User-generated content in marketing research: methods and applications

Inauguraldissertation
zur Erlangung des akademischen Grades eines Doktors
der Wirtschaftswissenschaften des Fachbereichs Wirtschaftswissenschaften
der Universität Osnabrück

vorgelegt
von

David Dornekott

Osnabrück,
Dezember 2022

Dekan/in	Prof. Frank Westermann, Ph.D.
Referenten:	Prof. Dr. Bernhard Baumgartner
	Prof. Dr. Valeriya Dinger
Tag der Disputation:	08.12.2022

"A good dissertation is finished, a great dissertation is published, a perfect dissertation is neither."

-

Origin unknown

Acknowledgements

I would like to thank my co-authors Philip Wollborn, Ulrike Holder and Lisardo Erman for their contributions and friendship, my supervisor Bernhard Baumgartner for the years of advice and support he provided and Valeriya Dinger for her ongoing mentorship. I would also like to thank my current and former co-workers at the University of Osnabrück, especially Lisa Wamhoff, Mechthild Brunzel and Florian Scholtysik, for their help over the years, as well as a number of anonymous reviewers from a variety of conferences and journals for their feedback on the papers that constitute this dissertation. I am further grateful to my wonderful wife Elena and to our two sons, Juri and Thees, for not being even worse sleepers and for reminding me of my priorities in life.

TABLE OF CONTENT

LIST OF TABLES

TABLE OF FIGURES

TABLE OF EQUATIONS

I. Introduction

There is perhaps no other technological development within the past half century as impactful and far-reaching as the introduction of networked computing. As human activity continues to move towards the digital realm, a growing amount of information is created in digital form. A key differentiator between digital content and its analog equivalents lies in just how much of it can be collected, stored and analyzed. While most of this data remains siloed, distributed over countless devices, cloud storage services, messaging platforms or customer databases, a non-negligible amount of it is freely shared on the open internet. This includes not only the content itself - in the form of text, photographs, audio and video - which has been shared actively, voluntarily and consciously, but also passive information in the form of metadata, such as the connection to other users and other by-products of the information sharing process such as time and location.

The totality of information submitted to digital platforms - such as social networks, microblogging services and review platforms - is summarily referred to as *user-generated content* and encompasses a broad variety of data types, ranging from text and images to audio and video recordings (Luca, 2015). A key characteristic of user-generated content is that it is both publicly available, as well as easily accessible and collectable in arbitrarily large amounts by anyone who has the necessary know-how and an interest in doing so. This glut of data has opened a multitude of potential new avenues for scientific inquiry, touching the fields of computer science, physics, political science and management, amongst many more, reflecting the huge range of potential applications and variety of methodological approaches involved. A broad distinction can be made between research concerned with the underlying data creation process itself and research that leverages user-generated content as a source of data. Research of the former type mostly comes from the direction of the natural sciences, wherein, for example, methods stemming from physics are used to investigate how social networks develop dynamically over time (Kleineberg & Boguñá, 2014). In contrast, the political and social sciences are more focused on exploring its informativeness of and impact on human thought and behavior, for example by investigating its connection to political polarization (Whittaker et al., 2021).

As a field that is majorly concerned with human behavior as it relates to the consumption of goods, marketing research has made pervasive use of user-generated content as well. Within the context of marketing, a subset of user-generated content representing consumers sharing their experience with and opinions about products, brands and services is a popular subject of investigation in marketing research, where it is referred to by a variety of names, including *electronic word of mouth* (i.e. Cheung et al., 2009; Park et al., 2019)*, online word of mouth* (i.e. King et al., 2014) or, more colloquially, *online buzz* (i.e. Houston et al., 2018; Khadjeh Nassirtoussi et al., 2014).

While in general, all communication surrounding the awareness of, expressed attitude towards, intention to buy or experience with a product, service or brand can be of interest for marketing research, the most common types of user-generated content investigated come in the form of user reviews (Archak et al., 2011; Kim et al., 2018; Tirunillai & Tellis, 2014), discussion forums (Cheung et al., 2009; Dellarocas, 2004; Netzer et al., 2012) and microblog postings (Culotta & Cutler, 2016; Ghiassi et al., 2013; Hennig-Thurau et al., 2015). Commonly, the goal is to use information derived from user-generated content to explain economic variables of interest, such as product sales (Babić Rosario et al., 2016). Within this framework, the focus often lies on two dimensions of user-generated content in the form of its volume (i.e. the amount of online chatter about a given product within a given timeframe) and its valence (the qualitative information about the subject of interest, i.e. the attitude expressed towards a given product by the originators of the messages), which are measured by a variety of methods and subsequently operationalized as explanatory variables within regression-type models (see, for example, Duan et al., 2008; Vujić & Zhang, 2018; Yang et al., 2012; Zhou & Duan, 2012).

The diversity and vastness of user-generated content presents a number of challenges and opportunities to marketing researchers. Accessing, managing and processing the large amounts of data involved requires methodological knowledge traditionally found in the realm of computer science and software engineering and the multitude of different platforms that act as data sources differ in their functioning, each requiring specific knowledge and experience in consequence. This dissertation builds upon data collected from a multitude of sources over several years, including reddit (studies one, two and four), twitter (study two) and Twitch (study three). Twitch, as a live-streaming platform, represents the ongoing development of new forms of user-generated content and the opportunities for marketing research they afford. Live-streaming exemplifies the ongoing development from online content that is strictly relevant to commercial interests in that it conveys commercially relevant information, towards user-generated content that itself creates commercial value. As such, it touches on a lot of aspects related to marketing, both in the way that streamers can act as communicators for brands and products (i.e. in the form of sponsorships and endorsements), as well as the way that streamers are marketing themselves and adjusting their activities to target their audience. Despite these aspects and a growing economic importance, live-streaming has not yet been extensively studied from the perspective of marketing, which the third study aims to address.

Since the type of data used in the aforementioned applications of user-generated content almost inclusively comes in the form of text, working with user-generated content usually involves the processing of large amounts of text data. While the attention of marketing researchers has recently shifted towards other content types, such as images (Klostermann et al., 2018), their processing involves currently involves labelling them algorithmically with the objects, people and situations they capture, all of which presently happens in the form of text, as it represents data that is both human readable and machine processable. As such, working with user-generated content of any kind currently necessitates the use of natural language processing and text mining techniques and consequently all of the studies that make up the present dissertation make use of it to varying degrees. The first study extensively relies on text-mining techniques to derive information on the degree to which brands in the motion picture industry are associated in the minds of consumers, whereas the second study evaluates new sentiment-classification techniques for their use in marketing research, which are in turn applied in the fourth study to quantify the qualitative valence of stock-related news reports and online discussions. Even the third study, which on the surface does not rely on text data at all, makes some use of text data to identify and subsequently filter out individual observations outside the scope of the investigation.

To conclude, the studies that make up this dissertation cover a wide range of topics and relate to different aspects of marketing in multiple ways and to varying degrees. All make use of the excessive amount of publicly available information that user-generated content represents and exemplify some of the ways that companies and marketing professionals can exploit it to better their understanding of the needs of their customers, develop more effective marketing measures and ultimately gain a competitive advantage. The content and contributions of the studies that form the chapters of this dissertation will be summarized more thoroughly in the following chapter.

References

Archak, N., Ghose, A., & Ipeirotis, P. G. (2011). Deriving the Pricing Power of Product Features by Mining Consumer Reviews. *Management Science*, *57*(8), 1485–1509. https://doi.org/10.1287/mnsc.1110.1370

Babić Rosario, A., Sotgiu, F., De Valck, K., & Bijmolt, T. H. A. (2016). The Effect of Electronic Word of Mouth on Sales: A Meta-Analytic Review of Platform, Product, and Metric Factors. *Journal of Marketing Research*, *53*(3), 297–318. https://doi.org/10.1509/jmr.14.0380

Cheung, M., Luo, C., Sia, C., & Chen, H. (2009). Credibility of electronic word-of-mouth: Informational and normative determinants of on-line consumer recommendations. *International Journal of Electronic Commerce*, *13*(4), 9–38. https://doi.org/10.2753/JEC1086-4415130402

Culotta, A., & Cutler, J. (2016). Mining Brand Perceptions from Twitter Social Networks. *Marketing Science*, *35*(3), 343–362. https://doi.org/10.1287/mksc.2015.0968

Dellarocas, C. (2004). Strategic Manipulation of Internet Opinion Forums: Implications for Consumers and Firms. *Ssrn*, *January 2019*. https://doi.org/10.2139/ssrn.585404

Duan, W., Gu, B., & Whinston, A. B. (2008). The dynamics of online word-of-mouth and product sales-An empirical investigation of the movie industry. *Journal of Retailing*, *84*(2), 233–242. https://doi.org/10.1016/j.jretai.2008.04.005

Ghiassi, M., Skinner, J., & Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with Applications*, *40*(16), 6266–6282. https://doi.org/10.1016/j.eswa.2013.05.057

Hennig-Thurau, T., Wiertz, C., & Feldhaus, F. (2015). Does Twitter matter? The impact of microblogging word of mouth on consumers' adoption of new movies. *Journal of the Academy of Marketing Science*, *43*(3), 375–394. https://doi.org/10.1007/s11747-014-0388-3

Houston, M. B., Kupfer, A. K., Hennig-Thurau, T., & Spann, M. (2018). Pre-release consumer buzz. *Journal of the Academy of Marketing Science*, 1–23. https://doi.org/10.1007/s11747-017-0572-3

Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. In *Expert Systems with Applications* (Vol. 41, Issue 16, pp. 7653–7670). Elsevier Ltd.

https://doi.org/10.1016/j.eswa.2014.06.009

Kim, S. J., Maslowska, E., & Malthouse, E. C. (2018). Understanding the effects of different review features on purchase probability. *International Journal of Advertising*, *37*(1), 29–53. https://doi.org/10.1080/02650487.2017.1340928

King, R. A., Racherla, P., & Bush, V. D. (2014). What we know and don't know about online word-of-mouth: A review and synthesis of the literature. *Journal of Interactive Marketing*, *28*(3), 167–183. https://doi.org/10.1016/j.intmar.2014.02.001

Kleineberg, K.-K., & Boguñá, M. (2014). Evolution of the Digital Society Reveals Balance between Viral and Mass Media Influence. *Physical Review X*, *4*(3), 031046. https://doi.org/10.1103/PhysRevX.4.031046

Klostermann, J., Plumeyer, A., Böger, D., & Decker, R. (2018). Extracting brand information from social networks: Integrating image, text, and social tagging data. *International Journal of Research in Marketing*, *35*(4), 538–556. https://doi.org/10.1016/j.ijresmar.2018.08.002

Luca, M. (2015). *User-Generated Content and Social Media*. *1*, 563–592. https://doi.org/10.1016/b978-0-444-63685-0.00012-7

Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine Your Own Business: Market-Structure Surveillance Through Text Mining. *Marketing Science*, *31*(3), 521–543. https://doi.org/10.1287/mksc.1120.0713

Park, M. S., Shin, J. K., & Ju, Y. (2019). Attachment styles and electronic word of mouth (e-WOM) adoption on social networking sites. *Journal of Business Research*, *99*(September), 398–404. https://doi.org/10.1016/j.jbusres.2017.09.020

Tirunillai, S., & Tellis, G. J. (2014). Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation. *Journal of Marketing Research*, *51*(4), 463–479. https://doi.org/10.1509/jmr.12.0106

Vujić, S., & Zhang, X. (2018). Does Twitter chatter matter? Online reviews and box office revenues. *Applied Economics*, *50*(34–35), 3702–3717. https://doi.org/10.1080/00036846.2018.1436148

Whittaker, J., Looney, S., Reed, A., & Votta, F. (2021). Recommender systems and the amplification of extremist content. *Internet Policy Review*, *10*(2), 2022. https://doi.org/10.14763/2021.2.1565

Yang, S., Hu, M. (Mandy), Winer, R. S., Assael, H., & Chen, X. (2012). An Empirical Study of Word-of-Mouth Generation and Consumption. *Marketing Science*, *31*(6), 952–963. https://doi.org/10.1287/mksc.1120.0738

Zhou, W., & Duan, W. (2012). Online user reviews, product variety, and the long tail: An empirical investigation on online software downloads. *Electronic Commerce Research and Applications*, *11*(3), 275–289. https://doi.org/10.1016/j.elerap.2011.12.002

II. Summary of Chapters and Contributions

The present dissertation is made up of four studies covering the range of research avenues laid out in the preceding chapter. Whereas the first two are more methodologically focused and concern themselves with ways to extract information from user-generated content through data mining and sentiment classification, the other two investigate the production of user-generated content in the form of live-streaming and whether and how it informs the behavior of economic actors, respectively. In the following, the premises, goals and contributions of these studies will be briefly summarized.

Chapter 1 – User-generated content as a source of data

The first chapter concerns the problem of deriving brand-associations from massive textual data in the context of illustrating and investigating the complex brand-arrangements underlying the contemporary motion picture industry. Movies are commonly jointly produced by a conglomerate of entities, ranging from large studios and smaller, specialized production companies to individual professionals, such as producers, directors, writers and actors. Such individuals commonly receive prominent placement within marketing campaigns, wherein referrals to their past successes are used to signal the qualities of the new product. They are thus considered as exhibiting properties of brands as well (Kupfer et al., 2018), with material impact on the economic performance of the motion picture products they are involved in. Consequently, an extensive literature exists on the aforementioned concept of "star power" and its impact on product success and actor salaries (see, i.e., Joshi, 2015; Mathys et al., 2016; Thomson, 2006; Treme, 2010).

Recent years have seen a rise in importance of franchise properties, most prominently in the form of Cinematic Universes, which represent overarching mega-franchises connecting multiple smaller franchises. The sheer variety of franchises and number of produced titles, combined with a limited number of bankable actors have resulted in a situation where actors routinely portray characters across competing franchises, or even multiple characters within the same franchise. At the same time, the dispersion of intellectual property rights, which is especially prominent for the comic book properties currently dominating the box office, has resulted in a situation where properties belonging to the same overarching franchise are at times produced by competing studios. These complex entanglements make the movie industry a prime candidate to investigate consumer associations between brands. A common method for mapping and measuring such associations comes in the form of consumer-associative models

(Henderson et al., 1998) and brand-concept maps (John et al., 2006), both of which rely on extensive surveys. Since movies represent a popular topic of online conversation, the vast amounts of user-generated content produced about them might represent an alternative source of data which could be leveraged to elicit brand associations based on a statistical approach. User-generated content has been used in this way for market structure surveillance, wherein large amounts of text data are mined for relationships between brands, products and product attributes on an industry-level (Culotta & Cutler, 2016; Lee & Bradlow, 2019; Netzer et al., 2012). The goal of the present study was thus to extend these approached to the examination of brand associations. This required the collection of large amounts of data, in this case taken from a movie-specific subsection of popular news-aggregation site reddit.com. Since the goal of the study was to answer questions related to known or speculative properties of the investigated industry, a pre-identified model of the motion picture industry was built based on known relationships between studios, franchises, titles, actors and characters.

The study contributes to the literature in several ways. Firstly, it establishes that combining the aforementioned bottom-up approaches with a top-down approach based on a pre-identified model of an entire industry extends its capability beyond purely exploratory analysis. Since the methodology is not limited to the motion picture industry, it could be applied to investigate other industries with similarly complex brand arrangements in the future, including the automobile, consumer electronics and fashion industries. The study is also among the first in the marketing literature to use reddit as a source of user-generated content and the first to thoroughly characterize its functioning and qualities relative to other established sources. Regarding branding issues in the motion picture industry more specifically, it provides confirmatory evidence for several previously established findings, such as the positive relationship between cast star power and box office revenue. It further establishes that consumers actively associate studios with titles they produce and that actors are heavily associated with characters they portray, which - given the previously described issue of franchise density – should be of increasing interest to brand managers in the industry. In a similar vein, preliminary evidence that franchises affect associations between studios was found, though the relationship cannot yet be considered conclusively identified.

Chapter 2 – Natural language processing for user-generated content analysis

The second study is more methodologically focused, as it investigates how recent advancements in the field of natural language processing may be applied to the processing of textual data in a marketing research context.

As previously argued, text data represents the most common type of user-generated content in the marketing literature and the pervasiveness of its use as a source of information could thus be taken to imply that methods to process it are mature and well-understood. In contrast, while natural language processing represents a field of study with a rich history, it is still rapidly developing. As a result, text-processing methods are diverse and plentiful and can be roughly placed on a spectrum between rule-based and statistical approaches. Wherein rule-based language models aim to create systems that encode pre-identified rules governing the grammar of naturally occurring languages and are thus more closely related with the field of linguistics (Clark et al., 2010, pp.28), statistical models aim to map the conditional distributions of vocabularies and are situated in the field of machine learning (Goodfellow et al., 2016, pp.448). Statistical language models in particular have seen especially rapid advancements in recent years and have enabled – and are in turn driven forward by – newly emergent applications involving natural language processing, such as text-to-speech and speech-to-text methods for voice assistants, conversational chat bots and automated content classification.

The year 2017 saw the introduction of a new class of statistical language models based on the transformer architecture (Vaswani et al., 2017), which universally outperformed previous methods in a variety of benchmarks and saw immediate adoption in industry. A key attribute of these models that enabled these gains in performance, which coincidentally makes them especially promising for applied marketing research, is their comparatively high degree of generalizability. In a marketing research context, text data is commonly classified based on the sentiment it conveys (i.e. to create valence measures), which requires choosing the correct classifier. This in turn is highly dependent on the investigated topic and properties of the data and thus necessitates a series of informed decisions on part of the researcher. Since transformer models work in a variety of contexts, they should be able to be applied to a variety of datasets with only minimal tweaking required, while maintaining a high level of reliability in their classifications. To test whether this is indeed the case and to investigate how transformer models fit into the marketing researcher's toolbox, was the goal of the present study.

The study contributes to the extant literature on text-classification methods for marketing research in multiple ways. First, multiple transformer model's performance was evaluated on real datasets, rather than resorting to common benchmark datasets. Based on these findings, differences in behavior between different types of transformer model and different kinds of data could be identified, which ultimately informed the formulation of preliminary guidelines for their use by marketing researchers.

Chapter 3 – Content as product: The case of live-streaming

The third chapter, written in conjunction with Philip Wollborn and Ulrike Holder, takes a look at the frontier of user-generated content and its complex relationship to marketing. It investigates the field of live-streaming, an emergent form of user-generated content that exists on a spectrum between consumption and content production, with no clear line differentiating the two extremes (Johnson & Woodcock, 2019). In contrast to the other studies, which either build upon user generated content as a source of information or investigate methods to do so, this study concerns itself with user's motivation to actively produce content. It does so by investigating the behavior of individuals broadcasting on live-streaming services and quantifying how sensitive the decision to actively supply content relates to external circumstances. Specifically, the economic and sociocultural changes brought by the COVID-19 pandemic are leveraged to find supportive evidence of a causal relationship between opportunity costs and entrepreneurial activity, a special case of which we classify professional live-streaming as.

The study contains multiple noteworthy contributions to the growing literature on professionalized live-streaming. It is, to the knowledge of its authors, the first to characterize live-streaming as entrepreneurial activity and to propose determining factors of streamers' professionalization efforts based on theoretical considerations, which it argues are dependent on a mixture of outside factors and individual characteristics. Empirical evidence suggests that a change in external factors affects the supply of live-streaming content in the short term, whereas whether these changes are sustained in the longer term is determined by subsequent market feedback informed by individual characteristics. It further identifies ways in which professional streamers can monetize their activities and provides empirical data on the way professional streamers market themselves. As streamers personify the increasingly blurry distinction between content creation and marketing activity, understanding their motivations and needs can help companies to leverage their talents and audience reach in order to promote their products more effectively.

Chapter 4 – User-generated content as a driver of human behavior

The fourth study adds to the literature on how user generated content influences economic decision making. User-generated content distributed via social media represents a form of social information, wherein individuals base their own behavior on the observed actions of others in their environment. This phenomenon has been thoroughly investigated in the context of financial markets (Li et al., 2014; Tirunillai & Tellis, 2012; Yu et al., 2013), where it is considered to play a significant role in informing investor behavior, for example in the form of attention induced trading, extreme cases of which can affect overall market functioning in the form of herding events (Eaton et al., 2021; Pagano et al., 2020). A string of such herding events has recently put the spotlight on a group of investors referred to as zero-commission-investors, which is comprised of individual private investors making small trades using commission-free trading apps. A growing literature concerns itself with these zero-commission investors (Barber et al., 2020, 2021; Welch, 2020) and the goal of this study was to add to this by investigating the extent to which they are influenced by content submitted to social media relative to traditional news media. We combine high-resolution stockholder data with large amounts of news reports, as well as user-generated content gathered from trading-related reddit communities. By implementing the zero-shot classification methods explored in study two, we are able to extract stock-specific sentiment from the news articles and reddit comments which we subsequently use as explanatory variables in a panel regression framework. We show that social media activity is informative of zero-commission investors' trading decisions and more so compared to traditional news media, which we find to be especially pronounced for stocks with smaller market capitalization and penny stocks. We further find media volume to be more informative of zero-commission investor's trades than valence, which is in line with previous findings that they increase holdings in reaction to strong market movements in either direction.

References

Barber, B. M., Huang, X., Odean, T., & Schwarz, C. (2020). Attention Induced Trading and Returns: Evidence from Robinhood Users. *SSRN Electronic Journal*, *November*. https://doi.org/10.2139/ssrn.3715077

Barber, B. M., Lin, S., & Odean, T. (2021). Resolving a Paradox: Retail Trades Positively Predict Returns but are Not Profitable. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3783492

Clark, A., Fox, C., & Lappin, S. (Eds.). (2010). *The Handbook of Computational Linguistics and Natural Language Processing*. Wiley-Blackwell. https://doi.org/10.1002/9781444324044

Culotta, A., & Cutler, J. (2016). Mining Brand Perceptions from Twitter Social Networks. *Marketing Science*, *35*(3), 343–362. https://doi.org/10.1287/mksc.2015.0968

Eaton, G. W., Green, T. C., Roseman, B., & Wu, Y. (2021). Zero-Commission Individual Investors, High Frequency Traders, and Stock Market Quality. *SSRN Electronic Journal*, *March*. https://doi.org/10.2139/ssrn.3776874

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Henderson, G. R., Iacobucci, D., & Calder, B. J. (1998). Brand diagnostics: Mapping branding effects using consumer associative networks. *European Journal of Operational Research*, *111*(2), 306–327. https://doi.org/10.1016/S0377-2217(98)00151-9

John, D. R., Loken, B., Kim, K., & Monga, A. B. (2006). Brand Concept Maps: A Methodology for Identifying Brand Association Networks. *Journal of Marketing Research*, *43*(4), 549–563. https://doi.org/10.1509/jmkr.43.4.549

Johnson, M. R., & Woodcock, J. (2019). 'It's like the gold rush': the lives and careers of professional video game streamers on Twitch.tv. *Information Communication and Society*, *22*(3), 336–351. https://doi.org/10.1080/1369118X.2017.1386229

Joshi, A. (2015). Movie Stars and the Volatility of Movie Revenues. *Journal of Media Economics*, *28*(4), 246–267. https://doi.org/10.1080/08997764.2015.1094079

Kupfer, A.-K., Pähler vor der Holte, N., Kübler, R. V., & Hennig-Thurau, T. (2018). The Role of the Partner Brand's Social Media Power in Brand Alliances. *Journal of Marketing*, jm.15.0536. https://doi.org/10.1509/jm.15.0536

Lee, T. Y., & Bradlow, E. T. (2019). *Automated Marketing Research Using Online Customer Reviews*. *48*(5), 881–894.

Li, Q., Wang, T., Li, P., Liu, L., Gong, Q., & Chen, Y. (2014). The effect of news and public mood on stock movements. *Information Sciences*, *278*, 826–840. https://doi.org/10.1016/j.ins.2014.03.096

Mathys, J., Burmester, A. B., & Clement, M. (2016). What drives the market popularity of celebrities? A longitudinal analysis of consumer interest in film stars. *International Journal of Research in Marketing*, *33*(2), 428–448. https://doi.org/10.1016/j.ijresmar.2015.09.003

Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine Your Own Business: Market-Structure Surveillance Through Text Mining. *Marketing Science*, *31*(3), 521–543. https://doi.org/10.1287/mksc.1120.0713

Pagano, M. S., Sedunov, J., & Velthuis, R. (2020). How did Retail Investors Respond to the COVID-19 Pandemic? The Effect of Robinhood Brokerage Customers on Market Quality. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3703815

Thomson, M. (2006). Human brands: Investigating antecedents to consumers' strong attachments to celebrities. *Journal of Marketing*, *70*(3), 104–119. https://doi.org/10.1509/jmkg.70.3.104

Tirunillai, S., & Tellis, G. J. (2012). Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Science*, *31*(2), 198–215. https://doi.org/10.1287/mksc.1110.0682

Treme, J. (2010). Effects of celebrity media exposure on box-office performance. *Journal of Media Economics*, *23*(1), 5–16. https://doi.org/10.1080/08997761003590457

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-Decem*(Nips), 5999–6009.

Welch, I. (2020). *The Wisdom of the Robinhood Crowd*. NBER Working Paper. https://doi.org/10.3386/w27866

Yu, Y., Duan, W., & Cao, Q. (2013). The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems*, *55*(4), 919–926. https://doi.org/10.1016/j.dss.2012.12.028

III. Chapter 1

The impact of multi-studio franchises on brand-associations in the motion-picture industry

Abstract

Using an extensive dataset of more than seven million comments submitted to social news aggregator reddit.com covering a timespan of five years, brand associations within the motion picture industry are investigated. A network model of brands is proposed, wherein production studios are connected to individual products via intermediate brands such as actors, franchises and sub-brands. It can be shown that consumers strongly associate titles with their producing companies, as well as other titles produced by those companies, indicating a high salience of producer's brands in the minds of consumers. Similarly, actors are highly associated with titles they appear in and the characters they portrayed, underlying their importance to the marketing of motion pictures. Studios that are highly connected with franchises are also more highly associated amongst each other, implying that a new challenge to brand management has been introduced by the recent emergence of multi-studio film franchises.

1. Introduction

The last decade has seen tremendous changes in the motion picture industry, characterized by a trend towards increased market concentration and consolidation among incumbents on one hand, as well as by the entrance of new players in the form of streaming services and smaller sized production houses on the other. Whereas up until 2013 the big five production studios divided box office receipts relatively evenly amongst themselves, Disney has gradually taken a dominating position since, culminating in a close to fifty percent share following the acquisition of its previously biggest competitor, 20th Century Fox, in 2019 (see Figure III-1). Franchises have played a key role in this development, as emphasized by the fact that the Marvel Cinematic Universe alone is responsible for 9.37 percent of American box office receipts within that timeframe, with the Star Wars franchise consolidating another 4.07 percent, both of which are owned by Disney following the prominent acquisitions of Marvel Comics in 2009 and Lucasfilm in 2012 (BoxOfficeMojo, 2022; TheNumbers, 2022).

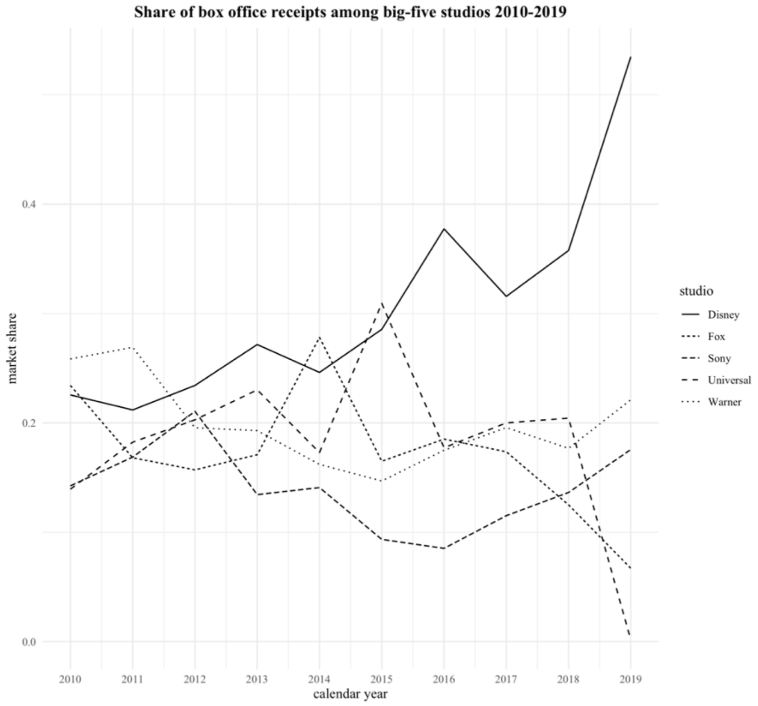


Figure III-1*: Share of U.S. box office receipts by studio. Source: Author, based on data from* BoxOfficeMojo (2022*).*

With the major studios increasingly focused on producing high-budget, tentpole releases aimed at the widest possible audience, the niche for small- to medium-sized productions was filled by more arthouse-oriented entrants such as Annapurna Pictures and A24. All the while, the landscape of movie distribution has itself changed, with the market for theatrical releases decreasing in (relative and absolute) importance in the wake of the advent of at-home streaming, itself accompanied by the entrance of new players in the form of Netflix, Amazon and Apple, amongst others (Esquire, 2021). In light of these developments, the contemporary motion picture industry is fertile ground for researchers interested in the role of brands in the marketing of experience goods.

Branding has long been considered a key aspect in the marketing of motion pictures, as brands play a crucial role in setting consumer expectations and signaling quality and familiarity. Motion pictures can be thought of as composite products that involve multiple cooperating brands, wherein aspects of the product itself - such as its placement within a franchise or the characters it features - and the companies and people involved in its production - such as studios, labels, actors or directors - can themselves be considered independent brands that together inform the product's overall identity (O'Reilly & Kerrigan, 2013). More broadly, they can thus be considered special cases of brand alliances (Kupfer et al., 2018). As previously mentioned, recent years have seen the emergence of cinematic universes, franchises-of-franchises, wherein each property is connected to other properties via shared themes, characters or locations and is marketed under an overarching brand. Their production and distribution routinely involves not only brands on different levels of the value chain, such as studios and actors, but also brands that are otherwise considered competitors. This can be a result of direct cooperation (e.g. Disney and Sony cooperating in the production of titles within the Spider-Man franchise) or indirect affiliation through secondary brands (e.g. Disney and Fox's indirect affiliation through the Disney-owned Marvel Comics, which is the overarching brand of the Fox-produced X-Men franchise).

The goal of this study is to shine a light on these complex arrangements by empirically measuring how brands in the motion picture industry are associated in the minds of consumers. Online media enable consumers to publicly communicate about products and brands and have become increasingly important platforms for brand management, both as a source of information and as a tool to engage consumers (Gensler et al., 2016). Lee & Bradlow (2011) leverage online product reviews to elicit product attributes and brand's relative positionings via text-mining techniques, while Netzer et al. (2012) similarly infer brand- and product-related associations from online discussion boards for automobiles and diabetes drugs. This study builds upon their work and applies it to the motion picture industry, while differing in some key ways. Most notably, the structure of the model is not inferred from the data in a bottom-up fashion, but pre-identified based on known industry data, thus enabling its use in the investigation of previously identified hypotheses rather than remaining purely exploratory. Rather than using product reviews or discussion forums as its source of text data, this study uses a large corpus of movie-related discussions gathered from the social-news aggregation platform *reddit*. Furthermore, the focus of the analysis is put squarely on associations between brands, rather than product attributes.

While a number of extant studies concern themselves with the mining of brand and product information from user generated content (such as the aforementioned Lee & Bradlow (2011) and Netzer, Feldman, Goldenberg, & Fresko (2012), as well as Culotta & Cutler (2016) and Klostermann, Plumeyer, Böger, & Decker (2018)), this study is, to the knowledge of its author, the first to do so within the context of the contemporary motion picture industry. Additionally, while the impact and dynamics of online user generated content has been intensively studied in the context of movies (see Babić Rosario, Sotgiu, De Valck, & Bijmolt (2016) for a recent meta-analysis), none of them have used data gathered from reddit, which has a number of unique properties that make it attractive for marketing research.

