CONTEÚDO PATROCINADO E DINÂMICAS DE INFLUÊNCIA NO INSTAGRAM

LUCAS MACHADO DE OLIVEIRA

CONTEÚDO PATROCINADO E DINÂMICAS DE INFLUÊNCIA NO INSTAGRAM

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

ORIENTADOR: OLGA NIKOLAEVNA GOUSSEVSKAIA

Belo Horizonte, Minas Gerais

Junho de 2020

LUCAS MACHADO DE OLIVEIRA

SPONSORED CONTENT AND INFLUENCE DYNAMICS ON INSTAGRAM

Thesis presented to the Graduate Program in Computer Science of the Universidade Federal de Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Olga Nikolaevna Goussevskaia

Belo Horizonte, Minas Gerais

June 2020

© 2020, Lucas Machado de Oliveira.
Todos os direitos reservados.

Oliveira, Lucas Machado de

O48s Sponsored content and influence dynamics on Instagram [manuscrito] / Lucas Machado de Oliveira. — 2020.
xxiv, 70. ; il.; 29cm.

Orientadora: Olga Nikolaevna Goussevskaia
Dissertação (mestrado) — Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação.
Referências: f. 67-70

1. Computação – Teses. 2. Redes sociais on-line – Teses. 3. Internet na publicidade – Teses. 4. Influenciador digital – Teses. 5. Instagram – Teses. I. Goussevskaia, Olga Nikolaevna. II. Universidade Federal de Minas Gerais; Instituto de Ciências Exatas; Departamento de Ciência da Computação. III. Título.

CDU 519.6*04 (043)

Ficha catalográfica elaborada pela bibliotecária Irénquer Vismeg Lucas Cruz CRB 6ª Região nº 819.

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Sponsored Content and Influence Dynamics on Instagram

LUCAS MACHADO DE OLIVEIRA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROFA. OLGA NIKOLAEVNA GOUSSEVSKAIA - Orientadora
Departamento de Ciência da Computação - UFMG

PROFA. ANA PAULA COUTO DA SILVA
Departamento de Ciência da Computação - UFMG

PROF. FABRÍCIO BENEVENUTO DE SOUZA
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 5 de Junho de 2020.

To all of you.

Acknowledgments

This has been a long journey. Longer than expected. The ending is great, but the entire adventure has been greater still. I could not have done all this work without the help and support of many people.

I would like to greatly thank my advisor, Prof. Dr. Olga Goussevskaia, for brilliantly guiding me throughout this project. I think it's safe to say that this whole Master's has been a fun, rewarding, and quite light experience (not many sleepless nights). This most definitely would not have been possible without the amazing orientations I received from Olga. Thank you a lot!

I also need to thank my mother Gisele, my father Paulo, and my sister Luísa for the unending support I've been receiving ever since I was born. You are my foundation, and if I'm here today, it's thanks to you.

There are also a few friends that have coped with me during moments I thought I would break out. Thank you for the long conversations, moments of leisure, and for making sure to keep my social life healthy. I won't write names, because I'm afraid I could forget some. But you know well who you are.

Finally, I could not finish these acknowledgments without thanking the whole body of staff (secretariat, professors, teaching assistants, cleaning staff, everyone) from UFMG and, more specifically, from the Department of Computer Science. In these dark times we are living in, especially for our public education, I am immensely proud to have had the chance of being part of a community that still believes that an empowering and inclusive education is the surest path to a more egalitarian society.

Thank you.

"The most dangerous phrase in the language is, "We've always done it this way.""

(Grace Hopper)

Resumo

Este trabalho analisa conteúdo patrocinado e estratégias de monetização no contexto de marketing de influência no Instagram, uma rede social do Facebook voltada para o compartilhamento de fotos e videos. Coletamos uma base de dados de larga escala, contendo cerca de 5 anos de conteúdo disponível publicamente de influenciadores digitais no Instagram.

Definimos métricas quantitativas para medir a performance de uma estratégia de patrocínio no engajamento com usuários, bem como apresentamos análises do comportamento dessas métricas ao longo do tempo em diferentes grupos de influenciadores, agrupados por número de seguidores. Para isso, definimos também um conjunto de características de estratégia semanais que descrevem o comportamento de postagem de um influenciador. Além disso, implementamos uma metodologia baseada em hashtags para detecção de tópicos em posts, de forma a entender o comportamento de tópicos e tendências no Instagram, e se influenciadores podem se utilizar de tendências globais para aumentar o seu alcance e, consequentemente, sua receita. Em particular, analisamos como a variação na quantidade de conteúdo patrocinado se correlaciona com o crescimento no número de curtidas e comentários semanalmente.

Um dos resultados das nossas análises diz que um aumento no número de propagandas por semana pode impactar negativamente o engajamento do usuário, independente da quantidade de seguidores de um influenciador. Além disso, algumas características que, tipicamente, estão fora do controle de um influenciador, como alinhamento com tendências globais, se mostrou altamente correlacionada com engajamento de audiência. Do lado positivo, nossa análise revelou que existem estratégias que podem ser controladas pelo influenciador e que são capazes de mitigar este impacto negativo, como o ajuste da frequência de postagem, e tamanho do texto. Ademais, o uso estratégico de hashtags específicas de um tópico, não necessariamente aqueles que são tendências globais, se mostraram tão importantes quanto o alinhamento com tendências globais. Nossos resultados esclarecem alguns pontos relacionados a estratégias para melhorar o engajamento do público e o desempenho de estratégias de

monetização em plataformas de mídia social online.

Palavras-chave: Redes sociais online, propaganda na internet, marketing de influência, Instagram.

Abstract

This dissertation focuses on the analysis of sponsored content and ad placement strategy in the context of influencer marketing on Instagram, a photo and video-sharing social network owned by Facebook Inc. We have collected a large scale dataset of Instagram influencer profiles, spanning 5 years, comprised of publicly available data from influencer's public profiles.

We define quantitative metrics to measure a strategy's performance on user engagement, as well as analyze the behavior of those metrics over time in different influencer groups, grouped by size. For this, we also define a set of distinct weekly strategy features that describes how an influencer is posting. Moreover, we also implement a simple hashtag-based methodology to detect topics on posts, in order to shed a light on topics and trends on Instagram and whether influencers can make use of global tendencies in order to boost its reach and, consequently, its income. In particular, we analyze how the variation in the amount of sponsored content correlates with the growth in the number of likes and comments on a weekly basis.

One of the negative results of our analysis is that an increasing number of ads per week can impact audience engagement negatively, no matter the size of an influencer's follower base. Moreover, some features that typically are beyond the influencer's control, such as alignment with global topic trends, has proven to be highly correlated with audience engagement. On the positive side, our analysis revealed that there are strategies within the control of the influencer that might mitigate this negative effect, such as adjustment of the posting frequency and text length, as well as smart usage of mentions. Moreover, the strategic usage of hashtags within specific topics, not necessarily those that are globally trending, has proven to be equally as important as the alignment with global trends. Our results shed light on strategies to boost audience engagement and performance of monetization strategies on online social media platforms.

Keywords: Online social networks, internet advertisement, influencer marketing, Instagram.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Internet advertising has been growing steadily for years. In 2020, around US$ 374 billion are expected to be spent on digital advertising, out of which US$ 105 billion are expected to be spent in social media advertising alone, a share of 28%.[1]

Online ads are more effective and easily measured than traditional media, and few businesses can now do without ads that are bought in an automated fashion using algorithms. The technique allows brands to follow internet-goers wherever they spend time and direct ads specifically at them.

Online social media platforms are especially attractive for advertising because they gather data on users' demographics, consumption patterns, and interests, which allows ads to be aimed at them with unprecedented accuracy. To bring the most out of the ability to target consumers precisely on social media, advertisers are changing their campaigns fundamentally. Instead of creating a single and broad message to be displayed on television, radio, or newspapers, they are producing many variations on a theme, matching each to the subset of consumers they judge most likely to respond to it. Online advertising in social media has become such a big phenomenon, that the Federal Trade Commission [2019] of the United States of America has created guidelines on how to comply with the law when endorsing brands or products. In this context, *influencer marketing* is a relatively new form of advertising, in which a social media user with a large and engaged following is paid to post on social media with or about a brand or products. Influencer marketing is especially popular on Instagram.

Instagram is a photo and video-sharing social networking service, launched in 2010, and acquired by Facebook Inc. in 2012. According to Klear Influencer Marketing [2019], in 2018 Instagram reached one billion users with 500 million monthly active

[1] `www.statista.com/outlook/216/100/digital-advertising/worldwide`

users[2], and influencer marketing grew by 40%. As of May/2020, there have been over 11 million posts with the *#ad* hashtag[3].

Brands seek to reach Instagram audiences as users of the social network show high engagement rates with the displayed content. Industries for Instagram branded partnerships span various segments, Lifestyle and Fashion being the leading ones. Examples of prominent brands on the social platform include Lego, a Danish line of plastic construction toys, with almost 4M followers, and Shiseido, a Japanese line of beauty products, with over 360K followers. Instagram's own user account is the most followed account on the platform, having over 344 million followers.

Although there isn't a strict guideline on how to calculate a price for posts on Instagram, the monetary value of a sponsored post is typically estimated based on user *engagement*, which is a function of the number of *interactions* (*likes* and *comments*) with the posts, and the number of *followers* of the poster. For example, a post by the footballer Cristiano Ronaldo, who is the most-followed individual on Instagram, with more than 214 million followers, has an estimated average value of US$735,386[4]. Analysis and management of Instagram business accounts has emerged as a business of its own, providing design tips about how to boost engagement[5].

Instagram allows influencers to post about virtually anything, ranging from fashion and beauty to sports, food, and lifestyle. Users align themselves with and follow influencers mostly by what they post. In this context, a topic can be defined as a major subject users post about. These topics can be identified by the visual contents of the posted content, as well as the textual content that accompanies it. Some topics are more explored than others, while some can be more popular, attracting more viewers. Klear Influencer Marketing [2019] has shown, for example, that Fashion, Travel, Fitness, and Beauty are the leading influencer categories, making up to 16% of the industry that has partnered with influencers on Instagram in 2019. There are also those topics that can be highly seasonal, like Christmas, New Year's Eve, and Summer, attracting users and influencers alike only during specific periods throughout the year.

Unlike some other online social networking platforms, such as Twitter, Instagram data is not entirely publicly available and notably hard to collect, requiring a long and expensive crawling process Segev et al. [2018]. Therefore, research studies based on Instagram data are relatively scarce. There have been a few characterization studies,

[2]As of January 2019, 120 million monthly active users are form the US, 75 million are from India, and 69 million from Brazil. These 3 countries are the top ranked with the most Instagram users

[3]`www.instagram.com/explore/tags/ad/`

[4]`www.statista.com/statistics/779263/most-followers-instagram-athletes-post-value`

[5]`klear.com`, `iconosquare.com`

such as the one conducted by Segev et al. [2018], some works applying image processing, presented by Jang et al. [2015]; and others addressing influence prediction, such as the work introduced by Pal et al. [2016].

1.1 Motivation

Ever since the introduction of social networks, companies and brands have begun using tools to better reach audiences and increase their revenues. Meanwhile, a whole new category of professions has emerged, such as influencers and social media celebrities. The growth of this new area of marketing has been highly organic and the understanding of user behavior is, sometimes, strongly based on empirical experience.

Our motivation is reaching a better understanding of sponsored content on social media, specifically, on Instagram. Whilst companies and brands are interested in reaching the higher possible audience keeping minimal costs, social media users should be aware of how this type of content affects their followers and to which extent it poses a threat to its overall audience.

1.2 Goals

In this work, we focus on the characterization of sponsored content and the analysis of its impact on user engagement in the context of influencer marketing. We seek answers to questions, such as:

- How can we measure the impact of sponsored content placement or monetization strategies on audience engagement with the profile of an influencer on Instagram?
- How can this (positive or negative) impact be optimized?
- Which factors are relevant to the success of an influencer over time? What kind of control does the influencer have over them?
- How do different topics and global trends impact an influencer's monetization strategy success over time? Can influencers leverage these trends to boost their audience engagement?

Monetization on Instagram is highly connected to the posting of sponsored content, recommendation of brands and products, and paid partnerships. By monetization strategies, we mean the different usages of these tools in order to not only boost user

engagement, but also potentialize the reach and consumption of the advertised product or service.

With these goals, we aim to understand how and when it is interesting for a given influencer to make use of advertisement content for monetization and when it is not. Our results are a first step towards understanding how (disclosed) sponsored content impacts user engagement on Instagram. They might provide guidelines for advertisement strategies for influencers to increase influence and income from sponsored content over time in a sustainable manner. Moreover, our study also sheds a light on topic posting behavior and subject trends on Instagram, and if and how influencers can make better use of global trends in order to boost reach and, consequently, profit with their profiles. In addition, there might be several ways to measure success of a monetization strategy. One obvious way is the direct return a sponsored content caused for a product or service. In this work, we focus on analyzing the success on a influencer's perspective, by looking at the numbers and variations of the engagement with the influencer's profile.

1.3 Contributions

Our contributions are the following. Firstly, we collected a dataset, comprised of over three million posts, created by over seven thousand influencers, that posted at least one sponsored post, disclosed by the use of specific ad-related hashtags, between December 2018 and May 2019. In the referred period, the total number of likes and comments was information publicly available for all (public) posts on Instagram, so we were able to collect and use them in our analysis.

We have performed a characterization of this dataset, in regards to aspects related to sponsored content. We analyzed when and how frequently advertisement posts appear in influencers' posting timelines and how different monetization strategies (which might include both sponsored and regular posts) impact audience engagement. We estimate audience engagement based on the number of likes and comments received by each post in a user's timeline.

We also analyze trends on Instagram. Based on hashtag usage, we group our dataset into different topic categories, such as fashion, beauty, lifestyle, sports, etc, as well as into seasonal categories, such as Christmas, Summer, Black Friday, Halloween, etc. We study popularity and audience interaction trends according to the topics to which each post belongs to. To do that, we present a methodology that uses hashtags for detecting topics on Instagram posts.

We also explore the distribution of topics posted about on Instagram and their behavior in time, as well as analyze the usage of specific hashtags and exploring of trendy topics and how they influence a strategy's performance. With this, we aim to understand how different trends and seasonal events are used in regards to sponsored content and how influencers might benefit from global, local, and seasonal trends to achieve higher reach and engagement.

Secondly, based on the information collected about each post and influencer profile, we propose quantitative metrics to measure the success of sponsored content placement strategies over time. To elaborate these metrics, we used information that was publicly available at the time of collection, such as time of posting and number of likes and comments of a post, as well as the number of followers of an influencer. With this, we were able to inspect the behavior of different posting strategies in time and compare their respective performance.

Thirdly, we analyze how these metrics correlate with different factors, such as posting volume, ad frequency, assignment of hashtags, time of posting, number of followers, and usage of trendy topics and hashtags. In particular, we analyze how the global popularity of each topic and hashtag on Instagram impacts the popularity of individual posts and influencer profiles. For example, fashion has proven to be the most posted-about topic, both in terms of number posts and the number of users. Other highly popular topics include fitness, beauty, food, and travel, which are popular throughout the entire timespan of our dataset. Nevertheless, some seasonal events boost interest in other topics, such as summer, winter, and specific holidays, such as Christmas, New Year's Even, and Mother's Day. Our analysis showed that leveraging those global trends can have an impact on strategy performance.

Lastly, we compare the relative importance of different features to the prediction of the performance metrics, i.e., whether it impacts user interaction positively or negatively, and whether the grows, decays, or remains in stagnation, in terms of audience engagement over time. One of the objectives of our analysis is to understand to which extent an influencer can influence the growth of their own influence, in terms of profile's popularity and audience engagement with their posts, over time. We distinguish between features controlled by the user, such as posting frequency, text length, and usage of hashtags of each post, from features beyond the user's control, such as the number of followers and the global popularity trends of different topics on Instagram. In particular, we analyzed if a post is aligned with some globally trending topic and if it uses some trending hashtags within the influencer's topic of choice. Our results revealed that both aspects are equally and significantly important to the success of a post, even though the former is typically hard to predict, while the latter can be more

easily adjusted and exploited by the influencer.

The results presented in this dissertation are part of the following papers:

- **Oliveira, L.**, Goussevskaia, O. (2020). Sponsored content and user engagement dynamics on Instagram. In *Proceedings of the 35th ACM/SIGAPP Symposium on Applied Computing (SAC'20).*

- **Oliveira, L.**, Goussevskaia, O. (TBD) Influencer marketing: topic trends and user engagement on Instagram. Submitted to the *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'20).* [**Under review**]

1.4 Dissertation outline

This dissertation is organized as follows:

Chapter 2 briefly reviews research efforts on influencer marketing and Instagram from different perspectives. First, we discuss efforts related to online advertisement over different media sources. Second, we present some studies addressing online advertisement perception. Third, we present a few results regarding Instagram data analysis. Fourth, we present some miscellaneous work on Instagram data. Then, we present some researches focusing on the problem of topic modeling in short-text and on Instagram. Finally, we present our conclusions and comparisons between previous efforts and our work.

Chapter 3 presents our methodology for data collection, preprocessing, and analysis. We also introduce an initial overview of the collected dataset along with two different subsets of the collected dataset: the *Posts* dataset and the *Influencers* dataset. We, then, present some limitations of our collected data.

In Chapter 4, we propose quantitative metrics to analyze user engagement and monetization strategy impact. We also present our methodology for user grouping. Moreover, we also introduce our methodology for defining the time granularity used throughout the entire work.

