010.

Glaeser, E., Kincaid, M. S., & Naik, N. (2018). Computer vision and real estate: Do looks matter and do incentives determine looks. *NBER Working Paper*, *27164*.

Goncu, C., Madugalla, A., Marinai, S., & Marriott, K. (2015). Accessible on-line floor plans. *Proceedings of the WWW*.

H. Ahmed, E., & Moustafa, M. (2016). House price estimation from visual and textual features. *Proceedings of the IJCCI*.

Hanazato, T., Hirano, Y., & Sasaki, M. (2005). Syntactic analysis of large-size condominium units supplied in the Tokyo metropolitan area. *Journal of Architecture and Planning*, *591*, 9–16.

Hill, R. C., Knight, J. R., & Sirmans, C. F. (1997). Estimating capital asset price indexes. *Review of Economics and Statistics*, *79*(2), 226–233.

Hoesli, M., & Reka, K. (2015). Contagion channels between real estate and financial markets. *Real Estate Economics*, *43*(1), 101–138.

Kato, N., Yamasaki, T., Aizawa, K., & Ohama, T. (2020). Users' preference prediction of real estate properties based on floor plan analysis. *IEICE Transactions on Information and Systems*, *E103.D*(2), 398–405.

Kolbe, J., & Wüstemann, H. (2014). Estimating the value of urban green space: A hedonic pricing analysis of the housing market in Cologne, Germany. *Acta Universitatis Lodziensis. Folia Oeconomica*, *5*(307).

Krainer, J., & Wei, C. (2004). House prices and fundamental value. *FRBSF Economic Letter*, *27*.

Li, M., Bao, Z., Sellis, T., Yan, S., & Zhang, R. (2018). Homeseeker: A visual analytics system of real estate data. *Journal of Visual Languages & Computing*, *45*, 1–16.

Lu, B., Charlton, M., Harris, P., & Fotheringham, A. S. (2014). Geographically weighted regression with a non-euclidean distance metric: A case study using hedonic house price data. *International Journal of Geographical Information Science*, *28*(4), 660–681.

Lutz, B., Adam, M. T. P., Feuerriegel, S., Pröllochs, N., & Neumann, D. (2023). Affective information processing of fake news: Evidence from NeuroIS. *European Journal of Information Systems*, *32*, 1–20.

Lutz, B., Pröllochs, N., & Neumann, D. (2019). The longer the better? The interplay between review length and line of argumentation in online consumer reviews. *Proceedings of the ICIS*.

Malpezzi, S. (2002). Hedonic pricing models: A selective and applied review. In T. O'Sullivan & K. Gibb (Eds.), *Housing economics and public policy* (pp. 67–89). John Wiley & Sons, Ltd.

Monson, M. (2009). Valuation using hedonic pricing models. *Cornell Real Estate Review*, *7*(1), 62–73.

Natividade-Jesus, E., Coutinho-Rodrigues, J., & Antunes, C. H. (2007). A multicriteria decision support system for housing evaluation. *Decision Support Systems*, *43*(3), 779–790.

Naumzik, C., & Feuerriegel, S. (2020). One picture is worth a thousand words? The pricing power of images in e-commerce. *Proceedings of the WWW*.

Nowak, A., & Smith, P. (2016). Textual analysis in real estate. *Journal of Applied Econometrics*, *32*(4), 896–918.

Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T., & French, N. (2003). Real estate appraisal: A review of valuation methods. *Journal of Property Investment & Finance*, *21*(4), 383–401.

Peterson, S., & Flanagan, A. (2009). Neural network hedonic pricing models in mass real estate appraisal. *Journal of Real Estate Research*, *31*(2), 147–164.

Pröllochs, N., Feuerriegel, S., & Neumann, D. (2018). Statistical inferences for polarity identification in natural language. *PLOS ONE*, *13*(12), e0209323.

Sharma, D., Gupta, N., Chattopadhyay, C., & Mehta, S. (2019). A novel feature transform framework using deep neural network for multimodal floor

plan retrieval. *International Journal on Document Analysis and Recognition (IJDAR)*, *22*(4), 417–429.

Sirmans, S., Macpherson, D., & Zietz, E. (2005). The composition of hedonic pricing models. *Journal of Real Estate Literature*, *13*(1), 1–44.

Sopranzetti, B. J. (2010). Hedonic regression analysis in real estate markets: A primer. In C.-F. Lee, A. C. Lee, & J. Lee (Eds.), *Handbook of quantitative finance and risk management* (pp. 1201–1207). Springer US.

Sun, D., Du, Y., Xu, W., Zuo, M., Zhang, C., & Zhou, J. (2014). Combining online news articles and web search to predict the fluctuation of real estate market in big data context. *Pacific Asia Journal of the Association for Information Systems*, *6*(4), 19–37.

Sun, G., Liang, R., Wu, F., & Qu, H. (2013). A web-based visual analytics system for real estate data. *Science China Information Sciences*, *56*(5), 1–13.

Wallace, N. E., & Meese, R. A. (1997). The construction of residential housing price indices: A comparison of repeat-sales, hedonic-regression, and hybrid approaches. *The Journal of Real Estate Finance and Economics*, *14*(1/2), 51–73.

You, Q., Pang, R., Cao, L., & Luo, J. (2017). Image-based appraisal of real estate properties. *IEEE Transactions on Multimedia*, *19*(12), 2751–2759.

Yuan, X., Lee, J.-H., Kim, S.-J., & Kim, Y.-H. (2013). Toward a user-oriented recommendation system for real estate websites. *Information Systems*, *38*(2), 231–243.

Zumpano, L. V., Johnson, K. H., & Anderson, R. I. (2003). Internet use and real estate brokerage market intermediation. *Journal of Housing Economics*, *12*(2), 134–150.

Declaration of Authorship

Hiermit erkläre ich, dass ich die vorgelegten Aufsätze selbstständig und nur mit den Hilfen angefertigt habe, die für den jeweiligen Aufsatz angegeben sind. In der Zusammenarbeit mit den angeführten Koautoren war ich wie angegeben anteilig beteiligt. Bei den von mir durchgeführten und in den Aufsätzen erwähnten Untersuchungen habe ich die Grundsätze guter wissenschaftlicher Praxis eingehalten, wie sie in der Satzung der Justus-Liebig-Universität Gießen zur Sicherung guter wissenschaftlicher Praxis niedergelegt sind.

Ort, Datum	Unterschrift