The study is structured as follows: In the second chapter, an in-depth analysis of the theoretical framework underlying brand-networks is used to derive the working hypotheses based on the previously laid-out research agenda. Chapter 3 covers the data collection process and gives a descriptive overview of the underlying reddit data, followed by the derivation of the brand-network model in chapter 4, which is subsequently analyzed in chapter 5 based on the previously identified hypotheses. The study concludes with a short discussion of its main results and an outlook on future research.

2. Theoretical framework and research agenda

Market structure analysis, wherein markets are systematically mapped to identify relationships between products, their attributes, brands and consumers, has been a staple tool for marketing research for the past several decades. More recently, established methods to elicit market structure such as multidimensional scaling and customer segmentation were joined by the large-scale analysis of user-generated content (Lee & Bradlow, 2011). While market structure analysis usually focusses on product characteristics, a distinct but related approach that focusses on brand-related issues can be found in consumer associative networks, which map qualities of products and brands that are associated in the minds of consumers (Aaker, 1996; Henderson et al., 1998; John et al., 2006). Whereas consumer associative models aim to assign qualities to relationships between network nodes, which aside from brand and product names and attributes can include aspects that are more conceptual in nature such as generic product categories and emotions (Henderson et al., 1998), the focus in this study is limited the strength of associations between nodes. In implementation, consumer-associative networks are derived from survey-based techniques such as repertory grids, making their creation a fairly involved and expensive undertaking. This invites the application of statistical language analysis as already established for market structure analysis (Lee & Bradlow, 2011; Netzer et al., 2012) to the elicitation of brand-associations. A fully text-mining based approach to this would be to look for salient word-pairs first and subsequently identifying brand relations from these candidates. However, a brand-network derived in this way will only ever include keywords that are co-mentioned often enough in order to become salient in the first place, ignoring those that fail to do so. The opposite approach would thus be to start with a known model of a given market or industry and subsequently measuring the strengths of connection between nodes from their co-occurrence, which in turn enables the model be used to investigate hypotheses about the relationship between nodes in a confirmatory, rather than strictly exploratory, fashion.

Regarding the contemporary motion picture industry, a brand-network derived in such way can be used to investigate a variety of as-of-yet uninvestigated lines of questioning, as long as it is sufficiently validated. The approach in this study is thus to first analyze the network model based on known attributes of the investigated industry. Contingent on the model's ability to capture these known properties, further research questions may then be investigated. One such well-established relationship is that of actors, who routinely partake in the marketing of titles they appear in and actively manage their own brands by associating themselves with certain roles (Kupfer et al., 2018). It would thus follow, that consumers should more strongly associate actors with those titles they appear in, compared to titles they do not appear in and that the same holds for characters they have portrayed. By analogous argument, titles should be more strongly associated with the studios involved in their production as compared to other studios, given the prominent placement of producing companies' logos in i.e. promotional material. Using these three proposed hypotheses as a baseline, newlines of questioning regarding the role of franchises may be pursued, such as the question of whether the complex arrangements of brands involved in the production of motion pictures can lead to spillover between brands engaged in such cooperative efforts, via their (direct or indirect) association in the minds of consumers. Brand spillover is commonly defined in the context of a weaker and stronger brand, wherein the weaker brand will emphasize its similarity to the stronger one in order to benefit from the association (Wu et al (2021) investigate such constellations before the background of common suppliers between the brands). In context of the motion picture industry, spillover could present itself in a different way. As previously argued, the pervasiveness and complexity of modern franchises routinely result in multiple studios producing different sub-franchises within an overarching mega-franchise. Within such a constellation, if both studio's brands are tied to the mega-franchise and the sub-franchise is itself tied to the mega-franchise, each studio's brands could be dependent on the other studio's operational decisions, since each ostensibly competing studio now has a vested interest in the other studio increasing the brand value of the common mega-franchise (as postulated in industry publications, this aspect has been assumed to be a driving factor in Disney seeking increased creative control over the Spider-Man franchise (The Hollywood Reporter, 2019)). To investigate whether such contrivances exist and to gage whether and how franchises mediate the association strength between brands of competing studios, it is first investigated whether studio brands are at all associated with franchises under their domain, followed by an exploration of the strength of association between studios' brands. Table III-1 presents a formal overview of the proposed working hypotheses.

Table III-1 *Overview of research hypotheses.*

Investigated relationship	**Working Hypothesis**
Actors and titles	**H1:** Actors are strongly associated with titles they starred in.
Actors and characters	**H3:** Actors are strongly associated with characters they portray.
Titles and studios	**H3:** Titles are strongly associated with studios involved in their production
Studios and the franchises	**H4:** Studios are strongly associated with franchises they produce.
Studios and other studios	**H5:** Studios are more strongly associated with other studios they share a franchise affiliation with.

Another aspect that is to be investigated is whether the network-model lends itself to the derivation of measures of brand equity. Aaker (1996) defines brand equity as the sum-total of brand awareness, customer loyalty and - positive or negative – brand associations that increase or decrease the value of a product marketed under a given brand. Within context of the present study, the quantity of mentions can be considered a measure of brand awareness, which has been found to be positively correlated with economic success for actor brands (Joshi, 2015; Treme, 2010). It is thus of interest whether these results can be replicated and whether a similar effect can be identified for producing studio's brands.

3. Data and descriptive analysis

The following chapter gives an overview of the data collection process and descriptively analyzes the collected data. Following an introduction to reddit as a data source (3.1), the data collection and processing is described in detail (3.2) and the collected data explored (3.2).

3.1 Reddit as a source of user-generated content

User generated content has become a staple source of data for marketing research over the past two decades. Popular targets for data collection have been twitter (Vujić & Zhang, 2018), Instagram (Klostermann et al., 2018) and static web forums (Netzer et al., 2012). A platform that has so far been underrepresented in the marketing literature compared to other fields, is reddit. As a social news aggregation and discussion platform, it allows registered users to submit content - such as links to other websites, images or videos, as well as self-authored text - which is subsequently voted on and discussed by other users. Users can *upvote* or *downvote* a given submission, affecting its ranking relative to other submissions. The ranking is determined by the score of the submission (downvotes subtracted from upvotes) weighted by its age as measured by the time in minutes passed since submission (Stoddard, 2015). Each submission to reddit includes its own comments section, wherein users can discuss the submission's content. The top 50 ranked submissions at any given moment appear on the *frontpage* of reddit, which commonly results in a large boost in exposure for submission that are featured this way. Apart from the general frontpage of reddit, users self-organize into specialized communities referred to as *subreddits*. This makes reddit very conductive to researchers investigating specific topics, such as motion pictures, as it narrows down the ground that needs to be covered to comprehensively reflect the discussion of a certain topic on reddit and ensures that discussions within a given subreddit are at least tangentially related to its theme. In a way, reddit thus shares the advantages of static discussion forums, which are highly topic specific, but limited to a certain audience that is unlikely to be representative of the general population, since participation requires people to sign-up to the specific forum. Similarly, it retains the advantages of microblog services such as Twitter, in the way that all content is in principle discoverable by all users on the platform via organic dissemination and propagation mechanisms (network effects in the case of Twitter and ranking in the case of reddit). Another aspect which makes reddit conductive to marketing research is that submissions are voted on and ranked by their score, giving additional valuable information on how a given posting has been received. The fact that submissions to reddit can also reach a negative score represents a marked advantage compared to other scoring mechanisms used on i.e., Twitter, which are based on favorites or retweets which cannot go negative. Thus, it is impossible to determine whether a tweet was actively disliked, or simply unseen. Reddit's scoring system, in contrast, gives a more fine-grained measure for the agreeableness of a given message, which can be leveraged to identify "loud minority" type effects, wherein expressions of opinion are not representative of the overall population. Lastly, comments posted to reddit are markedly longer compared to tweets, which are limited to 160 characters and thus allow for higher semantic complexity and information within a single body of text, whereas the subject of tweets must often be inferred from the surrounding discussion context.

3.2 Data collection and processing

Daily North American box office data starting from January 1st 1993 were collected on 15th February 2022 from boxofficemojo.com (BoxOfficeMojo, 2022). Additional information was taken from the IMDb data repository on 12th January, 2022 (IMDb, 2022). This included information on all movie titles, actors and their associated roles in IMDb's database.

This study uses two distinct sets of data collected from reddit. The first set was collected in real time over a period covering the full calendar year of 2018. The dataset consists of a total of 438,725 hourly observations of 76,231 unique submissions to the frontpage of reddit, as well as the movie-focused subreddit */r/movies*, collected via hourly snapshots. These snapshots provide information on all submissions visible at the time of collection on the respective front pages and their ranking relative to each other, thus enabling insights into the dynamics of reddit submissions. The postings' unique identifiers were subsequently used (via the official Python Reddit API Wrapper (Boe, 2012)) to retroactively retrieve all comments pertaining to each movie-related post on January 28th, 2019, totaling 4,116,944 unique comments by 546,110 authors. An auxiliary set of reddit submission data was collected in retroactive fashion in January of 2022 from the pushhift.io service (Baumgartner et al., 2020). This second dataset covers all submissions made to reddit (including those which did not reach the frontpage) for the calendar years of 2017, 2019, 2020 and 2021, for which again all available comments were subsequently retrieved via reddit's API. Figure III-2 represents a process diagram of the data collection and processing pipeline.

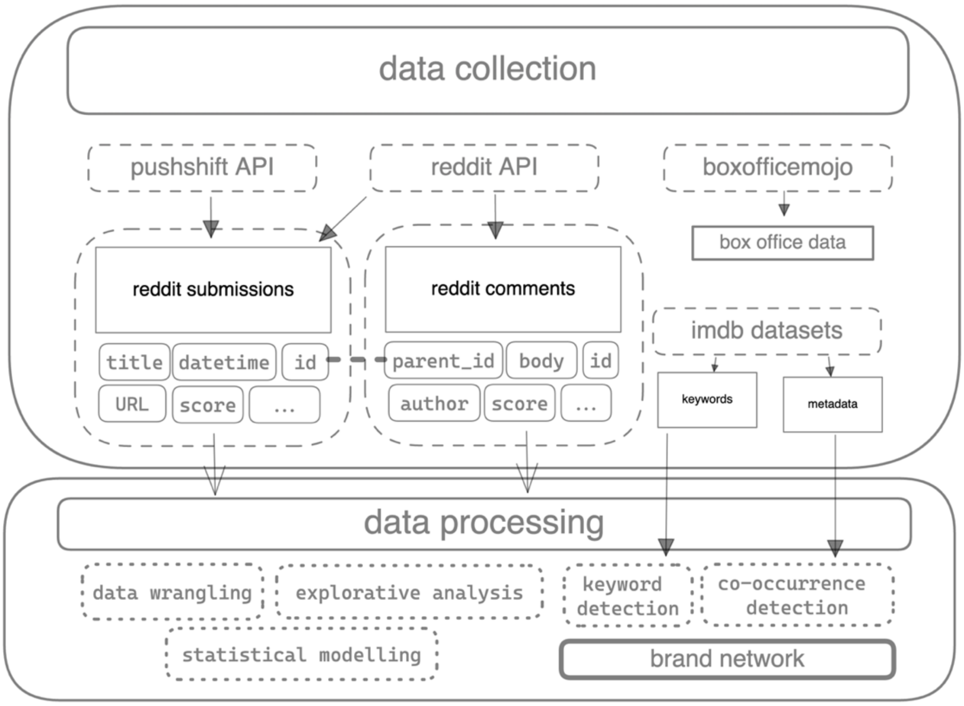


Figure III-2 *Data collection and processing diagram*

3.3 Descriptive statistics of reddit data

As briefly explained in section 3.1, two distinct types of reddit data were collected in the form of submissions and comments. *Submissions* refer to posts to reddit, which usually consist of a URL linking to a piece of content outside of or within reddit but can also be a self-authored submission text (so-called self-post), whereas comments make up the discussion section under each submission. Each submission and comment consists of its own URL, which can be used to retrieve its respective data and metadata. While the majority of these are technical in nature and immaterial for the analysis at hand, data points of interest include the date and time of submission, author name and identifier, submission headline, its score at time of collection and number of comments. Table III-2 gives a descriptive overview of all collected submissions and comments, as well as the distribution of some of their related measures. Most submissions made (69.94%) have a score of 1 (which is the default score) and a median comment count of 0, indicating no engagement by other users. In general, the distribution of both score and number of comments is extremely long-tailed, which is expected, since data collected from social media commonly follow power law distributions owing to preferential attachment, colloquially referred to as the Matthew effect or simply "rich-get-richer" (Johnson et al., 2014). In case of reddit, this is likely exacerbated by the underlying ranking algorithm, which gives already successful submissions higher visibility.

Table.III-2 *Descriptive statistics of reddit submissions and comments.*

year	obs	unique authors	score			comments		
			Min	Mean	Max	Min	Mean	Max
Submissions	160,752	60,629	0	293.7	118,554	0	46.33	69,494
Comments	7,663,214	943,172	-1465	19.82	43,654	(-)	(-)	(-)

The high correlation between a submission's score and its number of comments (Pearson's *r* of 0.526[0.525-0.532]) similarly indicates that visibility drives engagement and vice-versa. Looking at the score distribution for comments, it is immediately noticeable that negative scores are much smaller in magnitude compared to positive scores (minimum of -1465 vs. maximum of 43,654), which similarly is likely a result of the ranking system hiding downvoted posts while giving prominently displaying already popular posts. Regarding the actual content of submissions, more than a third of submissions (36,87%) were so-called self-posts, i.e., texts authored by the submitting user equivalent to discussion posts in a static web forum. Of the posts linking to off-site online content (90,700 URLs in total), YouTube URLs represent by far the most common destination (27,686 or 30.51% of all outbound links), followed by common film-related news sites Hollywoodreporter, Variety and Indiewire (1946, 1014 and 833 occurrences, respectively).

Of special interest to the analysis at hand are submissions related to marketing activity. These may include promotional material submitted either organically or through active engagement by marketers on reddit, such as trailer videos, posters and cast announcements. As presented in Table III-3, such submissions create markedly higher engagement compared to the average submission. For marketers within the motion picture industry, such submissions can provide valuable metrics on how promotional materials are perceived by (prospective) consumers and inform further marketing activity, while review and box office related discussions can provide early post-release feedback.

Table III-3 *Descriptive statistics for marketing-related submissions*

type	**obs**	**score**			**comments**		
		Min	Mean	Max	Min	Mean	Max
Trailer release	12,241	0	327.4	96,414	0	53.2	14,917
Poster release	4220	0	1,983	102,423	0	146.5	14,930
Cast announcement	11,080	0	627.2	97,895	0	78.25	22,636
Awards discussion	2,529	0	674.6	99,357	0	89.99	21.892
Box Office discussion	1,246	0	769.5	88,908	0	109.8	8,630
Review discussion	2,875	0	152.4	61,171	0	33.47	5,181
All submissions (for reference)	160,752	0	293.7	118,554	0	46.33	69,494

4. Construction of the network model

The following chapter covers the construction of the network model from underlying industry data. Since the industry data is extremely detailed and comprehensive (with more than 12 million potentially relevant keywords) and the number of user comments to be scanned for mentions is similarly large (at more than seven million observations), computability quickly becomes a concern, requiring a narrowing-down of keywords to those relevant to subsequent analysis (subchapter 4.1). Once relevant keywords are identified, the comment data is scanned for their occurrence (subchapter 4.2). Previously postulated relationships between keywords are then quantified based on co-occurrence within the same textual unit and further measures constructed that take baseline probabilities into account (subchapter 4.3).

4.1 Identification of relevant keywords

While the mere detection of mentions scales linearly with the number of keywords and processed comments, the computation of co-mentions grows exponentially in the number of keywords, with the exponent depending on the order of the relationship of interest (i.e. quadratic for dyadic relationships, cubic for triadic and so on). Thus, a contingency table of relationships between n keywords comprises a (usually sparse) matrix of dimension $(n\ x\ n)$ containing n^2 observations, of which only the upper or lower triangle (comprising $\frac{n^2}{2} - n$ observations) are informative if the measured relationships are symmetric in nature (as is the case for co-mentions). However, since the majority of dependencies between keywords is functionally irrelevant or of lessened interest to the research at hand (i.e. the relationship between different movie titles or among actors), the number of investigated relations can be significantly reduced a priori. Similarly, the number of keywords can be limited to only those related to products (in this case motion pictures) that require certain criteria, such as release within a given period or type of distribution. Since economic success was a factor to be investigated in downstream analysis and such data is not available for titles on i.e. streaming services, it was decided to limit analysis to titles that have received a theatrical release by a major studio or publisher. Thus, based on the box office data, the top 30 movie studios and publishers between the years 2000 and 2021, as measured by total theatrical gross, were first identified, along with all titles released by these top 30 studios. The resulting list of titles was then matched against the IMDb data to connect associated characters and actors. The relationship between characters and the titles they appear in enabled the inference of franchises, as movies that share (non-ambiguously named) characters can be considered likely to be part of a common franchise. This narrowed down the list of potential franchise-entries to a more manageable 431, for which connections to franchises and production labels were subsequently researched. The described approach minimizes the amount of manual research needed to construct a comprehensive set of relevant keywords, while providing a pre-structured model of the investigated industry to test previously formulated hypotheses against. The structure chosen for the model at hand features studios (which finance and distribute movies), sub-brands (usually direct subsidiaries of studios focused on the development of specific brands and franchises), titles (the motion picture products), franchises (overarching brands of multiple connected titles), as well as actors and the characters they portray. Overall, 7,884 unique keywords were identified, constituting 1888 titles, 29 studios, 6 sub-brands, 27 franchises, 1257 characters and 4687 actors.

4.2 Detection of keyword mentions and co-occurrences

As a first step in analyzing the comment dataset for mentions of the aforementioned keywords, the unit of observation that serves as the basis for analysis had to be chosen. Netzer et al. (2012), who obtained data from static discussion forums, identify discussion threads (sets of messages, ideally about a shared topic), messages (sets of sentences) and sentences (sets of words) as possible observational units, all of which have their functional equivalent within the reddit data. Thus, analogous to Netzer et al. (2012), the message level was chosen as the observational unit for analysis. All previously identified keywords were subsequently matched against all collected comments’ text bodies using case insensitive direct matching. The list was then manually checked for and cleaned of keywords that were likely to have triggered a high number of false positives due to their ambiguity or use as common idioms (i.e. “Anything”, “Wonder”, “Bigger”, “Zero”). Table III-4 shows the absolute and relative number of mentions for the 10 most-mentioned studios. When considering the share of mentions among studios as a rudimentary measure of mind share and comparing it to market share as measured by share box office, Disney is the only major studio with a higher mention share than market share (2.48x). Only independent and arthouse publishers A24 and Neon have higher ratios when excluding streaming services Amazon and Netflix, which is to be expected given that users that are active on a movie-specific discussion board will skew towards more niche titles.

Table III-4 *Absolute and relative mentions and box office share for Top 10 studios by number of mentions. *Cleaned for mentions of unrelated overlapping keywords **Streaming services; Amazon manages limited theatrical releases of its productions itself whereas Netflix partners with Paramount.*

Rank	Studio	Mentions	Mention share	Market share	Ratio
1	Disney	71,822	0.425	0.172	2.48
2	Netflix**	43,451	0.257	(-)	(-)
3	20th ct. Fox*	13,583	0.081	0.110	0.733
4	Sony	10,651	0.063	0.0993	0.635
5	Amazon**	9305	0.055	0.0001	550
6	Universal	6132	0.036	0.115	0.317
7	Warner Bros.	5146	0.031	0.131	0.232
8	Paramount	2371	0.014	0.0764	0.184
9	A24	2365	0.014	0.0023	6.16
10	Neon*	1454	0.009	0.001	11.1

4.3 Derivation of model weights from keyword co-occurrences

Co-occurrences within the same body of text were used to derive the degree to which pairs of keywords are associated. In principle, a variety of methods can be used to quantify these dyadic relationships, such as the Jaccard Index, Lift, term frequency–inverse document frequency (TF-IDF) or odds-ratio, since all of the above measures of direct co-occurrence are almost perfectly correlated (Netzer et al., 2012). Analogous to Netzer et al. (2012), the Jaccard Index was chosen as a measure of association, which was subsequently computed for all dyadic keyword relationships. In principle, the Jaccard index simply measures the ratio of the intersection of two sets (in this case, comments where two keywords are mentioned together) and their union (comments where either is mentioned):

$$J_{ii'} = \frac{\sum m_{ii'}}{\sum m_i + \sum m_{i'} - \sum m_{ii'}} \tag{III-1}$$

The left half of Figure III-3 shows a high-level overview of the network structure derived from these associations, pruned to only show nodes connected to studio nodes and visualized using a force-directed graph algorithm using the measured JI as edge weights. The right half of Figure 1 shows a magnified view of the central cluster. Red and green fill colors denote studio and sub-brand nodes, respectively. Titles are generally clustered around their producing studios, with franchises and sub-brands acting as bridges between studios. Actors are connected to titles they appear in, as well as other actors connected to those titles and the characters they portray. The major studios are heavily clustered, with Disney standing out as a major stand-alone structure. Streaming services (Netflix and Amazon) and arthouse-oriented studios (A24 and Annapurna Pictures) similarly form their own distinct clusters.

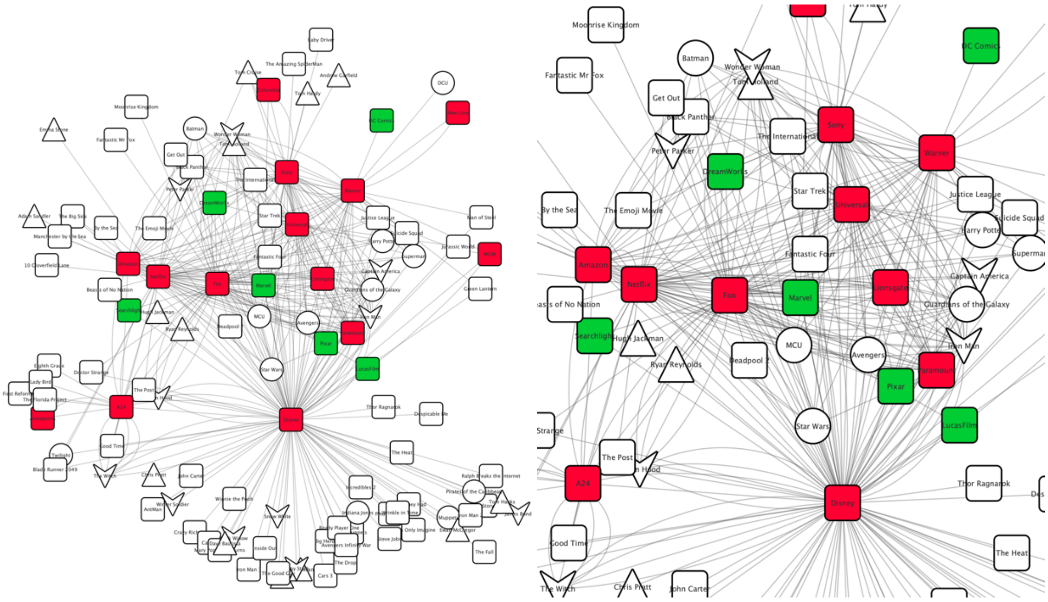


***Figure III-3** Illustrated network graph of the brand model. Left: Overall topology. Right: Zoomed portion of the core cluster.*

5. Analysis of the brand-associative model

The following chapter assesses the previously identified model's usefulness in answering a variety of research questions in the field of motion picture marketing. The first chapter validates the model based on *a piori* assumptions and briefly explores its topology (5.1). This is followed by an exploration of the role of franchises on studio brand associations (5.2), as well as of the relationship between of studio and cast brand awareness on economic variables (5.3).

5.1 Model validation and exploration

In order to ensure that the model is fundamentally qualified to capture associations between keywords, it is first evaluated based on face validity. As laid out in chapter 2, the first two Hypotheses act as validating hypotheses given that they represent well-established findings that the present model should thus be able to replicate. To test this assumption, associations are first grouped by the type of relationship they signify (i.e. actors and titles they appear in, actors and characters they portray, studios and titles they produced and so forth). Subsequently, the distribution of the association measures within these group-conditions is compared to the distribution of associations within the group of relationships forming the logical opposite condition (i.e. associations between actors and titles they appear in are compared to associations between actors and titles they do not appear in).

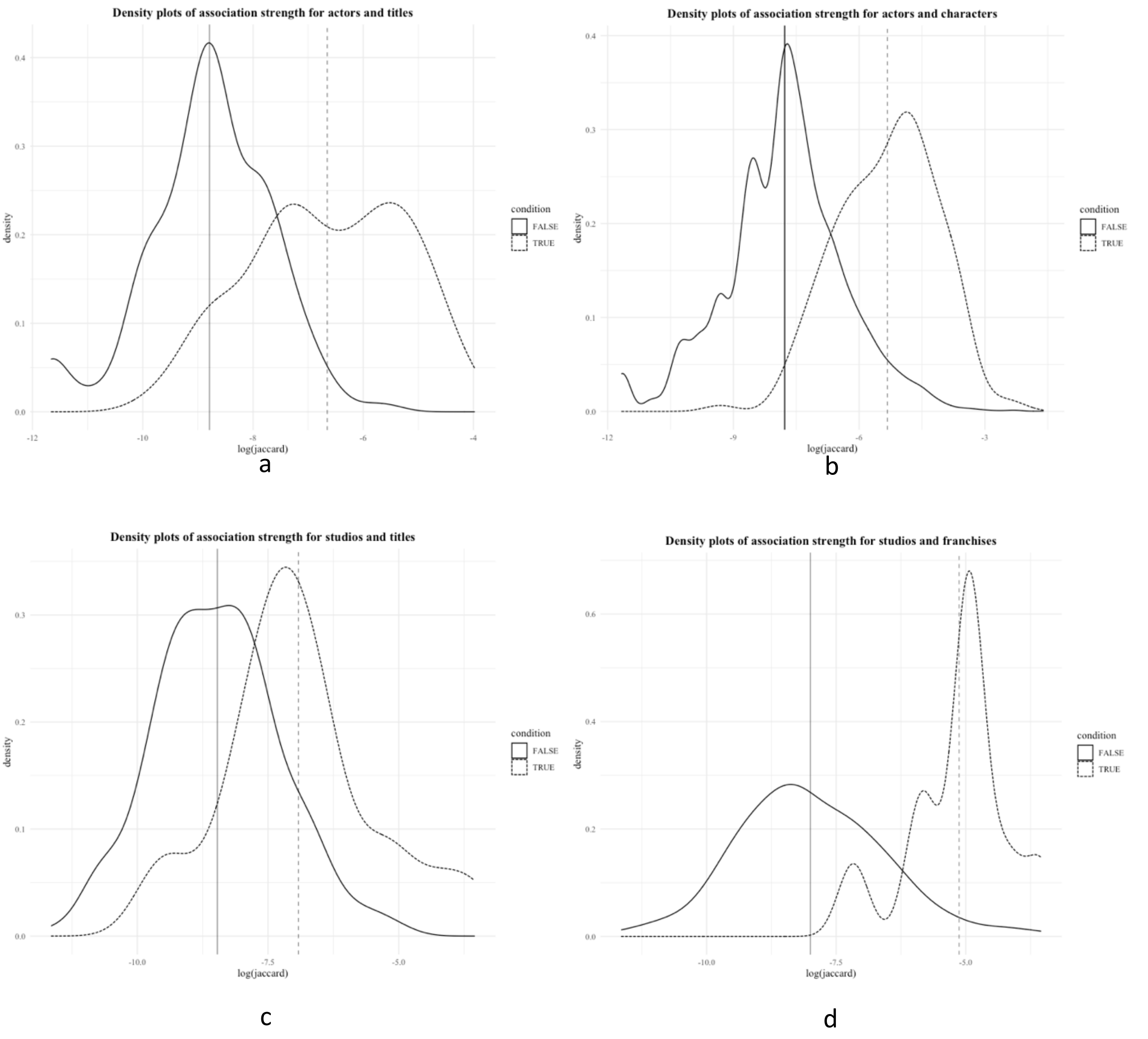


Figure III-4 *Density plots of relationship types, vertical lines represent means for each condition.*

Figure III-4 shows density plots for four such relationship types, wherein the dotted solid and dotted lines represent the distribution where a relationship condition was met or not, respectively. Figure 3a shows the density of (log) Jaccard associations for actors and titles they did not act in (solid line), as well for actors and titles they did act in (dotted line). While the both conditions overlap and are long-tailed, the latter is noticeably right-shifted (with an estimated location shift of 2.137[95% confidence interval: 1.79-2.48] as computed using Welch's two-sample t-test) and the majority of its probability mass lies to the right of the former. This provides evidence in favor of **H1**, which postulated that actors should be more highly associated with titles they act in compared to those they do not. The same is the case for actors and characters they do or do not portray, as presented in Figure 3b, with an estimated difference in means of 2.45[2.264-2.637], giving credence to Hypothesis **H2** which stated that actors are highly associated with characters they portray. As to the relationship between studios and titles they did or did not produce (**H3**), associations between titles and their producing studios are again significantly stronger, though the right shift is less pronounced compared to the previously compared relations (difference in means of 1.553[0.86-2.245]) The opposite is the case for studios and franchises: studios are significantly more heavily associated with franchises in whose production they are involved compared to those they are not, providing evidence in favor of **H4** (difference in means of 2.87[2.14-3.6]). Overall, the model seems to be able to accurately reflect relationships between brands in the motion picture industry, warranting deeper analysis.

Table III-5 *Top 10 co-occurrences of keyword-pairs, relationships by node type and degrees by node.*

Top co-occurrences				**Strongest relationship types**			**Highest degrees**	
keyword one	**keyword two**	**n**	**Jaccard**	**type one**	**type two**	**mean Jacc.**	**keyword**	**degree**
Star Wars	Disney	6032	0.047	subbrand	subbrand	7,26	Spider-Man	1158
Batman	Superman	5961	0.120	character	character	5,98	Iron Man	762
DC	Marvel	5326	0.061	actor	actor	4,01	Marvel	673
MCU	Marvel	4366	0.047	studio	studio	3,72	Disney	671
Marvel	Disney	4108	0.030	franchise	subbrand	3,44	X-Men	644
Avengers	Marvel	3480	0.040	subbrand	franchise	3,44	Batman	633
Star Wars	Marvel	3140	0.025	studio	subbrand	2,85	Netflix	620
Fox	Disney	2899	0.035	subbrand	studio	2,85	Wonder Woman	608
Batman	DC	2473	0.043	franchise	franchise	2,09	Harry Potter	598
Iron Man	Marvel	2296	0.0253	subbrand	title	2,04	Star Wars	565

Table III-5 gives an overview of the top ten dyadic relationships, as well as the ten most highly connected nodes in the network as denoted by their degree (the number of unique nodes they are connected to). The most commonly observed keyword-pairs are between studio Disney and its Star Wars franchise, followed by the characters of Batman and Superman. The dominance of comic book properties is immediately apparent, although it should be noted that the simplicity and lack of specificity of the keywords (most of which are both characters or franchises, as well as subsets of specific titles) likely plays an important role in this. When aggregating by type of relationship (i.e. by the classifications of keywords), it can be seen that associations between sub-brands, franchises, as well as franchises and sub-brands are particularly strong, as are those between studios and sub-brands and studios. The top ten highest-degree nodes in the network are again dominated by comic-book properties, with Spider-Man, Iron Man and Marvel as the most highly connected nodes.