Chapter 5 introduces our hashtag-based methodology and our efforts in detecting topics on Instagram. Likewise, we also introduce a few metrics to quantify topic engagement over time, as well as the detection of global and seasonal trends.

Chapter 6 presents a characterization of the collected dataset, following three main axes: the *Posts* dataset, presenting analysis related to the posts contained in the data, especially in regards to user engagement, the impact of hashtag usage and ad

placement strategies; the *Influencers* dataset, presenting our efforts in characterizing the influencers in the collected data, with information such as size, the number of posts, likes, comments and the performance metrics for each different user group; and a section analyzing *Hashtag usage, Topics, and Trends*, in which we try to quantify and identify trends on Instagram, by leveraging the set of hashtags of the posts dataset.

In Chapter 7, we analyze the correlation between different features present in our data with the proposed success and impact metrics. We also use machine learning techniques to quantify the relative feature relevance in the prediction of an ad placement strategy performance. Moreover, we also present a discussion about features that can be tweaked by the users versus features beyond the influencer's control.

Finally, Chapter 8 concludes the dissertation and discusses directions for future work.

Chapter 2

Related Work

In this chapter, we review the literature in five different contexts. First, we present research efforts in describing and understanding the online advertisement phenomenon. Secondly, we discuss some efforts in targeting online advertisement perception. Thirdly, we explore works focusing on Instagram data analysis. Next, we introduce some efforts in topic modeling on short texts on Instagram. Finally, we present the concluding remarks.

2.1 Online advertisement

Research addressing online advertisement has been conducted with different perspectives and approaches. An early study regarding online advertisement was performed by Shatnawi and Mohamed [2012]. In this work, the authors present an overview of online advertisement selection and summarize the main technical challenges and open issues in this field. The paper investigates most of the relevant existing approaches carried out towards this perspective and provides a comparison and classification of these approaches. More specifically, the researchers explore three categories of internet advertising (sponsored search, contextual matching, and shopping websites advertising), and provide an extensive comparative study of the existing approaches in these categories.

With the emergence of the internet, online advertisement on these platforms has become especially attractive because it made it possible to gather data on users' demographics and consumption patterns, allowing ads to be targeted with unprecedented accuracy. In this context, Cramer [2015] conducted a study on native advertisements, which are those types of ads that are highly cohesive with a platform's normal content. The authors conducted an experiment to analyze the perceived quality of native ads

in relation to the website's overall quality. The authors concluded that, even though native ads had high perceived quality, they still tend to hurt the perceived quality and credibility of a website.

Still in the context of native advertisement, Koutsopoulos and Spentzouris [2016] presented a methodology for optimal native ad selection and allocation in social media post feeds, by proposing a mathematical formulation for optimizing a metric which combines *(i)* platform expected revenue, and *(ii)* uncertainty in revenue. In this work, the authors argue that in native advertising, the ad click probability may depend on three things: *(i)* relevance (i.e. context similarity) of the ad to preceding posts, *(ii)* the distance between consecutively shown ads, and *(iii)* the position of the ad in the stream. Following the proposal of the model and mathematical formulation, the researchers show that the problem above is an instance of a resource-constrained minimum-cost path one on an appropriately defined directed acyclic graph. The solution path reflects the policy of selecting which ads to show in the feed (out of a given set of ads), and in which positions to place them.

Ever since the introduction of online advertisement, it has become the main source of income for many companies and professionals. Likewise, systems in which sponsored content is integrated into, are developing solutions to maximize advertisement efficiency and, consequently, ad revenue. Being so, some studies, like the one presented by Wang et al. [2019], propose strategies for revenue maximization in sponsored search. The authors present two innovative metrics to evaluate the ads ranking model's performance and prove that optimizing the proposed metrics is approximately equivalent to maximizing expected revenue. The paper presents some experiments on real-world platforms to show that the proposed methods perform better than the state-of-the-art methods.

Another research effort in understanding online advertisement was conducted by Loude [2017]. This work focused on analyzing two different factors to recognize influencer marketing on Instagram: Instagram use and types of disclosure; and the effects of unethical disclosure practices on the user's perceived value of the advertisement. The study found that Instagram use and ambiguous disclosure can impact the recognition of influencer marketing, while unethical disclosure practices may not have a direct effect on brand attitudes and advertising value.

Evans et al. [2017] examined the effect of disclosure language on ad recognition, brand attitude, purchase intent, and sharing intention among students. The results indicate that disclosure language containing the text "Paid Ad" affect positively the ad recognition, which subsequently interacted with participants' memory of disclosure and mediated the effect of disclosure language on attitude toward the brand and sharing

intention.

Online advertising is being used extensively by multiple industries and companies. Still, there are some drawbacks. Pinder [2017] conducted a study in understanding and stimulating a discussion about the need for and possible incarnations of anti-advert technology. The authors argue that advertisers are increasingly using pervasive and non-conscious routes to emotionally manipulate people, and present a discussion about design and ethical issues in giving users tools to counteract emotionally manipulative ads.

2.2 Online advertisement perception

As well as traditional advertisements, such as television, newspaper, and radio, online advertisement also has the purpose of targeting potential consumers and inducing a consumption intent. Given the increased reach of online advertisement, this poses some interesting opportunities, as well as some challenges. While, as stated before, online ads can be targeted with high accuracy and velocity, an early report by Lithium Technologies [2016] stated that about 56% of millennials were decreasing or stopping social media usage due to the advertisements in their news feed. In this context, some studies have been conducted in understanding online advertisement perception and its impact on users.

Mathisen and Stangeby [2017] studied how users perceive ad content on Instagram. They conduct two different studies to determine which attributes in an ad, users notice and favor. The first study is an exploratory study utilizing qualitative cognitive mapping to address the key attributes for ad evaluation, while the second study tests overall ad evaluation using conjoint analysis to determine which attributes have the largest positive and negative effects. The research found that brand, endorser, and ad type (native obvious) all predict ad effectiveness and purchase intention.

Some studies analyzed how online targeted advertising platforms can have a negative impact by being intrusive (Zhao et al. [2017]), offensive (Zhou et al. [2016]), or by discriminating against users belonging to sensitive groups, i.e., excluding users belonging to a particular race or gender from receiving their ads (Speicher et al. [2018]).

In Zhou et al. [2016], the researchers explored the notion of ad quality from the user's perspective. The authors argue that providing a good user experience with the served ads is crucial to ensure long-term user engagement. They developed a framework to predict the quality of native ads and concluded that to quantify ad quality, the offensive ad rate, as informed by the user, is more trustworthy than the commonly

used click-through rate metrics. They also show that the developed model is efficient in the detection of offensive advertisements.

Following the same line of work, Zhao et al. [2017] adopted a hedonic/utilitarian approach to explore which brand type is more popular on Instagram, and what kind of Instagram sponsored ad is more effective in terms of causing less perceived intrusiveness and driving higher consumer engagement intentions. Moreover, the study examined if the match between brand and sponsored ad with regard to hedonic/utilitarian attributes primacy could lead to more desirable marketing outcomes. The authors concluded that perceived advertising intrusiveness have a negative influence on consumers' engagement intentions on Instagram, and revealed that perceived intrusiveness was negatively associated with all the proposed positive consumer responses, while also being positively associated with consumers' avoidance of sponsored content.

Another research effort in evaluating the perception and recognition of sponsoring on social media was conducted by De Jans et al. [2020]. In this paper, the authors present an experimental study to examine whether or not the source of the sponsored content, which can be endorsed by the brand itself or by a social media influencer who has been compensated to promote the brand, plays a crucial role in determining advertising effectiveness among adolescents. The results revealed that influencer posts lead to higher brand liking, whereas brand posts lead to higher brand awareness. Furthermore, influencers are more greatly admired, whereas brands are perceived as more credible.

There have also been some efforts in understanding the downside of online advertisement, such as ad transparency (Andreou et al. [2018]) and discrimination in online targeted advertising (Speicher et al. [2018]).

Given the rise in privacy concerns that targeted advertising has been subject to, social media platforms, such as Facebook Inc., have developed some transparency mechanisms. In the context of Facebook, Andreou et al. [2018] investigate the levels of transparency provided by two of Facebook's mechanisms to increase transparency in its platforms. The researchers define several key properties of explanations and then evaluate empirically whether Facebook's explanations satisfy them. Their results have shown that ad explanations are often incomplete and sometimes misleading while data explanations are often incomplete and vague.

Apart from the privacy and transparency issues, the possibility of meticulously targeting advertisements to specific audiences has also risen other issues, such as the discrimination against sensitive groups. Facebook Inc., for example, was targeted with an intense media criticism and civil rights lawsuits, for allowing advertisers to discriminate against users belonging to sensitive groups, by excluding users belonging to a

certain race or gender from receiving their ads (Angwin and Parris Jr. [2016]). Such criticism has led Facebook to develop mitigations to prevent advertisers from creating discriminatory ads on its platform (Facebook [2017]).

This exact phenomenon, and its implications, were studied by Speicher et al. [2018]. The authors argue that discrimination measures should be based on the targeted population and not on the attributes used for targeting, and question whether the mitigation strategies applied by Facebook are, in fact, sufficient. For this, the paper presents an investigation of the different targeting methods offered by Facebook for their ability to enable discriminatory advertising. The researchers have shown that malicious advertisers can create highly discriminatory ads without using sensitive attributes.

2.3 Instagram data analysis

Instagram is a photo and video-sharing social networking service, launched in 2010. It has over 1 billion users with 500 million monthly active users. This platform is extensively used by companies and brands to advertise products and services. The amount of data produced daily on Instagram is huge, with over 100 million posts being uploaded daily[1]. This has led to several research efforts in understanding Instagram usage and impact, as well as studies using Instagram data.

An early quantitative analysis of Instagram data was performed by Araújo et al. [2014]. In this work, the authors investigate user practices in Instagram, based on a dataset comprised of 1.265.080 publicly accessible photos and videos posted by ordinary and popular Instagram users. Some of the results have shown that, for instance, users tend to concentrate their posts during the weekend and at the end of the day. Furthermore, people tend to endorse photos with many likes and comments, inducing the rich get richer phenomenon.

Another early work on Instagram data was presented by Ferrara et al. [2014], in which the authors collected a dataset comprised of all posts uploaded by 2100 users, and investigate three major aspects on this dataset: *(i)* the structural characteristics of its network of heterogeneous interactions, to unveil the emergence of self-organization and topically-induced community structure; *(ii)* the dynamics of content production and consumption, to understand how global trends and popular users emerge; *(iii)* the behavior of users labeling media with tags, to determine how they devote their attention and to explore the variety of their topical interests. Their work provides clues

[1] www.omnicoreagency.com/instagram-statistics/

to understand human behavior dynamics on socio-technical systems, such as users and content popularity, the mechanisms of users' interactions in online environments, and how collective trends emerge from individuals' topical interests.

A more recent and extensive analysis of Instagram usage was conducted by Klear Influencer Marketing [2019]. According to the report, in 2018 Instagram reached one billion users with 500 million monthly active users[2], and influencer marketing grew by 40%. As of May/2020, there are over 11 million posts with the *#ad* hashtag[3]. This study uses a dataset of 2 million Instagram sponsored posts that included #ad hashtags, between January and December 2018. Some of the presented statistics support that brands prefer micro-influencers, i.e., users with relatively few followers, but high engagement, over celebrities with large follower bases but relatively low audience engagement.

Another recent study focuses on the problem of scoring and ranking influential users on Instagram (Segev et al. [2018]). For the purpose of this study, a set of Instagram data, with a total of 940.439 posts by 115.044 Instagram users, was collected. Among the millions of users, this work has shown that photos shared by more influential users are viewed by more users than posts shared by less influential counterparts. The authors raise the question of how to identify those influential Instagram users. To address the issue, the paper discusses the lack of relevant tools and insufficient metrics for influence measurement, presents a network-oblivious approach, arguing that graph-based approaches used in other OSNs are a poor fit for Instagram, due to the absence of such graphs, and because building them for Instagram users requires a great deal of resources, e.g., crawling time and computing costs.

In Pal et al. [2016], the authors present a novel methodology for discovering topical authorities on Instagram. The authors affirm that while the large volume of user-generated content is the application's notable strength, it also makes the problem of finding authoritative users for a given topic quite challenging. The paper also argues that discovering topical authorities might lead to better and more relevant recommendations for users, as well as aid in building a catalog of topics and top topical authorities in order to provide a solution to the *cold-start* problem, which is a problem recommendation algorithms face when providing recommendations to new users. Since the algorithms have little to no inputs about new users, the recommendations are typically broad and generic. To tackle this issue, the researchers present the Authority Learning Framework (ALF) to find topical authorities on Instagram. This framework is

[2]As of January 2019, 120 million monthly active users are form the US, 75 million are from India, and 69 million from Brazil. These 3 countries are the top ranked with the most Instagram users

[3]`www.instagram.com/explore/tags/ad/`

based on the self-described interests of the follower base of popular user accounts. The authors also perform experiments that demonstrate that ALF performs significantly better at user recommendation task when compared to fine-tuned and competitive methods.

Another recent research effort went in a quite different direction than the aforementioned works. Song et al. [2020] have performed a study focusing on emotional analysis and classification model development. This study is based on a large-scale dataset of 120.000 images, reflecting posters' emotions. The researchers develop color and content-based emotion classification models by considering: *(i)* the dynamics of SNS, in terms of the volume and variety of images shared, and *(ii)* the fact that people express their emotions through colors and objects. The paper results demonstrate the comparable performance of the proposed model with models proposed in prior studies and discuss its applications and limitations.

2.4 Topic modeling on Instagram

Topic detection on social media has also been extensively studied by researchers. This task is especially challenging on short texts, due to their less semantic information and high sparseness.

An early effort has been made by Zhao et al. [2011], in which they propose the Twitter-LDA method for topic modeling on short texts. The researches aimed at empirically comparing the contents of Twitter with a traditional news medium, the New York Times, using unsupervised topic modeling. For this task, since standard LDA may not work well with Twitter because of the short size of tweets, the paper proposes a new approach, so-called Twitter-LDA, and showed its effectiveness in extracting topics from short texts, in comparison with previous models.

Another effort in short-text topic modeling and text classification has been made by Chen et al. [2016], where they develop a short text classification method based on the Latent Dirichlet Allocation model and K-Nearest Neighbor algorithm. The authors argue that the probabilistic topics generated by the LDA model help make the texts more semantic-focused and reduce its sparseness. They also present a novel topic similarity measure method with the topic-word matrix and the relationship of the discriminative terms between two short texts. The paper also presents some experiments showing the effectiveness of their proposed model.

Some studies also focused on topic modeling as an alternative way for image annotation on Instagram. Argyrou et al. [2018] have used Latent Dirichlet Allocation

(LDA) model to discover latent topics in a collection of Instagram hashtags of a specific subject and quantify the topic similarity by calculating topic coherence. The authors argue that since a topic is composed of a set of related terms, the identification of the visual topic of an Instagram image, through their proposed method, provides a plausible set of tags to be used in the context of training automatic image annotation methods.

In yet another study, Giannoulakis and Tsapatsoulis [2019] explored hashtag filtering on Instagram. The authors have shown that the application of a modified version of the hyperlink-induced topic search (HITS) algorithm, in a crowd-tagging context, provides an effective way for finding pairs of Instagram images and hashtags, leading to more representative and noise-free training sets for content-based image retrieval.

2.5 Concluding remarks

In this chapter, we have presented several studies addressing online advertisement, as well as Instagram data analysis.

Our work pursues a similar objective as the one performed by Mathisen and Stangeby [2017], but it differs because we focus not on the perception of ad content, but on the impacts of sponsored content on influencers' profile. For this task, we have used a similar strategy for data collection as the one presented by Klear Influencer Marketing [2019], in which posts were collected by using the *#ad* hashtag. In our work, we have included more ad-describing hashtags, which also included posts for Brazilian influencers. Our research also differs from the aforementioned related work, as we perform a quantitative analysis of how sponsored content and monetization strategies can affect audience over time, in the context of influencer marketing.

In terms of advertisement perception, we focus our efforts on analyzing the impact of monetization strategies on the engagement with the influencers' own follower base. While previous works have focused on the impact of advertisement content on consumer's perception and engagement, we focus our analysis on understanding the impact on the engagement with the poster's profiles.

Regarding topic detection on short-texts, more specifically, on Instagram, our proposed methodology differs greatly from the works presented in the previous section. We propose a hashtag-based methodology, combined with a hierarchical clustering algorithm.

We also present a discussion related to features within and beyond influencer's control, and how influencers can leverage seasonal trends to boost its monetization

strategy performance.

To the extent of our knowledge, this is the first in-depth study of the relationship among sponsored content placement, global topic trends, and audience engagement on Instagram. We hope our result might provide a first step towards understanding how (disclosed) sponsored content and topic trends impact user engagement on Instagram, as well as providing guidelines for advertisement strategies for influencers to increase influence and income from sponsored content over time in a sustainable manner.

Chapter 3

Instagram Sponsored Content Dataset

In this Chapter, we describe the dataset used for the analyses performed in this work. Section 3.1 presents the steps taken to collect e pre-process the dataset. In Section 3.2, we present an overview of the collected dataset, with details such as size and features. Finally, in Section 3.3, we present some limitations of the collected dataset.

3.1 Data collection

Due to API limitations, the public timeline of Instagram users is not easily accessible. Instagram's Private License, which is used to interact with a user's private account, does not provide access to any other user's profile data unless explicitly authorized by the profile's owner. Therefore, we built a web crawler in Python to access the data publicly available from Instagram's desktop web page.