5.2 The impact of franchises on studio brand-associations

As previously established, studios are more heavily associated with franchises they produce compared to franchises produced by competing studios. Within the context of the contemporary motion picture industry, which is characterized by a high degree of complexity in the production of large franchises, an open question remains whether studios that are involved in joint productions of franchise properties are also more highly associated with each other in the minds of consumers. A cursory look at specific instances of multi-studio cooperation as presented in Table III-6 seems to initially confirm this notion. For example, Sony is strongly associated with Disney both directly. Similarly, Fox's association is with Disney, though this is likely affected by the former's acquisition by the latter, which took place in the covered timeframe. However, it is also heavily associated with the Marvel sub-brand, as well as the X-Men property. The X-Men franchise, while most strongly associated with Fox, which is its producing studio, and Marvel, which is its overarching sub-brand, is a also highly associated with Disney, which has so far not been directly involved in the X-Men franchise. In summary, based on the limited amount of investigated instances of joint franchise production, studios that are indirectly affiliated via shared franchises are significantly more strongly associated compared to studios they do not share such affiliations with. Furthermore, the properties at the heart of these relationships are strongly associated with these studios, even when they had no active part in their production.

Table III-6 *Associations between studios, brands and franchises involved in multi-studio productions*

Brand one	**Brand two**	**Jaccard**	**Percentiles**	**Rank**
Disney	Sony	0.011	99/97	7/6
Disney	Fox	0.035	100/100	2/1
Marvel	Disney	0.03	100/100	4/3
Marvel	Sony	0.025	99/100	6/2
Marvel	Fox	0.015	99/99	10/4
Spider-Man	Disney	0.003	97/97	10/17
Spider-Man	Marvel	0.012	100/98	3/13
Spider-Man	Sony	0.014	100/98	1/4
X-Men	Disney	0.007	99/99	6/10
X-Men	Fox	0.029	100/100	1/2
X-Men	Marvel	0.017	100/99	3/8

To further quantify and generalize the relationship between franchise production and studio brand association strength, a new measure of franchise connectedness between studios was devised. The franchise association score (FS) was computed as the cosine similarity between the vectors of franchise associations for two studios, as given by the previously computed Jaccard index values for each studio and franchise in the sample. For each pairing of studios s and s' in the sample, the cosine similarity between vectors J_s^f (containing the associations between studio s and the vector of franchises f) and $J_{s'}^f$ (containing the associations between studio s' and the vector of franchises f) was computed as follows:

$$FS_{ss'} = \frac{J_s^f * J_{s'}^f}{\sqrt{\Sigma {J_s^f}^2} * \sqrt{\Sigma {J_{s'}^f}^2}} \qquad \text{(III-2)}$$

The measure takes on values between 0 and 1, wherein a higher value can either signify a high common franchise association (i.e. both studios are highly associated with the same franchises, for example Sony and Fox) or a low common franchise association (i.e. both studios share their disassociation with franchises, for example A24 and Annapurna). Overall, there is a positive relationship between the strength of association between studios (as measured by the Jaccard index) and their FS, as Figure III-5 illustrates.

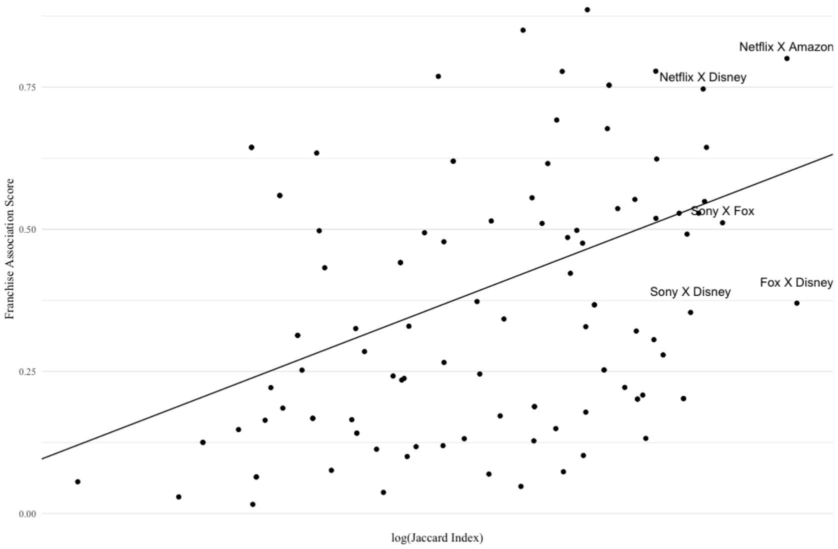


Figure III-5 *Plot of association strengths between studio pairs and their franchise association score*

In a log-linear specification the estimated slope coefficient for the FS is 2.301 ($p<0.001$), though a more robust sampling-based estimation based on STAN as implemented in the brms package for R (Bürkner, 2017) reveals a smaller estimate of 0.71 whose 95% highest-density interval does not exclude zero (Table III-7 and Figure III-6). In summary, while the evidence points towards the existence of a positive relationship between studio brand's relationships to franchises and their shared association with other studio brands (providing supportive, albeit weak, evidence in favor of **H5**), its strength could not be sufficiently determined owing to a small sample size.

Table III-7 *Regression results for the relationship between studio associations and FS*

Covariate	Estimate	Est. Error	l-95% HDI	u-95% HDI
Intercept	-7.16	0.28	-7.72	-6.61
FS	0.71	0.52	-0.32	1.74

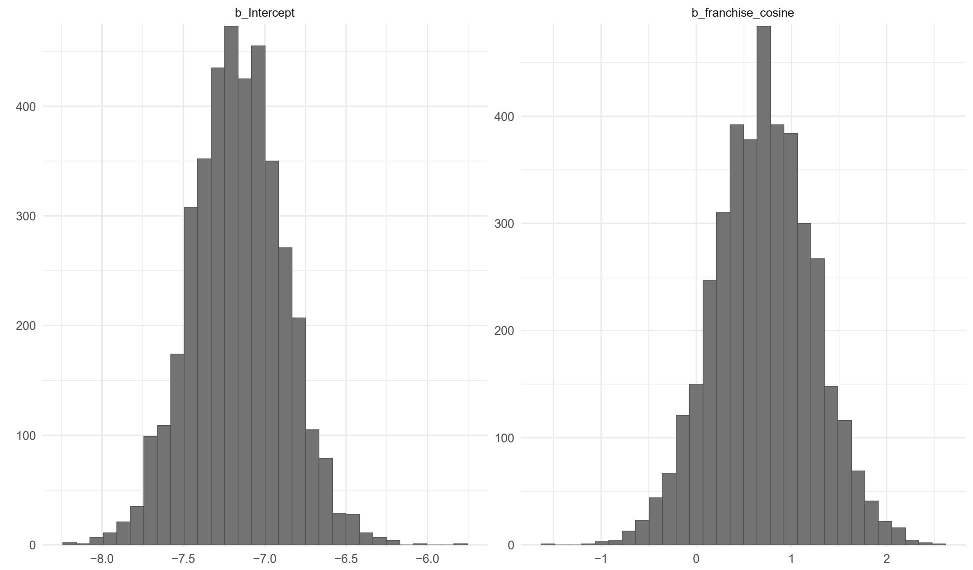


Figure 1.6 *Posterior density plots for model coefficients*

5.3 Studio and cast brand awareness and box office revenue

To further evaluate the usefulness of the brand-associative model presented in this study in answering questions in the field of motion picture marketing, it was explored how measures derived from the underlying reddit data in combination with the model's representation of industry structure can be leveraged in the modelling of box office success. Though the way motion pictures are distributed has seen significant changes over time with the advent of home video in the 80s and the more recent experimentation with streaming-first releases effected by the COVID-19 pandemic, theatrical releases still represent the default initial mode of distribution for high-budget productions and box office receipts remain the largest contributor to life-time sales for that category (Lang & Rubin, 2022). Due to this economic importance and availability of high-quality data, box office modeling has been very popular with researchers, who produced a variety of highly sophisticated modelling schemes (Ghiassi et al., 2015; Sawhney & Eliashberg, 1996; Swami et al., 1999). The express goal for this exploration was thus not to improve upon existing state-of-the-art box office forecasting models. Instead, the baseline specifications were chosen to reflect commonly included controls established in the literature, such as genre fixed effects, user rating and the number of screens at release. Since the number of screens a title is exhibited on is adjusted in reaction to the previous week's revenue, it has been found to be contemporaneously correlated with revenue (Elberse & Eliashberg, 2003), which is why, for example, Treme (2010) uses a two-stage approach to avoid the ensuing endogeneity. However, this should only be an issue in longitudinal designs, as otherwise the number of screens at first release is merely a (marginally more informative) metric for the type of release (limited or wide). The main variables of interest which were derived from the underlying graph model are measures of brand awareness in the form of mention counts, of which two distinct types were devised. The first type covers the studio brand awareness and is simply defined as the sum of mentions of a given studio in a given year. This measure replaces fixed effects for studios that are commonly used in the literature (Packard et al., 2016). The second type covers the brand awareness surrounding the cast of a given movie. It is well established that a movie's cast significantly influences its economic prospects (see, for example, Joshi, 2015; Mathys et al., 2016; Treme, 2010) and it would thus be of interest whether a measure of cast brand awareness derived from the present brand-network model reflects this relationship. To create this measure, all mentions of all actors connected to a given title within the year of release for said title are accumulated, reflecting the total number of mentions of a given title's cast in the year it was released.

Table III-8 *Summary statistics for regression variables*

Variable	**N**	**Mean**	**Std. Dev.**	**Min**	**Pctl. 25**	**Pctl. 75**	**Max**
Total box office gross	336	80,244,855	116,306,756	5,559	12,071,405	101,281,485	858,373,000
First week box office gross	336	34,508,929	54,602,211	5,559	4,207,571	37,593,127	473,894,638
Theaters at opening	336	2,504.16	1,490.95	2	1,484	3,654	4,725
Average IMDb user rating	336	6.44	0.98	2.2	5.8	7.1	8.4
Days in theaters	336	84.29	87.84	0	55	97	1,061
Studio mention count (year)	336	2675.14	6744.05	1	65	2214	41,138
Cast mention count (title)	336	336.07	777.25	1	29	353.5	10,516

Table III-8 shows summary statistics of all variables used in the subsequent regression models. In all, four models were computed, which differ in the choice of dependent variable (total box office gross vs. opening gross, both taken in logs) and the inclusion of the variables measuring brand awareness, the results of which are presented in Table III-9. Models (1) and (3) represent baseline models of their respective specification (model (1) includes the number of days a title was exhibited in theaters as a further control), whereas models (2) and (3) add the brand awareness measures. Total box office revenue is much harder to predict due to its higher variance compared to first week gross, as is reflected by the difference in model fit (adj. R^2 of 0.547 vs. 0.809 under the baseline specifications). The addition of the brand awareness variables barely improves model fit over the baseline, though the coefficient for cast mentions is positive and significant for both specifications, as well as economically meaningful (as one percent change in the number of cast mentions is associated with an expected 0.16 percent change in total box office revenue and 0.09 percent change in first week gross). The situation is less clear-cut in the case of studio mentions, which was found to have no significant effect on total box office gross, but a similarly sized effect compared to cast mentions for first-week gross.

Table III-9 *Regression summary table for all four specifications*

	Dependent variable:			
	log(total_gross)		log(first_week_gross)	
	(1)	(2)	(3)	(4)
log(opening_theaters)	0.541^{***}	0.513^{***}	0.845^{***}	0.819^{***}
	(0.036)	(0.036)	(0.026)	(0.026)
main_genreAdventure	0.032	0.164	-0.043	-0.032
	(0.243)	(0.246)	(0.177)	(0.178)
main_genreAnimation	-0.557	0.185	-0.761	0.054
	(1.362)	(1.352)	(1.002)	(0.982)
main_genreBiography	-0.055	-0.029	-0.444^{**}	-0.367^{*}
	(0.291)	(0.287)	(0.214)	(0.209)
main_genreComedy	-0.176	-0.010	-0.294^{*}	-0.176
	(0.233)	(0.231)	(0.171)	(0.168)
main_genreCrime	-0.144	-0.048	-0.154	-0.098
	(0.367)	(0.360)	(0.270)	(0.262)
main_genreDocumentary	-0.854	-0.543	-0.824^{*}	-0.652
	(0.571)	(0.567)	(0.420)	(0.411)
main_genreDrama	-0.193	-0.056	-0.271	-0.167
	(0.251)	(0.248)	(0.185)	(0.180)
main_genreHorror	0.111	0.490	0.247	0.481^{*}
	(0.338)	(0.345)	(0.243)	(0.247)
main_genreThriller	-0.227	0.352	-0.341	0.117
	(0.802)	(0.797)	(0.590)	(0.579)
averageRating	0.557^{***}	0.503^{***}	0.473^{***}	0.423^{***}
	(0.087)	(0.086)	(0.060)	(0.059)
log(studio_mentions)		0.045		0.085^{***}
		(0.033)		(0.024)
log(cast_mentions)		0.161^{***}		0.091^{***}
		(0.047)		(0.034)
in_theaters	0.004^{***}	0.004^{***}		
	(0.001)	(0.001)		
Constant	9.521^{***}	8.943^{***}	7.322^{***}	6.829^{***}
	(0.666)	(0.668)	(0.474)	(0.470)
Observations	336	336	336	336
R^2	0.563	0.584	0.815	0.829
Adjusted R^2	0.547	0.566	0.809	0.822
Residual Std. Error	1.354 (df = 323)	1.325 (df = 321)	0.996 (df = 324)	0.963 (df = 322)
F Statistic	34.658^{***} (df = 12; 323)	32.213^{***} (df = 14; 321)	130.049^{***} (df = 11; 324)	119.781^{***} (df = 13; 322)
Note:				*p<0.1; **p<0.05; ***p<0.01

6. Discussion and outlook

This study has shown that the combination of a structurally pre-identified graph model with weights determined via text-mining of a large corpus of user generated content can be a valuable tool for marketers and marketing researchers. It was shown that a highly comprehensive monitoring of an entire industry can be built in a highly automated fashion using well-established data mining techniques applied on abundantly available user-generated content collected from public sources. Such a monitoring infrastructure can – once established – be used for several years with minimal upkeep und can provide valuable feedback and market context to stakeholders in the investigated industry. From the perspective of a marketer or brand manager in the motion picture industry, the implementation demonstrated in this paper can be used to evaluate past promotional efforts within their historical and competitive context and adjust future activities accordingly, gage the brand value of potential cast members and monitor conversation on their own brands. From a marketing researcher's perspective, it was shown that a variety of inquiries regarding the role of brands in motion picture marketing can be addressed. In one application, it was found that consumers associate movies with their producing studios and actors with titles they appear in, providing confirmatory evidence for common assumptions about the roles of actors and studio brands in the marketing of motion pictures. Regarding its usefulness for marketing research in general and the identification of branding issues in the motion picture industry in specific, it could be established that consumers indeed associate franchises with their producing companies and that studios involved in joint production of franchise properties are more heavily associated with each other, though it should be kept in mind that the audience under investigation is likely comprised to a high degree of enthusiasts, limiting the ability to generalize results to the wider movie-going audience. Overall, this emphasizes the outsize role franchises play in the motion picture industry and how they can be both an asset and a liability to brands engaged in their production. To further elaborate on this issue, a more thorough exploration of what determines the strength of associations between brands should be addressed in future research. A possible approach to this could be to gather data on how specific titles were branded in their official marketing communication and whether this has a measurable effect on the association strength between involved brands in the minds of consumers.

Regarding the study's methodology, the model used in this paper is built upon a relatively simple text-mining approach, wherein only direct co-mentions of pre-identified keywords were measured and thus does not take into account the context in which co-mentions occur. As Culotta & Cutler (2016) argued, this issue severely limits the conclusions that can be drawn from the mining of user-generated content and becomes especially problematic when not only considering the quantity of co-mentions, but also their qualitative valence. While techniques exist that consider linguistic context, incorporation of implicit context provided by conversation structure between multiple units of text is as of now a largely unsolved issue. In principle, the threaded nature of reddit comments should lend itself to the inclusion of conversational context for indirect attribution and accurate quantification and attribution of valence, though implementation is likely to be complex. Since this issue is in no way limited to the present application, but rather common to a variety of applications of text mining on user-generated content, it should be a high priority for future research.

References

Aaker, D. A. (1996). Measuring Brand Equity Across Products and Markets. *California Management Review*, *38*(3), 102–120. https://doi.org/10.2307/41165845

Babić Rosario, A., Sotgiu, F., De Valck, K., & Bijmolt, T. H. A. (2016). The Effect of Electronic Word of Mouth on Sales: A Meta-Analytic Review of Platform, Product, and Metric Factors. *Journal of Marketing Research*, *53*(3), 297–318. https://doi.org/10.1509/jmr.14.0380

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The Pushshift Reddit Dataset. *Proceedings of the 14th International AAAI Conference on Web and Social Media, ICWSM 2020*, *Icwsm*, 840–847. http://arxiv.org/abs/2001.08435

Boe, B. (2012). *PRAW: The Python Reddit API Wrapper*. https://github.com/praw-dev/praw/

BoxOfficeMojo. (2022). *Box Office Mojo*. www.boxofficemojo.com

Bürkner, P. (2017). *Advanced Bayesian Multilevel Modeling with the R Package brms*. 1–18.

Culotta, A., & Cutler, J. (2016). Mining Brand Perceptions from Twitter Social Networks. *Marketing Science*, *35*(3), 343–362. https://doi.org/10.1287/mksc.2015.0968

Elberse, A., & Eliashberg, J. (2003). Demand and Supply Dynamics for Sequentially Released Products in International Markets: The Case of Motion Pictures. *Marketing Science*, *22*(3), 329–354. https://doi.org/10.1287/mksc.22.3.329.17740

Esquire. (2021). *How Has the Unstoppable Rise of Streaming Platforms Impacted Film? We Asked the Experts*. https://www.esquire.com/uk/culture/tv/a36842312/streaming-platforms/

Gensler, S., Völckner, F., Egger, M., Fischbach, K., & Schoder, D. (2016). Listen to Your Customers: Insights into Brand Image Using Online Consumer-Generated Product Reviews. *International Journal of Electronic Commerce*, *20*(1), 112–141. https://doi.org/10.1080/10864415.2016.1061792

Ghiassi, M., Lio, D., & Moon, B. (2015). Pre-production forecasting of movie revenues with a dynamic artificial neural network. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2014.11.022

Henderson, G. R., Iacobucci, D., & Calder, B. J. (1998). Brand diagnostics: Mapping branding effects using consumer associative networks. *European Journal of Operational Research*, *111*(2), 306–327. https://doi.org/10.1016/S0377-2217(98)00151-9

IMDb. (2022). *IMDb datasets*. https://datasets.imdbws.com

John, D. R., Loken, B., Kim, K., & Monga, A. B. (2006). Brand Concept Maps: A Methodology for Identifying Brand Association Networks. *Journal of Marketing Research*, *43*(4), 549–563. https://doi.org/10.1509/jmkr.43.4.549

Johnson, S. L., Faraj, S., & Kudaravalli, S. (2014). Emergence of power laws in online communities. *Mis Quarterly*, *38*(3), 795--A13.

Joshi, A. (2015). Movie Stars and the Volatility of Movie Revenues. *Journal of Media Economics*, *28*(4), 246–267. https://doi.org/10.1080/08997764.2015.1094079

Klostermann, J., Plumeyer, A., Böger, D., & Decker, R. (2018). Extracting brand information from social networks: Integrating image, text, and social tagging data. *International Journal of Research in Marketing*, *35*(4), 538–556. https://doi.org/10.1016/j.ijresmar.2018.08.002

Kupfer, A.-K., Pähler vor der Holte, N., Kübler, R. V., & Hennig-Thurau, T. (2018). The Role of the Partner Brand's Social Media Power in Brand Alliances. *Journal of Marketing*, jm.15.0536. https://doi.org/10.1509/jm.15.0536

Lang, B., & Rubin, R. (2022). How Movie Theaters Fought to Survive (Another) Year of Turbulence and Change. *Variety*. https://variety.com/2021/film/news/movie-theaters-box-office-2021-pandemic-omicron-1235142992/

Lee, T. Y., & Bradlow, E. T. (2011). Automated marketing research using online customer reviews. *Journal of Marketing Research*, *48*(5), 881–894. https://doi.org/10.1509/jmkr.48.5.881

Mathys, J., Burmester, A. B., & Clement, M. (2016). What drives the market popularity of celebrities? A longitudinal analysis of consumer interest in film stars. *International Journal of Research in Marketing*, *33*(2), 428–448. https://doi.org/10.1016/j.ijresmar.2015.09.003

Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine Your Own Business: Market-Structure Surveillance Through Text Mining. *Marketing Science*, *31*(3), 521–543. https://doi.org/10.1287/mksc.1120.0713

O'Reilly, D., & Kerrigan, F. (2013). A view to a brand: Introducing the film brandscape. *European Journal of Marketing*, *47*(5), 769–789. https://doi.org/10.1108/03090561311306868

Packard, G., Aribarg, A., Eliashberg, J., & Foutz, N. Z. (2016). The role of network embeddedness in film success. *International Journal of Research in Marketing*, *33*(2), 328–342. https://doi.org/10.1016/j.ijresmar.2015.06.007

Sawhney, M. S., & Eliashberg, J. (1996). A Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures. *Marketing Science*, *15*(2), 113–131. https://doi.org/10.1287/mksc.15.2.113

Stoddard, G. (2015). Popularity and Quality in Social News Aggregators: A Study of Reddit and Hacker News. *Www*, 815–818. https://doi.org/10.1145/2740908.2742470

Swami, S., Eliashberg, J., & Weinberg, C. B. (1999). SilverScreener: A Modeling Approach to Movie Screens Management. *Marketing Science*, *18*(3), 352–373. https://doi.org/10.1287/mksc.18.3.352

The Hollywood Reporter. (2019). *"Spider-Man" Shocker: Disney, Sony Striking a Deal for One More Movie*. https://www.hollywoodreporter.com/heat-vision/spider-man-shocker-disney-sony-striking-deal-new-movie-1243777

TheNumbers. (2022). *Box Office History for Marvel Cinematic Universe Movies*. mojohttps://www.the-numbers.com/movies/franchise/Marvel-Cinematic-Universe

Treme, J. (2010). Effects of celebrity media exposure on box-office performance. *Journal of Media Economics*, *23*(1), 5–16. https://doi.org/10.1080/08997761003590457

Vujić, S., & Zhang, X. (2018). Does Twitter chatter matter? Online reviews and box office revenues. *Applied Economics*, *50*(34–35), 3702–3717. https://doi.org/10.1080/00036846.2018.1436148

IV. Chapter 2

Text classification for marketing research using pre-trained general language models

Abstract:

Marketing research increasingly involves the use of unstructured text data, such as user-generated content collected from social media platforms. A common task in rendering these data useful for analysis is classification of the text units, most commonly by their expressed sentiment towards brands and products. Pre-trained general language models based on the transformer architecture have recently shown very promising results in these problem domains but have so far not seen extensive use by marketing researchers. The goal of this short study is to evaluate these models on common marketing research data such as microblog postings and product reviews and compare them to other methods commonly encountered in the marketing literature. Transformer models are found to perform well on a variety of datasets, as long as some basic rules are followed which are identified in this study. Overall, the findings motivate further exploration of their capabilities and development of best practices for different marketing use cases.

1. Introduction

Text-classification tasks are becoming increasingly ubiquitous in marketing research, most commonly in the form of sentiment classification of user-generated content (i.e. Liu, 2006; Tirunillai & Tellis, 2012; Hennig-Thurau, Wiertz & Feldhaus, 2015). Quantification of sentiment involves a series of mutually dependent choices on behalf of the researcher, ranging from how to preprocess the data to how the resulting measures are ultimately aggregated and operationalized. In a relatively recent study, Hartmann, Hupperts, Schamp & Heitmann (2019) provide a thorough comparison of commonly used text classification methods in the marketing literature, covering a plethora of dictionary- and machine-learning-based methods. An issue that arises with this high methodological diversity, especially in the case of sentiment classification, is that it comes at the cost of comparability between publications. This is especially pronounced for studies who rely on bespoke classifiers, such as artificial neural networks trained on labeled subsets of training data. These are generally found to exhibit the best classification performance, though at a high cost of requiring large amounts of human-labeled data and come at a high risk of overfitting to the training data, which is a common pitfall even for experienced machine learning practitioners. In an ideal situation, classifiers would exist which match these bespoke solutions, while maintaining enough flexibility to enable their use in a variety of contexts. This would allow researchers to choose from a smaller pool of well-understood models, enable the development of clear best practices and ultimately reduce the burden of choice for researchers that look to apply them. Over the past five years, a new class of statistical language models has emerged that might fulfill these requirements. Characterized by extremely large model sizes and the employment of the transformer architecture (Vaswani et al., 2017), these models represent general language models that have been found to perform remarkably well on a wide variety of natural language processing tasks, such as text summarization, classification, question-answering and named-entity recognition (Radford et al., 2018). As a consequence, they found almost immediate industry adoption for marketing-related purposes. For example, Salesforce has trained and put into production its own transformer model, called CTRL (Keskar et al., 2019) and - starting in October of 2019 - Google gradually started employing its BERT model (Devlin, Chang, Lee & Toutanova, 2018) within its search algorithm. Despite this quick adoption by businesses, they have so far been largely absent from the marketing literature, which is surprising given its ubiquitous use of text data and resulting need for text processing methods. An explanation for this absence might lie in the relatively high barriers to entry in the form of programming knowledge requirements, as well as the quantity of computational resources needed not only for training models, but also running them. Multiple recent developments have, however, increased their usability for applied marketing research, namely the development of smaller models via so-called model distillation (Sanh, Debut, Chaumond & Wolf, 2019) and - perhaps more importantly - the introduction of simpler interfaces and infrastructure, such as the Huggingface library for Python (Wolf et al., 2019).

As transformer have not yet seen extensive use or evaluation within the marketing literature, the goal of this study is to reduce the resulting gap. Since the performance advantage of large language models over previous methods is already well established both by general benchmarks such as GLUE (Wang et al., 2018), as well as in applied settings (Heitmann et al., 2020), the focus of this study instead specifically lies on their applicability and usability for marketing research. To do so, a brief, high-level overview of state-of-the-art transformer models is given, followed by a preliminary assessment of their usefulness for marketing research via comparison to established sentiment classification methods. The approach is twofold: after a short characterization and contextualization of transformer models, multiple such models are compared on their predictive accuracy using pre-labeled datasets (that is, datasets containing observations classified by human coders). In a second step, the strengths and weaknesses of the different methods when applied to multiple raw, unlabeled datasets are explored on aggregate, as well as using example data comprised of representative sequences. The results are then discussed within the context of their implications for marketing research and guidelines for their use are proposed. The paper concludes with a short summary of the findings and a forward-looking assessment of future applications of the technology.

2. A brief introduction to transformer models

The transformer architecture was developed by researchers at Google Brain and first released to the public in the form of a paper by Vaswani et al. (2017). While its initial application was in natural language processing, the transformer architecture has seen pervasive and highly successful use within other machine learning problem domains as well, such as computer-vision (Chen et al., 2020) and protein folding, where it enabled a well-publicized leap in capability in the form of Deepmind's AlphaFold 2 (Ronneberger et al., 2021). In the following, a brief explanation of transformer model's functioning and their advantages in comparison to previous methods will be given, followed by a historic timeline and overview of prominent language models that make use of it.

2.1 The transformer architecture and the attention mechanism

Conceptually, transformers build upon sequence-to-sequence models which pioneered the encoder-decoder architecture (Cho et al., 2014; Sutskever et al., 2014) and were later enhanced by an attention mechanism (Bahdanau et al., 2015; Luong et al., 2015), which they make pervasive use of. Attention enables the parallel processing of a sequence in its entirety rather than sequentially, while still taking context between tokens into account. While a thorough explanation of the underlying mechanism and how it fits into the overall transformer architecture exceeds the scope of this study, a highly simplified one will be given based on the example sentence *“Transformers are great.”*. With previous methods such as n-gram models, the sentence would be processed in sequence. This implies that to determine the probability of observing the word “great” after observing “Transformers are”, the conditional probability of observing the word “are” after the word “Transformers” needs to first be computed (for a more mathematically sound description see Goodfellow et al., 2016, pp.449). While feasible for very short sequences such as the given example, this quickly becomes computationally unfeasible for longer sequences, though over the years a multitude of approaches were developed to address this problem, such as long short-term memory models (Hochreiter & Schmidhuber, 1997). In a transformer model, the sentence would first be tokenized, that is split into parts of pre-determined size. Every token would then be run through an *embedding* layer, wherein each is assigned a vector of continuous values representing its relationship to every other word in the vocabulary. The sentence can then be represented by a matrix of dimension Nx3, wherein N is the size of the vocabulary. The values are then augmented to encode the position of the word in the sentence and subjected to a process that Vaswani et al. (2017) dubbed *multi-head attention*. This involves a series of steps which will be skipped in this simplified explanation, but would for each attention head result in a square matrix of the following form, wherein values represent the relative importance of words for other words in the sentence (note that the values themselves are merely illustrative, though they do add up to 1 on the margins since the last step applies the logistic function over all values):

	Transformers	are	great
Transformers	0.8	0.1	0.1
are	0.1	0.6	0.3
great	0.1	0.3	0.6

Figure IV-1 *Example of dot-product attention matrix.*

Note that all values in this matrix are context dependent. For example, changing the word "great" to "toys" would result in completely different values for the other words in the sentence. In this example, this would allow for differentiation between i.e. electrical transformers and the Transformers brand of children's toys. The key point is that the matrix produced by the attention mechanism still contains contextual information for each word while removing the need for sequential processing. When employed in conjunction with subsequent layers of deep neural networks, the attention mechanism thus enables a fully parallel training process which, compared to earlier architectures, allows for the use of larger training corpuses and/or larger models for any given amount of computational resources.

2.2 Timeline and overview of transformer-based language models

Language models based on the transformer architecture first emerged in 2018 in the form of OpenAI's GPT (Radford & Salimans, 2018) and Google's BERT (Devlin et al., 2018) models, with many more models to follow in subsequent years. Table IV-1 gives a brief and inexhaustive overview of a number of prominent transformer models that can be applied to common text classification tasks. Following the initial introduction of the transformer architecture, two parallel trends began to emerge. First was a move towards ever larger models. As the architecture allowed for an orders of magnitude increase in the number of parameters, efforts were initially directed towards higher parameterization while keeping training data size fixed, since this was considered the most effective way to increase overall capability (Kaplan et al., 2020). In 2021, OpenAI's GPT-3 was the first of such extremely large models, at then unprecedented 175 billion parameters. Even larger models quickly followed, such as Gopher (Rae et al., 2022) at 280B parameters and Megatron-Turing at 530 billion parameters (Smith et al., 2022), culminating in Google AI's Switch Transformer at over 1.6 trillion parameters (Fedus et al., 2021). As of 2022, new developments point towards models benefiting more strongly from more training data and longer training times, with the smaller Chinchilla model (Hoffmann et al., 2022) outperforming the previously described models at 70 billion parameters. Parallel to changes in model size were architectural improvements of existing architectures, commonly on the basis of the original BERT model (i.e. T5, CTRL, BART, roBERTa). Furthermore, specialized versions of each model often exist that either use the same architecture trained from the ground up on new data (i.e. language specific BERT models such as CAMEMbert (Martin et al., 2019) for French) or are versions of an existing model that were fine-tuned on new, application-specific data (Barbieri et al., 2020).

Table IV-1 *Inexhaustive overview of major transformer models. M refers to million, B to short billion, T to short trillion.*

Model	Paper	Year	Size
GPT-2	Radford et al., 2018	2018	124M (*small*), 1.5B (*full*)
BERT	Devlin et al., 2018	2018	110M (*base*) 340M (*large*)
T5	Raffel et al., 2019	2019	~220M
BART	Lewis et al., 2019	2019	~130M
CTRL	Keskar et al., 2019	2019	1.6B
roBERTa	Yinhan Liu et al., 2019	2019	355M
GPT-3	Brown et al., 2020	2020	175B
Switch	Fedus et al., 2021	2021	1.6T
Megatron-Turing	(Smith et al., 2022)	2022	530B
Gopher	Rae et al., 2022	2022	280B
Chinchilla	Hoffmann et al., 2022	2022	70B

A unique aspect of transformer models is that they can be set up for so-called "zero-shot" classification tasks, which is the case for Facebook AI Research's BART model. In a zero-shot setting, the classifier has not been engineered based on a fixed set of classes. Instead, a list of label candidates (which can vary in number and complexity) are passed to the model, which classifies inputs into these categories, without any retraining or calibration. This gives the classifier an unprecedented degree of flexibility and potentially enables topic classification on new datasets without the need for human labeling, which represents a common bottleneck in applied research.

3. Evaluating Transformer models for marketing research

The goal of this chapter is to evaluate transformer-based language models on their usefulness for marketing research. Since their possible application is broad and the scope of this study is limited, the focus is put on sentiment classification, which is the identification and measurement of the qualitative valence of a given body of text. Furthermore, while larger models would be expected to outperform smaller ones (and are available as cloud hosted services), the goal of this study was to evaluate classifiers that marketing researchers would be able to use interchangeably with established methods, that is on a single, local machine. As such, all the evaluated models were chosen to be able to run on a relatively modern (as of 2022) laptop computer. In application, transformer models are equivalent to other established text-classification algorithms. The to-be-scored text passages are passed to the classifier (along with a list of candidate labels in the case of *zero-shot* classification) and a list of predictions will be returned alongside confidence scores for each prediction.

3.1 Benchmark datasets and models

This study employs a total of five datasets to evaluate the performance of the aforementioned models, as summarized in Table IV-2. The first dataset was sourced from the extant literature and featured in Hartmann et al. (2019). This was done to facilitate comparison between the extant text classification methods covered by Hartmann et al. and new, transformer-based methods. It is comprised of movie reviews gathered from the internet movie database (imdb) that were specifically selected for conveying strong positive or negative sentiment, with equal sample sizes for both classes. As such, it should be comparatively easy to score compared to more noisy datasets and thus be less representative of real datasets a researcher might encounter. To represent the other extreme, another dataset was kindly provided by the author of Wollborn (2020), which contains a random sampling of messages posted to twitter on the subject of indie videogames that were labeled by a human coder. Both datasets will be used to assess the predictive performance of the investigated models, whereas another three unlabeled datasets, comprised of original data from the author's own research, will be used to further investigate differences in model behavior. Table 2 gives a descriptive overview of the used datasets. Dataset (I) is the only dataset that features equal sample sizes for each class, whereas all other datasets represent random samplings from their respective distributions. Random samples of 10,000 observations were drawn from each dataset except (II), which only comprises 5000 observations. Each dataset was scored by a total of three transformer models, of which BERT (Devlin et al., 2018) forms the baseline, as it is comparatively easy to use and ubiquitous (more specifically, distilBERT is used, which is half the size of BERT while retaining 97% of its performance, see Sanh et al., 2019). It is compared to roBERTa, which was specifically fine-tuned on tweets (Barbieri et al., 2020) and BART, which was fine-tuned for topic classification in a zero-shot context.

Table IV-2 *Descriptives for benchmark datasets.*

Dataset ID	Data type	Data source	lang	Avg seq length	# obs	# classes
(I)	movie reviews (IMdB)	Maas et al.(2011)	en	76.96	10,000	2
(II)	twitter posts (indie video games)	Wollborn, (2020)	en	17.28	5,000	3
(III)	reddit comments (movies)	Author, via reddit API	en	26.29	10,000	-
(IV	twitter posts (video games)	Author, via twitter API	en	18.34	10,000	-
(V)	financial news (summaries)	Author, via finnhub.io API	en	32.63	10,000	-

Four classifiers based on the BART model were devised. The first two replicate other basic sentiment classification schemes (i.e. as used by roBERTa) in that they use either two (“POSITIVE”, “NEGATIVE) or three („NEGATIVE“, „NEUTRAL“, „POSITIVE“) candidate labels representing polarity, whereas the other two, „enhanced“ models make use of the zero-shot-classifier’s inherent flexibility to use context-specific labels for both specifications. For both labeled datasets, the given product category was used to create naïve, custom labels (i.e. “positive for movie” / ”positive for game”). The intent was to investigate whether supplying the model some additional contextual information would change the conditional probabilities the models compute in a way that improves overall predictive performance. Additionally, as a point of comparison, sentimentr (Rinker, 2019), a dictionary-based method, was included as well. Sentimentr employs a sophisticated scoring mechanism that considers valence-shifters and (de-)amplifiers, as well as emojis, making it a strong representative of non-statistical language classification methods.

3.2 Performance on labeled datasets

Table IV-3 reports the performance of each model on the movie review dataset (dataset I) as measured by a variety of performance metrics. These include the recall (the ratio of true positives to true positives and false negatives, also commonly referred to as sensitivity), specificity (ratio of true negatives to true negatives and false positives), accuracy (overall percentage of correct predictions) and precision (the ratio of true positives to true positives and true negatives). Since a trade-off exists between recall and precision, a common measure for overall classification performance is the F1 score, which represents a weighted average of precision and recall and is given as well. The transformer models significantly outperform the dictionary-based method, as BERT beats sentimentr's baseline on all measures, with an overall predictive accuracy of 78,91%, the highest among all models tested. In comparison, for the same dataset, Hartmann et al. 2019 reported a very similar 62.5% for dictionary-based methods, whereas the highest-scoring method they employed, an artificial neural network, topped out at 77.6% accuracy. Keeping in mind that the latter was specifically trained on the dataset, whereas BERT is a general model trained on Wikipedia and books, the latter outperforming the former by more than 1.2 percentage points is rather noteworthy.

Table IV-3 *Classification performance on dataset I, as measured by sensitivity, specificity, accuracy, precision and F1 score.*

Classifier:	**BERT**		**Performance metrics:**				
Class	POSITIVE	NEGATIVE	Sensitivity:	Specificity:	Accuracy:	Precision:	F1:
POSITIVE	5711	1963	74,42%	83,33%	78,91%	81,42%	77,76%
NEGATIVE	1303	6512					
	Sentimentr		**Performance metrics:**				
Class	POSITIVE	NEGATIVE	Sensitivity:	Specificity:		Precision:	F1:
POSITIVE	5519	2155	71,92%	57,59%	64,69%	62,48%	66,87%
NEGATIVE	3314	4501					
	BART (three Classes)*		**Performance metrics:**				
Class	POSITIVE	NEGATIVE	Sensitivity:	Specificity:	Accuracy:	Precision:	F1:
POSITIVE	5697	1977	74,24%	82,50%	78,40%	80,64%	77,31%
NEGATIVE	1368	6447					
	BART (two classes)		**Performance metrics:**				
Class	POSITIVE	NEGATIVE	Sensitivity:	Specificity:	Accuracy:	Precision:	F1:
POSITIVE	5145	2529	67,04%	85,66%	76,43%	82,11%	73,82%
NEGATIVE	1121	6694					
	enhanced BART (two classes)		**Performance metrics:**				
Class	POSITIVE	NEGATIVE	Sensitivity:	Specificity:	Accuracy:	Precision:	F1:
POSITIVE	6874	800	89,58%	66,36%	77,86%	72,34%	80,04%
NEGATIVE	2629	5186					
	enhanced BART (three classes)*		**Performance metrics:**				
Class	POSITIVE	NEGATIVE	Sensitivity:	Specificity:	Accuracy:	Precision:	F1:
POSITIVE	6874	800	89,58%	66,36%	77,86%	72,34%	80,04%
NEGATIVE	2629	5186					

A deeper look at the zero-shot model BART, which was used in four different specifications, reveals a number of interesting observations. The inherent flexibility of a zero-shot classifier allows both for variations in the number of classes, as well as in the choice of classes. In the first two cases, a two-class specification ("negative"/"positive") was compared to a three class specification (negative"/"neutral"/"positive"), wherein neutral classifications were discarded in favor of the second-most highly scored alternative label ("positive" or "negative"). Interestingly, the latter approach results in a slightly higher overall performance driven by more accurate detection of positive valence, with it coming close to matching BERT in overall performance despite not being specifically fine-tuned on sentiment classification. The other two specifications are of the “enhanced” variety, wherein the classification labels were changed to include information about the general topic of the data, both in a two class specification (“negative for movie”/”positive for movie”), as well as a three class specification (“negative for movie”/”neutral for movie”/”positive for movie”), wherein again neutral classifications were reclassified into the second-most highly scored alternative. Interestingly, in this case both classifiers produced identical results. In this “enhanced” form, they significantly outperform BERT as measured by the F1 statistic, owing to a better discrimination of positive sentiment. Overall, this signifies that passing additional information on the product category can indeed increase classification performance, which marks a new development that should be more thoroughly investigated in future research.

The second dataset (dataset II) tells a very different story, as presented in Table IV-4. While again a transformer model (roBERTa) presents the highest accuracy, it is by a very small margin compared to the simplest, dictionary-based method. However, a deeper look reveals that roBERTa is much better at correctly identifying instances of positive and negative valence, which are underrepresented in the dataset compared to the neutral category. Sentimentr’s relatively strong performance at face value can thus be fully explained by its bias towards assigning neutral valence, which makes sense given that the method cannot distinguish between true neutral valence (complete absence of polarized words) and mixed valence (positive and negative words cancel each other out). The other transformer-based models notably underperform on this dataset, with the BERT classifier exhibiting the worst performance. This was expected however, since as a binary classifier, it is severely mismatched with the dataset which is largely made up of the neutral category - which it has no label for - and the classifier will always assign one of the classes it has been given, in this case “positive” and “negative”. However, since the model assigns each of its predictions a confidence score representing uncertainty, this could potentially be used to turn a binary classifier into one with three classes, by re-labeling low-confidence results as belonging to the “neutral” class. This was attempted in the BERT* specification, wherein the optimal re-classification threshold that maximizes the predictive accuracy was found by iterating through possible cut-off values (note that doing this likely heavily overfits the classifier on the benchmark dataset, which would constitute bad practice in real applications). Doing so does significantly increase overall performance, but this is entirely due to an increase in neutral classifications, at the cost of correctly identified instances of positive and negative valence. Overall, this demonstrates the importance of including a “neutral” category for datasets that are not known or pre-selected to be mostly made up of observations exhibiting strong valence, which is the case for most real-world datasets. Interestingly, the zero-shot BART classifier in its enhanced specification exhibits worse performance compared to the regular BART classifier, as the additional information on the topic of the tweets effected a severe decrease of neutral classifications. This did, however, lead it to more accurately classify instances of positive and negative valence.

Table IV-4 *Classification performance on dataset II, as measured by sensitivity, specificity, accuracy, precision and F1 score. *Classificaiton with confidence lower than 99% re-classified as neutral.*

Classifier:	**BERT**			**Performance metrics:**			
Class	POSITIVE	NEUTRAL	NEGATIVE	Recall:	Accuracy:	Precision:	F1:
POSITIVE	702	0	402	63,59%	15,51%	39,95%	49,07%
NEUTRAL	1051	0	2649	0,00%		0,00%	0,00%
NEGATIVE	4	0	52	92,86%		92,86%	92,86%
Classifier:	**BERT***			**Performance metrics:**			
Class	POSITIVE	NEUTRAL	NEGATIVE	Recall:	Accuracy:	Precision:	F1:
POSITIVE	519	376	209	47,01%	47,96%	57,73%	51,82%
NEUTRAL	378	1768	1554	47,78%		82,08%	60,40%
NEGATIVE	2	10	44	78,57%		2,43%	4,72%
Classifier:	**Sentimentr**						
Class	POSITIVE	NEUTRAL	NEGATIVE	Recall:	Accuracy:	Precision:	F1:
POSITIVE	395	662	47	35,78%	66,85%	40,47%	37,98%
NEUTRAL	577	2839	284	76,73%		80,24%	78,45%
NEGATIVE	4	37	15	26,79%		4,34%	7,46%
Classifier:	**tweet-roBERTa**						
Class	POSITIVE	NEUTRAL	NEGATIVE	Recall:	Accuracy:	Precision:	F1:
POSITIVE	888	163	53	80,43%	66,95%	45,38%	58,02%
NEUTRAL	1067	2319	314	62,68%		93,17%	74,94%
NEGATIVE	2	7	47	83,93%		11,35%	20,00%
Classifier:	**BART (three classes)**						
Class	POSITIVE	NEUTRAL	NEGATIVE	Recall:	Accuracy:	Precision:	F1:
POSITIVE	940	79	85	85,14%	34,55%	27,57%	41,65%
NEUTRAL	2465	696	539	18,81%		88,89%	31,05%
NEGATIVE	5	8	43	76,79%		6,45%	11,89%
Classifier:	**enhanced BART (three classes)**						
Class	POSITIVE	NEUTRAL	NEGATIVE	Recall:	Accuracy:	Precision:	F1:
POSITIVE	999	9	96	90,49%	23,54%	25,36%	39,61%
NEUTRAL	2937	94	669	2,54%		90,38%	4,94%
NEGATIVE	4	1	51	91,07%		6,25%	11,70%

The main take away of this exercise is that the chosen classifier has to at least roughly fit the data it is going to be used on. This means that it should at the very least account for neutral valence and in the case of data that differs significantly from the data the classifier was trained on (as is the case for i.e. tweets), a model should be chosen that was specifically retrained on this datatype, a variety of which are shared by other researchers and practitioners. Should no such model be available, re-training an existing model on a labeled subsample of the to-be-scored dataset should be considered.

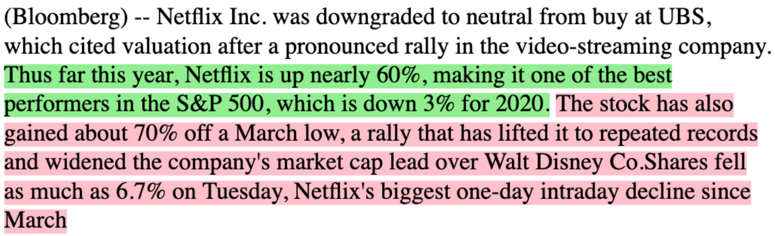
3.3 Aggregate model behavior using unlabeled data

The unlabeled datasets III through V were further used to investigate the general behavior of the different methods. As illustrated in Table IV-5, there are large discrepancies both between different models, as well as across datasets. While it is not possible to identify clear consistencies, BERT seems to exhibit a pronounced negativity bias. Furthermore, both roBERTa and BART are less likely to classify observation as neutral in contrast to the dictionary-based Sentimentr, which tends to default to neutral as previously discussed. Interestingly, roBERTa, which was trained on twitter data, is more likely across datasets to classify observations as neutral compared to BART, signifying a higher prior on neutral valence informed by its re-training on tweets, for which neutral valence would be the most probable class.

Table IV-5 *Percent classified as positive (BERT), positive/neutral/negative (all others) for each method.*

Dataset	Model	BERT		roBERTa			Sentimentr			BART		
	Class	POS	NEG	POS	NEUT	NEG	POS	NEUT	NEG	POS	NEUT	NEG
III (reddit)		4061	6617	2435	4620	3622	2112	7083	1483	3761	2264	4653
		38,03%	61,97%	22,81%	43,27%	33,92%	19,78%	66,33%	13,89%	35,22%	21,20%	43,58%
IV (twitter)		2857	7113	3500	5537	962	1794	7288	918	6391	1603	2006
		28,66%	71,34%	35,00%	55,38%	9,62%	17,94%	72,88%	9,18%	63,91%	16,03%	20,06%
V (news)		3505	6069	2299	5964	1311	2307	6401	866	4101	3195	2278
		36,61%	63,39%	24,01%	62,29%	13,69%	24,10%	66,86%	9,05%	42,83%	33,37%	23,79%

In order to gain deeper insight into the different classifier's individual strengths and weaknesses, a sample of observations on which they disagreed was subsequently subjectively evaluated. As had already been the case with the labeled datasets, performance for each model again varies between the different datasets. This is especially pronounced for the BERT model and the twitter dataset, as it tends to classify clearly positive tweets as negative with high confidence, while simultaneously performing comparatively well on the reddit and news datasets. This points to sequence length being a key determinant of its classification performance, which is in line with the results for the labeled datasets, where it also performed worse on tweets, which are generally very short in length. In contrast, roBERTa performs well on twitter data, as is to be expected given that this particular model was fine-tuned on tweets. The zero-shot BART model, on the other hand, seems to perform well in the identification of weak polarization or cases where the polarity is expressed implicitly, rather than explicitly through word choice. This is especially pronounced in the financial news and reddit datasets. Regarding language content, transformer models generally seem to handle profanity and slang better compared to dictionary-based methods, which usually give profanity highly negative valence scores, resulting in erroneous classification in cases where expletives are used as amplifiers, as is very common in the English language. In one such case, they actually outperformed the human coder on dataset (II), who misclassified the message *"[this product] is the shit"* as having negative valence, although the idiom conveys a strongly positive sentiment, which the transformer models seemingly take into account.



(Bloomberg) -- Netflix Inc. was downgraded to neutral from buy at UBS, which cited valuation after a pronounced rally in the video-streaming company. Thus far this year, Netflix is up nearly 60%, making it one of the best performers in the S&P 500, which is down 3% for 2020. The stock has also gained about 70% off a March low, a rally that has lifted it to repeated records and widened the company's market cap lead over Walt Disney Co.Shares fell as much as 6.7% on Tuesday, Netflix's biggest one-day intraday decline since March

Figure IV-2 *Financial news summary, highlights show polarity as computed by Sentimentr (overall score slightly positive).*

As illustrated by Figure IV-2, taken from dataset (V), positive or negative historical developments often accompany the positive or negative current message in news coverage, making it particularly challenging to score and attribute. In this particular case, the dictionary-based Sentimentr scores the first sentence as neutral, the second sentence as positive and the third and fourth sentences as negative, with an overall slightly positive polarity score. Both BERT and BART correctly identify the news as overall negative, as finance-specific jargon such as "downgraded" is likely part of their implicit vocabulary. This large vocabulary is one of the transformer model's key advantages, as they can be deployed on data pertaining to a variety of topics without the need for bespoke dictionaries.

The main takeaway is that the more semantically complex the input, the bigger the performance delta between transformer models and dictionary-based methods, at least on the type of data investigated.

4. Discussion & Conclusion

This study explored the utility of transformer models for sentiment classification tasks by applying them to real datasets commonly encountered in marketing research. Overall, when even roughly matching the data and task at hand, transformer models generally outperformed a benchmark dictionary-based approach by a significant margin. This is especially impressive when considering that the models used are comparatively small, not fine-tuned on the specific datasets and, with the exception of the roBERTa model, not even trained on the specific data type. Another notable aspect that bears mentioning is that no pre-processing of data took place prior to scoring. Whereas prior methods required the input data to be cleaned to varying degrees (i.e. by removing URLS or parts of product names that would induce bias because they are considered polarized in a lexicon), the transformer models tested in this study seem to have no such requirements.

This study does, however, include a number of limitations that need to be kept in mind. First of all, only two labeled datasets were compared, and performance varied widely between the two. The results can thus not yet be generalized until further investigation. Secondly, all models were used in their standard configuration, even though each of the models used features has a number of parameters that may be tweaked to increase performance on specific tasks. This was a conscious decision, since the goal was to evaluate their baseline performance based on a standard configuration, to both impede the implied combinatorial explosion of tested configurations and to represent the results a mostly naïve user can expect. Lastly, only sentiment classification was covered in this study, even though transformers are reported to show impressive performance on other tasks, such as named-entity recognition and topic classification, as well. Since both are common problems in marketing research, transformer's performance on these types of tasks should be evaluated as well.

To summarize, transformer models show very promising performance for sentiment classification on a variety of datasets without any prior data processing. In the future, this could make them the default method for sentiment classification in applied marketing research, tough further study is needed to better understand the determinants of their performance and to gather more detailed "best-practices" for their use. As it stands, transformer models' advantages over bespoke solutions in performance and usability seems high enough that they should be strongly considered for application contexts in which model-interpretability is not of concern, for example when the output is used in aggregate within a regression model.

References

Bahdanau, D., Cho, K. H., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., & Neves, L. (2020). *TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification*. 1644–1650. https://doi.org/10.18653/v1/2020.findings-emnlp.148

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *ArXiv*.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Dhariwal, P., & Luan, D. (2020). Generative pretraining from pixels. *International Conference on Machine Learning*.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, *4*(January), 1724–1734. https://doi.org/10.3115/v1/d14-1179

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. *Mlm*. http://arxiv.org/abs/1810.04805

Fedus, W., Zoph, B., & Shazeer, N. (2021). *Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity*. 1–31. http://arxiv.org/abs/2101.03961

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Hartmann, J., Huppertz, J., Schamp, C., & Heitmann, M. (2019). Comparing automated text classification methods. *International Journal of Research in Marketing*, *36*(1), 20–38. https://doi.org/10.1016/j.ijresmar.2018.09.009

Heitmann, M., Siebert, C., Hartmann, J., & Schamp, C. (2020). More than a Feeling: Benchmarks for Sentiment Analysis Accuracy. *Ssrn*, Working Paper. https://dx.doi.org/10.2139/ssrn.3489963

Hennig-Thurau, T., Wiertz, C., & Feldhaus, F. (2015). Does Twitter matter? The impact of microblogging word of mouth on consumers' adoption of new movies. *Journal of the Academy of Marketing Science*, *43*(3), 375–394. https://doi.org/10.1007/s11747-014-

0388-3

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. D. Las, Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G. Van Den, Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., … Sifre, L. (2022). *Training Compute-Optimal Large Language Models*. https://doi.org/https://doi.org/10.48550/arXiv.2203.15556

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling Laws for Neural Language Models*. http://arxiv.org/abs/2001.08361

Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., & Socher, R. (2019). CTRL: A conditional transformer language model for controllable generation. *ArXiv*, 1–18.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *arXiv*. https://doi.org/10.18653/v1/2020.acl-main.703

Liu, Yinhan, Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. *1*. http://arxiv.org/abs/1907.11692

Liu, Yong. (2006). Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue. *Journal of Marketing*, *70*(3), 74–89. https://doi.org/10.1509/jmkg.70.3.74

Luong, M. T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, 1412–1421. https://doi.org/10.18653/v1/d15-1166

Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142–150.

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D., & Sagot, B. (2019). *CamemBERT: a Tasty French Language Model*. *2*. http://arxiv.org/abs/1911.03894

Radford, A., & Salimans, T. (2018). Improving Language Understanding by Generative Pre-Training. *OpenAI*, 1–12. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2018). *Language Models are Unsupervised Multitask Learners*.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Guy, A., Jones, C., Bradbury, J., Johnson, M., Hechtman, B., Weidinger, L., & Gabriel, I. (2022). *Scaling Language Models : Methods , Analysis & Insights from Training Gopher*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Peter, W. L., & Liu, J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, *21*, 1–67.

Rinker, T. W. (2019). *{sentimentr}: Calculate Text Polarity Sentiment*. http://github.com/trinker/sentimentr

Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Ballard, A. J., Cowie, A., Romera-paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., … Kavukcuoglu, K. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(August). https://doi.org/10.1038/s41586-021-03819-2

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. *NeurIPS*, 1–5. http://arxiv.org/abs/1910.01108

Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V., Zhang, E., Child, R., Aminabadi, R. Y., Bernauer, J., Song, X., Shoeybi, M., He, Y., Houston, M., Tiwary, S., & Catanzaro, B. (2022). *Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model*. 1–44. https://doi.org/https://doi.org/10.48550/arXiv.2201.11990

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, *4*(January), 3104–3112.

Tirunillai, S., & Tellis, G. J. (2012). Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Science*, *31*(2), 198–215. https://doi.org/10.1287/mksc.1110.0682

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-Decem*(Nips), 5999–6009.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). *GLUE: A*

Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. http://arxiv.org/abs/1804.07461

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J. (2019). Transformers: State-of-the-art natural language processing. *ArXiv*.

Wollborn, P. (2020). Quantifying the effects of video live streaming on the video game industry-Substitute or complement? *DRUID Academy Conference 2020*.

V. Chapter 3

Entrepreneurial Efforts and Opportunity Costs: Evidence from Twitch Streamers

with Philip Wollborn and Ulrike Holder

Abstract

Recent years have seen substantial growth in the number of individuals deriving income through digital platform work, including live streaming platforms. Before this background, we define and contextualize live streaming as a form of digital entrepreneurship and examine whether reduced opportunity costs, in the form of an increase in free time and reduced employment opportunities, affect people's willingness to take up or intensify professionalized streaming activity. To do so, we use an extensive longitudinal dataset gathered from the live-video streaming service Twitch.tv and exploit the changes and restrictions brought about by the first wave of the COVID-19 pandemic to measure individuals' responses to a sudden change in external factors, accounting for individual differences in initial conditions derived from individual's standing with the platform leading into the pandemic. We observe intensified efforts across the spectrum of streamers, but also find that this effect was particularly strong for both fully professionalized streamers, as well as those who were not yet generating income from the platform prior to the pandemic. The ex-post analysis of such newcomers' success shows that once the - for newcomers previously unknown - income potential through live streaming was revealed, streamers with low income potential reduced their activities back to pre-COVID-19 levels, whereas the most successful newcomer streamers sustained their intensified streaming activities compared to their pre-pandemic efforts, implying that these individuals were able to transform their initial efforts into a longer-term commitment. Our results are consistent with the initial assumption that the uptake of platform work is positively correlated with a reduction in outside alternatives and that opportunity costs thus factor into an individual's calculus on the uptake or intensification of entrepreneurial efforts on digital platforms.

1. Introduction

In the last decade, the number of digital platforms, as well as individuals deriving income through them, has grown substantially. As these platforms provide individuals with highly flexible and independent work opportunities (Hall & Krueger, 2018), extant research on platform work has focused on its potential to generate income in reaction to fluctuations in employment (Fos et al., 2021; Jackson, 2020; Koustas 2018), as well as its ability to enable entrepreneurial activity outside of the platform work itself (Burtch et al., 2018; Barrios et al. 2020). A related but in many ways distinct aspect of digital platform work can be found in digital content creation, which in recent years has increasingly shifted towards the production of live video. An early entrant to this growing sector and – as of this writing – still market leader amongst live-streaming platforms, is Twitch. Twitch offers mass distribution of user-generated live video content via the internet and to date, more than one hundred million viewers regularly use the platform, with more than two million people regularly broadcasting their own material.[1] Compared to linear television or on-demand video providers such as Netflix, Twitch combines aspects of both social media and entertainment, wherein in principal every interested individual can act as content provider, but consumers tend to concentrate on a small number of highly successful producers. A 2021 leak of payouts by Twitch (Twitch.tv, 2021) to the top ten thousand streamers by revenue revealed median monthly payouts of 1,665$, rising to 49,821 $ for the 100 largest streamers and 155,323$ for the top ten.[2] In contrast to more transaction-oriented digitally-enabled platform work such as ride sharing or food delivery, which provide immediate monetary return proportional to time and effort invested, professional live-streaming thus represents an opportunity for income generation that is characterized by a very large, albeit highly uncertain, potential payoff. Moreover, becoming a successful live streamer requires an unknown investment of time and resources, as well as an initial endowment with abstract characteristics related to ability and personality traits, possession of which a potential streamer has limited ability to ascertain without first attempting their luck. In our view, this characterizes professional live streaming as a form of entrepreneurship, wherein initially high outcome uncertainty is gradually reduced through market feedback (in this case resonance with viewers) and attempts at achieving product-market-fit (i.e. through adapting content to accommodate the

[1] Historically, video game content has been the focus of the platform; however, over time it has widened its range of broadcast content to include general content such as arts and crafts or talk shows and has more recently entered partnerships with recording artists for concerts and sports events organizers such as Formula 1 Racing.

[2] direct payments from Twitch represent only a fraction of total income for most streamers, with off-platform donations and sponsorship deals routinely making up a far greater share (Wired, 2021)

demands of certain types of viewers). Within this framework, a streamer considering professionalization will base their decision to increase efforts both on positive signals of potential future *benefit* (i.e. in the form of increased viewership or a step-up in partnership status with the platform), as well as the *costs* associated with such efforts. As streaming's capital requirements, which in principle only consist of a stable internet connection and PC or smart devise, are negligible for residents of industrialized nations, the main cost associated with increased streaming activity should be of opportunity, mainly in the form of time that could have been invested towards other ends, such as employment or educational attainment. As previously argued in the literature, lower opportunity costs are more likely to encourage individuals to engage in entrepreneurial activities (Amit, Muller, & Cockburn, 1995) and we would thus expect the same to be true for streamers. As such, the purpose of this paper is to quantify the effect of reduced opportunity costs on the amount of entrepreneurial effort people put into live streaming on Twitch.

To do so, we exploit the changes and restrictions brought about by the first wave of the COVID-19 pandemic. In spring 2020, COVID-19 was spreading rapidly, placing economies and labor markets all over the world in a state of uncertainty. People reduced their mobility by staying home and changed their consumption behavior, especially regarding leisure activities (van Leeuwen et al., 2020). Concurrently, shutdowns in certain business areas and supply-chain disruptions distressed the economy, resulting in dramatic short-term effects on employment, while schools and universities temporarily closed. As such, we posit that the COVID-19 pandemic and related containment measures were an unexpected, positive shock to individuals' available time, while simultaneously decreasing other opportunities to derive income, including gig work (Ivaldi & Palikot, 2020), resulting in overall lowered opportunity costs for entrepreneurial activity on Twitch that in turn should effect an overall increase of entrepreneurial efforts on the platform (Douglas & Shepherd, 2000). Furthermore, the demand side of the market for live streaming was affected as well, characterized by a stark increase in viewership that should similarly lead to increased supply-side activity.

As streamers are highly heterogeneous in the intensity of their activity and aspirations for economic success, ranging from pure hobbyists with no expectations of monetary reward to fully professionalized individuals deriving the entirety of their income from streaming, we expect streamers to also be heterogeneous in their reaction to the previously described changes. We thus categorize streamers based on their status on Twitch, which determines whether and to which degree an individual can receive direct compensation from the platform, ranging from *standard* users with no ability to be compensated to *partners,* who negotiate a contractual agreement with the Twitch platform, with *affiliates* as an intermediate step. We argue that the status of partner is a necessary condition to fully professionalize as a streamer, whereas the affiliate status acts as a delineator between amateur and professional streamers that acts as both a steppingstone for those who possess the potential to truly professionalize, as well as a ceiling for those who do not. This is emphasized by the fact that a clearly defined and comparatively easy to achieve set of conditions needs to be met in order to reach affiliate status, whereas attainment of partner status is much more involved. In light of this, we would thus expect the previously described circumstances to lead to an increase in activity for all three groups, but to different degrees. Amateur users with as-of-yet unrealized ambitions to professionalize are likely more heavily constrained by other commitments and would thus be expected to more strongly react to reductions in opportunity cost. Conversely, we would expect partners, who are already professionalized to a high degree, to be more sensitive to the demand-side effect of increased viewership and would further expect their reaction to be much more limited in scope due to an already high baseline, with affiliates landing somewhere in-between the two. To test these assumptions empirically, we construct a panel on individual streaming behavior starting in calendar week 5 of 2020 until week 29. As we will describe in more detail later on, streamers can intensify their entrepreneurial activity on Twitch in a variety of ways. For the empirical approach in this study, we use measures of activity that are measurable for a large sample of streamers in high temporal resolution. These cover the decision of when to stream (as measured by the share of streams on weekends) and for how long (as measured by the weekly average stream length and total minutes streamed). Within a difference-in-differences (DiD) design, we then use the pandemic lockdown reactions (starting in week 11) as demarcating an exogenous positive shock to available leisure time and negative shock to income opportunities and estimate how these measures changed on an individual level. Our findings show that the COVID-19 pandemic resulted in an immense gain in both Twitch viewership on the demand side and an increased influx of new streamers on the supply side, coinciding with reduced mobility attributable to lockdown measures. While we find that streamers among all groups intensified their efforts during this period, the reactions were particularly strong the less streamers were able to monetize their streams before the pandemic began. A post-hoc analysis revealed that the strongest and most lasting response was seen in those streamers who managed to achieve an upgrade in status (to affiliate and partner), indicating that this subset of individuals took successful steps towards professionalization. In contrast, users that did not manage to upwardly mobilize quickly returned to their pre-pandemic behavior.