Our data collection was comprised of three main phases:

1. Collection of the most recent posts containing ad-related hashtags: `#ad`, `#sponsored`, `#publi`, and `#patrocinado` (the latter two are commonly used in Brazil to disclose ads): in this step, we crawled the search result's page of Instagram, when searching for the 4 different ad-related hashtags mentioned above[1]. This page presents the top and most recent posts containing that specific hashtag.

2. Selection of influencer profiles, associated with the posts collected in phase (1): in this phase, we extracted the set of influencers that created the posts collected

[1]The search result page for the `#ad` search, for example, can be found on `https://www.instagram.com/explore/tags/ad/`

on phase (1). Out of those, we selected only those with at least 1000 followers, resulting in a total of **7,583 distinct influencers**

3. Collection of the **500 most recent posts** of the selected influencers in phase (2): in this final phase, we again crawled the public profiles of the 7,583 influencers selected in phase (2), collecting the information of the 500 most recent posts by each influencer, resulting in **3,450,733 posts**.

The collection took place between **December 2018 and May 2019**. It is interesting to note that, since July 2019, the information about the number of likes and comments received by a post has become no longer publicly available on Instagram in selected countries. Therefore, it has become significantly harder to collect a dataset of this kind.

3.2 Dataset overview

After collection and pre-processing, we obtained two datasets, which we refer to as *Influencers* and *Posts*. In the subsections below, we describe both datasets:

3.2.1 Influencers dataset

The *Influencers* dataset is comprised of 7,583 entries, with the following attributes. For each influencer, we also collected the sequence of $k \leq 500$ most recent posts.

- **username**: a unique name of the user profile on Instagram;
- **biography**: a free text field with a user provided short bio;
- #**followers**: the number of followers;
- **mediacount**: the number of posts ever posted by the user;
- **posts**: the sequence of $k \leq 500$ most recent posts actually collected for the user profile.

Based on these attributes, we also computed success metrics, as defined in Section 4.2, for each collected user profile $u \in \mathcal{U}$.

Note that the influence dynamics metrics were calculated using only the $k \leq 500$ most recent posts that were actually collected for each user.

We partitioned the set of collected influencers in three categories, according to the size of their follower base:

- **Beginners**: $1{,}000 \leq \#followers \leq 10{,}000$: These user profiles can be seen as up and coming influencers, with a limited reach;
- **Micro-influencers**: $10{,}000 < \#followers \leq 100{,}000$: User profiles with a follower base between 10.000 and 100.000 followers; Micro-influencers have a decent number of followers and, thus, a high reach.
- **Celebrities**: $\#followers > 100{,}000$: User profiles with a follower base higher than 100.000 followers. Celebrities are those with a consistent and huge follower base. Celebrities are, usually, involved in the entertainment industry, movies/television, or sports.

Out of the 7,583 entries in the *Influencers* dataset, **3,253 were tagged as beginners**, **3,323 as micro-influencers**, and **1,007 as celebrities**.

3.2.2 Posts dataset

The *Posts* dataset is comprised of the $k \leq 500$ most recent posts owned by the influencers in the *Influencers* dataset, having a **total of 3,450,733 posts**, with the following attributes:

- **caption**: the post's textual content;
- **caption_hashtags**: hashtags included in the caption;
- **caption_mentions**: references to other user profiles mentioned in the post's caption;
- #**likes**: number of likes;
- #**comments**: number of comments;
- **date_local**: local date of posting;
- **location**: location of the post (not always available);
- **owner_username**: the owner of the post;
- **url**: URL of the post.

Based on these attributes, we computed the following features for each collected post $p \in \mathcal{P}$:

- **is_ad**: flag to identify AD (sponsored) and NAD posts.
- $\mathcal{I}(p)$: interaction number, as defined in Section 4.2.

To classify a post as being sponsored (AD) or not (NAD), we searched, in every post, for at least one of the ad-disclosing hashtags, listed in Section 4. We ended up with **264.518 AD posts and 3.186.215 NAD posts**.

It is interesting to point out that, in selected countries, the information about the number of likes and comments received by a post has become no longer publicly available on Instagram since July 2019, making a dataset of this kind even harder to collect.

3.3 Dataset limitations

Due to the characteristics of our collection process, we were unable to obtain the number of followers of a user profile at the time of publishing of a given post. Because of this limitation, we were unable to use a more widespread metric, called *engagement*, which is a function of the number of likes and comments of the post and the number of followers of the poster, at the time of posting. Instead, we propose an alternative metric of audience engagement (Sec. 4.2).

To characterize the ad placement strategy of an influencer, we analyzed how the number of sponsored posts changed between consecutive weeks in an influencer's timeline. We chose a time granularity of one week because, using this time window, we observed the most pronounced variation in the number of ads between consecutive time windows (Sec. 4.1). Therefore, to filter out irregular activity, our analysis included only influencer profiles with a minimum activity of three posts per week.

Our methodology for topic detection (as described in Sec. 5.1) is based on hashtag usage and clustering methods. There are several other ways of detecting topics, such as image processing and topic modeling. Consequently, our approach was able to detect only a subset of all topics contained within Instagram, which include a wide range of different topics and subjects that might not be included in the set of topics we have detected.

Finally, we would like to point out that, when categorizing a post as a disclosed sponsored post (AD) or a regular post (NAD), we certainly introduced a lot of false negatives, given that we checked for a limited number of hashtags that typically disclose an advertisement (Chapter 4). Also due to the characteristics of our collection, we were not able to identify posts that uses Instagram's native *paid partnership* functionality, since this information was not available in our crawler. As a consequence,

many sponsored posts in our dataset have not been labeled as such, due to the existence of several alternative ways of disclosure, like *paid partnerships* or assignment of other hashtags.

Chapter 4

Metrics

The problem of measuring ad impact can be addressed from different perspectives. In the literature, metrics such as user satisfaction surveys and click-through-rate have been used (Cramer [2015]). In this dissertation, we propose metrics, which can be computed from publicly available information.

Consider the set of all collected posts, $\mathcal{P}$, hashtags, $\mathcal{H}$, and influencers, $\mathcal{U}$. Each influencer $u \in \mathcal{U}$ is associated with a sequence of posts $posts(u) = (p_1, \ldots, p_k)$, $p_i \in \mathcal{P}$, $k \leq 500$. Each post $p \in \mathcal{P}$ is associated to a timestamp, number of likes, $\mathcal{L}(p) \geq 0$, comments, $\mathcal{C}(p) \geq 0$, a set of hashtags, $\mathcal{H}(p) \subset \mathcal{H}$ and a set of mentions $\mathcal{M}(p)$ (i.e., references to other users' profiles). If the set $\mathcal{H}(p)$ contains one or more hashtags from the set

$$\#ADs = \{\#ad, \#sponsored, \#publi, \#patrocinado\},$$

then we say that $type(p) = AD$; otherwise, $type(p) = NAD$, i.e., not a disclosed advertisement. In our dataset, we obtained a total of **264,518 AD posts** and **3,186,215 NAD posts**.

4.1 Time granularity

To analyze the context and impact of an ad placement strategy on influencer engagement with the content generated on an OSN platform, such as Instagram, we need to choose an appropriate time granularity, i.e., the length of time intervals into which to partition the posting timelines in our dataset. We group posts that belong to a particular time window, e.g., a day, a week, or a month, and study the relationship between the ad placement strategy and the average number of interactions in that time window.

Essentially, we want to set a time window length to optimally capture strategic behavior in terms of ad placement of each user. Since ad frequency is an essential part of an ad placement strategy, the time window of our analysis should be at least one week-long, as 86% of users post at most one AD post per week (see Figure 6.12). On the other hand, the time window cannot be too long, so that changes in posting behavior are not averaged out of the analysis.

On one extreme, the highest-granularity analysis would look at individual posts, i.e., how an AD post affects the number of interactions with the following post in an influencer's timeline. A coarser granularity analysis would look at higher time intervals, such as a quarter. Using too high or too coarse granularities, might not completely capture an influencer's posting strategy. Thus, the definition of an ideal granularity is important for the upcoming analyses.

We used the following methodology to choose the size of the time window for our analysis. For six different time windows $w \in \mathcal{W}$,

$$\mathcal{W} = \{1week, 2weeks, 1month, 2months, 3months, 4months\},$$

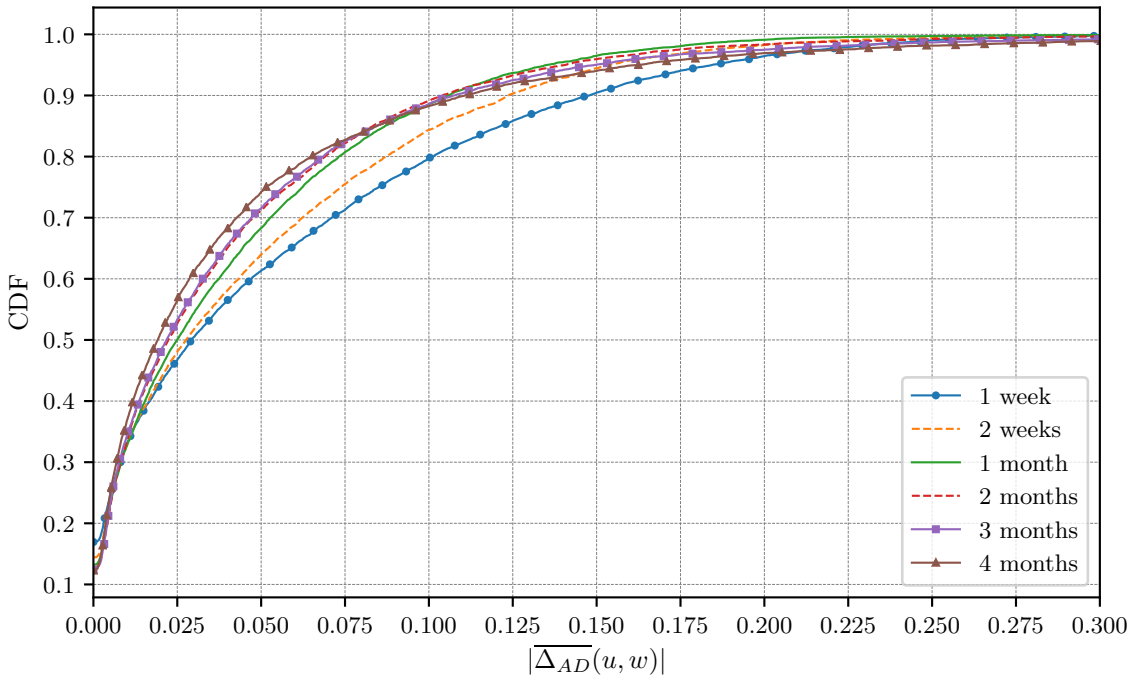
we analyzed the variation in the number of AD posts within a time window over time. More precisely, for each posting sequence $posts(u), u \in \mathcal{U}$ and each time window $w \in \mathcal{W}$, we considered a sequence of post subsets $\mathcal{T}(u, w) = (t_1, \ldots, t_f)$, with posts in $posts(u)$ grouped by the time window of their posting, in chronological order. Therefore, if two posts $\{p_x, p_y\} \in t_i \in \mathcal{T}(u, w)$, then p_x and p_y were posted in the same time window, e.g., January 2019 if $w = 1month$. Moreover, if $p_x \in t_i \in \mathcal{T}(u, w)$ and $p_y \in t_{i+1} \in \mathcal{T}(u, w)$, then p_x was posted in the month preceding that of p_y's posting.

Furthermore, for each time granularity $w \in \mathcal{W}$, we computed the relative variation in the number of AD posts in each pair of consecutive time windows $(t_i, t_{i+1}) \in \mathcal{T}(u, w)$, and the average AD variation of an influencer, $\forall u \in \mathcal{U}$, as follows:

$$\begin{aligned} \Delta_{AD}(t_i) &= \frac{|\{p|p \in t_i, type(p) = AD\}|}{|\{p|p \in t_{i-1}, type(p) = AD\}|}, 2 \leq i \leq |\mathcal{T}(u, w)| \\ \overline{\Delta_{AD}}(u, w) &= \frac{1}{|\mathcal{T}(u, w)| - 1} \sum_{i=2}^{|\mathcal{T}(u,w)|} \Delta_{AD}(t_i) \end{aligned} \tag{4.1}$$

In Figure 4.1 we can see the Cumulative Distribution Function (CDF) of the absolute value of $\overline{\Delta_{AD}}(u, w)$ over all influencers $u \in \mathcal{U}$ and time windows $w \in \mathcal{W}$. We observe that the time window that maximizes the average variation in the number of ads for most users is $w = 1week$ (the bottom curve). In particular, 40% of users

Figure 4.1: CDF of absolute value of $\overline{\Delta_{AD}}(u, w), u \in \mathcal{U}, w \in \mathcal{W}$



presented an average weekly ad variation of $\geq 5\%$, 20% of users presented a variation of $\geq 10\%$, and 10% of users had a $\geq 15\%$ average increase or decrease in the number of ad posts in a week. Since we are interested in analyzing the context and impact of ad placement strategies on audience engagement, we decided to perform our analysis based on the time granularity of one week.

Having fixed the time granularity of our analysis, we refer to $\mathcal{T}(u, 1week)$ as the *timeline* $\mathcal{T}(u)$ of influencer $u \in \mathcal{U}$ and partition each such timeline into three subsets of weeks:

- $AD^{+}(u) = \{w \in \mathcal{T}(u) | \Delta_{AD}(w) > \mu_u + \sigma_u\}$: subset of weeks with a significant relative increase in the number of ad posts;
- $AD^{-}(u) = \{w \in \mathcal{T}(u) | \Delta_{AD}(w) < \mu_u - \sigma_u\}$: subset of weeks with a significant relative decrease in the number of ad posts;
- $AD^{=}(u) = \{w \in \mathcal{T}(u) | \mu_u - \sigma_u \leq \Delta_{AD}(w) \geq \mu_u + \sigma_u\}$: subset of weeks with no significant changes in the ad posting strategy of user $u \in \mathcal{U}$;

where σ_u and μ_u are the standard deviation and mean values of the set $\{\Delta_{AD}(w) | w \in \mathcal{T}(u), u \in \mathcal{U}\}$, respectively.

4.2 Impact of ads on influence dynamics

To quantify the success of an individual post $p \in posts(u)$, $u \in \mathcal{U}$, we introduce the following definition of *post interaction number* $\mathcal{I}(p)$[1]:

$$\mathcal{I}(p) = \mathcal{L}(p) + \mathcal{C}(p), \forall p \in \mathcal{P}. \tag{4.2}$$

Similarly, in order to measure the (estimated) success of an influencer $u \in \mathcal{U}$, we define the *average*, $\overline{\mathcal{I}}(u)$, *influencer number of interactions* as follows:

$$\overline{\mathcal{I}}(u) = \frac{1}{|posts(u)|} \times \sum_{p \in posts(u)} \mathcal{I}(p), \forall u \in \mathcal{U}, \tag{4.3}$$

In order to measure the success of an influencer in terms of the number of comments and likes, we define $\overline{\mathcal{C}(u)}$ and $\overline{\mathcal{L}(u)}$, replacing $\mathcal{I}(p)$ by $\mathcal{C}(p)$ and $\mathcal{L}(p)$ in (4.3), respectively.

Similarly, we define $\overline{\mathcal{I}}(h)$, the average number of interactions with a hashtag $h \in \mathcal{H}$, by computing the average number of interactions with posts $p \in \mathcal{P} | h \in \mathcal{H}(p)$.

To measure the influence dynamics of an influencer over time, we define the *average weekly influencer interaction number* as follows:

$$\overline{\mathcal{I}}(u, w) = \frac{1}{|w|} \sum_{p \in w} \mathcal{I}(p), \forall w \in \mathcal{T}(u), u \in \mathcal{U}. \tag{4.4}$$

To account for the number of likes and comments separately, we define $\overline{\mathcal{L}}(u, w)$ and $\overline{\mathcal{C}}(u, w)$, analogously.

To measure user engagement variation in a particular week $w_i \in \mathcal{T}(u)$, in the context of the timeline of an influencer $u \in \mathcal{U}$, we define the *weekly interaction variation*, as follows:

$$\Delta^{\mathcal{I}}(u, w_i) = \frac{\overline{\mathcal{I}}(u, w_i) - \overline{\mathcal{I}}(u, w_{i-1})}{\max(\overline{\mathcal{I}}(u, w_{i-1}), \overline{\mathcal{I}}(u, w_i))}, 2 \leq i \leq |\mathcal{T}(u)| \tag{4.5}$$

where w_{i-1} is the previous week, relative to $w_i \in \mathcal{T}(u)$, ordered chronologically. To measure weekly variation in terms of the number of likes and comments separately, we

[1]Due to the lack of information about the exact number of followers of a user $u \in \mathcal{U}$ at the time of posting a particular post $p \in posts(u)$ or the number of post impressions, we do not use the widely used measure of *engagement*, defined as the number of likes and comments received by p, divided by the number of followers of u at the time of posting (https://theonlineadvertisingguide.com/ad-calculators/instagram-engagement-rate-calculator/).

use the notation $\Delta^{\mathcal{L}}(u, w_i)$ and $\Delta^{\mathcal{C}}(u, w_i)$, respectively. Based on eq. (4.5), we define whether a week has been successful for influencer $u \in \mathcal{U}$ or not, as follows:

- Successful week $w \in \mathcal{T}(u)$, if $\Delta^{\mathcal{I}}(u, w) > 0$;
- Unsuccessful week $w \in \mathcal{T}(u)$, if $\Delta^{\mathcal{I}}(u, w) < 0$.