While many existing studies have investigated different aspects of live streaming and have explored the motivations of streamers and viewers (e.g., Gros, Wanner, Hackenholt, Zawadzki, & Knautz, 2017; Johnson & Woodcock, 2019b, 2019a; Sjöblom & Hamari, 2017; Wulf, Schneider, & Beckert, 2020) as well as the way streamers cultivate their fanbases (e.g., Gandolfi, 2016; Sjöblom, Törhönen, Hamari, & Macey, 2019), we explore how aggregate changes to labor market conditions and overall societal function affect the market of live streaming supply. Thus, our paper offers several contributions to the literature. First, our findings contribute to the emerging literature on the short-term impact of COVID-19 on the labor market by analyzing the possibilities on the digital platform labor market Twitch. Second, this paper relates to existing literature on the relationship between opportunity costs, gig work and entrepreneurship (Agrawal et al., 2015; Burtch et al., 2018; Fos, Hamdi, Kalda, & Nickerson, 2021; Jackson, 2020). While these earlier studies more directly evaluated the effect of unemployment for the gig economy, we study the supply side of the platform economy and offer insight into why individuals supply labor to these platforms. Third, we add to the literature on live streaming and especially the path into professionalized streaming. Besides non-monetary factors such as occupational enjoyment and social interactions, our results show that time opportunity is a key driver of participation on the platform.

The next section provides a theoretical background for our considerations and assumptions. Section 3 then provides more details on the live streaming platform Twitch and streamers' range of action and monetization options. In Section 4, we describe our data and empirical strategy, the results of which we present in Section 5. We conclude with a discussion and final remarks in Sections 6 and 7.

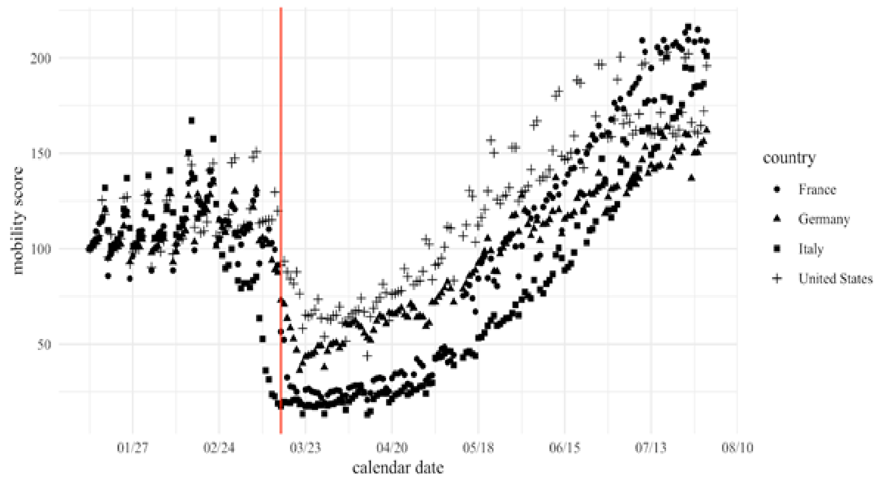
2. Theoretical Background – Opportunity Costs and the Pandemic as External Enabler

As we are interested in the behavioral differences of platform users that diverge in their entrepreneurial ambition and resource endowment, we base our research on literature covering the interplay of platform work and entrepreneurial activity. Furthermore, we argue that the pandemic-induced changes in opportunity costs, combined with the existence of a platform like Twitch, can be viewed as a set of *external enablers* (EE). EE are defined as circumstances that affect individuals' venture creation processes, typically by enabling entrepreneurial options that were not viable for the respective actors without these circumstances (Davidsson et al., 2020). On the other hand, under- and unemployment are often found to be positively correlated with entrepreneurship or self-employment (Fossen, 2020; Thurik et al., 2008). In times of economic hardship, people tend to increase their entrepreneurial activity due to lower opportunity costs, like an increase in available time or a lack of other employment options (Block & Koellinger, 2009; Burtch et al., 2018; Storey, 1991). From the perspective of entrepreneurship as a utility-maximizing response, people should increase (lower) their entrepreneurial activities when opportunity costs are low (high) (Amit et al., 1995; Douglas and Shepherd, 2000).

A common denominator of the current research is that it attaches entrepreneurial activity and platform work to the "physical world." As Nambisan (2017) argues, entrepreneurship research needs to include a perspective on digital entrepreneurship, as digital technologies have profound implications for the entrepreneurial process. In an entirely digital market, a producer is largely free of local demand limitations, since the potential audience is global and thus only limited by language and time zone differences between streamer and potential audience. Consequently, success on live streaming platforms and the associated earnings potential are potentially extremely large. Whereas income from gig work is limited by either time (Uber, TaskRabbit, MechanicalTurk) or capital endowment (Airbnb and to a lesser degree Uber), streaming enables brand building, which can subsequently be further leveraged through additional monetization opportunities in the form of, e.g., secondary content distribution or merchandizing. It further fosters innovation, as novelty can be decisive for increasing one's success on the platform. Uber drivers, for example, are much more limited in the ways they can innovate to attract customers or elicit good reviews (some drivers provide snacks or entertainment options during the ride) and are ultimately much more dependent on the platform, e.g., through top-down pricing. With streaming, however, these dependencies can actually reverse when streamers reach a degree of popularity in which they become an attractor to the platform (Bloom, 2019). Thus, with live streaming, entrepreneurial potential is much higher compared to the other aforementioned platforms.

To measure how entrepreneurial efforts by platform incumbents and newcomers change in reaction to abrupt and unpredictable changes in opportunity costs, we make use of a natural experiment in the form of the first wave of the COVID-19 pandemic. The mobility data presented in Figure V-1 shows that the emerging pandemic and related containment measures (i.e., "shelter-in-place" and "stay-at-home") started to measurably affect people's behavior in the USA, Germany, France and Italy in March 2020. Furthermore, the COVID-19 crisis had dramatic effects on employment, which were met with a variety of relief efforts designed to mitigate the ensuing short-term impact on income for those affected (Adams-Prassl, Boneva, Golin, & Rauh, 2020; Alstadsæter et al., 2020; Bauer & Weber, 2020; Bick & Blandin, 2020; Juranek, Paetzold, Winner, & Zoutman, 2020).

Figure V-1 *Mobility Score (January – August 2020).*



Note: To quantify changes in mobility over the observed time frame, we retrieved mobility data provided by Apple and Google via the *covdata* package for R (Healy, 2020).

We posit that the COVID-19 pandemic and related containment measures were an unexpected, exogenous shock to individuals' available income and amount of idle time spent confined to their home. On the demand side, Twitch is a highly interactive entertainment platform. In contrast to linear video content, live-streaming prominently incorporates aspects of social media such as interaction with peers (Diwanji et al., 2020). With people staying at home and experiencing fewer in-person social interactions, meeting the need for social interactivity may be one reason that the live-streaming suddenly saw a strong increase in viewership during the pandemic (see Figure 3). In turn, this increase in demand likely presented itself as an opportunity for both incumbent streamers looking to attract new viewers, as well as for new streamers looking to enter the market. Taken together, these key factors – namely the abrupt changes in employment, available free time, earnings potential, demand for (live) entertainment as well as the existence of a platform like Twitch that provides easy access to viewers – can in our view be considered a set of external enablers as described in Davidsson et al. (2020; 2021). While they originally define EE as a set of circumstances of external and distinctive nature that are theorized to affect individuals' entrepreneurial efforts by triggering, shaping or enhancing the outcome of the venture creation process (Davidsson, 2015, Davidsson et al., 2020), the authors further argue that the EE framework provides

> *"structure and terminology for analyzing the enabling effects of different types of external change for entrepreneurial initiatives, such as technological breakthroughs, regulatory*

reforms, macroeconomic shifts, demographic sociocultural trends, and changes to the natural environment" (Davidsson et al. 2021, p. 2).

Many of these external changes can be found within the context of the pandemic and Twitch. While the platform Twitch represents a *technological breakthrough*, the decrease in mobility is both an indicator of a *sociocultural trend* (decrease in demand for physical social interactions; increase in demand for digital alternatives) as well as *regulatory reforms* (curfews and other lockdown policies). Other regulatory reforms are found in the policies that lead to a decrease in productivity (manufacturing) and other business areas being shut down completely (e.g., live entertainment, gastronomy, commercial sports activities). As a result, the drastic increase in under- and unemployment that especially affected younger generations (Cho & Winters, 2020) represents a *macroeconomic shift*.[3]

Following the EE framework, we argue that these changes in the necessities and demand of individuals overall lowered the opportunity costs for increasing entrepreneurial efforts. Within the EE framework, this means that through the mechanics of *combination* (e.g., leveraging a technology platform like Twitch), *uncertainty reduction*, *demand expansion* as well as *substitution*, the venture creation process of streamers was affected in what the framework calls *roles*: first, by *triggering* the initial process of earning money with streams and, further, by *enhancing the outcome* of this process compared to an environment where these mechanics would not have been existent.

3. Understanding Twitch and Streamers' Possibilities

3.1 Streaming Platform Twitch

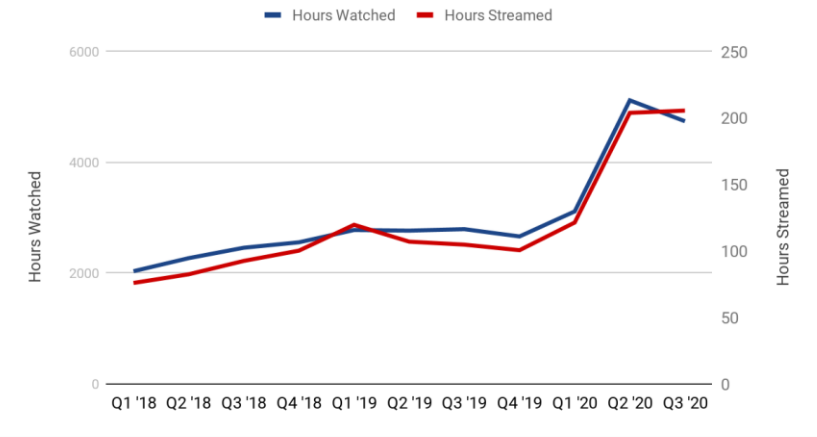
Platforms are not just digital marketplaces for established businesses and services: Platforms not only "enter or expand markets," but they can "replace (and rematerialize) them" (Cohen, 2017, p. 133) and, hence, provide new economic opportunities. As a platform, Twitch provides the core framing conditions for providing and consuming content on its platform. In the following, we describe some insights into the supply side of live streaming and streamer's on- and off-platform monetization options.

Twitch currently dominates the market for live streaming.[4] From February to March 2020, Twitch recorded an increase of total hours watched by almost 23% and an enormous increase

[3] See Davidsson et al. (2021) for an extensive view on the COVID-19 pandemic as an external enabler.
[4] See Appendix A for an overview of live streaming platforms and their development.

of the broadcasting user base, by about 64% (Streamlabs, 2020a). Figure V-2 shows the overall development of streaming hours watched and hours streamed on Twitch from 2018 until 2020. Poignantly, viewing behavior on Twitch during that period seems to be inversely correlated with the mobility data as provided in Figure V-2. With several countries going into lockdown in March 2020, hours watched on Twitch increased, reaching their peak in April, after which they stabilized at a slightly lower level in May and June, when mobility again increased in most countries. These stabilized viewer levels were still roughly 50% higher than at the beginning of 2020, suggesting that Twitch has benefited from the COVID-19 pandemic.

Figure V-2 *Twitch User Metrics (2018 – 2020).*



Note: Data derived from Streamlabs (2020a, 2020b).

The growth in live streaming during a global pandemic was likely abetted by its unique mixture of features. In addition to the purely consumptive act of watching live-streamed content, its interactive features add social elements. Twitch embodies a socially enjoyable experience and an easily accessible space to interact with a community (Wulf et al., 2020; Diwanji et al., 2020), which might be more highly sought after in times of social distancing. Furthermore, live streaming services allow users to not only passively consume but to actively create content by becoming streamers themselves, be it for the pure enjoyment of performing in front of an audience or simply as a potential source of income. Besides, low entry barriers in terms of costs and risks combined with high flexibility facilitate the uptake of and increase in activities on Twitch, which is what we are interested in in our study.

3.2 Deriving Income on Twitch

Relying on a range of different monetization techniques, hundreds of thousands of streamers are capable of generating some earnings from their activities on Twitch, while a few thousand have even managed to build full-time careers out of their streaming activities (Johnson & Woodcock, 2019b, 2019a). As previously mentioned, opportunities to generate income on

Twitch strongly depend on an individual's status: *standard*, *affiliate* or *partner*. These stages are passed through gradually, starting with the standard status as the default and no (on platform) monetization options. In contrast, the affiliate status opens up monetization opportunities, and within the partner status these opportunities are expanded even further. While the step to become an affiliate depends solely on measurable criteria (such as the number of followers or broadcasted minutes), receiving the partner status also includes subjective elements (twitch.tv, 2020b).[5] Hence, achieving affiliate status is the first step toward generating income from live streaming activities. Affiliates are comparable to semi-professionals (those who augment their otherwise-derived income with revenues from streaming), while partners tend to spend their entire working time on streaming and derive the majority of their income from it. The different revenue sources for streamers, either directly on Twitch or through external sources, are illustrated in Table V-1. The most prominent revenue sources are subscriptions, donations and advertisement. In addition, and with rising popularity of individual streamers, other options such as sponsored content and merchandising can provide further external revenue sources.

Table V-1 Revenue Sources for Streamers.

	Internal	External
One-time payments	virtual currency	money transfer
	gifted subscriptions (donations)	coupon codes
	referral links	referral links
	merchandise	merchandise
Recurring payments	subscriptions	digital patronage
Sponsoring	paid content	paid content
	advertisement revenue share	promotional activities
		event/appearance salaries

3.3 Professionalization on Twitch

The achievements of partner and affiliate status are usually accompanied by increased professionalization of all streaming activities and the streamer's appearance. Typical examples of professionalization we identified are an individualized logo (to display oneself as a brand), timetables for regular streams, merchandise offerings, providing features that encourage

[5] As of today, to be invited for affiliate status, streamers need to have at least 50 followers as well as to have broadcasted more than 500 minutes in total, on seven unique days, with an average of three or more concurrent viewers in the last 30 days (Twitch.tv, 2020a). Before applying for partner status, streamers have to fulfill further requirements (Twitch.tv, 2020b).

donations (such as automated donation messages, donation rankings and individual reactions to donations), embedding of social media profiles, and secondary use of content (e.g., on-demand videos on YouTube). To illustrate the interconnectedness of status and professionalization, we took a random sample of 100 channels from the top 100, 101 – 1,000, and 1,001 – 10,000 channels to get an impression of how these professionalization signals are distributed.

Table V-2 Overview of Professionalization and Income Generation.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Streamer Tier	Professional Appearance		External Revenue Sources				Social Media Presence	
	Logo	Time table	Sponsors	Exclusive Discounts	Merch-andise	Sec. Content distr.	Facebook, Twitter, Instagram	Other Platforms
Top 100	0.61	0.21	0.33	0.39	0.39	0.91	0.97	0.22
101 – 1,000	0.39	0.36	0.30	0.24	0.33	0.88	0.97	0.18
1,000 – 10,000	0.55	0.30	0.30	0.19	0.24	0.67	0.97	0.09

Note: Secondary content distribution refers to rebroadcasting via, e.g., YouTube.

As depicted in Table V-2, columns (1) and (8), common professionalization signals are personal logos and links to social media profiles. Whereas logos are found almost universally among all tiers, the differences in external revenue options are more pronounced. As live broadcasts typically last multiple hours, secondary use of this content requires editing; since editing is a time-consuming task, popular streamers often employ content editors (Graham, 2020). Less popular streamers, on the other hand, are less likely to be able to afford such employees or do the editing themselves, as this would likely outweigh benefits in the form of additional revenue, and part-time streamers might simply lack time or capability to do the editing themselves. Many professionalization signals are virtually available for anyone, while others (such as sponsors or merchandising) require a certain level of success. As all income-generating opportunities on Twitch are heavily promoted during live streams and some, like advertisement revenue sharing, are directly tied to the amount of time streamed, the expected income per stream should generally increase with its duration, making available time an important variable for a streamers' success.

3.4 Identification Strategy

Using the above information and taking the COVID-19 pandemic as an exogenous shock to streamers' opportunity costs, our empirical strategy builds on the idea that certain sets of streamers differed in their reactions to these changes in environment. We consider streamers' status levels on Twitch and suggest that professional streamers, who in large part already worked entirely from home, were largely unaffected by the pandemic relative to semi-professionals and amateurs (newcomers, as well as standard and affiliate users).

More precisely, we consider established professional streamers to be those that had already achieved the partner status on Twitch by the end of January 2020. We consider these streamers as our control group, while the first treatment group consists of established streamers that had not (yet) reached partner status at that point in time (standard users and affiliates). We further define new users who entered the platform during the observed period as "newcomers", which we hypothesize to be more eager to adapt their streaming behavior to their newly gained popularity and thus consider our second treatment group. We expect that the established professional streamers (control group) were able to continue streaming in the same manner as they did before the COVID-19 pandemic and, thus, should have remained (largely) unaffected by the lockdown. The treatment groups, in contrast, may have already been able to earn a small amount of money through streaming (more so for treatment group 1), but likely earned most of their income from other sources and, therefore, should have been relatively more heavily affected. We base this on the assumption that non-fully-professionalized streamers were, to varying degrees, dependent on off-platform employment. Due to the aforementioned demographic makeup of streamers on Twitch, this employment likely concentrated in sectors that were heavily affected by lockdown policies; typical student jobs are found in the service/gastronomy sector, where layoffs and furloughs were especially prominent during the initial lockdown. Thus, while the control group likely remained (largely) unaffected by the lockdown and should not have adjusted its behavior, the treatment groups would have been affected more heavily and are expected to show a stronger change in their streaming behavior. However, between the two treatment groups, the more established users (treatment group 1) might already been closer to their individual optimal streaming behavior, while newcomers (treatment group 2) would not yet be settled and might try to take full advantage of their new situation and thus show a stronger reaction (see Table V-3).

Table V-3 Definition of Our Setting – Control, Treatment Groups and Periods.

Group/Period	**Composition**	**Explanation**
Control	Established streamers with partner status based on our selection period (August 2019 – January 2020)	Partners have an array of monetization possibilities on Twitch and are expected to derive major parts of their income through their activity on the platform; we assume that they spend most of their working hours on Twitch, work from home and, thus, that their streaming activities were not/least affected by the COVID-19 pandemic.
Treatment 1	Established streamers with either affiliate or standard status as of January 2020	Standard users and affiliates have, compared to partners, have limited possibilities to earn money through their activity on the platform. As such, we assume that they spend their leisure time on Twitch. As the COVID-19 pandemic affected the time people spent at home, their streaming activities would be comparably more affected than those of our control group.
Treatment 2	Newcomer streamers, that newly appeared in in our data collecting process after January 2020	Streamers whose streams first appeared in our data collecting process after January 2020. As we started our data collecting in August 2019, we assume that streamers who first appeared in our data set in 2020 were newcomers to the platform that previously had not been able to extract any meaningful income and had not yet optimized their streaming behavior to a professional level. Especially in comparison to our first treatment group, they would not have been able to assess their income potential on the platform. Thus, we assume that the streaming behavior of newcomers was most heavily affected by the COVID-19 pandemic, as they most likely had to rely on other income sources before and their income potential on Twitch was not yet revealed.
Pre-pandemic period	Week 5 (starting January 27) until week 10 (ending March 8) 2020	As we first observed streamer status during week 4, 2020, week 5 was the first complete week where we had data to categorize streamers into the aforementioned groups. Further, in the first weeks of 2020, available leisure time should have been affected by the end of the holiday seasons. As the strongest global pandemic reactions came into effect on or around March 14, 2020, which was at the end of week 11, we excluded that week from our regression sample.
Post-pandemic period	Week 12 (starting March 16) until week 29 (ending July 19) 2020	Week 12, 2020, marked the first complete week where societies and politics fully reacted to the pandemic (see Fig. 1). During July, for the first time in 2020 mobility levels reached those above pre-lockdown levels in all countries in Fig. 1. Since the pandemic situation started to show stronger differences between countries in summer, especially regarding work opportunities, our sample ended July 19, 2020.

To measure activity on the platform, we focus on three indicators: streaming length, percentage of streams that occur on the weekend and total minutes streamed. These different components of a stream and, thus, of a streamer's behavior can be hypothesized to indicate the degree of an individual's professionalization (i.e., streamers with day jobs or in education should mostly stream in their spare time, which is on the weekends, whereas professional streamers are more flexible in their chosen hours).

4. Data and Descriptives

4.1 Data and Variables

We base our analysis on a panel of streamers that were active in the first seven months of 2020. We first collected hourly, stream-level data from Twitch via its official API, covering the time range of August 2, 2019, and July 30, 2020. The resulting 3,329,042 observations include the top 500 live streams by current viewers at the moment of collection. We used this dataset as a starting point to obtain additional streamer-level data for each streamer who reached the hourly top 500 live streams at least once over the observed period, which required a minimum amount of 21 concurrent viewers. Thus, although at first glance using only the top 500 live streams might give the impression of only covering the most successful streamers and that the resulting sample could thus suffer from survivorship bias, the low minimum viewer threshold of 21 shows that many streamers in our sample are far from what one would call successful. Furthermore, one single hourly top 500 placement of one single live stream sufficed to be included in the sample, resulting in a final sample of 18,467 individual streamers. Additionally, the goal of this study was to gain insights on professionalizing behavior of streamers, which in the first place requires streamers to possess the intention for professionalized streaming and at least some potential to be able to do so. Thus, while the Twitch platform boasts millions of regular streamers, since the entry barriers to initiating a stream are so low (i.e. Twitch streaming has been supported on every single PlayStation console starting in 2013 and can be enabled via a single button press), active streaming alone does not signify any ambition to do so professionally. As such, including these casual streamers would likely result in a more strongly biased sample.[6]

Subsequently, data covering every unique stream initiated by each streamer in the panel, when available, was extracted from data aggregator Sullygnome (Sullygnome, 2020) with permission from the service provider. This approach enabled us to retroactively collect the full activity history of streamers, even when their viewership numbers were too small to regularly enter the top 500 streams and allowed us to gather more accurate data on stream length and viewership.[7] Additionally, we twice collected each observed streamer's partnership status with Twitch, once on January 23rd and again on June 25th, 2020, to account for changes in partnership status.

[6] The behavior of standard users in Figure 8 most likely gives an impression on how this bias would affect our sample probably millions of even smaller streamers were included in our sample.

[7] In Appendix B, we provide an overview of the initial sample composition.

As our interest lies in individual streamers, and many channels are run by a group of individuals or organizations who operate under different constraints, we performed further data cleansing. Using the given self-descriptions of each channel, we removed every channel that contained the words “studio” or “official,” as we found these words to reliably indicate that the channel is operated by an organization. As Twitch does not extend partner status to channels specializing in gambling, even when all other requirements are fulfilled, we also excluded channels with the keywords “slots,” “casino” and “gambling.” Furthermore, we removed each stream with a duration longer than 24 hours. As calendar week 5 of 2020 marks the first complete week where data on the streamer status is present, and to prevent biases from the holiday season of 2019/2020, we removed the first four weeks of 2020. Thus, the first observations started on January 27, 2020. Relying on the previously described changes in mobility and unemployment, we consider mid-March (calendar week 11) as the beginning of the post-COVID-19 outbreak period, and we take six weeks before and 19 weeks after this as the period of study. Additionally, we ensured that all streamers in our sample streamed at least once in the immediate pre- and post-COVID-19 outbreak period, which we consider the six weeks leading up to and following the aforementioned cutoff.

4.2 Descriptive Analysis

In total, the panel covers 1,936,528 unique streams initiated by one of the streamers within the panel. Table V-4 provides the summary statistics of streams and stream characteristics for the whole period before and after the determined post-COVID-19 period in total and broken down by treatment and control groups. As presented in Panel A, except for the variable *percentage weekend*, all considered variables increased in the post-COVID-19 period. Among the different stream components, the highest average growth was recorded by the variable total views (growth of 48.52%), while the growth of the other variables was below 20%. In Panel B, we also differentiate the considered variables by control and treatment groups. By definition, the mean values for treatment and control groups are different, with on average lower values for the treatment groups than the control group, despite the weekend variable. While streaming length hardly changed among the groups and between pre-post time points, there were remarkable increases in total views and total time streamed, especially for treatment group 2. Unsurprisingly, partners had the highest absolute growth in total views and time streamed. In relative terms, however, the newcomers (treatment group 2) showed the highest growth rates, with 108.18% in total views and 30.13% in time streamed.

Table V-4 *Descriptive Statistics of Streamer-level Measures of Total Sample and by Control and Treatment Groups.*

Panel A: Total sample

	Pre-COVID-19				Post-COVID-19			
	Mean	Median	Min	Max	Mean	Median	Min	Max
N streams	25.46	24	1	314	28.93	28	1	284
Stream length	259.7	240.4	14	1,436.7	272.3	255.2	1	1,388.1
Total views	9,374	1,864	0	1,530,190	13,907	2,603	0	2,815,572
Time streamed	7,110	5,694	14	91.854	8,343	7,030	14	91,960
Percentage Weekend	0.279	0.27	0	1	0.278	0.27	0	1
N	18,467				18,467			

Panel B: Sample means by control and treatment groups

	Control			Treatment 1			Treatment 2		
	pre	post	%diff	pre	post	%diff	pre	post	%diff
N streams	27.07	31.19	15.22	25.38	27.32	7.64	22.05	26.22	18.91
Stream length	286.1	298.2	4.22	256.7	264.3	2.97	205.9	226.5	10.0
Total views	17226	25010	45.19	2392	3624	51.5	1845	3841	108.18
Time streamed	8259	9716	17.64	6875	7581	10.27	4906	6384	30.13
Percent Weekend	0.2688	0.2719	1.15	0.2831	0.2805	-0.9	0.2956	0.2862	-3.1
N	8,839			5,615			4,013		

Note: Stream length and time streamed in minutes.

Figure V-3 *Plot of Weekly Viewership.*

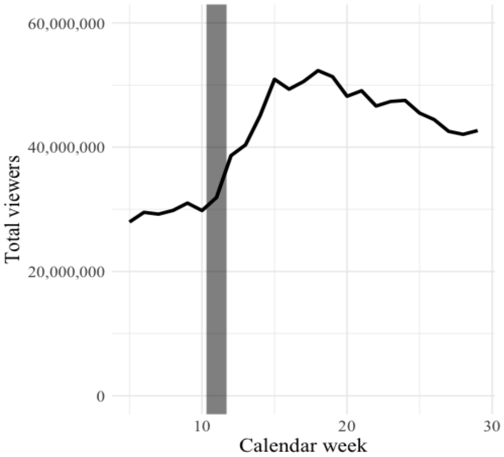
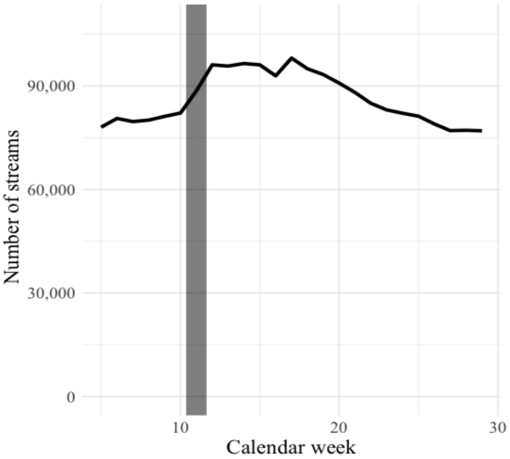


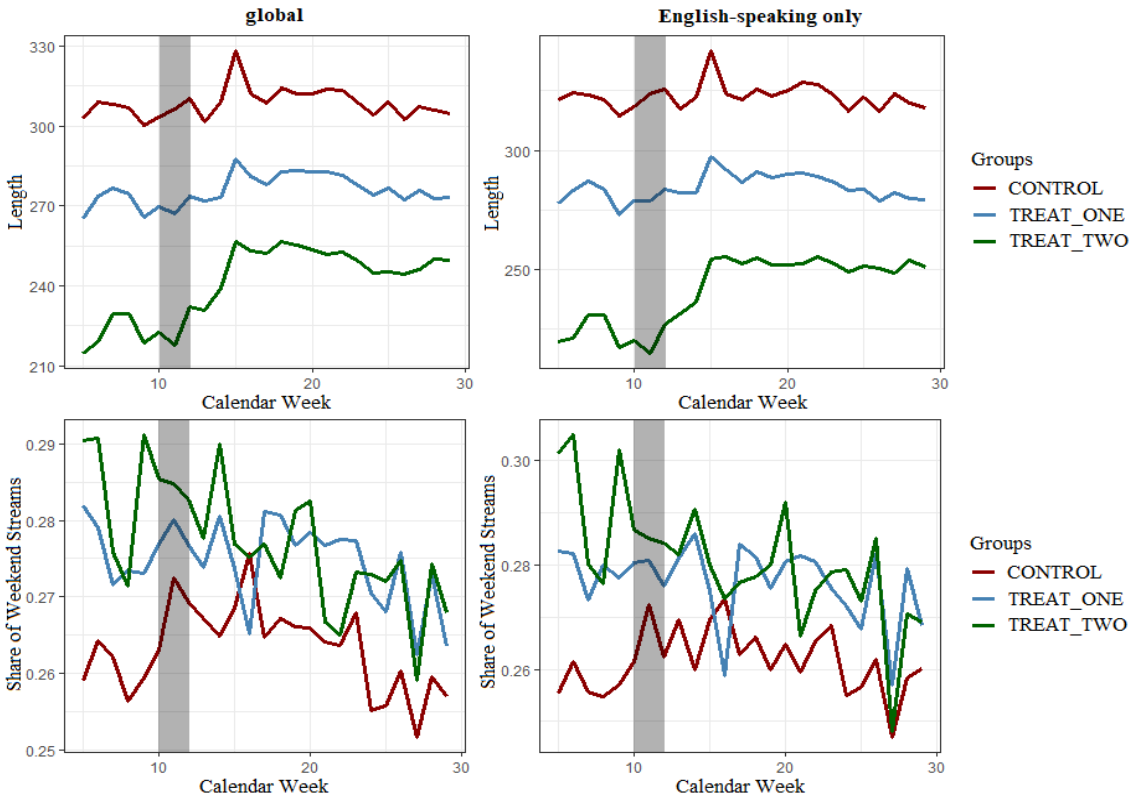
Figure V-4 *Plot of Weekly Stream Activity.*



In Figures V-3 and V-4, we show the viewership and stream activity (each aggregated by calendar week) over the observed time frame. Both time series show a stark increase after week 11, which marked the most drastic decrease in global mobility, as previously described. Although viewership remained high after the initial increase and seemingly stabilized at a higher level compared to the initial condition, the number of streams slowly returned to its previous level.

The upper graphs in Figure V-5 display the average stream length for the global sample and a subset consisting of English-speaking only streamers, which we considered in order to exclude possible differences between various countries. Clearly, while all groups showed an initial positive reaction to the lockdown measures, the reaction of treatment group 2 was the strongest and stabilized at a much higher level than the other established groups. We assume that the established groups (control group and treatment group 1) had already found their utility-maximizing stream length, while newcomers were able to use the increasing viewership and available time to increase their professionalization and, hence, benefited more greatly from a sustainable increase in streaming time. The lower graphs in Figure V-5 show how the share of streams on weekends changed over time. At first glance, the assumption that weekend percentage can act as a proxy for professionalization holds, as the control group (partnered streamers) showed a consistently lower weekend percentage compared to treatment group 1 (consisting of affiliate and standard users), which, in turn, had a lower weekend percentage compared to treatment group 2 (newcomers). Overall, the weekend percentage decreased over time, though the change was gradual rather than sudden. The degree of change was negatively correlated with the degree of professionalization, with treatment group 2 having the biggest change compared to the other groups, while the control group showed an unexpected initial increase in weekend streaming.

Figure V-5 *Stream Length and Share of Weekend Streams over Time.*



Note: Graphs on the left-hand side show the complete dataset, those on the right-hand side show the subset of English-speaking only streamers. Grey bars mark week 11, when lockdown measures set in.

5. Empirical Analysis

5.1 Methodology

We estimate the differences between streamers that responded more strongly to the exogenous shock (treatment groups) and streamers that were largely unaffected (control group), as explained in detail in Section 2.3. Our formal DiD specification is written as follows:

$$Y_{it} = \alpha + \beta_1 * Post_t + \beta_2 * TreatOne_i + \beta_3 * Treat\,Two_i + \beta_4 * (Post_t * Treat\,One_i) + \beta_5 * (Post_t * Treat\,Two_i) + \Gamma_i + \Theta_t + \varepsilon_{it} \tag{V-1}$$

where the dependent variable Y_{it} reflects streaming activity in terms of streaming length and total number of streams and stream percentage during weekends by individual i in week t.

$Post_t$ denotes a binary variable that equals one for post-lockdown periods and zero otherwise. The treatment dummies $Treat\ One_i$ and $Treat\ Two_i$ indicate whether a given individual belongs to one of the treatment groups and captures the differences between the three groups that exist irrespective of the lockdown. The interaction terms between the treatment and post-lockdown variables measure the difference between the groups in the post-lockdown period. If we assume that entrepreneurial ambition was higher for the treatment groups but followed a parallel trend prior to lockdown, then β_4 and β_5 capture the causal impact of these individuals experiencing a change in their time constraints and/or income compared to individuals (established professionals) who were largely unaffected by the pandemic on these points; these values should have a positive sign. Γ_i controls for geographic fixed effects (with language acting as a proxy), and Θ_t captures time fixed effects.

Our DiD approach will provide empirical evidence if the parallel trend assumption holds. This assumption assumes that without treatment, treatment and control groups follow similar trends in the dependent variable; i.e., that without the exogenous shock, the outcome variables of our three groups would have followed the same trend. We verify this assumption by illustrating the trend in our samples before the exogenous shock in Figure 6: Similar trends before the treatment are indicated. When comparing the trends between the groups, one must keep in mind that the underlying dataset covers a global sample on a relatively frequent basis (weeks) in a highly dynamic market. If an overall common trend is visible despite differences in weather, time zones and potentially other country-specific factors, we argue that this suffices to fulfill the parallel trends assumption. For our first dependent variable, mean weekly streaming length, the parallel trends before treatment are visible. For the share of weekend streams, the graphical evidence is less clear. Nonetheless, all groups show a significant downwards and upwards movement before treatment, though less pronounced in the subset of English-speaking only streamers.

5.2 Main Results

Table V-5 provides the results of our DiD estimations. The coefficients of interest are the DiD estimators in rows four and five and confirm our graphical evidence. Column (1) shows the change in streaming length for treatment group 1 and treatment group 2 relative to the change in streaming length for the control group around the 25-week window surrounding week 11. The coefficients are positive and statistically significant, indicating that the post-COVID-19 period had an immediate effect on streaming length relative to partners. In particular, for treatment group 1 the estimate of 0.017 suggests that length was on average 1.7% higher in the post-pandemic period relative to the streaming length in the control group, whereas for treatment group 2, the post-period growth in streaming length a much higher 13%. We find comparable results in the more homogeneous subset of English-speaking only streamers (Columns 3 and 4).

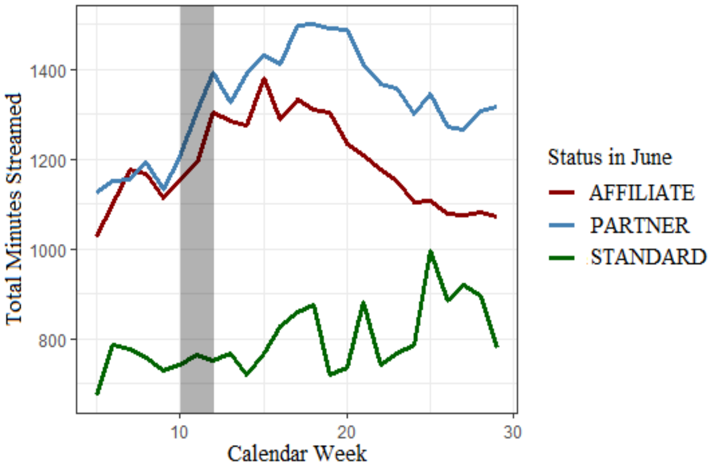
Table V-5 *Effect of COVID-19 Pandemic on Streaming Behavior.*

Dep. Variables	(1) Log(length)	(2) Pct Weekend	(3) Log(length)	(4) Pct Weekend
Language Selection	Global	Global	English	English
TREAT_ONE	-0.179**	0.015**	-0.183**	0.022**
	(0.002)	(0.002)	(0.003)	(0.002)
TREAT_TWO	-0.470**	0.024**	-0.526**	0.034**
	(0.003)	(0.002)	(0.005)	(0.003)
POST	0.049**	-0.007**	0.028**	-0.002
	(0.004)	(0.003)	(0.005)	(0.004)
POST × TREAT_ONE	0.017**	-0.005*	0.017**	-0.008**
	(0.003)	(0.002)	(0.004)	(0.003)
POST × TREAT_TWO	0.130**	-0.012**	0.131**	-0.019**
	(0.004)	(0.003)	(0.005)	(0.004)
Constant	5.161**	0.280**	5.566**	0.259**
	(0.007)	(0.004)	(0.004)	(0.003)
Time FE	Yes	Yes	Yes	Yes
Language FE	Yes	Yes	No	No
Observations	1,936,528	375,391	1,033,674	209,060
R-squared (adj.)	0.047	0.003	0.044	0.002
F-Statistic	1,664.019**	21.907**	1,772.634**	15.912**

Note: Control group is defined as the subset of streamers who had the status of partner in January 2020. TREAT_ONE is defined as streamers who had the status of affiliate or standard in January 2020. TREAT_TWO is defined as streamers who entered the sample after January 2020 and thus had no pre-pandemic status. POST is defined as every week since week 12, 2020. Time fixed effects = Week FE, · p<0.1; * p<0.05; ** p<0.01. Robust standard errors in brackets.

We showed that newcomers in particular adapted their streaming behavior by increasing their average streaming length per broadcast and reducing the share of weekend streams in the post-COVID-19 period. To gain more insights into this subsample, we further investigate changes in their total streaming time. It seems plausible that, within this subgroup, the most successful streamers who subsequently saw and increase in status (to affiliate or partner), adapted their streaming behavior in a sustainable manner that led to long-lasting changes, which is to be expected insofar as initial success should act as a strong signal to continue the activity that effected it. Figure V-7 shows that those who reached at affiliate status by June 2020 initially increased their total streaming time to a similar degree as those who reached partner status. However, when lockdown policies were lifted again and mobility increased, new affiliates quickly reduced their streaming time to pre-COVID-19 levels, while streamers who gained partner status stabilized their streaming time at a higher level. In contrast, users who remained standard users seemed to be completely unaffected (and are not considered in the following).

Figure V-6 *Total Minutes Streamed among Newcomers.*



Note: Newcomers were grouped by their acquired streamer status in June 2020. The gray bar marks week 11, when lockdown measures set in.

To assess whether the difference between newcomers who gained affiliate and partner status was significant, we estimated another DiD specification wherein the most successful group (new partners) was the control group, while new affiliates made up the treatment group. Results are provided in Table V-6. As expected, the POST dummy indicates that after the lockdown measures, both groups increased their total streaming time by 22.7% and that, overall, the newly-minted affiliates streamed 9.8% less compared to those who reached partner status. The coefficient of the DiD estimator shows that, in total, our treatment group of new affiliates streamed 15.4% less than newcomers who made the jump to partner status in the global sample (Column 1) and 11.8% in the English-speaking only sample.

Table V-6 *Effect of COVID-19 Pandemic on Streaming Behavior among Newcomers Who Reached Affiliate or Partner Status by June 2020.*

Dep. Variables	(1) Log(total minutes streamed)	(2) Log(total minutes streamed)
Language Selection	Global	English
TREAT	-0.098**	-0.101**
	(0.020)	(0.031)
POST	0.227**	0.177*
	(0.037)	(0.053)
POST × TREAT	-0.154**	-0.118**
	(0.023)	(0.035)
Constant	6.303**	6.536**
	(0.051)	(0.041)
Time FE	Yes	Yes
Language FE	Yes	No
Observations	66,786	35,910
R-squared (adj)	0.025	0.009
F-Statistic	34.087**	13.564**

Note: Subsample of streamers who entered the sample after January 2020 (newcomers). Control group is defined as streamers who had reached the status of partner by June 2020. TREAT is defined as streamers who had reached the status of affiliate by June 2020. POST is defined as every week since week 12, 2020. Time fixed effects = Week FE, · p<0.1; * p<0.05; ** p<0.01. Robust standard errors in brackets.

6. Discussion and limitations

The first wave of the COVID-19 pandemic led to a significant boost to activity on Twitch by both viewers on the demand side and streamers on the supply side. Regarding its differential effect on streamers by level of professionalization prior to the pandemic, we first confirmed our prior assumption that, on average, a higher degree of professionalization (as measured by a streamer's standing with the platform) is associated with a higher frequency and duration of broadcasts, as well as a larger share of streams conducted on weekdays, which we thus considered indicative of professionalization efforts. We then investigated how both measures changed in reaction to lockdown-measures for streamers grouped by their status prior to the pandemic. We find that, by increasing the length of their broadcasts and shifting stream activity from the weekends towards weekdays, less professionalized streamers adapted their behavior to converge towards that of more strongly professionalized streamers, though this effect was only lasting for the least-professionalized group. Interestingly, the most and least professionalized streamers initially increased their total streaming time more strongly than semi-professional streamers did. This is in contrast to our initial expectations, which were that streamers would be the less affected – and thus the less strong to react – the higher their degree of professionalization was pre-pandemic. An explanation for this might lie in the high degree of heterogeneity amongst even the most professionalized group (partners). As previously described, while partner status can be considered a necessary condition for fully professionalized streaming, it is not a sufficient condition in and of itself. As such, a sizable share of partnered streamers are likely on a level of professionalization we had initially assumed for affiliate streamers.

Going back to our theoretical model of streaming as entrepreneurship, wherein an individual's decision to increase professionalization efforts is responsive to both (time) opportunity, as well as market feedback that reduces uncertainty about the individual's success potential, it makes sense that whether increased efforts are sustained at a higher level or revert to pre-pandemic levels is dependent on whether these efforts were met with positive market feedback (i.e. in the form of increased viewership or an upgrade in status on the platform). The interviewees in Johnson and Woodcock (2019b) stressed the large amount of time that needs to be invested into building and maintaining a professional streaming career. Thus, marginal costs for any additional hour invested into professional streaming should be relatively high and need to be compensated adequately (through direct earnings or potential future earnings) in order to be worth the investment. We thus took a deeper look at the group of streamers we dubbed "newcomers", which are those users who were either completely new to streaming or could be considered strictly casual streamers prior to the observational period. Among these users, newcomers who went on to become affiliates initially reacted in the same way as those who became partners. After four to eight weeks, however, most newcomers were likely able to better assess whether their increased efforts were indeed paying off and adapted their subsequent efforts accordingly, as newcomer affiliates returned to their pre-pandemic levels whereas newcomer partners remained on a much higher level of time spent streaming. This can also be considered evidence that the affiliate status group largely encompasses streamers who had some ambition to professionalize but failed to reach partner status due to their unwillingness or inability to put in the needed effort needed or otherwise lacking the necessary ability or talent. Whereas we initially assumed the affiliate status to be a steppingstone towards partner status, an analysis of the newcomer sample reveals that the majority of those streamers that achieved an increase in status, became partners either directly or within a short timeframe. In addition, newcomers that remained standard users throughout our observation period showed virtually no change in streaming behavior in the first place, leading us to assume that these users are pure hobbyists who lack any intention for professionalization and that the achievement of at least affiliate status is within reach of most users who are actively looking towards professionalization (which makes sense, given the clear conditions that need to be met).

The question remains why streamers with entrepreneurial ambition increased their efforts during the pandemic rather than immediately starting with maximized efforts in the first place. As we showed in Section 3, most professionalization signals on Twitch are comparatively easy to achieve. Those signals that are not are either very time consuming (secondary content distribution) or require a certain degree of success (merchandise, sponsoring). Our assumption is that before the COVID-19 pandemic, opportunity costs for those streamers were too high to increase their efforts. With the restrictions and constraints induced by the pandemic, we argue that these opportunity costs suddenly and unexpectedly changed for these streamers. The COVID-19 pandemic (or its containment measures) thus represents a set of *external enablers* that put these streamers into position to increase their efforts, which otherwise would not have been a viable option for them (Davidsson et al., 2020).

This work is subject to several simplifications and, thus, limitations that we want to address. First, as we assigned individuals to the various treatment groups based on fairly unspecific, observable characteristics (their status on Twitch), there likely remains a high degree of heterogeneity within these broad groups. As previously argued, the partnered streamers which we initially considered as close to fully professionalized likely consists to a significant degree of semi-professionals who we expected to be covered by the affiliate condition. The estimated differences in reaction between users on different professionalization levels should thus be considered closer to the lower bound of the true effect. Secondly, we do not account for specific lockdown policies in different countries in our analysis, nor do we consider different time zones. However, at least the former concern can be mitigated by looking at our fairly comparable subsample results for English-speaking streamers. While lockdown policies in the USA, Canada and the United Kingdom still differed in many aspects, we can assume that most streamers in our sample were situated in the USA, which was severely hit by the pandemic in terms of infections as well as unemployment.

Lastly, we are not able to differentiate whether the measured changes in streaming behavior are attributable to changes in either income or available time, as both factors were affected simultaneously. In order to differentiate between the two factors, more data on individual streamers' employment background would have been necessary.

Concluding Remarks

In this study, we classify professional live-streaming as a form of entrepreneurial activity and subsequently evaluate whether and how a sudden disruption to external constraints impact its intensity. We consider the first COVID-19 lockdown period that started in March 2020 as an exogenous shock to available time and income (which we together consider to make up the opportunity costs to professionalized streaming activity) and use the standing with the platform preceding the intervention to differentiate between control and treatment groups. In summary, we find an increase in professional streaming activities after the primary lockdown measures. Our DiD estimates indicate that an increase in available time was followed by an increase of up to 13% in average stream length and a 1.2% reduction in the share of weekend streams, both of which we previously argued to be characteristic of professional streaming activity. As such, professionalization efforts expanded when opportunity costs were lowered, especially for individuals with larger initial uncertainty of success, as the differences between our treatment groups show. Further, the ex-post analysis of newcomers' success shows that even when mobility increased beyond pre-lockdown levels and unemployment rates had again fallen strongly, streamers who were met with success in their initial streaming efforts sustained their efforts on a higher level, indicating that initial reactions are indeed indicative of the intention to pursue entrepreneurial efforts. Future research could thus take a deeper look at streamer-level data in order to provide additional insights into the motivations of streamers as well as the determinants of success on the platform.

Overall, low opportunity costs may be advantageous for starting and professionalizing entrepreneurial activities in the form of digital content production. Digital platforms have the continued potential to supplement and displace more established forms of employment and self-employment, an issue that will continue to be key for future policy and research. As digital content creation has very low entry barriers, initial investment risks are relatively low. Instead, as we have shown, it is more restricted by other resources such as available time. Thus, it could be a viable option for under- and unemployed individuals, and policymakers should carefully consider these implications, for example in the design of unemployment benefits, by affording recipients the flexibility to generate supplemental income from digital content creation, including live streaming.

References

Adams-Prassl, A., Boneva, T., Golin, M., & Rauh, C. (2020). Inequality in the impact of the coronavirus shock: Evidence from real time surveys. *Journal of Public Economics, 189*, 104245. https://doi.org/10.1016/j.jpubeco.2020.104245

Agrawal A., Catalini C., Goldfarb A. (2015) Slack time and innovation. *NBER Working Paper No. 21134*. https://doi.org/10.1287/orsc.2018.1215

Alstadsæter, A., Bratsberg, B., Eielsen, G., Kopczuk, W., Markussen, S., Raaum, O., & Røed, K. (2020). The first weeks of the coronavirus crisis: Who got hit, when and why? Evidence from Norway. *NBER Working Paper No. 27131*. doi: 10.3386/w27131

Amit, R., Muller, E., & Cockburn, I. (1995). Opportunity costs and entrepreneurial activity. *Journal of Business Venturing, 10*(2), 95–106. https://doi.org/10.1016/0883-9026(94)00017-O

Bauer, A., & Weber, E. (2020). COVID-19: How much unemployment was caused by the shutdown in Germany? *Applied Economics Letters*, 1–6. https://doi.org/10.1080/13504851.2020.1789544

Barrios, J., Hochberg, Y., and Hanyi, Y. (2020). Launching with a Parachute: The Gig Economy and New Business Formation. *NBER Working Paper No. 27183*. doi: 10.3386/w27183

Bick, A., & Blandin, A. (2020). Real-time labor market estimates during the 2020 coronavirus outbreak. *Available at SSRN 3692425*. http://dx.doi.org/10.2139/ssrn.3692425

Block J., Koellinger P. (2009). I Can't Get No Satisfaction—Necessity entrepreneurship and procedural utility. *Kyklos, 62*(2):191–209. https://doi.org/10.1111/j.1467-6435.2009.00431.x

Bloom, D. (2019). Ninja Jumps From Twitch To Microsoft's Mixer in Exclusive Deal. Retrieved July 27, 2021, from https://www.forbes.com/sites/dbloom/2019/08/01/twitchs-biggest-star-ninja-joins-microsofts-mixer-in-exclusive-programming-deal/

Cho, S. J., Winters, J. V. (2020). The Distributional Impacts of Early Employment Losses from COVID-19. *Available at SSRN 3602755*. http://dx.doi.org/10.2139/ssrn.3602755

Cohen, J. E. (2017). Law for the platform economy. *UCDL Rev., 51*, 133-204.

Cunningham, S., & Craig, D. (2019). Creator governance in social media entertainment. *Social Media and Society, 5*(4), 1–11. https://doi.org/10.1177/2056305119883428

Davidsson, P. (2015). Entrepreneurial opportunities and the entrepreneurship nexus: A re-conceptualization. *Journal of Business Venturing, 30*(5): 674–695. https://doi.org/10.1016/j.jbusvent.2015.01.002

Davidsson, P., Recker, J., and von Briel, F. (2020). External Enablement of New Venture Creation: A Framework. *Academy of Management Perspectives, 34*(3): 311–332. https://doi.org/10.5465/amp.2017.0163

Davidsson, P., Recker, J., and von Briel, F. (2021). COVID-19 as External Enabler of entrepreneurship practice and research. *BRQ Business Research Quarterly, 24*(3): 214–223. doi: 10.1177/23409444211008902

Diwanji, V-., Reed, A., Ferchaud, A., Seibert, J., Weinbrecht, V., Sellers, N. (2020). Don't just watch, join in: Exploring information behavior and copresence on Twitch. *Computers in Human Behavior, 105*: 106221. doi: 10.1016/j.chb.2019.106221

Douglas, E. J., & Shepherd, D. A. (2000). Entrepreneurship as a utility maximizing response. *Journal of Business Venturing, 15*(3), 231–251. https://doi.org/10.1016/S0883-9026(98)00008-1

Fos, V., Hamdi, N., Kalda, A., & Nickerson, J. (2021). Gig-labor: Trading safety nets for steering wheels. *Available at SSRN 3414041.* http://dx.doi.org/10.2139/ssrn.3414041

Fossen, F.M. (2020). Self-employment over the business cycle in the USA: a decomposition. *Small Business Economics.* https://doi.org/10.1007/s11187-020-00375-3

Gandolfi, E. (2016). To watch or to play, it is in the game: The game culture on Twitch. Tv among performers, plays and audiences. *Journal of Gaming & Virtual Worlds, 8*(1), 63–82. https://doi.org/10.1386/jgvw.8.1.63_1

Goodman-Bacon, A., & Marcus, J. (2020). Difference-in-differences to identify causal effects of COVID-19 policies. *DIW Berlin Discussion Paper No. 1870.*

Graham, S. (2020). Confronting Pokimane | Inside The Million Dollar Empire. Retrieved December 11, 2020, from https://www.youtube.com/watch?v=wvl05CQKkY0.

Gros, D., Wanner, B., Hackenholt, A., Zawadzki, P., & Knautz, K. (2017). World of streaming. Motivation and gratification on Twitch. In G. Meiselwitz (Eds), *Social Computing and Social Media. Human Behavior* (Vol. 10282, pp. 44-57). Cham: Springer.

Hall, J. V., & Krueger, A. B. (2018). An Analysis of the labor market for Uber's driver-partners in the United States. *ILR Review, 71*(3), 705–732. https://doi.org/10.1177/0019793917717222

Healy, K. (2020). *Rpackage (covdata)—COVID19 Case and Mortality Time Series*.

Ivaldi, M., & Palikot, E. (2020). Sharing when stranger equals danger: Ridesharing during Covid-19 pandemic. *CEPR Discussion Paper No. 15202*.

Jackson, E. (2020). Availability of the gig economy and long run labor supply effects for the unemployed. *Stanford University Working Paper.*

Johnson, M. R., & Woodcock, J. (2019a). "And today's top donator is": How live streamers on Twitch.tv monetize and gamify their broadcasts. *Social Media + Society, 5*(4), 1–11. https://doi.org/10.1177/2056305119881694

Johnson, M. R., & Woodcock, J. (2019b). 'It's like the gold rush': The lives and careers of professional video game streamers on Twitch.tv. *Information, Communication & Society, 22*(3), 336–351. https://doi.org/10.1080/1369118X.2017.1386229

Juranek, S., Paetzold, J., Winner, H., & Zoutman, F. (2020). Labor market effects of Covid-19 in Sweden and its neighbors: Evidence from novel administrative data. *CESifo Working Paper Series 8473*. http://dx.doi.org/10.2139/ssrn.3660832

Kavanagh, D. (2019). Watch and learn: The meteoric rise of Twitch. Retrieved January 2, 2021, from https://blog.globalwebindex.com/chart-of-the-week/the-rise-of-twitch/

Koustas, D. (2018). Consumption insurance and multiple jobs: Evidence from rideshare drivers. *Working Paper*, University of Chicago, Chicago.

Morse, A. (2015). Peer-to-peer crowdfunding: Information and the potential for disruption in consumer lending. *Annual Review of Financial Economics, 7*, 463–482. https://doi.org/10.1146/annurev-financial-111914-041939

Nambisan, S. (2017). Digital Entrepreneurship: Toward a Digital Technology Perspective of Entrepreneurship. *Entrepreneurship: Theory and Practice, 41*(6): 1029–1055. https://doi.org/10.1111/etap.12254

RND. (2020). Umfrage: 40 Prozent der Studenten haben wegen Corona-Krise Job verloren. Retrieved January 2, 2021, from https://www.rnd.de/politik/studenten-in-corona-krise-40-prozent-haben-ihren-job-verloren-R6IJD2ROBQVQEB6FH5KFJ3EHKM.html

Sjöblom, M., & Hamari, J. (2017). Why do people watch others play video games? An empirical study on the motivations of Twitch users. *Computers in Human Behavior, 75*, 985–996. https://doi.org/10.1016/j.chb.2016.10.019

Sjöblom, M., Törhönen, M., Hamari, J., & Macey, J. (2019). The ingredients of Twitch streaming: Affordances of game streams. *Computers in Human Behavior, 92*, 20–28. https://doi.org/10.1016/j.chb.2018.10.012

Spilker, H. S., Ask, K., & Hansen, M. (2020). The new practices and infrastructures of participation: How the popularity of Twitch.tv challenges old and new ideas about television viewing. *Information, Communication & Society, 23*(4), 605–620. https://doi.org/10.1080/1369118X.2018.1529193

Storey, D.J. (1991). The birth of new firms – Does unemployment matter? A review of the evidence. *Small Business Economic, 3*(3): 167–178. https://doi.org/10.1007/BF00400022

Streamlabs. (2020a). Streamlabs & Stream Hatchet Q1 2020 Live streaming industry report. Retrieved December 22, 2020, from https://blog.streamlabs.com/streamlabs-stream-hatchet-q1-2020-live-streaming-industry-report-9630bc3e0e1e

Streamlabs. (2020b). Streamlabs & Stream Hatchet Q3 2020 Live Streaming Industry Report. Retrieved December 22, 2020, from https://streamlabs.com/content-hub/post/streamlabs-and-stream-hatchet-q3-2020-live-streaming-industry-report

Sullygnome. (2020). Twitch statistics & analysis—Games / Channels—SullyGnome. Retrieved June 30, 2020, from https://sullygnome.com/

Thurik, A., Carree, M., van Stel, A., Audretsch, D. (2008). Does self-employment reduce unemployment? *Journal of Business Venturing, 23*(6): 673–686. https://doi.org/10.1016/j.jbusvent.2008.01.007

Twitch.tv. (2020a). Am Affiliate-Programm teilnehmen. Retrieved December 11, 2020, from https://help.twitch.tv/s/article/joining-the-affiliate-program?language=de

Twitch.tv. (2020b). Twitch.tv—Partners. Retrieved December 11, 2020, from https://www.twitch.tv/p/de-de/partners/

Twitch.tv. (2021). *Updates on the Twitch Security Incident*.

https://blog.twitch.tv/en/2021/10/15/updates-on-the-twitch-security-incident/

U.S. Bureau of Labor Statistics. (2021). Unemployment rate—16-24 Yrs. Retrieved January 2, 2021, from https://fred.stlouisfed.org/series/LNS14024887

van Leeuwen, M., Klerks, Y., Bargeman, B., Heslinga, J., and Bastiaansen, M. (2020). Leisure will not be locked down – insights on leisure and COVID-19 from the Netherlands, *World Leisure Journal, 62*(4): 339-343. https://doi.org/10.1080/16078055.2020.1825255

Wired. (2021). *Millionaire Twitch Streamers React to Their Leaked Earnings*.

https://www.wired.com/story/twitch-streamers-earnings-exposed-now-its-a-meme/

Wulf, T., Schneider, F. M., & Beckert, S. (2020). Watching players: An exploration of Media enjoyment on Twitch. *Games and Culture, 15*(3), 328–346. https://doi.org/10.1177/1555412018788161

Zervas, G., Proserpio, D., & Byers, J. W. (2017). The rise of the sharing economy: Estimating the impact of Airbnb on the hotel industry. *Journal of Marketing Research, 54*(5), 687–705. https://doi.org/10.1509/jmr.15.0204

Appendix

Appendix A. Overview of Live Streaming Platforms

Period	February 2020		June 2020	
Platform	Total hours watched (in hours)	Market share (in %)	Total hours watched (in hours)	Market share (in %)
Twitch	979,968,010	63,59	1,553,283,931	66,82
YouTube Gaming Live	351,522,493	22,81	471,578,917	20,29
Facebook Gaming	183,858,030	11,93	268,801,339	11,56
Mixer	25,611,913	1,66	30,997,472	1,33

Appendix B. Composition of Initial Sample

	Count			Movement to			
Status	January	June	Net Change	Standard	Affiliate	Partner	Dropout
Standard	801	2,667	1,866	451	177	118	55
Affiliate	5,473	8,362	2,889	4	4,331	1,073	65
Partner	9,514	12,076	2,562	50	9	9,418	37
Newcomer	12,295	4,978	-7,317	2,162	3,845	1,467	4,821
∑	28,083	28,083	0	2,667	8,362	12,076	4,978

Note: “Movement to” refers to changes in partnership status during January and June 2020.

Appendix C. Robustness Test

	(1)	(2)	(3)	(4)
Dep. Variables	Log(length)	Pct Weekend	Log(length)	Pct Weekend
Language Selection	Global	Global	English	English
TREAT_ONE	-0.178**	0.015**	-0.183**	0.022**
	(0.003)	(0.002)	(0.003)	(0.002)
TREAT_TWO	-0.469**	0.024**	-0.526**	0.034**
	(0.003)	(0.002)	(0.004)	(0.003)
POST	0.055**	0.005**	0.034**	0.007·
	(0.004)	(0.003)	(0.006)	(0.004)
POST × TREAT_ONE	0.005	-0.008**	0.013**	-0.011**
	(0.004)	(0.002)	(0.005)	(0.003)
POST × TREAT_TWO	0.098**	-0.012**	0.090**	-0.019**
	(0.004)	(0.003)	(0.006)	(0.004)
Constant	5.157**	0.279**	5.566**	0.259**
	(0.008)	(0.005)	(0.004)	(0.003)
Time FE	Yes	Yes	Yes	Yes
Language FE	Yes	Yes	No	No
Observations	1,004,536	193,528	537,166	107,850
R-squared (adj.)	0.052	0.003	0.052	0.002
F-Statistic	1,233.679**	13.861**	1,9595.902**	15.854**

Note: Control group is defined as streamers who had the status of partner in January 2020. TREAT_ONE is defined as streamers who had the status of affiliate or standard in January 2020. TREAT_TWO is defined as streamers who only entered the sample after January 2020. POST is defined as six weeks since week 12, 2020. Time fixed effects = Week FE, · p<0.1; * p<0.05; ** p<0.01. Robust standard errors in brackets.

Appendix D. Robustness Test

Dep. Variables	(1) Log(total minutes streamed)	(2) Log(total minutes streamed)
Language Selection	Global	English
TREAT	-0.100**	-0.101**
	(0.021)	(0.031)
POST	0.416**	0.308**
	(0.037)	(0.053)
POST × TREAT	-0.084**	-0.035
	(0.028)	(0.043)
Constant	6.401**	6.536**
	(0.064)	(0.040)
Time FE	Yes	Yes
Language FE	Yes	No
Observations	35,106	18,985
R-squared (adj)	0.024	0.007
F-Statistic	22.674**	11.676**

Note: Subsample of streamers who joined the top 500 on Twitch after January 2020 (newcomers). Control group is defined as streamers who had the status of partner in June 2020. TREAT is defined as streamers who had the status of affiliate in June 2020. POST is defined as every week since week 12, 2020. Time fixed effects = Week FE, · p<0.1; * p<0.05; ** p<0.01. Robust standard errors in brackets.

VI. Chapter 4

Buy the rumor, or the news? Quantifying mass and social media influence on retail investors' trading Decisions

with Lisardo Erman

Abstract

We use a longitudinal data set of aggregate stockholder data gathered from zerocommission trading platform Robinhood to measure the degree to which retail investors are influenced by news coverage and social media activity in their decision to open or close stock positions. In line with previous findings, we observe that zero-commission retail traders commonly engage in herding behavior and attention-induced trading by preferentially increasing their positions in stocks with high volume and price volatility. We further find that aggregate trading behavior is strongly influenced both by activity on social media platforms, as well as general news coverage, with the former dominating the latter in effect size. Social media coverage is highly informative of changes in stockholdership in the hours following their dissemination. This effect is especially pronounced for lower market cap equities and penny stocks.