Finally, to measure the progress of the influencer interaction number over time, we introduce the following definition:

$$\Delta^{\mathcal{I}}(u) = \frac{1}{|\mathcal{T}(u)|} \sum_{w \in \mathcal{T}(u)} \Delta^{\mathcal{I}}(u, w), \forall u \in \mathcal{U}. \tag{4.6}$$

Assuming that an Instagram post's success, or performance, can be inferred by looking at its interaction number, the proposed metrics capture the variation in the average interactions with posts between consecutive weeks in an influencer's timeline. With these metrics, we seek to understand if there is a correlation between shifts in ad placement strategy and the performance of an influencer's posting week.

Chapter 5

Topic detection and trend measurement metrics

Instagram allows users to post about a variety of different subjects and topics. Among influencers, some topics are more explored, such as fashion, beauty, and fitness. Different topics might attract different users with different consumption behaviors.

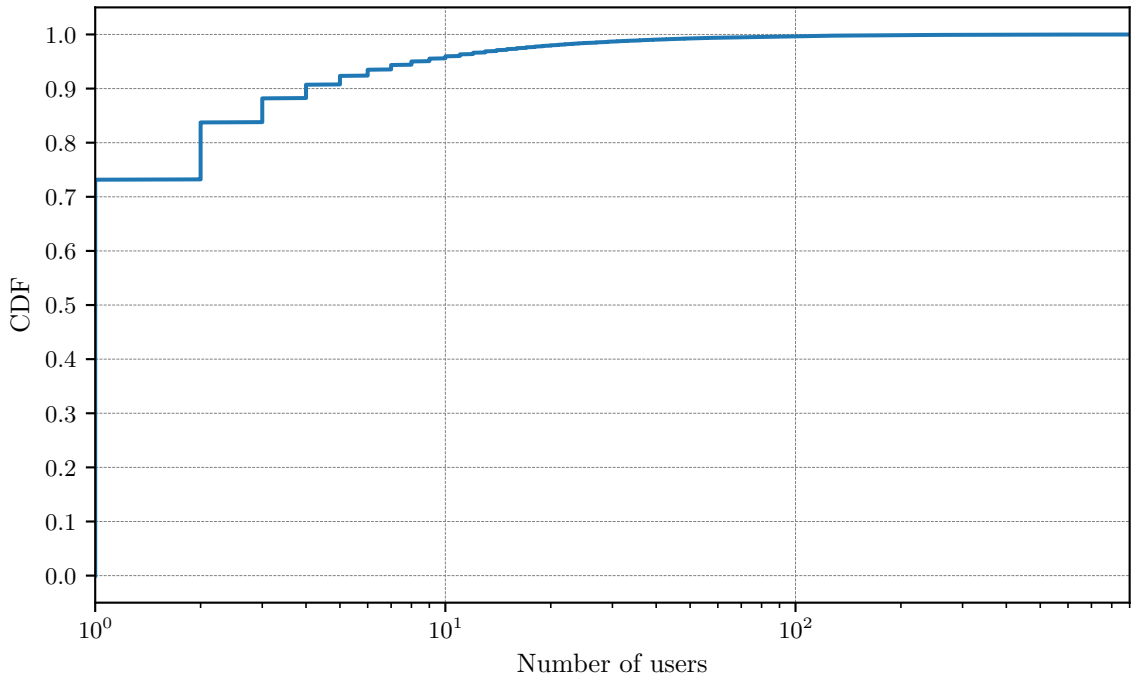
Topic modeling is a well-known problem in natural language processing. Perhaps the most widely used topic modeling technique is the Latent Dirichlet Allocation (LDA), proposed by Blei et al. [2003], which is used to find hidden topics in documents. A topic might be a subject, like "arts" or "education", that is discussed in the documents. The original setting in LDA, in which each word has a topic label, may not work well with short texts, such as Instagram posts and Twitter tweets. Thus, several other approaches have been explored.

In the literature, different efforts have been made in order to detect topics in short texts (Argyrou et al. [2018]; Giannoulakis and Tsapatsoulis [2019]; Zhao et al. [2011]; Chen et al. [2016]). For this task, we used a simpler hashtag-based clustering approach, which we describe in this section.

5.1 Topic detection

As stated previously in Section 3.2, our dataset is comprised of 3,450,733 different posts. On every post, there is the possibility to include one or more hashtags, usually intended to describe the contents of the post. Out of all posts, our data set contains 1,956,374 distinct hashtags. Since this is a free text field, there is a lot of noise in this data. In fact, Argyrou et al. [2018] have concluded that only 20% of the Instagram hashtags describe the actual content of the image they accompany. Figure 5.1 further explores

Figure 5.1: CDF of number of users per hashtag



this phenomenon. It is possible to see that around 70% of all hashtags contained in our dataset are used by only one user.

To remove too specific hashtags and those with irregular usage, we filtered out only those hashtags used by at least 100 users, throughout the entire time span of the dataset. This filtering returned a total of 5,119 distinct hashtags, around 0.2% of the original amount.

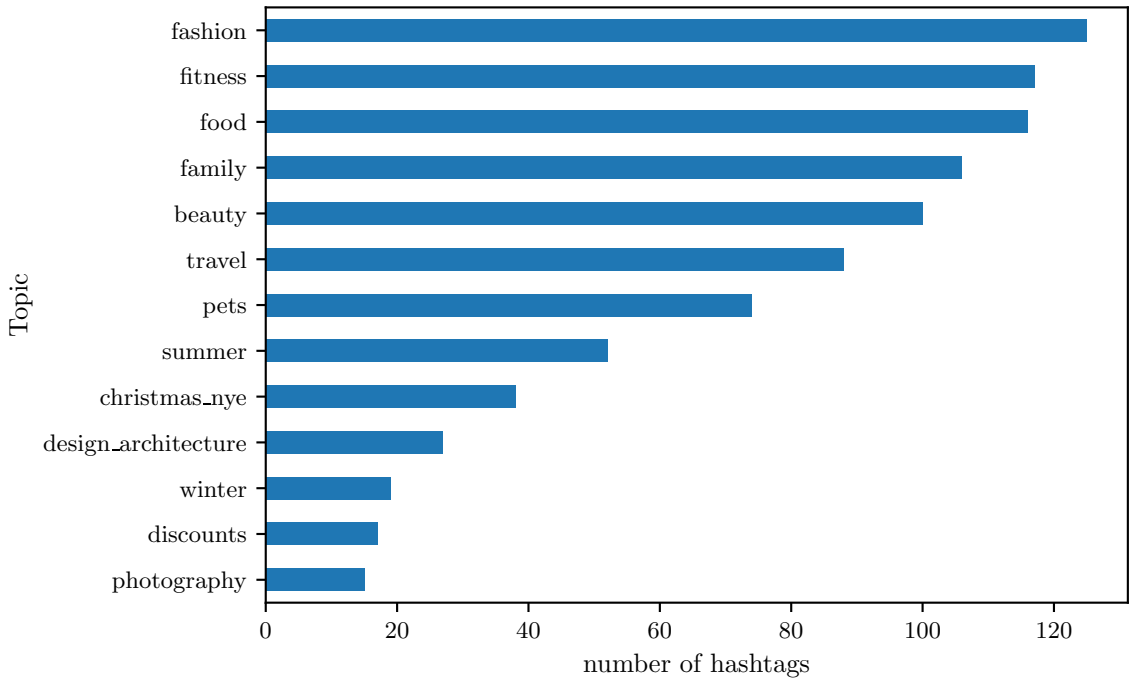
To identify similar hashtags, we used a clustering approach. For every hashtag $h \in \mathcal{H}$, we first calculated the co-occurrence between hashtag h_i and every other hashtag $\{h_j \in \mathcal{H} | i \neq j\}$. After this procedure, we ended up with n vectors of co-occurrence, where n is the number of distinct hashtags (5,119). We, then, calculated the cosine similarity between every pair of vectors, defined as:

$$cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}|| \, ||\mathbf{B}||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{5.1}$$

where A_i and B_i are the components of vectors A and B, respectively.

The final step in our topic detection methodology was to cluster similar hashtags together, in order to identify groups of hashtags defining of topics. For this task, we fed our similarity matrix to an agglomerative hierarchical clustering algorithm. According to Rokach and Maimon [2005], in this method "each object initially represents a cluster of its own. Then clusters are successively merged until the desired cluster structure is

Figure 5.2: Hashtags per topic



obtained". The merging of clusters is done according to a linkage criterion, that determines the distance between sets of observations. Some examples of linkage criterion are: single-linkage, complete-linkage, Ward's criterion, etc. We used Ward's criterion, which minimizes the increase in variance for the cluster being merged, as proposed by Ward and Joe [1963].

The result of a hierarchical clustering is represented as a tree (or dendrogram), in which each leaf is a hashtag. We, then, chose a similarity threshold of 0.65 to cut the tree at, returning the resulting clusters.

This procedure returned a total of 14 distinct and disjoint sets of hashtags. Out of the 14 clusters, 13 were identified as a topic and 1 was tagged as garbage. The 13 topics were, then, manually tagged with respect to the hashtags contained in the cluster. The set of 13 distinct topics is, thus, defined as $\mathcal{TOPICS}$, which is a set of sets of hashtags, grouped by similarity. Likewise, each $topic \in \mathcal{TOPICS}$ is associated with a set of hashtags $\mathcal{H}(topic)$. Figure 5.2 shows the size of each topic, by number of hashtags and its name. It is possible to see that the biggest groups are related to fashion, fitness, and food, while the smallest ones are related to photography, discounts, and winter.

Finally, we also identified the set of posts belonging to each topic, based on hashtag usage. That is, for every $topic \in \mathcal{TOPICS}$, we say that a post $p \in \mathcal{P}$ belongs to that topic if $|\mathcal{H}(p) \cap \mathcal{H}(topic)| > 0, \forall p \in \mathcal{P}, topic \in \mathcal{TOPICS}$. Note that there

might be more than one topic per post. Of all 3.450.733 distinct posts, 1.000.356 were tagged with at least one topic (28,9%).

Table 5.1 presents the number of posts and users after each processing step (removal of unpopular hashtags and topic definition).

Table 5.1: Overview of the data cleaning results

	Number of posts	Number of users
Initial dataset	3.450.733	7.583
Dataset containing only popular hashtags	1.926.445	7.569
Dataset containing only posts with at least one topic	1.000.356	7.263

5.2 Topic performance measurement metrics

After defining the topics contained in our dataset, we can, then, identify which topics are contained within each post and, similarly, we can find, for each topic, the set of users posting about it. Thus, we define $users(topic)$ and $posts(topic)$ as the set of users $u \in \mathcal{U}$ and posts $p \in \mathcal{P}$ that posted about this specific $topic \in \mathcal{TOPIC}$, respectively. We can also define $topics(u)$ as the set of topics a user $u \in \mathcal{U}$ has posted about.

With these definitions, we introduce the following metrics:

$$\overline{\mathcal{I}}(topic) = \frac{1}{|posts(topic)|} \times \sum_{p \in posts(topic)} \mathcal{L}(p), \forall topic \in \mathcal{TOPICS}, \tag{5.2}$$

as the average number of likes of a specific topic.

Following the time window analysis presented in Section 4.1, we can also define $users(topic, w_i)$ as the set of users posting about a specific topic on week $w_i \in \mathcal{T}(topic)$. Likewise, we can also define the same metric for individual hashtags, by replacing *topic* by $h \in \mathcal{H}$. Alternatively, we can also define $topics(u, w)$ as the set of topics posted by user $u \in \mathcal{U}$ in week $w \in \mathcal{T}(u)$.

We also wanted to identify trends in postings on each topic. A trend, which can be positive or negative, can be defined as a change in the number of users posting about a specific topic or hashtag, relative to a previous time window. Thus, we define the following metric:

$$\Delta^U(topic, w_i) = \frac{|users(topic, w_i)| - |users(topic, w_{i-1})|}{|users(topic, w_{i-1})|}, 2 \leq i \leq |\mathcal{T}(topic)| \tag{5.3}$$

as the topic weekly variation in number of users.

Based on Eq. (5.3), we can also define whether a topic is trending positively or negatively in a particular week, as follows:

- Positively trending $topic \in \mathcal{TOPICS}$ in $w \in \mathcal{T}(topic)$, if $\Delta^U(topic, w) > 0$;
- Negatively trending $topic \in \mathcal{TOPICS}$ in $w \in \mathcal{T}(topic)$, if $\Delta^U(topic, w) < 0$;

Finally, in order to measure the overall performance of a topic over time, we introduce the following definition:

$$\Delta^U(topic) = \frac{1}{|\mathcal{T}(topic)|} \sum_{w \in \mathcal{T}(topic)} \Delta^U(topic, w), \forall topic \in \mathcal{TOPIC}. \tag{5.4}$$

Equations (5.2), (5.3) and (5.4) can also be defined for individual hashtags, by replacing *topic* by $h \in \mathcal{H}$.

Chapter 6

Data characterization

In this section, we present a characterization, focused on sponsored content, of our dataset, which we partition into two subsets: *Posts* (Section 6.1) and *Influencers* (Section 6.2). We analyze our dataset in regards to different characteristics, such as engagement with sponsored content (Section 6.1.1), and ad placement strategy and engagement dynamics (Section 6.1.2).

We also analyze different aspects of *Influencers* on Instagram, such as follower base size, amount of posts, and frequency of posting, as well as the percentage of ad content and performance of monetization strategies, by user group.

Finally, we present a characterization of hashtag usage, topics, and trends on Instagram (Sec. 6.3.2). First, we investigate the behavior of different hashtags and their popularities, and the usage of hashtags as means of boosting audience engagement. We, then, characterize topic dynamics in time and its distribution among influencers.

6.1 Posts

The *Posts* dataset is comprised of 3,450,733 posts, posted by 7,583 distinct Instagram influencers. It was collected by crawling the public timelines of the selected influencer profiles, between December 2018 and March 2019, and downloading the 500 most recent posts of each profile. The collected set of posts spans a period of 5 years, distributed between 2015 to 2019, as illustrated in Figure 6.1. As we can see, the majority of posts were created in 2018 and 2019 (69.61%), whereas a smaller fraction of the content was posted in earlier years (20.42% in 2017, 7.23% in 2016, and 2.73% in 2015).

Figure 6.1: Posts per year

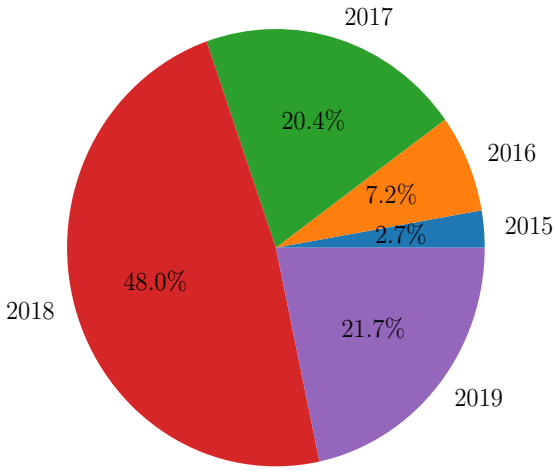
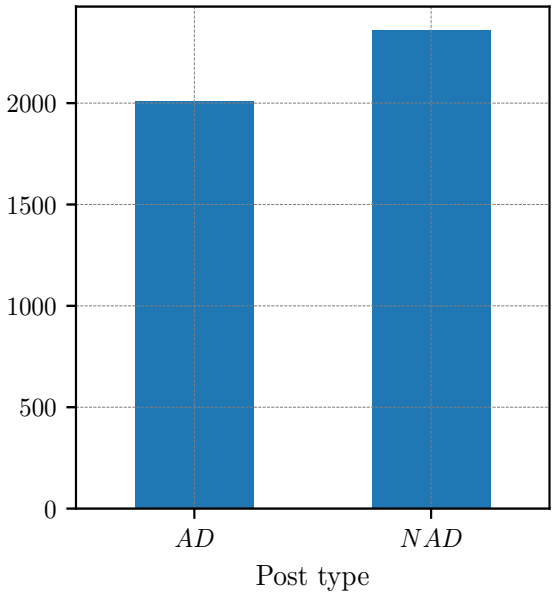


Figure 6.2: $\mathcal{I}(p)$ per post type



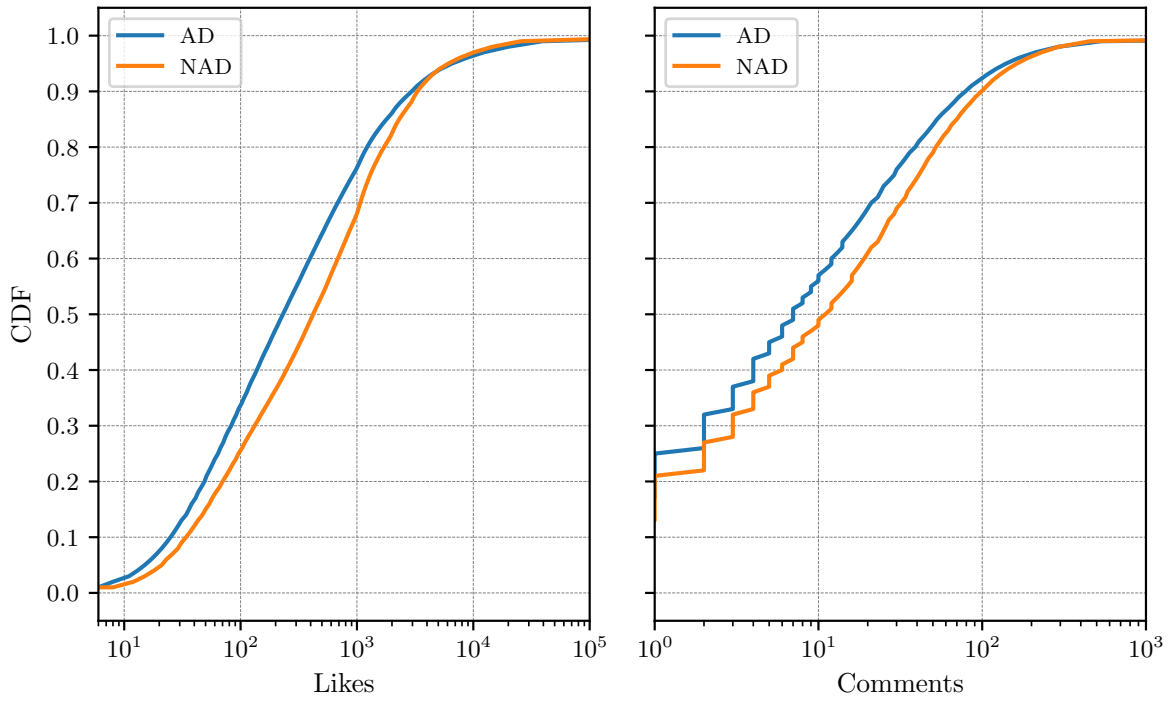
6.1.1 Engagement with sponsored content

For influencers, the usage of sponsored content on Instagram is one of the main monetization possibilities. Being so, it is important to understand how users interact and engage with advertisements, as opposed to regular content. In our dataset, 7.6% of posts were classified as sponsored content (AD).