1. Introduction

Recent years have seen the rise of *appified* financial services, which through an accessible interface and low or even non-existent trading fees have attracted a fast-growing user base of predominantly young, inexperienced investors. The concurrent growth in the share of transactions enacted by retail investors [8]has brought with it the increased attention of researchers regarding how retail investors make investment decisions and whether and how this may affect general market functioning and outcomes. A widelyassumed characteristic of ”zero-commission” traders is their affinity for social media. The ”meme stock” frenzy of early 2021, in which high numbers of retail traders coordinated over social media to drive up the price of heavily shorted stocks such as Gamestop and AMC in the hopes of triggering a short squeeze, has received widespread international media attention and has put a spotlight on the degree to which social media hype can influence the decision-making of retail investors. Bradley, Jr., et al., 2021 investigated buyer-side analysis on the *wallstreetbets* subreddit and found that “buy” recommendations on average predicted positive abnormal returns in between those predicted by sell-side analysis and buyer-side analysis published on Seeking Alpha. They further found that retail trading increases in the hours immediately following the posting of investment advice on reddit, indicating that retail investors actively look for and act upon signals on social media.

What — to our best knowledge — is still missing from the literature is the identification of attention triggers that works with data sufficiently disaggregated to assess their relative effect size. Of particular interest in this context is the relative importance of social media versus traditional mass media in triggering the attention of such ”zero commission” investors. It is widely agreed upon that social media significantly differ in function from traditional media because of their democratic and interactive nature. The question to what extent social media have already overtaken traditional media among the young and tech-savvy users of zero-commission trading platform Robinhood could in turn shed light on how the transmission mechanisms of news to the financial market will change in the future. Yu, Duan, and Cao, 2013 investigated the respective effect on stock prices of news and social media and find social media to have a stronger impact on stock returns compared to news media, but its direct impact on trading decisions (which stock prices are a product of), has so far not been investigated. We intend to contribute to closing this gap by quantifying the degree to which both traditional news

[8] As reported by Citadel Securities via BusinessInsider, the share of transactions made by retail investors rose to 25% in July of 2020, from just 10% in 2019. (BusinessInsider, 2020)

media, as well as user-generated content submitted to social media platforms, affect the propensity of retail investors to open and close stock positions.

To do so, we compiled a data set comprised of very large numbers of news articles and social media submissions pertaining to publicly traded stocks. The hype surrounding the aforementioned ”meme stocks” has been heavily attributed to social news site reddit (more specifically the high-risk investment-focused subforum r/wallstreetbets) of which we use an extensive data set covering more than 10,000 submissions made over a timeframe of 7 months. We combine these data with a longitudinal data set comprised of aggregate stockholdership data provided by trading platform Robinhood. The resulting data set is both long (in the time dimension) as well as wide (in the number of assets covered) and allows us to model investor reactions to media triggers with high frequency.

Overall, we find that Robinhood investors are most strongly influenced by high price volatility, which corroborates the findings of (Welch, 2020) and is in line with the attention-induced trading model of retail investor behavior. We find social media volume to be the most informative non-price variable, above trade volume and general news coverage. Reactions to social media coverage are strongest in the hours following increases in social media activity (peaking around 6 hours), which suggests that Robinhood investors actively look for and act upon signals from social media sources, whereas the reaction to news coverage is both quantitatively smaller and more delayed (peaking at 12 hours post event). While the quantity of both news coverage and social media activity have significant and large explanatory value for changes in stockholdership, the impact of their respective qualitative valence is much weaker. This phenomenon has been documented in other studies (i.e. Boehmer et al., 2021.) and is likely explained by the fact that retail traders increase their holdings both during upward, as well as downward market movements, which should coincide with positive, respectively negative, overall market sentiment, thus obfuscating the true effect.

We further investigated multiple possible mediators of the aforementioned relationships and found that when conditioning on market capitalization of the underlying stocks, the results represent those of the baseline model, albeit more extreme. Within this specification, news coverage only significantly influences stocks with high market capitalization, whereas social media coverage is even more influential for small-cap stocks. Overall, zero-commission investors seem to favor stocks with smaller market capitalization, which is in line with previous findings by Beck and Jaunin, 2021 and is further evidence that zero-commission traders are characterized by their interest in volatile stocks with high future return potential.

To further investigate this phenomenon, we specifically investigated penny stocks, which exhibit high volatility due to their low price and relative trade volume. For this specification, the results are again more extreme, as we find that the economic significance of social media coverage is more than twice as large for penny stocks compared to the baseline model. Overall, we find further evidence that zero-commission investors are majorly influenced by social cues in their investment decisions and that this is more strongly driven by social media activity rather than general news coverage, especially so for smaller stocks.

2. How Media Affects Zero-Commission Investors

Based on a comprehensive review of the extant literature with the subject of retail investor reactions to media exposure, we identified four main hypotheses guiding our analysis. The first two concern the absolute and relative impact of both social and traditional mass media on investment decisions of zero-commission investors. First of all, it has generally been established that zero-commission investors are driven by social media activity, with a number of recent studies investigating its role in price discovery (Hu et al., 2021), as well as its ability to predict future abnormal returns and drive retail purchases (Bradley, Jr., et al., 2021; Dim, 2021; Farrell et al., 2021). We thus expect to measure a positive correlation between the volume of social media activity a given stock generates and the change in holders of that stock:

Hypothesis 1: *Due to historic hype events and zero-commission investor's demographic makeup we expect social media to play a significant role in gathering investor attention and thereby inducing trades.*

We similarly expect traditional mass media to significantly influence trades enacted by retail investors, though, extrapolating from studies simultaneously investigating both media variables' influence on stock returns (which generally correlate with increases and decreases in stock positions), we expect this effect to be smaller in size compared to social media (Yu, Duan, and Cao, 2013):

Hypothesis 2: *In line with previous findings for stock returns, we also expect traditional media to stimulate trade by zero-commission investors, but expect its effect to be smaller than that of social media.*

A common finding in the literature has been that zero-commission investors tend to ”buy the dip”, which means they increase positions both during strong upward and downward market movements (Welch, 2020). We thus expect the volume of social media activity to be more informative of zero-commission investor’s behavior than its qualitative valence and thus expect to measure a smaller effect for the latter:

Hypothesis 3: *We know that zero-commission investors react to large price movements in both directions. Similarly, we expect the strength of a media signal to be more important than its valence, i. e. its underlying sentiment.*

A further aspect that specifically characterizes zero-commission investors is their preference for small-cap stocks with potential for large, hype-induced upwards movements, such as penny stocks (Welch, 2020). We therefore expect both the absolute and relative (compared to traditional mass media) importance of social media activity to be even higher for this asset class:

Hypothesis 4: *Based on prominent past herding events, we expect social media’s relative importance to be higher for small-cap and penny stocks.*

We use these four working hypotheses as a basis for our empirical strategy and as a backdrop from which to interpret the resulting model, which we will describe in depth in the following chapters.

3. Data and Methodology

Uniquely amongst trading firms, Robinhood provides a relatively permissive and easily accessible API that allows insight into activities on the platform. Until August of 2020, the number of users holding each asset traded on Robinhood was publicly available via this API, which in turn allowed third parties to aggregate and serve these data. This data is fairly unique in that it covers all assets traded on the platform, which is usually proprietary information and not available in such scale and scope. As a result it has already seen extensive use by researchers investigating retail investors Eaton et al., 2021; Welch, 2020, which in absence of such data had to either limit analysis to smaller proprietary data sets of single trading firms (i.e. Kumar and Lee, 2006; Odean, 1999) or deduce retail trades from order flow data (Boehmer et al., 2021; Bradley, Jr., et al., 2021). Robinhood stopped providing the data on which our model is based in August 2020. Data from news and social media could be sourced from January 2020, giving us 7 months of data. We regard this time period as especially valuable from a research perspective as it precedes the ”meme stock” frenzy of early 2021. Deriving any kind of trading-relevant information from social media is complicated by problems of endogeneity. As it is already generally assumed that social media activity influences stock buying decisions, social media platforms are both already heavily monitored for signals, as well as the target of manipulation efforts. While investment focussed subreddits in general and the Wallstreetbets subreddit in particular have seen steady growth in activity and new users over the past several years, the ”meme stock” frenzy of early 2021 has put an international media spotlight on WSB, resulting in an exponential influx of new users and inorganic content with manipulative intent (Bradley, Jr., et al., 2021. As investment-focused subreddits were still comparatively obscure in the first half of 2020, we argue that submissions within that timeframe still represent organic activity by retail investors, thus allowing us to measure the unadulterated effect of peer-information. In addition, we are interested in the decision making process of ”zero-commission” investors, which have been found to congregate on commission-free trade platforms, such as Robinhood. While we could have used order imbalances to identify periods of net-buying and -selling of retail investors (see i.e. Bradley, Jr., et al., 2021), the ability to directly measure changes in holdings of Robinhood investors allows us to more accurately assess the behavior of that particular type of retail investor. Boehmer et al., 2021’s method mainly identifies retail trades based on order size and frequency, thus trades attributed to retail investors likely subsume a variety of retail investor types, which are likely to differ in their information seeking behavior.

3.1 Data collection

We retrieved holdings data from robintrack.com, covering the period between of May 2nd 2018 to August 13th 2020, which we combined with data on stock prices gathered from the Yahoo Finance API via the *tidyquant* package for R. We further collected additional information on the stocks and ETFs covered in the holdings data from financial data service finnhub.io via its API, including general information (such as industry and outstanding shares, among others), as well as news reports (headlines and short summaries) attributed to each stock covering the period of January 1st 2020 to July 31st 2020. The headline and summary columns contain extensive information, averaging 874 and 460 characters respectively. We further collected social media data (submissions and user comments pertaining to each submission) from social news aggregator reddit via the *pushshift* API (Baumgartner et al., 2020). The data, which covers the four largest investment-focused **subreddits** (topic-specific subforums) r/wallstreetbets, r/investing, r/stocks and r/Robinhood, over the aforementioned period, was subsequently scanned for mentions of stock tickers. To do so, every submission headline and comment text body was matched against each ticker abbreviation both on its own was well as in combination with a cashtag (i.e. ”AAPL” and ”$AAPL”), with the exception of tickers consisting of less than three characters for which only the latter was used to reduce erroneous attribution.

3.2 Measuring stock-specific sentiment

We know that Robinhood traders are prone to participate in episodes that are characterized by hype, which we suspect to be transmitted through social media to great extent. Our aim is thus to distil the potential effect a text has on its reader by classifying news articles and social media posts into positive and negative conveyed sentiment. Extant studies commonly incorporate measures of sentiment towards given stocks by either scoring all text resources on their polarity (i.e. Yu, Duan, and Cao, 2013) or by considering only a subset of observations such as those containing expert analysis and deriving i.e. buy or sell recommendations from them (i.e. Bradley, Jr., et al., 2021; Farrell et al., 2021). An issue arising with the former is that of attribution, as for example financial news routinely contain references to multiple stocks and sentiment is often implicit and heavily context dependent. The latter approach is limited to measuring the impact of expert analysis and does not take into account wider market sentiment and peer information, which we are interested in. For the scoring of news articles and social media posts, we leverage the power and versatility of zeroshot classification using large language models, as implemented in the Huggingface libary for Python (Wolf et al., 2019). We specifically use a pipeline built on a pre-trained large language model called BART (Lewis et al., 2019), which enables the classification of text on arbitrary classes without model retraining. It allows us to tune the wording of our classes in a way that takes into account that the text embedded sentiment is potentially stock specific. Being able to grasp stock specific sentiment is particularly important for information that affects more than one stock as it is, for instance, the case in the title of this article:

”*Tim Cook reveals Mac computers to transition away from Intel-designed chips*”
(www.ft.com, June 23rd 2020)

The zero-shot classifier gets assigned a set of classes and returns a corresponding list of probabilities for each piece of text. In the context of sentiment analysis, often the generic classes ”positive” and ”negative” are used. For the text above, we obtain scores of 0.59 and 0.41, respectively. So the sentiment based on the calculated probabilities is sightly positive. But these classes don’t do justice to the complexity of the underlying sentiment. For the human reader it is quite obvious that it is bad news for Intel as the company loses an important customer and rather positive for Apple as greater parts of the value chain are developed in-house. If we consequently score the article with the two classes ”positive for Intel/Apple”, ”negative for Intel/Apple”. we obtain scores of of 0.08 and 0.92 in Intel’s case and 0.89 and 0.11 in the case of Apple. We decided to apply this more fine-grained scoring technique to all social media posts and news articles, using the models’ confidence level as a score, thus taking uncertainty into account. Consequently, every social media post and news article is assigned a sentiment value between 0 and 1, where a value of 1 denotes a clearly positive sentiment.

3.3 Descriptive Statistics

Our analysis starts on January 1st 2020 and ends on July 31st 2020. The units of observation are hourly stock ticker values for the number of holders on Robinhood (Holders), stock prices (Price), trading volume (Volume), market capitalization (MC), total number of submissions (Reddit) and published news articles (News). All sources report their data in UTC time. For Price, Volume and Holders we attribute the last measured observation in a given hour to that particular hour and drop the remaining observations if existent.[9] We do not want to restrict our sample to trading hours as news articles and social media posts can be emitted at anytime. We therefore include pre-market and after-hours prices and volume in the data set. Additionally, in line with Mu¨ller et al. (1990), we linearly interpolate Price between the last measured closing value and the first value at opening. Similarly, Volume is assumed to be 0 during these periods. The degree of social media and news intensity is measured as the sum of posts or articles emitted during an hour. As all of the described variables exhibit quite a skewed distribution, we apply the logarithm to each. To account for the fact that Volume, News and Reddit contain zero entries, we add 1 to these variables before applying the logarithm. Welch (2020) found Robinhood traders to increase their stock holdings in response to large price movements. Interestingly, this relationship is not restricted to price increases but also applies to situations when shares loose value. We account for this finding by entering the maximum absolute price spread (PriceSpread), i.e. the absolute difference between the maximum and minimum log price in a given hour, into the analysis. We further match the derived media related sentiment scores for news articles (NewsSentiment) and social media posts (RedditSentiment) to the data set by averaging them for each ticker and hour. Whenever no data exist, we assume a sentiment score equal to the stock specific mean sentiment score. If no news

[9] Price and Volume are obtained at a 5 minute interval. Holders mostly have one observation per hour.

articles are available at all for a stock, we assume sentiment being equal to the global mean sentiment score.

Panel A of Table 1 contains summary statistics for our variables of interest. In total, the data set contains 2071 unique stocks and 6,090,966 hourly stock values. Both media related variables (Reddit, as well as News) have a high percentage of zero entries. This is especially the case for Reddit, which exhibit a percentage of non-zero entries of only 0.2%. In contrast, News have a non-zero entry in 1.9% of cases. This is not only due to the fact that reddit posts happen more rarely, but also because they are more concentrated around specific hours — a fact illustrated by the variation coefficient of Reddit that is 22.4 versus 7.9 for news articles. The high sparsity of the media variables naturally also affects the sentiment scores, as these can only be calculated on the basis of existing news articles and submissions to social media. This sparsity is the consequence of working with data of relatively high frequency: the higher the data frequency, the higher the degree of sparsity of media variables. To reduce this sparsity, we could aggregate our data and work with daily or weekly observations. However, sparsity in itself is not an obstacle to identification as long as the total number of non-zero entries is large enough. On the contrary, working with hourly data is a central pillar of our identification strategy, as we will explain later on.

Panel B of Table VI-1 contains pairwise correlations between all variables of interest. As can be seen from Column (1), all variables except of Price and NewsSentiment exhibit a positive correlation with Holders. The negative correlation between Holders and Price has its roots in the well-known fact that Robinhood traders have a preference for so called penny stocks. Although the positive correlations are similarly in line with the theory of attention induced trading, it is probably also driven by the fact that, unsurprisingly, Robinhood traders are more likely to invest in stocks that have a relatively high market capitalization (30% correlation) and news media focuses over-proportionally on larger stocks as well (50% correlation). To identify a causal relationship from Price, Volume, Reddit, RedditSeniment, News and NewsSentiment on retail investor's trading behaviour, we will explore how movements in these variables change the number of holders of a particular stock. By focusing on the within stock difference, we are able to address the described endogeneity problem.

Table VI-1 *Descriptive statistics and Correlations.*

Panel A: Descriptive Statistics

Metric	Holders	Price	PriceSpread	Volume	MC	Reddit	RedditSentiment	News	NewsSentiment
N	6604425	6604425	6604425	6604425	6604425	6604425	6604425	6604425	6604425
#Tickers	2104	2104	2104	2104	2104	2104	2104	2104	2104
Mean	7.04	2.73	0.01	1.14	14528.00	0.00	0.64	0.02	0.61
St. Dev	1.97	1.57	0.02	2.50	84938.19	0.04	0.24	0.16	0.05
VC	0.28	0.58	3.39	2.20	5.85	22.46	0.38	8.09	0.08
Min	0.00	-2.53	0.00	0.00	-99999.99	0.00	0.00	0.00	0.00
Q_05	3.85	-0.07	0.00	0.00	24.85	0.00	0.11	0.00	0.56
Q_25	5.69	1.76	0.00	0.00	212.05	0.00	0.52	0.00	0.61
Q_50	7.02	2.74	0.00	0.00	1149.90	0.00	0.64	0.00	0.61
Q_75	8.30	3.83	0.00	0.00	6016.87	0.00	0.82	0.00	0.61
Q_90	9.55	4.72	0.02	5.99	25300.31	0.00	0.95	0.00	0.61
Q_95	10.37	5.16	0.03	7.17	53046.86	0.00	0.98	0.00	0.66
Q_99	12.08	6.08	0.07	8.93	210661.30	0.00	1.00	0.69	0.79
Max	13.50	8.35	4.10	13.81	2134525.00	2.94	1.00	5.02	1.00

Panel B: Correlations

	Holders	Price	PriceSpread	Volume	MC	Reddit	RedditSentiment	News	NewsSentiment
Holders	1.00	-0.07	0.09	0.17	0.24	0.03	0.04	0.20	-0.02
Price	-0.07	1.00	-0.05	0.16	0.25	0.01	-0.08	0.10	-0.04
PriceSpread	0.09	-0.05	1.00	0.40	-0.01	0.05	0.01	0.02	-0.00
Volume	0.17	0.16	0.40	1.00	0.07	0.03	-0.01	0.09	-0.03
MC	0.24	0.25	-0.01	0.07	1.00	0.04	-0.01	0.35	0.01
Reddit	0.03	0.01	0.05	0.03	0.04	1.00	-0.00	0.04	-0.00
RedditSentiment	0.04	-0.08	0.01	-0.01	-0.01	-0.00	1.00	-0.01	0.05
News	0.20	0.10	0.02	0.09	0.35	0.04	-0.01	1.00	-0.06
NewsSentiment	-0.02	-0.04	-0.00	-0.03	0.01	-0.00	0.05	-0.06	1.00

Panel A contains descriptive statistics for different metrics: **N** (the number of day-ticker data points available), **#Tickers** (the number of unique tickers), **Mean**, **St. Dev** (Standard Deviation), **VC** (variation coefficient), **Min** (Minimum), **Q_05** (5 percent quantile), **Q_25** (25 percent quantile), **Q_50** (50 percent quantile), **Q_75** (75 percent quantile), **Q_90** (90 percent quantile), **Q_95** (95 percent quantile), **Q_99** (99 percent quantile) and **Max** (Maximum). The remaining columns show the metric values for **Holders** (the number of holders on Robinhood), **Price** (stock price in $), **PriceSpread** ($\frac{Price_{max}-Price_{min}}{Price_{max}}$, where $Price_{max}$ and $Price_{min}$ are the highest and lowest prices observed at a given day), **Volume** is the total amount of traded shares in million, **MC** (Market capitalization in million $), **Reddit** (number of submission Titles) and **News** (number of news articles). Panel B displays correlations of all pairwise variables..

4 Modelling approach

Previous studies have found that Robinhood traders are influenced more by attention (Barber et al., 2020; Welch, 2020) than the average investor. However, none of these studies exploit the high frequency characteristic of the Robintrack data set to investigate the relative importance of different sources of attention. We want to fill this gap by modelling intraday trading behaviour with a focus on the influence of news and social media activity. Our identification strategy relies on the following properties. First, we estimate within stock time differences in holdings and thereby mitigate the chances of unobserved heterogeneity. Second, we include all potential attention triggers identified by the literature into our set of control variables. Third, we make a longitudinal data set with hourly frequency the basis of our analysis. In conjunction with our identification assumption that retail traders on average, unlike trading bots, react with a certain lag to signals conveyed by prices and media channels, we are able assess the relative importance of our explanatory variables. The assumption that Robinhood investors do not react contemporaneously, i.e. within minutes, to new information is motivated by the fact that the main occupation of retail investors is not stock trading and they are therefore bound by time constraints. This identification assumption is also consistent with Schroff and Siering (2013). Finally, to prevent estimates being distorted by non-stationarity, we estimate the model in first differences. The panel regression specification takes the following form:

$$\begin{aligned}\Delta log(Holders_{i,t}) = {} & \sum_{s=0}^{24}\beta_{1_s}\ \Delta log(Price_{i,t-s}) + \sum_{s=0}^{24}\beta_{2_s}\ \Delta PriceSpread_{i,t-s} + \\ & \sum_{s=0}^{24}\beta_{3_s}\ \Delta log(Volume_{i,t-s}+1) + \sum_{s=0}^{24}\beta_{4_s}\ \Delta log(Reddit_{i,t-s}+1) + \\ & \sum_{s=0}^{24}\beta_{5_s}\ \Delta log(News_{i,t-s}+1) + \sum_{s=0}^{24}\beta_{6_s}\ \Delta RedditSentiment_{i,t-s} + \\ & \sum_{s=0}^{24}\beta_{7_s}\ \Delta NewsSentiment_{i,t-s} + \alpha + \lambda_i + \gamma_t + \epsilon_{i,t},\end{aligned} \qquad \text{(VI-1)}$$

where Δ denotes the first difference of a variable and the subscripts *i* and *t* correspond to ticker and hour, respectively. *s* is the lag level and spans the intraday period 1−24. To test our assumption that holders do not react contemporaneously to attention variables, we also include $s = 0$ into the specification. α is the global bias, whereas λ_i and γ_t are ticker and hour specific fixed effects. We estimate the model with the *lfe* package for R that efficiently computes fixed effects on multiple groups even for large data sets. Robust standard errors are clustered on both ticker and hourly level.

The first set of explanatory variables have performance related character stemming from the Yahoo Finance API. The log change in prices can be interpreted as stock market return.[10] We do not have a clear-cut hypothesis regarding the sign of its coefficient, β_1, as Pagano, Sedunov, and Velthuis (2021) found Robinhood traders to engage in both momentum and contrarian trading strategies. In contrast, the effect of intraday price spreads on the number of holders, β_2, is expected to be positive as Welch (2020) found Robinhood traders to prefer large price movers. We also include trading volume as control variable as it incentivizes retail trading activity (Reyes, 2019; Barber et al., 2020), which is why we hypothesise β_3 as being positive.

The second set of explanatory variables are media related variables based on news articles and social media posts. Barber et al. (2020) found Robinhood traders to be influenced by the number of news articles published at a given day. Since the majority of Robinhood investors are young and tech-savvy, we suspect that they are not only influenced by online news articles but also by social media posts. We therefore calculate media intensity measures as the count of published news articles (News) and social media posts (Reddit), following the approach of Barber et al. (2020). The coefficients β_4 and β_5 will reveal the relative importance of traditional news versus social media channels for Robinhood traders' decision to buy stock. Another dimension we want to look at is the relative importance of strength and valence of media signals. In other words, we want to investigate whether Robinhood traders are more influenced by the raw intensity of media exposure or by informational content. To achieve this, we add the hourly sentiment scores derived in the previous section to the model.

Finally, through the lag structure of our specification, we will learn how long it takes Robinhood traders to react to changes in the variables. All of these points will help to better understand the attention characteristics of Robinhood traders – a group that has recently proven to be of considerable interest to market makers, regulators and the wider financial services industry.

4.1 Baseline Results

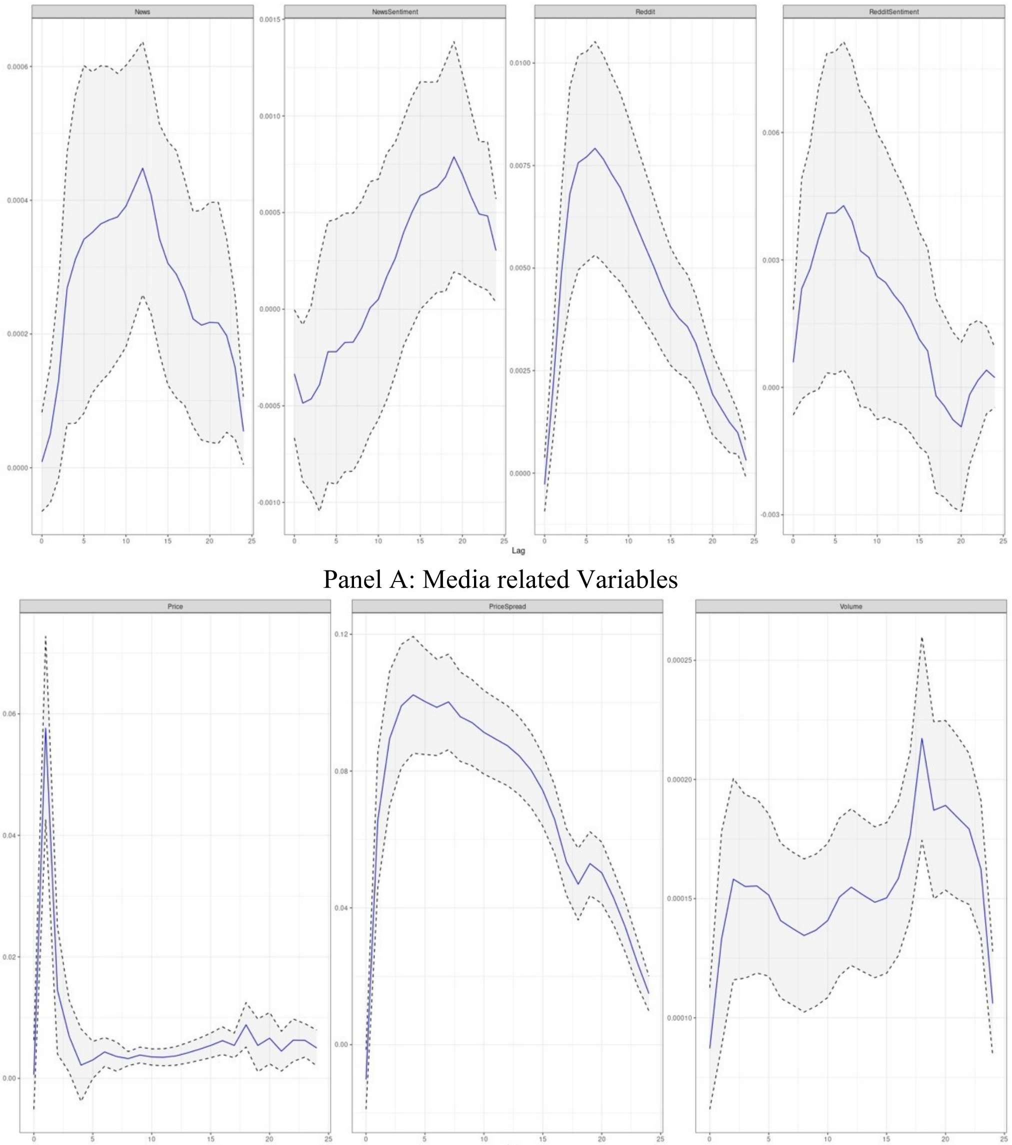
Estimation results of Equation VI-1 visualized in Figure VI-1. Whereas the contemporaneous effects of both performance and media related variables are relatively small and most of the time not significant, we see significant increases of Holders in response to lagged values of all variables. When it comes to the timing of responses, they happen faster after increases in performance related variables with peak estimates at lag values 1, 3 and 4 hours for Price, Volume and PriceSpread, respectively. Media related variables, in contrast,

[10] For simplicity, dividents are ignored in the analysis.

need more time to transmit, with peak lag values of 6, 6, 12 and 17 hours for Reddit, RedditSentiment, News and NewsSentiment, respectively. The corresponding estimates of these peak lags are 0.05764 (Price), 0.10229 (PriceSpread), 0.00016 (Volume), 0.00792 (Reddit), 0.00428 (RedditSentiment), 0.00045 (News) and 0.00063 (NewsSentiment). All estimates except for the sentiment variables can be interpreted elasticities. The overall effect, however, is not limited to a particular hour but is the sum of estimates over the day. The resulting cumulative effects are 1.7386 (PriceSpread), 0.1787 (Price), 0.0067 (Volume), 0.1096 (Reddit), 0.0423 (RedditSentiment), 0.0067 (News) and 0.0050 (NewsSentiment). These differences in effect sizes can partly be explained by different standard deviations of the variables. Multiplying the effect sizes with a variable's standard deviation, therefore, leads to a more accurate metric to assess relative effect sizes. Figure VI-2 shows the result of this multiplication. We harmonize the scale within the groups of performance and media related variables to highlight differences in the effect magnitudes between variables. The y-axes show the percentage increase of Holders after a variable increases by one standard deviation. For media related variables, it becomes apparent that Reddit is the most dominant determinant explaining the decision of Robinhood investors to buy stock. Its peak lag value is by the factor of 6 higher than the corresponding value of News meaning that Robinhood traders not only react faster to social media posts but also with a greater response. Figure VI-2 also reveals that raw media exposure (Reddit, News) has a greater impact on Holders than their sentiment related counterparts (RedditSentiment, NewsSentiment). This is especially true for Reddit, with a peak lag value by the factor of 4.5 greater than the peak lag value of RedditSentiment. For performance related variables, the largest peak effect on retail investment behaviour can be attributed to PriceSpread, a finding that is in line with Welch (2020) who observe a strong positive correlation between absolute returns and the percentage increase in Robinhood holders of a stock. Comparing the effect sizes across groups, we calculate the cumulative effect based on Figure 2. The corresponding numbers are 3.18 (PriceSpread), 0.24 (Price), 0.67 (Volume), 0.50 (Reddit), 0.08 (RedditSentiment), 0.09 (News) and 0.002 (NewsSentiment). These numbers reveal that PriceSpread is by far the most important variable influencing the decision of Robinhood traders to buy stock. The effect of Reddit is comparable to the other two performance related variables, Price and Volume, with an effect that is 50 % larger than the effect for Price and 30 % smaller than that for Volume. The effects of News and NewsSentiment are negligible in comparison to all other variables.

The Baseline results, in summary, confirm that the performance related variables discussed by the literature being important attention triggers to have a significant impact on the numbers of holders. Retail traders on Robinhood react in particular to large movements in prices, irrespective of whether these are positive or negative. Moreover, we find social media to have a cumulative economic effect more than 5 times larger than the cumulative effect of news articles. The baseline results are largely consistent with the attention-induced trading hypothesis: strong price movements in either direction have the strongest impact on trading. For a price signal to have an impact, it apparently has to be strong enough to attract attention, and it matters less whether it conveys positive or negative information. The same applies to media related variables: what attracts attention is the sheer number of news articles and posts published on social media, while the embedded sentiment has a smaller impact.