In Figure 6.2 we can see that, on average, AD posts are less appealing to the audience, with an average number of interactions 14.7% lower than that of NAD (not sponsored) posts. This characteristic is confirmed in Figure 6.3, which shows the CDF of the number of likes and comments received by AD and NAD posts. We can see that 74.5% of NAD posts received ≥ 100 likes, while only 66.4% of AD posts received ≥ 100 likes. Likewise, 21.2% of NAD posts received ≥ 50 comments, while only 16.1% of AD posts received ≥ 50 comments.

Figure 6.4 illustrates the seasonal distribution of each type of post. More precisely, it shows what percentage of each post type (AD and NAD) was posted in each month of the year. We can see that, of all NAD posts, around 40% were posted in the first 4 months of the year (Jan - Apr), followed by a relative decrease between May and September, and a relative increase during the final months of the year (Oct - Dec). The distribution of AD content follows a similar pattern, having a higher percentage of posts during the first trimester, followed by a relative decrease in the summer months. Interestingly, we observe a steep rise in AD posts in the second semester, peaking in

Figure 6.3: CDF of likes and comments for AD and NAD posts



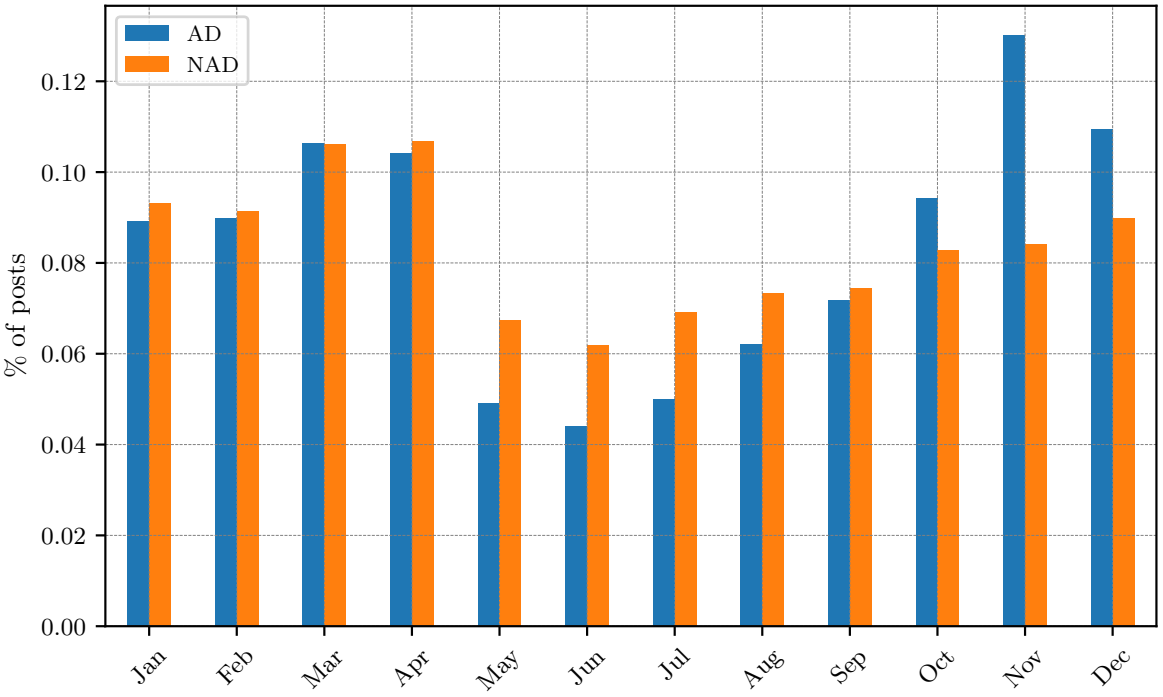
November and December. This behavior towards the end of the year might be related to popular seasonal events, such as Halloween, Black Friday, Christmas, and New Year's Eve, producing an elevated appeal for the purchase and consumption of products and services.

6.1.2 Ad placement strategy and engagement dynamics

In Section 4.2, we defined metrics to measure the performance of an ad placement strategy on a weekly basis. We categorized each week in an influencer's timeline based on the variation of the number of AD posts. Using these definitions, we seek to analyze the impact on audience engagement of different posting strategies.

Figure 6.5 presents the CDF of the weekly interaction variation $\Delta^{\mathcal{I}}(u, w)$ for $w \in AD^{+}(u)$, $w \in AD^{-}(u)$, and $w \in AD^{=}(u)$. This metric, as defined in (4.5), is the variation in the number of interactions between consecutive weeks and, thus, can be seen as a measurement of (weekly) performance. Of all 446,080 weeks in our dataset, 226,443 weeks were tagged as **successful** ($\Delta^{\mathcal{I}}(u, w_i) > 0$), while 219,022 weeks were tagged as **unsuccessful** ($\Delta^{\mathcal{I}}(u, w_i) < 0$), and 615 weeks were tagged as neutral, with no change in the number of interactions, relative to a previous week ($\Delta^{\mathcal{I}}(u, w_i) = 0$). It is possible to see that weeks with an increase in the number of AD posts tend to have lower $\Delta^{\mathcal{I}}(u, w_i)$. In fact, around 52.9% of AD^{+} weeks had a negative $\Delta^{\mathcal{I}}(u, w_i)$,

Figure 6.4: Distribution of posts per month of the year



i.e., were unsuccessful, while 49.1% of $AD^{=}$ and 45.3% of AD^{-} weeks had negative $\Delta^{\mathcal{I}}(u, w_i)$, respectively.

This analysis shows that, indeed, an increase in the weekly number of advertisements is associated with a decrease in audience engagement. However, more than 47% of AD^{+} weeks were successful, despite containing significantly more ad content than the previous week. We hypothesize that the usage of other resources, such as hashtags, mentions, and seasonal postings, might mitigate this negative effect.

Seasonal information is another feature that can be used when planning an ad placement strategy. In Figure 6.6, we show how successful and unsuccessful AD^{+} weeks (as defined in Section 4.2) are distributed over different months of the year. We can see that the ratio of successful weeks with an increased ad strategy is significantly higher in the final two months of the year. Interestingly, although the percentage of AD posts between May and September is the lowest (Figure 6.4), the percentage of successful AD^{+} weeks during this time of the year is higher than the percentage of unsuccessful weeks.

Figure 6.5: CDF of $\Delta^{\mathcal{I}}(u, w)$

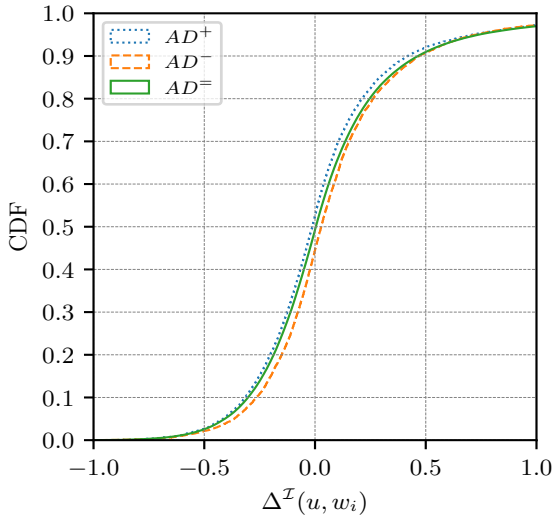
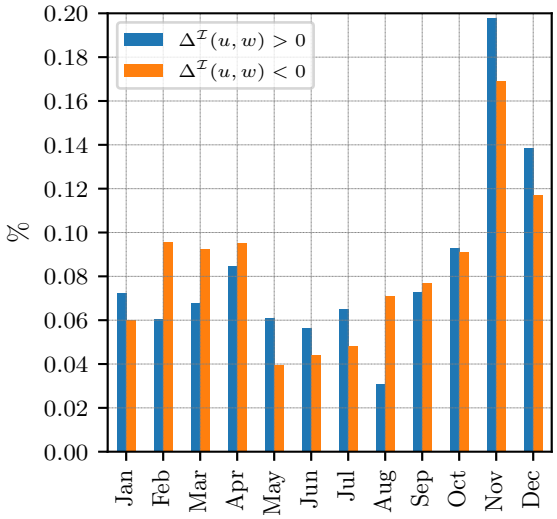


Figure 6.6: $AD^+(u)$ per month



6.2 Influencers

In this section, we characterize our data from the *Influencers'* dataset perspective. As we have previously described in Chapter 6, the *Influencers* dataset contains information regarding **7583** distinct Instagram influencers and their most recent 500 posts. In order to understand the influencer scenario of our dataset, we perform an analytical characterization of our data. First, we describe the *Influencer* dataset concerning the number of posts and followers. We, then, investigate the distribution of the average number of likes and comments. Afterward, we discuss our results regarding the usage of ads. Finally, we present an analysis of the impact of monetization strategies on different user groups.

6.2.1 Overall analysis of influencer's numbers

Figures 6.7 and 6.8 show the CDF of the number of followers and the total number of posts per influencer[1], respectively. As we can see, around 43.8% of influencers have $\leq$ 10,000 followers (**beginners** group), around 43% have between 10,000 and 100,000 followers (**micro-influencers** group), around 13% have $\geq$ 100,000 followers (**celebrities** group). We can also observe that approximately 54% of influencers have published $\leq 1,000$ posts; 41% published between 1,000 and 5,000 posts, and just 5% own more than 5,000 posts in their timelines.

[1]Recall that, even though there are influencers with over 12,000 posts in their posting timeline (Figure 6.8), our dataset contains only the $k \leq 500$ most recent posts of each user.

Figure 6.7: CDF of the number of followers per influencer

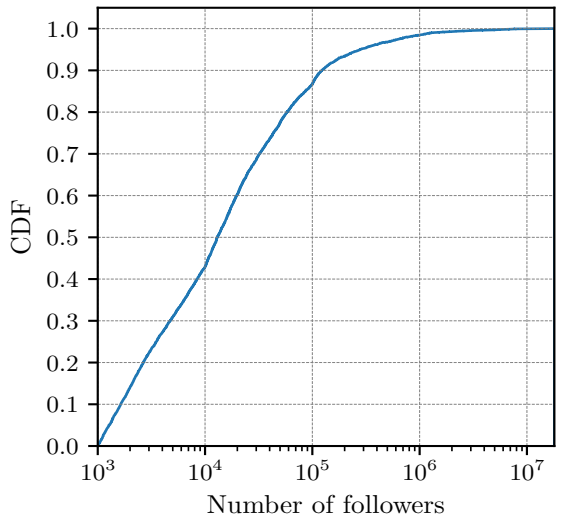
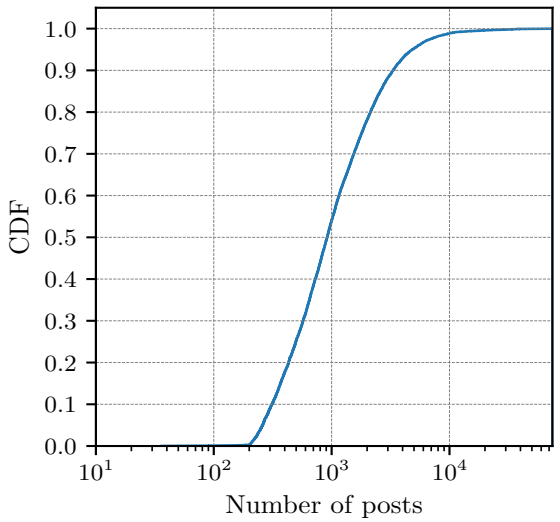


Figure 6.8: CDF of the number of posts per influencer



Figures 6.9 and 6.10 shows the CDF of the average number of likes and comments received by all posts of an influencer, respectively. We can see that, on average, the number of likes received by posts of an influencer is significantly higher than the number of comments. As an example, while over 71% of influencers have, on average, ≥ 100 likes, only 8.6% have on average ≥ 100 comments. As a matter of fact, this is somehow expected, since it is easier to cast likes, which can be done with a single click of a button, than it is to post comments.

Figure 6.11 gives us a better overview of influencer activity. As we can see, the posting frequency differs greatly within the influencers in our dataset. Around 85% of users post at most 10 posts per week, which gives us a daily activity of around 1.5 posts per day. Around 4.4% of influencers create more than 20 posts per week. Only 0.6% of all influencers within our dataset post with a frequency higher than 70 posts per week, which would give an average of 10 posts per day. Those profiles with a very high frequency of posting are, in general, profiles related to news outlets and online stores.

6.2.2 Advertisement usage and distribution

Figure 6.12 shows the ratio of sponsored content in an influencer's (collected) timeline. Around 90% of influencers have at most 20% of advertisement content in their timelines, and less than 4% contain more AD than NAD content. Only 0.3% of influencers have more than 90% of ad content in their timelines. Those profiles are usually related

Figure 6.9: CDF of the average number of likes per influencer $\overline{\mathcal{L}(u)}$

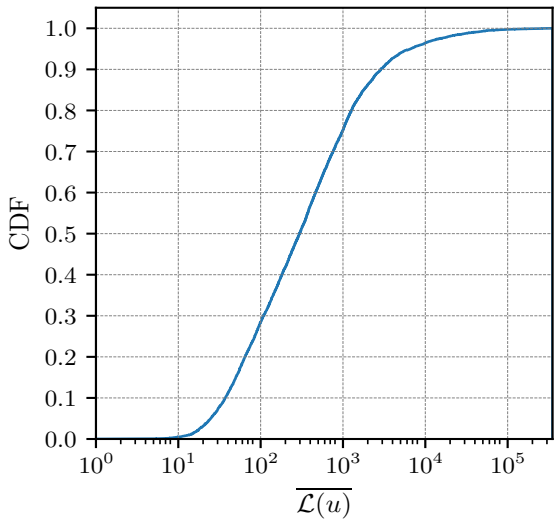


Figure 6.10: CDF of the average number of comments per influencer $\overline{\mathcal{C}(u)}$

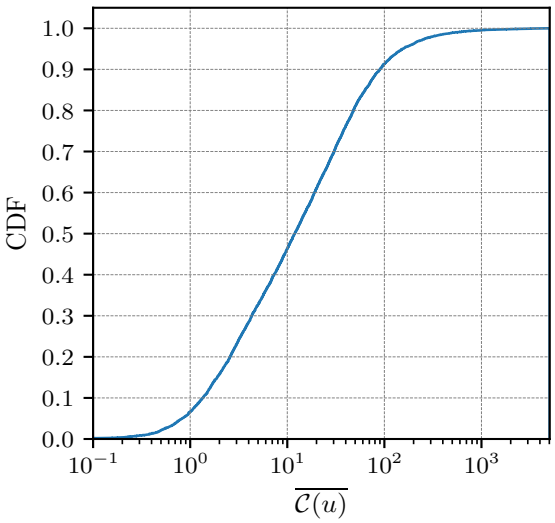


Figure 6.11: CDF of the weekly frequency of posting per user profile

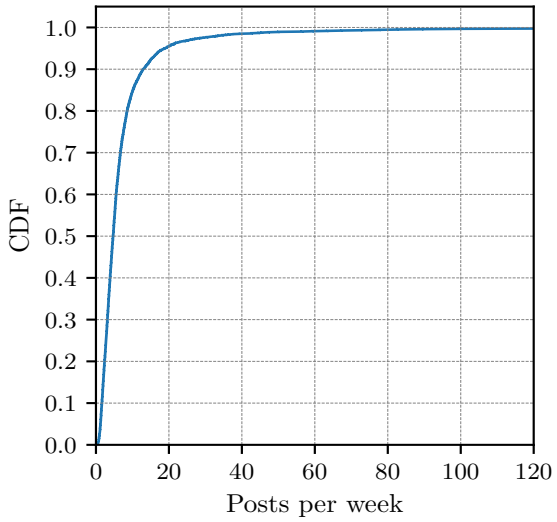
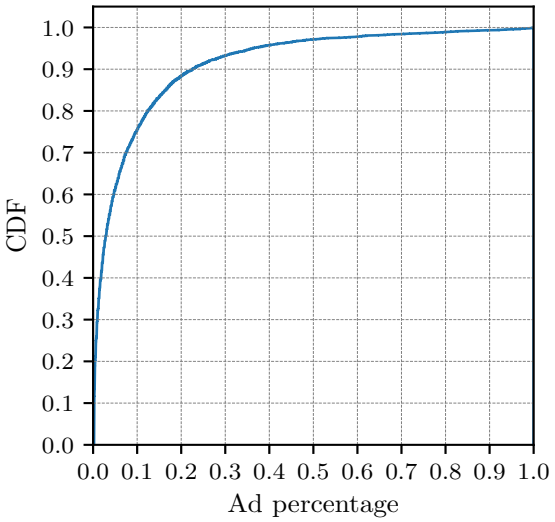
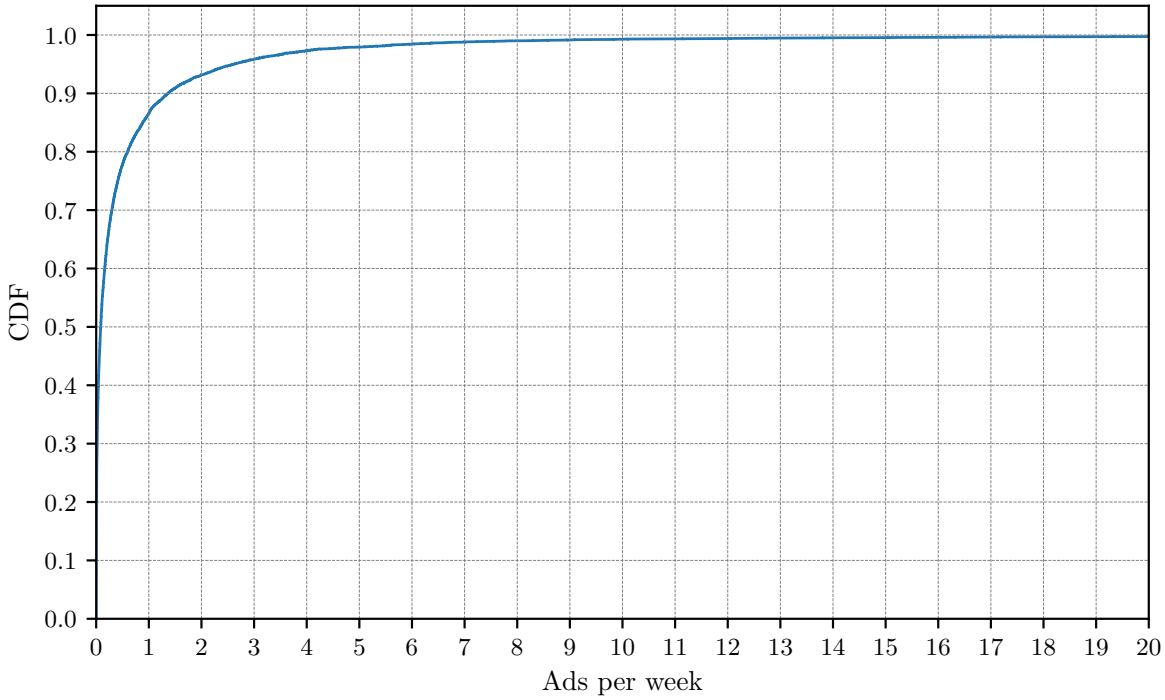


Figure 6.12: CDF of the percentage of ad content



to online stores and news outlets. Alternatively, in Figure 6.13, we can see that the frequency of ad posting is not too great. As an example, of all influencers, only 14% post more than one AD per week.

Figure 6.13: Number of ads per week



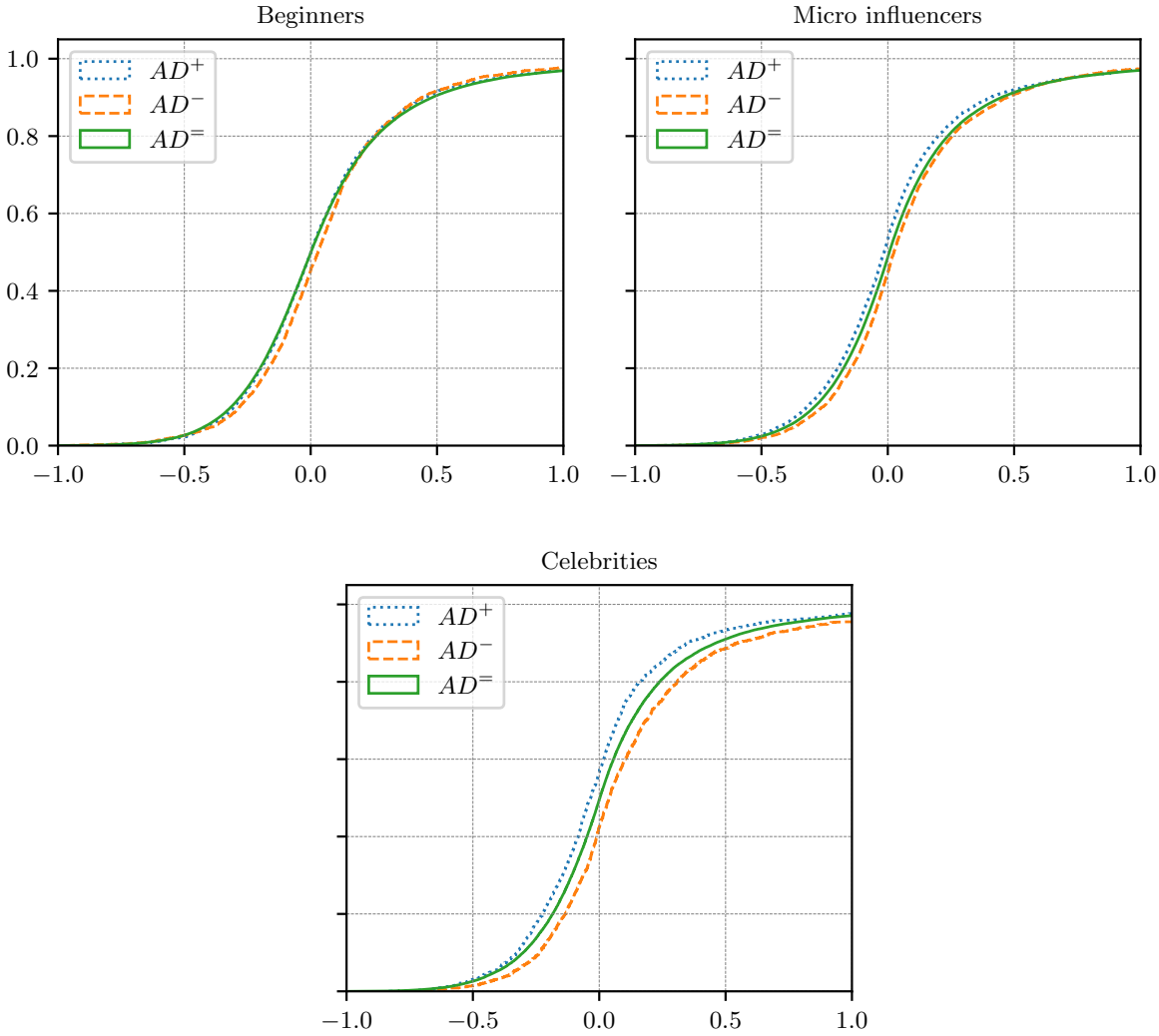
6.2.3 Monetization strategy performance

Figure 6.14 shows the CDF of the weekly interaction variation $\Delta^{\mathcal{I}}(u, w)$, $w \in AD^{+}(u)$, $w \in AD^{-}(u)$, and $w \in AD^{=}(u)$ (as defined in Section 4.1), for each of the influencer groups defined in Section 3.2.

The behavior is similar to the one shown in Figure 6.5, in which AD^{+} weeks have, on average, lower $\Delta^{\mathcal{I}}(u, w)$ than AD^{-} and $AD^{=}$ weeks. This shows us that, indeed, an increasingly aggressive ad placement strategy, usually tend to have more negative effects that lowering or maintaining the amount of sponsored content in a week's monetization strategy.

However, the discrepancy in impact between successful and unsuccessful weeks varies across distinct influencer groups. We can see that for beginners, around 51% of AD^{+} weeks have a negative $\Delta^{\mathcal{I}}(u, w)$, while for micro-influencers, around 53% of AD^{+} weeks have negative $\Delta^{\mathcal{I}}(u, w)$. For celebrities, the unsuccessful AD^{+} weeks comprise around 56%. This suggests that, even though weeks with a more aggressive ad placement strategy tend to have a decrease in audience engagement in general, celebrities are the ones that are affected the most, while beginners and micro-influencers tend to have a milder negative impact.

Figure 6.14: CDF of $\Delta^{\mathcal{I}}(u, w)$, $w \in AD^{+}(u)$, $w \in AD^{-}(u)$, $w \in AD^{=}(u)$



6.3 Hashtag usage, topics, and trends

In this Section, we present a preliminary analysis of hashtag and topic usage, as defined in Section 5.1. We also analyze topic and hashtag trends over time.

6.3.1 Hashtag usage

Hashtags are extensively used on Instagram. They are keywords used to describe the content of a post. Users insert one or multiple tags in the body of the description of a picture, preceded by the # symbol. Instagram also gives its users the possibility of searching for hashtags, which yields the most recent and most popular posts with that specific hashtag.

As stated before, our dataset contains over 1.9 million distinct hashtags. Out of those, some are more widespread than others. Table 6.1 presents the most popular hashtags, by the weekly average number of users (Table 6.1a), $\Delta^U(h, w)$ (Table 6.1b), and interactions (Table 6.1c).

In Table 6.1a, we can see that the most popular hashtags by the weekly average number of users are closely related to fashion, fitness, and beauty. Indeed, those topics are the ones with the highest amount of posts. There are also some quite regular and generic hashtags, such as *#love*, *#blogger*, *#instagood*, and *#repost*. Those hashtags are extensively used together with others, in order to boost a post's reach, since they are always very popular.

Table 6.1b shows a quite different list. This table presents the hashtags with the highest average weekly variation in number of users. Interestingly, we can see that all hashtags are closely related to highly seasonal events, spanning just one day. For example, *#8m*, *#iwd2019*, *#internationalwomensday*, and *#womensday* are hashtags related to the International Women's Day, which happens every year on the 8th of March. We can also see hashtags related to Father's Day, Mother's Day, Easter, and also some less known events, such as Pizza Day, Pet Day, and Doughnut Day. What all those events have in common, is a steep increase followed by an equally steep decrease in the number of users posting about this event.

Finally, Table 6.1c presents the most popular hashtags by average weekly interactions. While this list is interesting, showing the most liked and commented hashtags, it might be biased towards hashtags used by only a few influencers with a high follower base, like celebrities, for example. Still, we can see, for example, the *#10yearchallenge* hashtag appearing in both Table 6.1b and 6.1b, meaning that this hashtag is not only positively trending in regards to user variation, but also achieving great performance in likes and comments.

By analyzing hashtag usage among AD and NAD posts, Figure 6.15 presents the CDF of the number of hashtags per post, grouped by post type. It is possible to see that AD posts use a lot more hashtags than NAD posts. For example, 30% of NAD posts use ≥ 10 hashtags, whereas, for 30% AD posts, the number of hashtags is ≥ 25. Since it is possible to search for posts with specific hashtags on Instagram, the usage of this feature can be seen as a way to boost the reach of a post.

In Figure 6.16, we analyze the impact of the hashtag assignment strategy on audience engagement with a post. We can see that, as the number of hashtags assigned to a post increases, the interactions tend to decrease, suggesting that overusing hashtags can damage a post's performance. This suggests that, whereas AD posts tend to have more hashtags than NAD posts, possibly with the intent to reach a larger audi-

Table 6.1: 15 most popular hashtags

hashtag	$\overline{\lvert\mathcal{U}(h,w)\rvert}$
#ad	315.44
#love	234.42
#sponsored	218.31
#fashion	178.41
#ootd	174.44
#instagood	144.01
#tbt	142.40
#style	120.18
#fitness	116.28
#blogger	108.29
#beauty	108.28
#photooftheday	107.55
#photography	104.01
#repost	101.76
#liketkit	101.35

(a) Average influencers per week

hashtag	$\Delta^U(h,w)$
#8m	82.00
#nationalpizzaday	50.00
#happyfathersday	41.70
#nationaldogday	41.16
#iwd2019	39.44
#internationalwomensday	32.51
#felizdiadospais	28.91
#nationalicecreamday	24.79
#diadoamigo	24.11
#felizdiadasmaes	20.08
#nationalpetday	19.50
#womensday	15.74
#happyeaster	15.07
#nationaldonutday	15.02
#10yearchallenge	14.51

(b) Average user variation (Δ^U) per week

hashtag	$\overline{\lvert\mathcal{I}(h,w)\rvert}$
#yay	17958.25
#inmyfeelingschallenge	17421.19
#winterishere	10951.23
#mycalvins	10378.12
#friendshipgoals	10259.17
#diadasmulheres	9937.31
#always	9241.81
#yes	9022.67
#amormaior	8580.52
#23	8489.62
#aerolook	7743.60
#boohoo	6866.43
#bbb	6755.13
#10yearchallenge	6643.04
#slowmotion	6579.92

(c) Average interactions per week

ence through search engines, such a hashtag assignment strategy should be balanced carefully, since it might negatively impact the audience engagement with a post.

Figure 6.15: Hashtags per post

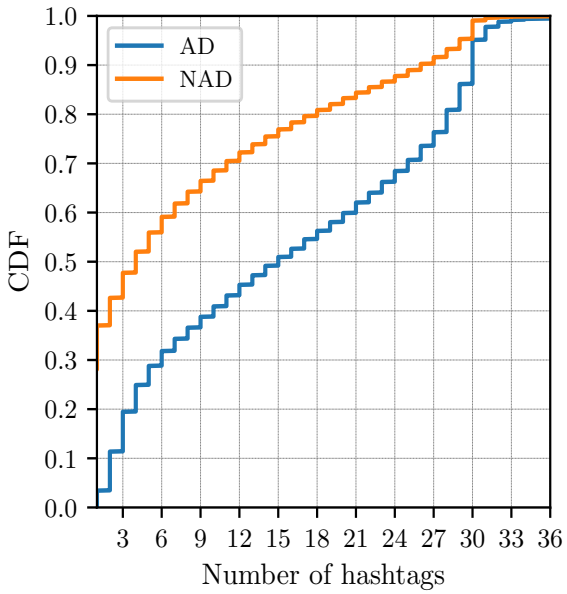
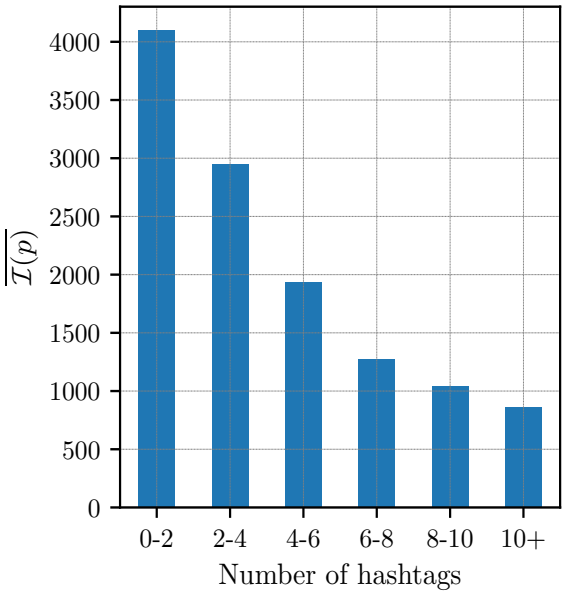


Figure 6.16: $\mathcal{I}(p)$ per $|\mathcal{H}(p)|$



6.3.2 Topics and trends

In Section 5.1, we described our efforts in detecting topics for each post in our dataset, based on hashtag similarity. We detected 13 distinct topics, with themes such as Fitness, Fashion, Christmas, and Beauty. Out of the 13 distinct topics, some are more popular than others, while some can be seen as highly seasonal. Figure 6.17 shows the number of users posting about each topic, in time. Some interesting patterns can be seen. First, we can see that the most popular topic is Fashion. It has a steady positive trend throughout the years and it tops every other topic almost every time of the year. After Fashion, the most popular topics are Fitness and Beauty.

It is interesting to see that some topics have steep spikes in usage during the year. Those topics can be seen as highly seasonal, with a hugely positive trend followed closely by an equally great negative trend. Among those topics, we. can see, for example, the Christmas and NYE topic, which, during the final days of the year, has a great increase in the number of users, topping even the Fashion topic.

Some other interesting seasonal topics include Discounts and Family. The Discounts topic includes many posts related to Black Friday and Cyber Monday, which are seasonal events happening in late November, traditionally known for great discounts on products and services. Likewise, the Family topic, which includes posts about childhood, babies, families, etc., also spikes in May, which is usually when Mother's Day happens.

One exciting result is the behavior of the Winter and Summer topics. Throughout the year, summer and winter happen alternatively, so it is possible to see that the positive trends in the summer topic usually happen when there is a negative trend in the winter topic, and vice-versa. It is also interesting to notice that the summer topic is more popular than its winter counterpart.