Figure VI-1 *Baseline Results.*

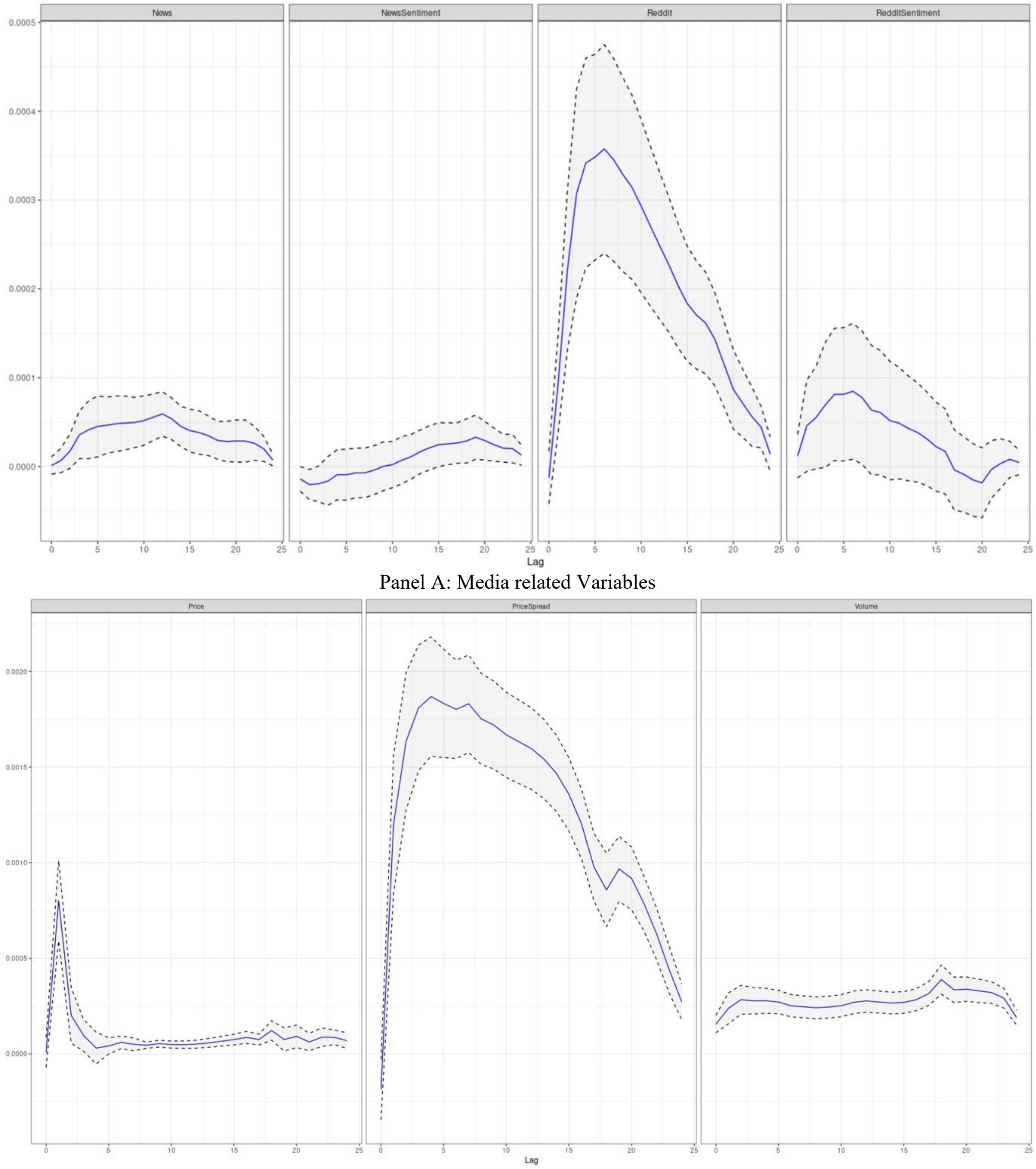


Panel A: Media related Variables

Panel B: Performance related Variables

This chart visualizes the panel regression results from Table Table A1. The blue measures the coefficient’s size. The area between the line above and below the blue line contain the 95% interval. On the x-axis the lag number is depicted.

Figure VI-2 *Baseline Results (Economic Significance)*



Panel A: Media related Variables

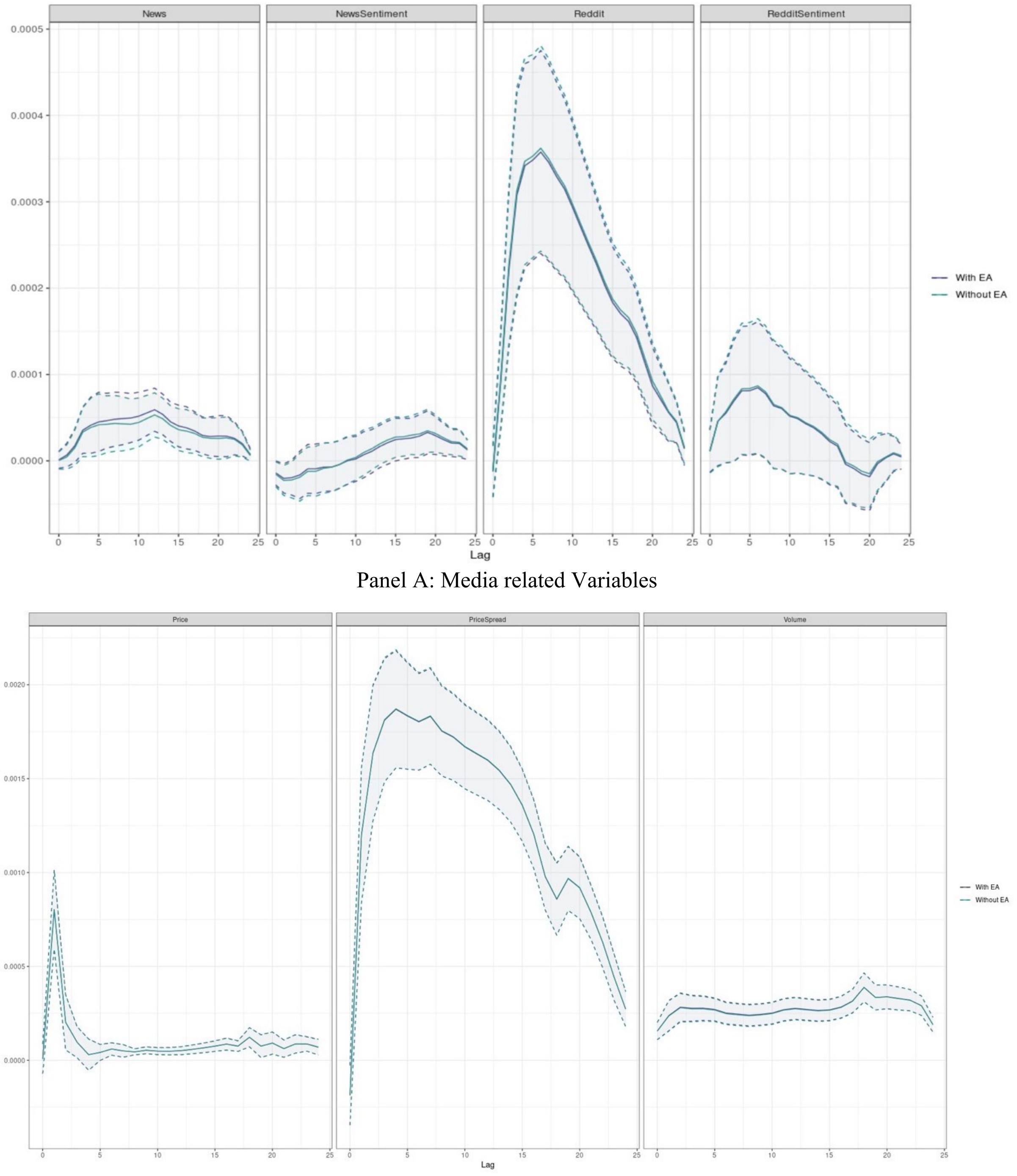
Panel B: Performance related Variables

This chart visualizes the panel regression results. The blue line is the coefficient multiplied by the standard deviation of the variable. The area between the line above and below the blue line contain the 95% interval. On the x-axis the lag number is depicted, whereas the y-axis shows the expected increase in holders after the respective variable increased by one standard deviation.

4.2 Earnings Announcements

The relationship between earnings announcements and investors' reactions to them has been a longstudied topic in the empirical finance literature (e.g., Beaver, 1968; May, 1971; Bernard, 1992; Bartov, Radhakrishnan, and Krinsky, 2000). In recent years this nexus has been complemented by studies focusing explicitly on the influence of investor attention on heterogeneous stock price reactions after earnings announcements (Peress, 2008; Li et al., 2019). As the information contained in earnings announcements is, on average, not expected by investors to some extent, it can lead to timely intraday stock market reactions and therefore potentially immediate movements in the number of Robinhood investors holding a stock that has had an earnings announcement. In this section we want to assure that we do not confuse this effect with the influence of one of our other explanatory variables. For this purpose, we collect information on earnings announcements from the Alpha Vantage API. Since we only receive the date of an earnings announcement and not the timestamp, we cannot include it as a control variable in our intraday analysis. But as our focus is not on studying the intraday reactions to earnings announcements but only to exclude its influence from our analysis, it is sufficient to exclude all observations that relate to a date and stock with earnings announcement. The baseline model data set contains matches to 822 earnings announcements relating to 357 different stocks. These number translate into 19,728 hourly observations that belong to a date with earning announcement, which is about 0.2% of all observations. Given this small percentage share we do not expect our result being distorted by earnings announcements. To test this hypothesis we exclude days with earnings announcement from the estimation. In Figure 3 we visualize estimation results for the subset without earnings announcements against the results from the baseline model. Curves in both graphs are so similar that it is even hard to spot a difference at all. All variables follow an almost identical course in both groups. We can therefore rule out the possibility that our results from the previous section are in any way distorted by the publication of earnings announcements.

Figure VI-3 *Earnings Announcements Split (Economic Significance).*



Panel A: Media related Variables

Panel B: Performance related Variables

This chart visualizes the panel regression results from the earnings announcements sample split. The blue line is the coefficient multiplied by the standard deviation of the variable. The area between the line above and below the blue line contain the 95% interval. On the x-axis the lag number is depicted, whereas the y-axis shows the expected increase in holders after the respective variable increased by one standard deviation. The full regression results are available on request.

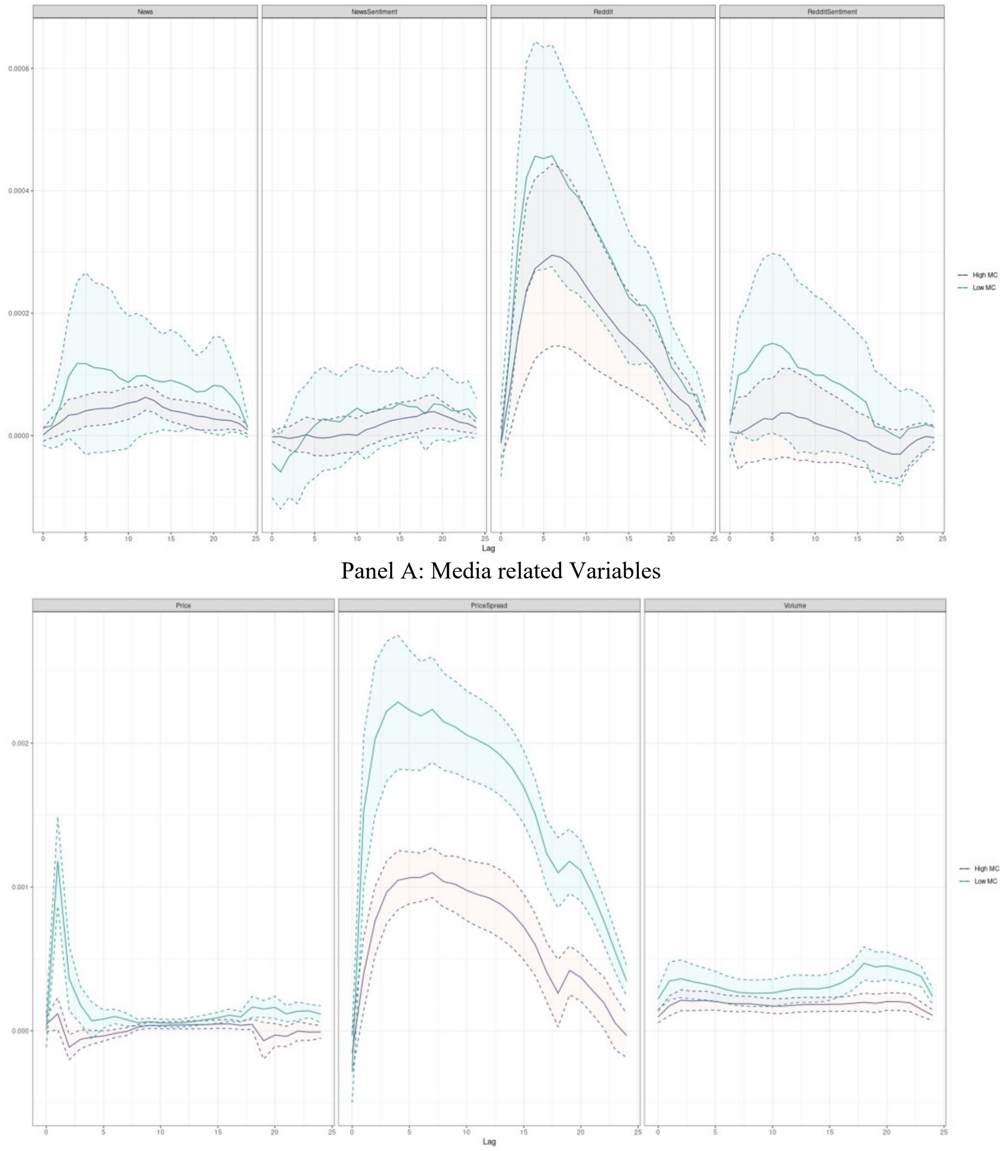
5 Stock Characteristics

In the last section, we examined the impact of news articles and social media posts on Robinhood traders' decision to buy stocks. In this section, we attempt to condition our results on variables that are likely to influence this relationship by dividing our sample into subgroups. Specifically, we divide the sample according to a stock's market capitalisation, its share price, and whether an earnings announcement was released on a given day. The analysis of different subgroups will show whether stock characteristics play a role in transmitting the effects on Robinhood trading decisions. To compare the effects across subgroups, we will look at the economic significance of a variable, i.e. the coefficient multiplied by its standard deviation, taking into account that our variables have subgroup-specific distributional properties. In addition to elucidating the transmission mechanisms of attention-induced trading, this section also aims to test the robustness of our baseline results.

5.1 Market Capitalization

The first candidate to potentially alter the effect of media channels is market capitalisation, as Robinhood investors could choose their source of information based on a company's size and popularity. The notion is that Robinhood traders might rely on traditional news for information about large, established companies, while preferring social media for niche stocks that may receive less attention from reporters working for traditional news outlets. To test this hypothesis, we split our sample at the median market capitalization value of 1.02 billion USD and estimate Equation 1 for each subgroup. This split leads to two almost equal groups of 2,334,881 observations above median market capitalization and 2,304,132 observations below median market capitalization, respectively. Figure 4 visualizes estimation results scaled by a variable's standard deviation. In both subgroups, responses of Holders resemble the main findings from the baseline specification. The most striking difference is the stronger reactions of Robinhood traders in the small-cap subgroup following an increase in Reddit, Price and PriceSpread. While we find a stronger reaction of Robinhood investors to social media posts of small-cap stocks, consistent with our hypothesis, the inverse relationship does not hold for News: Although the News coefficient is significant only in the high-cap group, its economic significance is not larger than in the baseline model. We observe that the main attention triggers from the baseline model play an even greater role in the small-cap sample. Robinhood traders simply seem to be more interested in small-cap stocks — a finding that complements the results of Beck and Jaunin (2021) who show that Robinhood investors are the main drivers of small-cap stocks accounting for 25% of their market capitalisation. They attribute this influence to demand characteristics of Robinhood investors, which, unlike those of institutional investors, are not inelastic. Figure VI-4 suggests that this is not the only explanation, but that the large influence of Robinhood investors on small-cap stocks is also based on their preferences.

Figure VI-4 *Market Capitalization Split (Economic Significance).*



Panel A: Media related Variables

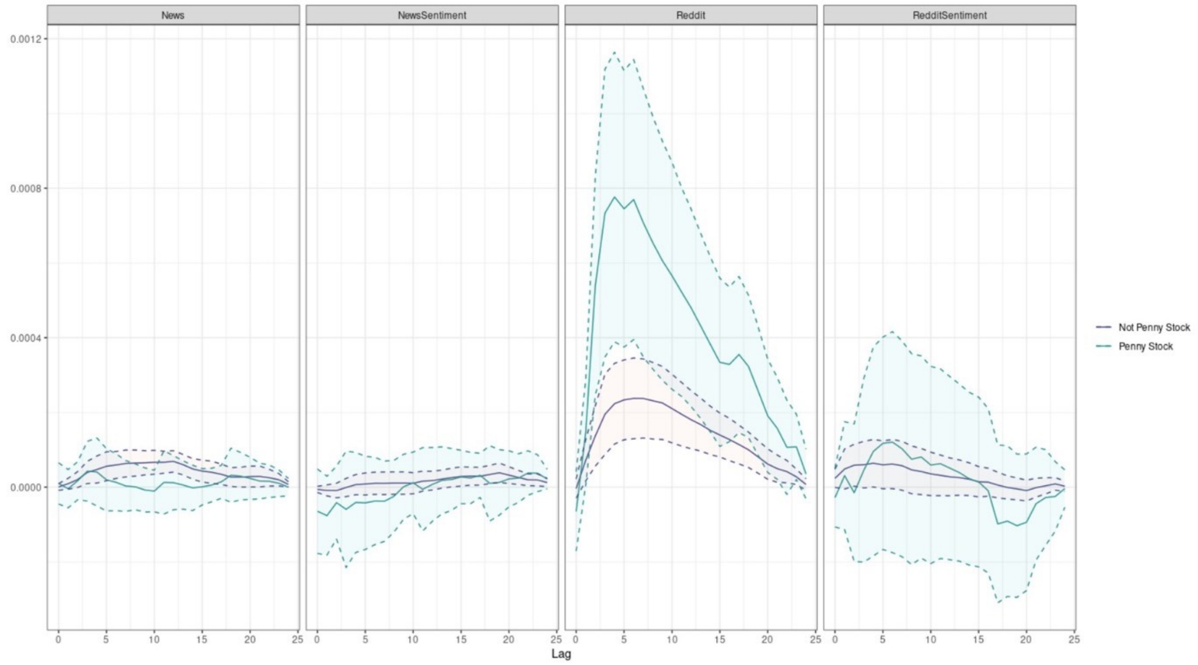
Panel B: Performance related Variables

The blue line is the coefficient multiplied by the standard deviation of the variable. The area between the line above and below the blue line contain the 95% interval. On the x-axis the lag number is depicted, whereas the y-axis shows the expected increase in holders after the respective variable increased by one standard deviation. The full regression results are available on request.

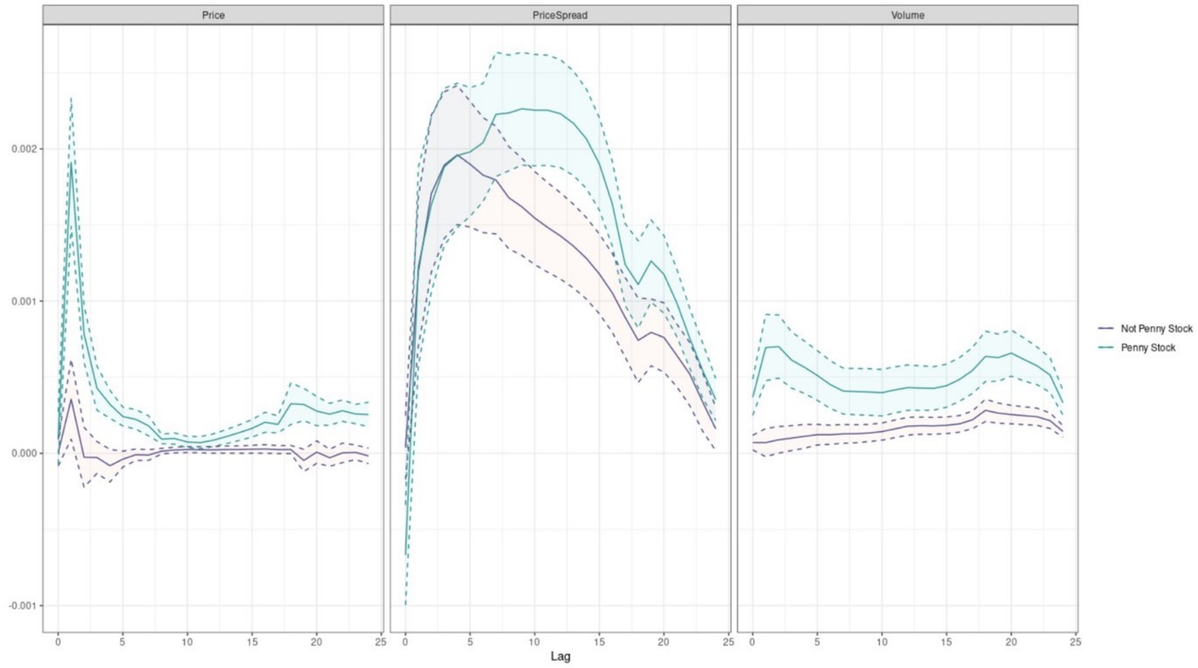
5.2 Penny Stocks

Another variable on which we want to condition the relationship of media exposure and stock holdings, is stock price. Specifically, if a stock can be classified as penny stock. It is well know that Robinhood investor's have a weakness for low priced stocks as they require minimum investment amounts and exhibit lottery features as explained by Kumar (2009). Also the negative correlation between Holders and Price in Table 1 points towards this direction. Penny stocks are often obtained for gambling purposes and are well suited to be targeted in ”pump-and-dump” schemes as their relatively low liquidity leads to high price increases in the pump phase (Leuz et al., 2017). We expect that social media play a key role in initialising purchases of penny stocks, as their democratic and interactive nature facilitates herd behaviour and the shifting of attention to relatively unknown stocks. Following Bradley, Cooney Jr, et al. (2006) we classify stocks with a share price of less than 5$ as penny stocks. Similar to the last subsection, we split the sample at this threshold and re-estimate the model for both subsamples. The penny stocks subgroup consists of 883,796 observations, while the non-penny stocks group comprises 3,755,217 observations. Results are presented in Figure 5. In line with our hypothesis we see a much stronger social media component in the penny stock subgroup. The economic significance of Reddit at its peak value is more than twice as great as the corresponding value in the baseline model. The cumulative effect is 1.02 in comparison to 0.50 in the baseline model. News, in contrast, is not a significant driver of Holders in the penny stock subgroup. The response of Robinhood investors to Price in the penny stock subset is also notable, with a peak economic significance value almost double that in the baseline model. A similar percentage increase of peak economic significance in the penny stock subgroup compared to the baseline model is observed for Volume. Our results reveal that for penny stocks, often traded in the context of gambling and herding, attention of Robinhood investor's shifts towards Reddit, Price and Volume. PriceSpread remains being the strongest influence factor, but its economic significance is comparable across groups.

Figure VI-5 *Penny Stock Split (Economic Significance).*



Panel A: Media related Variables



Panel B: Performance related Variables

The blue line is the coefficient multiplied by the standard deviation of the variable. The area between the line above and below the blue line contain the 95% interval. On the x-axis the lag number is depicted, whereas the y-axis shows the expected increase in holders after the respective variable increased by one standard deviation. The full regression results are available on request.

6 Discussion and contribution

In this paper, our goal was to quantify the degree to which general news coverage and social media activity influence the trading behavior of zero-commission investors. Using a very large data set covering every equity shared on a the largest zero-commission trading platform over a long time period and careful modelling including all major previously identified attention triggers, we find that zero-commission investors are heavily influenced by both types of media coverage, which is in line with and corroborates the extant literature. However, since we simultaneously investigate both variables, we are able to quantify their relative importance and find that social media coverage is significantly more informative compared to financial news and that this effect is especially stark in case of lower market cap equities and penny stocks, which are also relatively more popular with zero-commission investors. Regarding the dynamics of the panel regression models, it can be seen that investors' reaction time differs among the respective variables. While investors take time to react for each variable (as would be expected), the reaction is much more immediate for the social media variables compared to the news variables (peaking at 6 and 12 hours post emission, respectively). While the quantity of coverage is heavily significant both in the statistical and economic sense, the effect of its qualitative valence is much less so. Though not unexpected, as similar findings are reported in the literature, it is impossible to determine whether this is truly indicative of sentiment's comparative irrelevance or an artifact of measurement and attribution errors. To minimize this issue as best as possible, we deliberately used raw data for both news articles as well as social media submissions and manually processed them using state of the art large language models, rather than relying on a black-box aggregate from a financial data provider. Regarding limitations to this study, it bears mentioning that the observed time frame included the first wave of the COVID-19 pandemic, coinciding with increased news coverage of stock market developments and high temporary unemployment, which might limit the representativeness of the observed sampling period. All in all we are confident that our model is well specified to capture the atmosphere surrounding retail trading decisions within the observed period.

To summarize our contribution, we add onto the growing literature on zero-commission trading and its impact on retail investor participation and behavior, by investigating the degree to which attentionbased drivers in the form of general news and social media coverage inform the trading decisions of zerocommission investors. Though whether retail activity positively or negatively impacts market quality is still controversial, they are likely to remain a topic of interest for market participants, makers and - as the congressional hearing triggered by the Gamestop frenzy has shown - regulators and legislators, for the foreseeable future. As we have shown trades to lag behind social media signals by several hours, monitoring the latter might yield crucial information for predicting periods of increased retail activity.

References

Barber, Brad M., Xing Huang, Terrance Odean, and Christopher Schwarz (2020). “Attention Induced Trading and Returns: Evidence from Robinhood Users”. *SSRN Electronic Journal* November. doi: 10.2139/ssrn.3715077.

Bartov, Eli, Suresh Radhakrishnan, and Itzhak Krinsky (2000). “Investor sophistication and patterns in stock returns after earnings announcements”. *The Accounting Review* 75.1, pp. 43–63.

Baumgartner, Jason, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn (2020). “The pushshift reddit dataset”. *Proceedings of the international AAAI conference on web and social media*. Vol. 14, pp. 830–839.

Beaver, William H (1968). “The information content of annual earnings announcements”. *Journal of accounting research*, pp. 67–92.

Beck, Philippe van der and Coralie Jaunin (2021). “The Equity Market Implications of the Retail Investment Boom”. *SSRN Electronic Journal*. issn: 1556-5068. doi: 10.2139/ssrn.3776421. url: https://www.ssrn.com/abstract=3776421.

Bernard, Victor Lewis (1992). “Stock price reactions to earnings announcements: A summary of recent anomalous evidence and possible explanations”.

Boehmer, Ekkehart, Charles M. Jones, Xiaoyan Zhang, and Xinran Zhang (2021). “Tracking Retail Investor Activity”. *Journal of Finance* 76 (5), pp. 2249–2305. issn: 15406261. doi: 10.1111/jofi.13033.

Bradley, Daniel, Jan Hanousek Jr., Russell Jame, and Zicheng Xiao (2021). “Place Your Bets? The Market Consequences of Investment Advice on Reddit’s Wallstreetbets”. *SSRN Electronic Journal*. issn: 1556-5068. doi: 10.2139/ssrn.3806065. url: https://www.ssrn.com/abstract=3806065.

Bradley, Daniel J, John W Cooney Jr, Steven D Dolvin, and Bradford D Jordan (2006). “Penny Stock IPOs”. *Financial Management* 35.1, pp. 5–29.

BusinessInsider (2020). “Retail traders make up nearly 25% of the stock market following COVID-driven volatility, Citadel Securities says”. Accessed: 2022-02-16. url: https:

//markets.businessinsider.com/news/stocks/retail-investors-quarter-ofstock-market-coronavirus-volatility-trading-citadel-2020-7-1029382035? op=1.

Dim, Chukwuma (2021). “Should Retail Investors Listen to Social Media Analysts? Evidence from Text-Implied Beliefs”. *SSRN Electronic Journal* (May). doi: 10.2139/ ssrn.3813252.

Eaton, Gregory W., T. Clifton Green, Brian Roseman, and Yanbin Wu (2021). “Zero-Commission Individual Investors, High Frequency Traders, and Stock Market Quality”. *SSRN Electronic Journal* (March). doi: 10.2139/ssrn.3776874.

Farrell, Michael, T. Clifton Green, Russell Jame, and Stanimir Markov (Sept. 2021). “The democratization of investment research and the informativeness of retail investor trading”. *Journal of Financial Economics*. issn: 0304405X. doi: 10.1016/j. jfineco.2021.07.018. url: https://linkinghub.elsevier.com/retrieve/pii/ S0304405X21004050.

Hu, Danqi, Charles M. Jones, Valerie Zhang, and Xiaoyan Zhang (2021). “The Rise of Reddit: How Social Media Affects Retail Investors and Short-sellers’ Roles in Price Discovery”. *SSRN Electronic Journal*. issn: 1556-5068. doi: 10.2139/ssrn.3807655. url: https://www.ssrn.com/abstract=3807655.

Kumar, Alok (2009). “Who gambles in the stock market?” *The Journal of Finance* 64.4, pp. 1889–1933.

Kumar, Alok and Charles M.C. Lee (2006). “Retail investor sentiment and return comovements”. *Journal of Finance* 61 (5), pp. 2451–2486. issn: 00221082. doi: 10.1111/j. 1540-6261.2006.01063.x.

Leuz, Christian, Steffen Meyer, Maximilian Muhn, Eugene Soltes, and Andreas Hackethal (2017). *Who falls prey to the wolf of wall street? investor participation in market manipulation*. Tech. rep. National Bureau of Economic Research.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer (2019). “Bart: Denoising sequenceto-sequence pre-training for natural language generation, translation, and comprehension”. *arXiv preprint arXiv:1910.13461*.

Li, Ruihai, Xuewu Wang, Zhipeng Yan, and Yan Zhao (2019). “Sophisticated investor attention and market reaction to earnings announcements: Evidence from the SEC’s EDGAR log files”. *Journal of Behavioral Finance* 20.4, pp. 490–503.

May, Robert G (1971). “The influence of quarterly earnings announcements on investor decisions as reflected in common stock price changes”. *Journal of Accounting Research*, pp. 119–163.

Mu¨ller, Ulrich A, Michel M Dacorogna, Richard B Olsen, Olivier V Pictet, Matthias Schwarz, and Claude Morgenegg (1990). “Statistical study of foreign exchange rates, empirical evidence of a price change scaling law, and intraday analysis”. *Journal of Banking & Finance* 14.6, pp. 1189–1208.

Odean, Terrance (Dec. 1999). “Do Investors Trade Too Much?” *American Economic Review* 89.5, pp. 1279–1298. issn: 0002-8282. doi: 10.1257/aer.89.5.1279. url: https://pubs.aeaweb.org/doi/10.1257/aer.89.5.1279.

Pagano, Michael S, John Sedunov, and Raisa Velthuis (2021). “How did retail investors respond to the COVID-19 pandemic? The effect of Robinhood brokerage customers on market quality”. *Finance Research Letters*, p. 101946.

Peress, Joel (2008). “Media coverage and investors’ attention to earnings announcements”. *Available at SSRN 2723916*.

Reyes, Tomas (2019). “Negativity bias in attention allocation: Retail investors’ reaction to stock returns”. *International Review of Finance* 19.1, pp. 155–189.

Schroff, Sebastian and Michael Siering (2013). “Media Sentiment and Leveraged Retail Investor Trading”. *26th Australasian Finance and Banking Conference*.

Welch, Ivo (Sept. 2020). *The Wisdom of the Robinhood Crowd*. Working Paper 27866. National Bureau of Economic Research. doi: 10.3386/w27866. url: http://www.nber.org/papers/w27866.

Wolf, Thomas et al. (2019). “Transformers: State-of-the-art natural language processing”. *arXiv*. arXiv: arXiv:1910.03771v5.

Yu, Yang, Wenjing Duan, and Qing Cao (Nov. 2013). “The impact of social and conventional media on firm equity value: A sentiment analysis approach”. *Decision Support Systems* 55.4, pp. 919–926. issn: 01679236. doi: 10.1016/j.dss.2012.12.028.