Figure 6.17: Number of users per topic in time

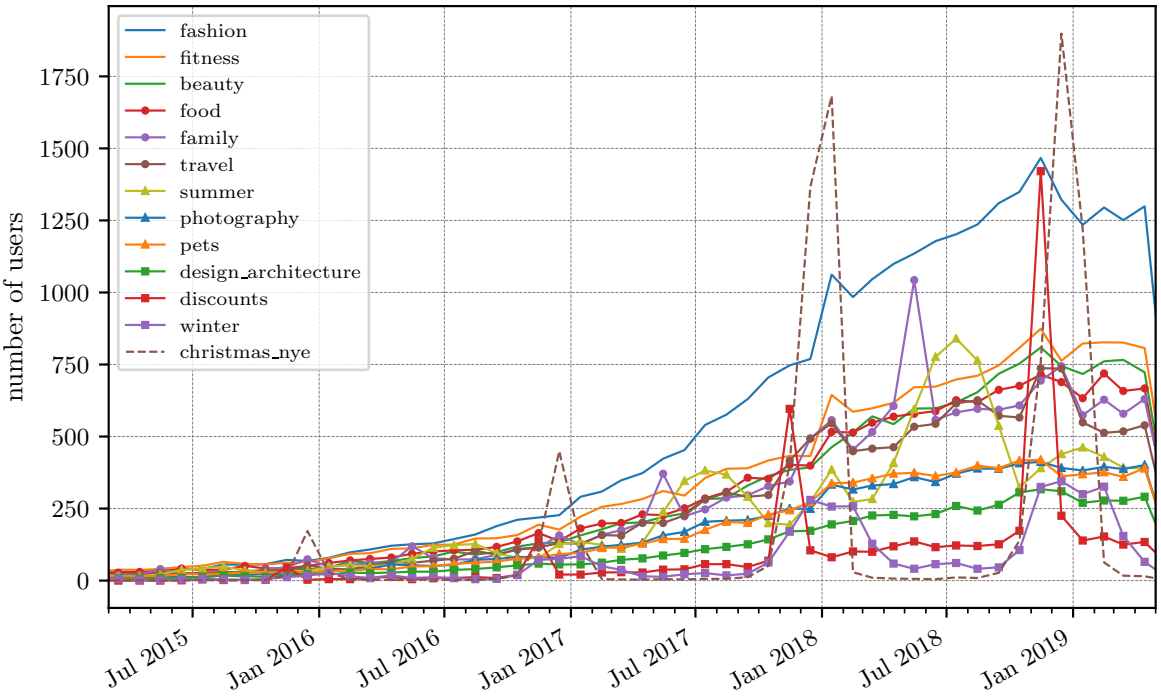


Figure 6.17 has given some insights about the posting behavior of each topic throughout time. It has shown that Fashion is the most popular topic, while also showing some interesting patterns in other topics. Indeed, as Figure 6.18 presents, the most posted-about topic is Fashion, with almost 350.000 distinct posts. It is followed by the Fitness and Beauty topic, with around 200.000 and 150.000 posts, respectively. The least posted about topics are, in order, Christmas and NYE, Winter and Discounts. Those topics are highly seasonal, having shorter timespans, which might explain the smaller amount of posts.

On the other hand, by analyzing the number of distinct users by topic, the discrepancy is less evident. Figure 6.19 shows that, even though Fashion is still the topic with most users posting about, the discrepancy from the second place is not as big as in the number of posts. We can also see the Summer topic rising to second place, meaning that even though it is only the 8th topic in number of posts, those posts are distributed between more users.

One interesting result is the positions of the Christmas and NYE topic by the number of posts and users. When analyzing the number of posts, it is possible to see that there aren't many posts related to Christmas and NYE, but almost 4500 users (60% of all users) have posted about this topic, meaning that even though this topic is not greatly explored or posted about, many users at least mention it somehow.

Figure 6.18: Number of posts per topic

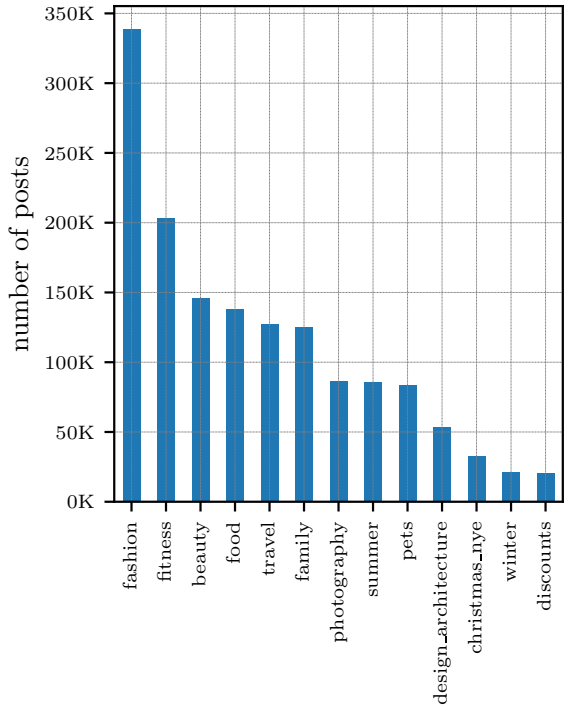
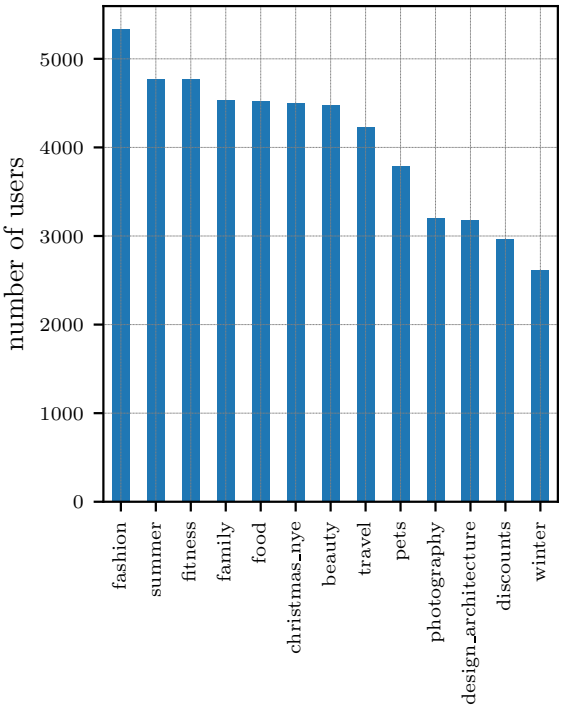


Figure 6.19: Number of users per topic



The dynamics of ad posting in each topic are explored in Figures 6.20 and 6.21, which shows the percentage of ad posts in each topic and the CDF of number of topics per post, respectively. It is possible to see that the topic with the highest percentage of ad posts is Discounts, with around 28% of ad posts. This is somehow expected since this topic includes posts with discounts, giveaways, and contests, also including content posted during Black Friday and Cyber Monday, which are seasonal events related to the consumption and purchase of products and services. The Discounts topic is followed by Fitness, with around 20% of ad posts, and Beauty, Design and Architecture, and Fashion, all with around 14% of ad posts.

Christmas and NYE, which is the topic with the highest variations in number of users, does not seem to have much advertisement in our database, with just a little above 10% of ad content. In fact, most of the Christmas-themed posts on our database do not seem to explore advertising, being mostly focused on the festivities and best

wishes.

Figure 6.20: Percentage of ads per topic

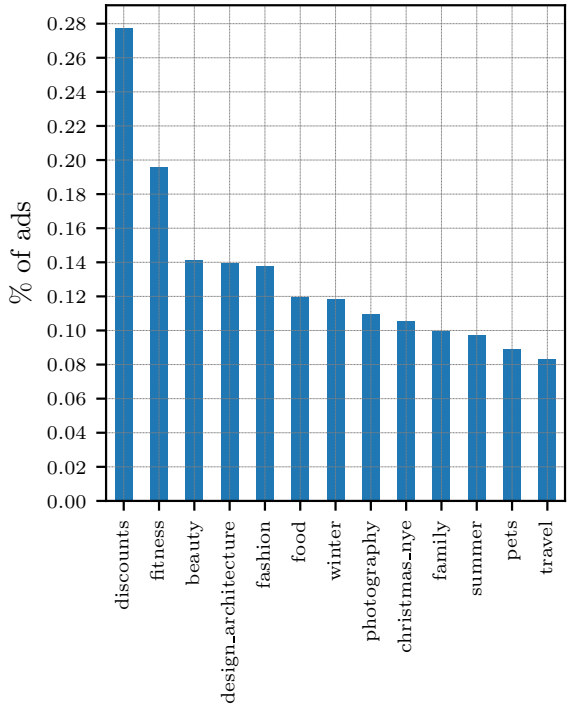


Figure 6.21: CDF of number of topics

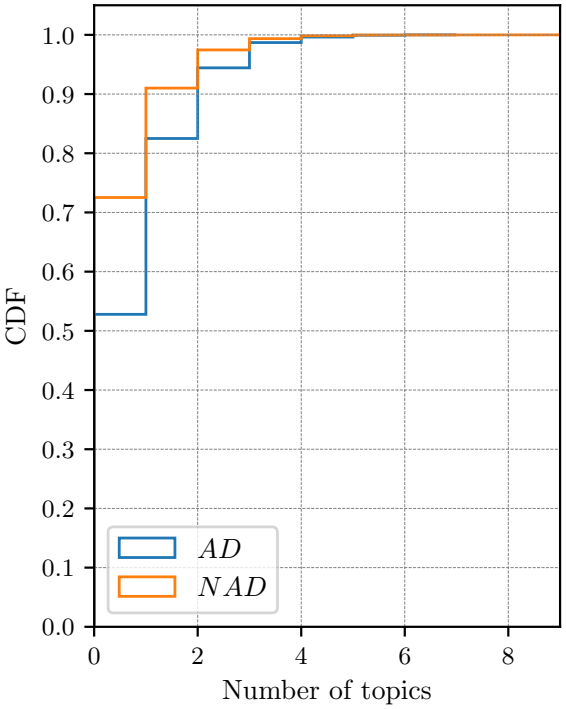


Figure 6.21 also shows us some interesting insights. NAD posts usually tend to be more specific, having fewer topics in their description than AD posts. As an example, we can see that, while AD posts with 2 or more topics are around 17.5% of all AD posts, only 9% of all NAD posts contain 2 or more topics. Since we also have shown in Figure 6.15 that AD posts usually have more hashtags than NAD posts, it is possible to imply that AD posts would potentially also have more topics.

Chapter 7

Analysis of impact and success

The data characterization presented in Section 6 revealed that sponsored content (AD) and regular content (NAD) are different in terms of audience engagement (Figures 6.2, 6.3), distribution in time (Figure 6.4), hashtag assignment strategy (Figures 6.15, 6.16), and distribution among topics (Figures 6.20, 6.21). Moreover, we could see that weekly ad posting strategies have different impacts on audience engagement in different influencer groups (Figures 6.5, 6.6).

Although anyone can exploit ad placement for monetization on Instagram, each influencer has a distinct audience and practices. It is important to understand how the usage of (disclosed) sponsored content affects each of these different groups.

In this section, we aim to analyze:

1. How sponsored content affects weekly interaction in an influencer's timeline;
2. Which posting strategies correlate with the success or failure of a week;
3. What are the main differences between influencer groups, i.e.: *beginners*, *micro-influencers*, and *celebrities*.

We hypothesize that the impact of ad posting strategies differs between influencer groups, and some posting strategies might be useful in mitigating the negative effect of advertisement content on audience engagement, which we observed in Figures 6.2 and 6.3. We also hypothesize that, while several strategies can be fine-tuned to act in favor of an influencer's monetization strategy, there are aspects beyond their control that make up a significant portion of their profile's success.

It is also important to note that, while we aim to understand if there is a correlation between usage of ads and user engagement, there might be many other factors into play that define an influencer's performance and success. We are analyzing only

one factor, while also trying to measure how this factor (usage of sponsored content and ad placement strategy) correlates with our proposed performance metrics.

For the purposes of this analysis, we used the grouping of users defined in Section 3.2.

7.1 Feature correlation analysis

To analyze how different features from our dataset are correlated, we compute the Spearman correlation coefficient defined as

$$r_s = \rho_{rg_X, rg_Y} = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X}\sigma_{rg_Y}},$$

where ρ_{rg_X, rg_Y} is the Pearson correlation coefficient between the rank variables, $cov(rg_X, rg_Y)$ is the covariance of the rank variables, and σ_{rg_X} and σ_{rg_Y} are the standard deviations of the rank variables. It has values in $[-1, 1]$, 1 indicates a total positive correlation, 0 means no correlation, and -1 indicates a total negative correlation. We also calculated the *p-values* for every correlation. All *p-values* were below a significance level of 0.05, which indicates that the correlation coefficients are statistically significant.

It is also important to note that correlation does not imply causation. The correlations only indicate trends between pairs of variables related to sponsored content on Instagram.

We considered the following **weekly features** for each influencer $u \in \mathcal{U}$ and week $w \in \mathcal{T}(u)$:

1. ***Total post count***: $|posts(u, w)|$;
2. ***Average hashtag count per post***: weekly average hashtags per post, defined as $\frac{\sum_{|posts(u,w)|}^{i=1} |\mathcal{H}(p_i)|}{|posts(u,w)|}$;
3. ***Average mention count per post***: weekly average mentions per post, defined as $\frac{\sum_{|posts(u,w)|}^{i=1} |\mathcal{M}(p_i)|}{|posts(u,w)|}$;
4. ***Ad ratio***: weekly percentage of AD content, defined as $\frac{|\{p|p \in posts(u,w), p \in AD\}|}{|posts(u,w)|}$;
5. ***Ad count***: weekly absolute AD count, defined as $|\{p|p \in posts(u, w), p \in AD\}|$;
6. $\Delta_{AD}(u, w)$: weekly variation in AD content, as defined in eq. (4.1);
7. ***Post length***: weekly average post length;

Table 7.1: Ad placement strategy correlation analysis and feature relevance

	$\overline{\mathcal{I}(w)}$	$\overline{\mathcal{C}(w)}$
$\lvert posts(u,w)\rvert$	-0.21	-0.20
Average hashtags per post	-0.16	-0.14
Average mentions per post	**0.20**	**0.25**
Ad ratio	-0.11	-0.12
Ad count	-0.17	-0.18
$\Delta_{AD}(u,w)$	-0.02	-0.02
Post length	0.12	**0.20**
Hashtag diversity	0.16	0.22
Trend alignment	0.11	0.13
Topic diversity	0.09	0.10
Intra-topic hashtag trend alignment Δ	**0.23**	**0.27**

(a) Beginners

	$\overline{\mathcal{I}(w)}$	$\overline{\mathcal{C}(w)}$
$\lvert posts(u,w)\rvert$	-0.06	-0.14
Average hashtags per post	-0.13	-0.11
Average mentions per post	0.04	**0.24**
Ad ratio	-0.02	-0.04
Ad count	-0.03	-0.08
$\Delta_{AD}(u,w)$	-0.02	-0.02
Post length	0.06	**0.21**
Hashtag diversity	0.03	0.13
Trend alignment	0.05	0.11
Topic diversity	0.04	0.09
Intra-topic hashtag trend alignment Δ	**0.15**	**0.19**

(b) Micro-influencers

	$\overline{\mathcal{I}(w)}$	$\overline{\mathcal{C}(w)}$
$\lvert posts(u,w)\rvert$	-0.10	-0.15
Average hashtags per post	**-0.43**	**-0.35**
Average mentions per post	-0.05	0.12
Ad ratio	-0.11	-0.07
Ad count	**-0.14**	**-0.13**
$\Delta_{AD}(u,w)$	-0.02	-0.02
Post length	**-0.35**	-0.11
Hashtag diversity	-0.14	-0.06
Trend alignment	0.01	0.05
Topic diversity	-0.03	0.00
Intra-topic hashtag trend alignment Δ	-0.02	0.00

(c) Celebrities

8. ***Hashtag diversity***: weekly number of distinct hashtags used;

9. ***Topic diversity***: weekly number of distinct topics posted about;

10. ***Global trend alignment***: weekly average topic user variation. For every $topic \in \mathcal{TOPICS}$ in our dataset, we calculated the weekly variation in the number of users $\Delta^U(topic, w)$, as defined in Eq. (5.3). We, then, selected the topics posted about by user u in week w, $topic \in topics(u,w)$, and computed the average value of $\Delta^U(topic), topic \in topics(u,w)$. The objective of this feature is to quantify the influencer's alignment with global topic trends: the more aligned an influencer is with global topic trends in a given week, the higher the $\Delta^U(topic)$ will be for each $topic \in topics(u,w)$, resulting in a higher value for this feature;

11. ***Intra-topic hashtag trend alignment*** Δ: weekly average user variation of the top 10% hashtags of each topic. As explained in Section 5.1, each $topic \in \mathcal{TOPICS}$ is composed of a set of hashtags $\mathcal{H}(topic)$, and for each week $w \in \mathcal{T}(topic)$, there is also a set of hashtags $h \in \mathcal{H}(topic, w)$. This feature was calculated by selecting the top 10% hashtags with the highest user variation $\Delta^U(h,w)$ from each $topic \in topics(u,w)$, selecting the intersection between those top hashtags with the influencer's weekly hashtags $\mathcal{H}(u,w)$, and computing the

average user variation $\Delta^U(h, w)$ of the hashtags in the intersection. With this feature, we aim to quantify the influencer's weekly performance in hashtag selection within the topics posted about by user u in week w. Note that an influencer that uses hashtags that are trending within some topic will have higher values for this feature.

Note that the two last features described above differ in a crucial aspect. While the alignment with global trends is typically beyond the influencer's control since it is difficult to predict arbitrary weekly trends in advance, the strategic allocation of hashtags within topics of choice can be fine-tuned by the influencer. By comparing these two features, we aim to understand to which extent topic trend alignment can be strategically exploited by an influencer.

The correlations were calculated between the features listed above and the following **performance metrics**, for each influencer $u \in \mathcal{U}$ and week $w \in \mathcal{T}(u)$:

1. $\overline{\mathcal{I}(u, w)}$ and $\overline{\mathcal{C}(u, w)}$: weekly average interactions and comments, as defined in (4.4);

Tables 7.1a, 7.1b, and 7.1c present the correlations between the weekly features and performance metrics for beginner, micro-influencer, and celebrity influencer groups, respectively. Firstly, we observe that all ad-related features (Ad ratio, Ad count, and $\Delta_{AD}(w)$) have negative correlations with the performance metrics for every influencer group, confirming our hypothesis that sponsored content can indeed be harmful to audience engagement. Among the influencer groups, micro-influencers have a relatively less negative impact of ad-related features on user engagement, whereas celebrities tend to suffer the most negative impact.

Furthermore, we can observe that the weekly average number of hashtags is also negatively correlated with all metrics among all three influencer groups. Celebrities, though, are the ones with the highest negative correlation, suggesting that the hashtag assignment strategy should be more cautious for that group of influencers. At the same time, the weekly number of posts is also negatively correlated with performance. Again, micro-influencers tend to suffer less from increasing the number of posts per week, while beginners had the most negative correlation. Contrariwise, it is possible to see that using a diversified set of hashtags might be beneficial, since there is a positive correlation between the number of distinct hashtags and the performance metrics, for both beginners and micro-influencers. This indicates that a balanced hashtag strategy, using fewer but more diverse hashtags, might have a positive impact on the performance of a week's ad placement strategy. For celebrities, though, the correlations are negative.

One remarkable result is the correlation between topic-related strategy features and performance metrics. There is a positive correlation between the topic diversity and the performance metrics, as well as between the trend alignment, meaning that planning well which topics to post about could be used to mitigate the negative effect of ad placement. Not only this, but also the existing correlation between the performance metrics and the intra-topic hashtag trend alignment features, shows us that it is important not only to post about trending topics but, more importantly, to wisely plan which hashtags within each topic will be used since only belonging to a trending topic is not as highly correlated as the correct usage of topic hashtags. This affirmative holds for both beginners and micro-influencers, beginners having the highest correlations, indicating that for up and coming influencers following trends is important. Celebrities, on the other hand, do not seem to benefit much from topic-related features, having correlations close to 0. Information regarding weekly topic trends are typically unknown in advance unless it is a predictable seasonal trend, and switching topics simply for the sake of global trend alignment might hurt an influencer's authenticity. These results could be seen as a "luck" component of the success of a week's posting activity.

The correlations shown in Table 7.1 which refer to ad-related features (more specifically, *Ad ratio*, *Ad count*, and $\Delta_{AD}(u, w)$) suggest that an increasingly aggressive ad placement strategy tends to harm the weekly audience engagement for all influencers. Nevertheless, some strategies might mitigate this effect. The number of mentions, for example, can be used to attract different audiences for an influencer's channel. It is possible to see that both beginners and micro-influencers might benefit from this tool, given the high positive correlation with the weekly performance metrics. The results also show that beginners and micro-influencers might profit from having longer post texts. This impact is even higher on the average number of comments, for both influencer groups. Unlike the features related to global topic trends, which cannot be easily predicted, these features can be adjusted by the influencer and can, therefore, be seen as strategic. Nevertheless, the high correlations seen for topic trend-related features can also be exploited when planning for easily predictable trends, such as seasonal events like Christmas and New Year's Eve.

Also, both beginners and micro-influencers can benefit from a smart hashtag assignment within the topic choice, as can be seen by the significant correlations between performance metrics and the *Intra-topic hashtag trend alignment* Δ feature. The same strategy for celebrities did not show significant correlations.

Celebrities had almost all feature correlations negative or close to zero. This can be seen as an indication that the bigger the follower base, the harder it becomes to keep growing. Still, some strategies should probably be avoided by celebrities. More

specifically, as can be seen in Table 7.1c, writing posts with longer texts (*post length*), using too many hashtags (*average hashtags per post*), and posting too many ads (*ad count*) is negatively correlated with the performance metrics.

Our analysis revealed that the same posting strategy could have a different impact, depending on the size of the follower base of an influencer. There are some evident differences. For example, while increasing the number of mentions in a post might be a good strategy to boost interaction numbers for a beginner or a micro-influencer, celebrities tend to have decreased engagement, when the number of mentions is high. Moreover, while having longer post texts can be beneficial for beginners and micro-influencers, for celebrities, this (highly) correlates with lower numbers of interactions and comments. Alternatively, while both beginners and micro-influencers can have increased performance by posting about trending topics, and by doing smart intra-topic hashtag selection, celebrities do not seem to benefit from the same strategy.

The results regarding topic trends presented above bring an interesting discussion into play. It is evident that there is a correlation between an influencer's posting strategy success and whether this influencer is aligned with global trends. But to what extent should an influencer focus on strategies exploiting these results? Does it make sense to change the subject about which they post to follow global tendencies? Based on these results, we conjecture that two major groups of features define an influencer's monetization success. The first group is comprised of features that can be controlled by the influencer, such as the post length and total amount of posts, usage of mentions and hashtags, and quantity of ads. The second group is related to those features that can be seen as the luck component of success and cannot be controlled by the user, such as global topic trends prediction and alignment with trending topics.

7.2 Relative feature relevance

To further understand what influences the performance of a post or an ad placement strategy of an influencer, we used our dataset to create a classification model, using machine learning techniques, to calculate the relative importance of different features. We modeled our data as a classification problem, with the week's success status as the target (as defined in Sec. 4.2, a week w is considered successful if $\Delta^{\mathcal{I}}(u, w) > 0$, and unsuccessful if $\Delta^{\mathcal{I}}(u, w) < 0$).

To train the classifier, we used the Extremely Randomised Trees algorithm Geurts et al. [2006]. This algorithm uses the Gini impurity index for the calculation of divisions during training. According to Breiman et al. Breiman [2001]:

> Every time a split of a node is made on variable M, the Gini impurity criterion for the two descendant nodes is less than the parent node. Adding up the Gini decreases for each individual variable over all trees in the forest, gives a fast variable importance that is often very consistent with the permutation importance measure.

To evaluate our model, we used the Area Under the ROC Curve (AUC). The AUC is a measurement of separability, having values in [0, 1]. The closer to 1, the better the model. Our model achieved an AUC of 0.85.

We considered the following *weekly features*, $\forall u \in \mathcal{U}, w \in \mathcal{T}(u)$:

1. ***Average text length***: average length of the post's textual content, in week w;
2. ***Average hashtag count***: average number of hashtags assigned to a post. in week w;
3. ***Average mention count***: average number of mentions in a post, in week w;
4. ***Month***: the month in which week w occurred;
5. ***Year***: the year in which week w occurred;
6. ***Total post count***: $|posts(u, w)|$;
7. ***Week type***: $w \in \{AD^{+}(u),\ AD^{-}(u),\ AD^{=}(u)\}$;
8. ***AD count variation***: $\Delta_{AD}(w)$ (defined in eq. (4.1));
9. ***AD percentage***: percentage of AD posts in week w;
10. ***Influencer group***: $u \in \{\textit{beginners}, \textit{micro-influencers}, \textit{celebrities}\}$;
11. ***Hashtag diversity***: number of distinct hashtags used in week w;
12. ***Topic diversity***: number of distinct topics posted about in week w;
13. ***Trend alignment***: weekly average topic user variation, as defined in Section 7.1;
14. ***Intra-topic hashtag trend alignment*** Δ: weekly average user variation of the top 10% hashtags of each topic, as defined in Section 7.1;

Table 7.2: Feature relevance on predicting the performance of a week $w \in \mathcal{T}(u), u \in \mathcal{U}$

Feature relevance					
post length	0.123		$\Delta_{AD}(w)$	0.056	
average hashtags	0.111		*topic diversity*	0.053	
month	0.096		*ad percentage*	0.044	
hashtag diversity	0.092		*week type*	0.039	
intra-topic hashtag trend alignment Δ	0.082		*influencer group*	0.038	
trend alignment	0.081		*year*	0.026	
$\lvert posts(u, w)\rvert$	0.080				
average mentions	0.074				

Even though we used our model simply to extract the feature relevances, it could be used to predict a given weekly posting strategy, for a given influencer, in a real world setting.

Table 7.2 analyzes the features relative importance when predicting the week's success status. As we defined in Section 4.2, a week is considered to be successful if it's variation in the number of interactions relative to the previous week $\Delta^{\mathcal{I}}(u, w) > 0$, and it's considered to be unsuccessful otherwise ($\Delta^{\mathcal{I}}(u, w) < 0$). We can see that the average text length is the most important feature, followed closely by the average number of hashtags. One remarkable result is the relatively high importance of the topic-related features, more specifically, of the feature indicating trend alignment. It is possible to see that posting about trending topics can be beneficial, as evidenced by the importance of the *trend alignment* feature. Moreover, a clever selection of topic-related hashtags (*intra-topic hashtag trend alignment* Δ feature) is also important, indicating that not only it is interesting to post about subjects having a positive growth tendency, but it is also important to use trending hashtags within the topics of choice.

It is interesting to point out that, even though the relative importance of both *trend alignment* and *intra-topic hashtag trend alignment* Δ are similar (with relevances close to 0.08), it is good news for influencers that they do have control of at least one of the two features. The former is typically beyond the influencer's control since it is difficult to predict weekly global trends; while the latter can be more easily exploited by an influencer, given that within their topic of choice, the selection of trending hashtags can rely on their expertise in that particular subject.

One unexpected positive result that is worth pointing out is the low relative relevance of the AD-related features, such as the weekly *ad percentage*, the $\Delta_{AD}(w)$, and *week type*. This is quite encouraging because even though sponsored content tends

to generate less audience engagement than regular content and increasing the number of weekly ads is negatively correlated with our performance metrics in all influencer groups, as shown in Sections 6.1 and 7.1, these ad-related features are less relevant in predicting weekly success than almost every other feature. This result suggests that, while the ad placement strategy should be taken into account, several other features are more important for the prediction of the success of a weekly monetization strategy and, thus, could be used by an influencer in order to mitigate the negative effects of sponsored content placement.

Another interesting result is the relatively high relevance of the *month* feature, which suggests that the time of the year in which the ads are posted might be relevant for the success of a week's monetization strategy. This goes hand in hand with the high relevance of the topic-related features, which might indicate that exploiting seasonal events, such as Christmas, NYE, and Summer, could be used to boost a week's performance. As opposed to arbitrary weekly trends, many seasonal trends are, in fact, easily predictable and the relative importance of trend-related features indicates that taking seasonal trends into account when defining ad posting strategies can be beneficial for boosting a week's performance.

Near the bottom of the feature relevance list is the *influencer group* feature. The fact that this feature has low relative relevance in our model is another motivating result for influencers in the beginners' group, suggesting that the size of their follower base does not determine the success of their ad placement strategies.

Chapter 8

Conclusions

In this work, we presented a characterization and analysis of sponsored content and its impact on user engagement on Instagram. We collected and characterized a dataset of over 7,500 influencer profiles and over 3 million posts. In order to better understand the different strategies used for sponsored content placement on Instagram, we divided the influencer profiles into three categories, based on the size of their follower base: *beginners*, *micro-influencers*, and *celebrities*. We proposed metrics to measure the success of ad placement strategies. More specifically, we analyzed how the variation in the number of sponsored posts in a week correlates with audience engagement within an influencer's profile and showed that, for each of these groups, the performance of ad content is influenced by distinct factors.

Our analysis showed that an increasing amount of sponsored content can negatively affect an influencer's audience engagement. Surprisingly, celebrities suffer the most negative impact when increasing the number of ad posts in a week. Nevertheless, our feature correlation analysis showed that different user groups could adopt different posting strategies to mitigate the negative effect of ad placement. In particular, for beginners, it is important to strengthen social ties within the platform, e.g., by increasing the number of mentions of other user-profiles and investing in fewer posts with longer texts. Celebrities, on the other hand, do not need to reinforce social ties within the platform, but rather should invest in more succinct posts with fewer hashtags. Finally, the month of the year has shown to be of importance for a successful week in terms of user engagement, suggesting that season-specific strategies can be used to increase the growth and reach of an influencer's profile.

Although we have shown that advertisement can, indeed, harm an influencer's performance, there seem to be a variety of ways to mitigate the negative effect of advertisements, such as strategic usage of mentions and hashtags, the textual content

of a sponsored post, as well as the season of the year, when planning an advertisement placement strategy. We also observed that whereas sponsored posts tend to have more hashtags than non-sponsored posts, possibly with the intent to reach a larger audience through search engines, such a hashtag assignment strategy should be balanced, since it might negatively impact the audience engagement with a post.

One of the negative results of our analysis is that some relatively important features are beyond influencer's control, such as alignment with arbitrary weekly trends, since it is difficult to predict global tendencies. Either way, even though some features like alignment with trending topics on a weekly basis cannot be controlled by an influencer, many features that shape an ad placement strategy can, in fact, be controlled and are relevant to the success of the influencer in terms of audience engagement with regular and sponsored posted content. As an example, the size of the text, the usage of mentions, and the hashtag diversity all correlated highly with our proposed success metrics. Furthermore, even though alignment with global trends is not within the reach of an influencer's control, the smart usage of hashtags within chosen topics has proven to be equally important and, unlike the later, can be controlled by an influencer and could be leveraged in order to positively boost weekly monetization strategies.

Interestingly, one of the least relevant features in our analysis turned out to be the influencer's group, i.e., the size of their follower base. This is a positive result for those influencers that do not have large numbers of followers, suggesting that the (current) size of their follower base must not be a limiting factor for the success of their monetization strategies on Instagram.

We also analyzed seasonal trends. We could detect that the usage of specific hashtags during specific times of the year, such as summer and Christmas, can boost ad performance. Even though arbitrary weekly trends are hard to predict, seasonal trends are easily predictable, and the high relevance of the trend alignment features might indicate that leveraging seasonal trends in monetization strategies might be beneficial. This strategy can be used to increase the performance of advertised content during selected periods of the year. Even so, the use of trending hashtags does not benefit everyone. *Celebrities*, for example, did not benefit from this strategy, when it comes to likes and interactions and *beginners* even had a negative impact.

To the extent of our knowledge, this is the first in-depth study of the relationship among (disclosed) sponsored content placement, global topic trends, and audience engagement on Instagram. Whilst companies and brands are interested in reaching the higher possible audience keeping minimal costs, social media users should be aware of how this type of content affects their followers and to which extent it poses a threat to their overall success. Our results revealed several strategies that can be exploited

to boost audience engagement and performance of monetization strategies on online social media platforms.

Finally, there are some future directions of work that might be interesting to point out. Instagram is a photo and video-sharing social network, and, thus, one exciting direction would be to study how the visual content of a post can impact a monetization strategy. Another direction could be in understanding how different posting tools can influence engagement, such as the difference between images, videos, Instagram stories, etc. In another sense, future efforts could also aim to further analyze the textual content of posts, by applying natural language processing in the posts textual content. All things considered, there are plenty of different approaches that might be interesting fields of study in the influencer marketing phenomenon.

Finally, we intend to release the dataset for the community at a later date.

Bibliography

Andreou, A., Venkatadri, G., Goga, O., Gummadi, K. P., Loiseau, P., and Mislove, A. (2018). Investigating ad transparency mechanisms in social media: A case study of Facebook's explanations. In *NDSS 2018, Network and Distributed Systems Security Symposium 2018, 18-21 February 2018, San Diego, CA, USA*, San Diego, USA.

Angwin, J. and Parris Jr., T. (2016). Facebook Lets Advertisers Exclude Users by Race. `https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race`. Online; May 2020.

Araújo, C. S., Corrêa, L. P. D., d. Silva, A. P. C., Prates, R. O., and Meira, W. (2014). It is Not Just a Picture: Revealing Some User Practices in Instagram. In *2014 9th Latin American Web Congress*, pages 19–23.

Argyrou, A., Giannoulakis, S., and Tsapatsoulis, N. (2018). Topic modelling on Instagram hashtags: An alternative way to Automatic Image Annotation? In *2018 13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, pages 61--67. IEEE.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993--1022.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5--32.

Chen, Q., Yao, L., and Yang, J. (2016). Short text classification based on LDA topic model. In *2016 International Conference on Audio, Language and Image Processing (ICALIP)*, pages 749--753. IEEE.

Cramer, H. (2015). Effects of Ad Quality & Content-Relevance on Perceived Content Quality. In *Proceedings of the ACM CHI'15 Conference on Human Factors in Computing Systems*, volume 1, pages 2231--2234, New York, New York, USA. ACM Press.

De Jans, S., Van de Sompel, D., De Veirman, M., and Hudders, L. (2020). # sponsored! how the recognition of sponsoring on instagram posts affects adolescents' brand evaluations through source evaluations. *Computers in Human Behavior*, page 106342.

Evans, N. J., Phua, J., Lim, J., and Jun, H. (2017). Disclosing Instagram Influencer Advertising: The Effects of Disclosure Language on Advertising Recognition, Attitudes, and Behavioral Intent. *Journal of Interactive Advertising*, 17(2):109--123. ISSN 1525-2019.

Facebook (2017). Improving Enforcement and Promoting Diversity: Updates to Ads Policies and Tools. `https://about.fb.com/news/2017/02/improving-enforcement-and-promoting-diversity-updates-to-ads-policies-and-tools`. Online; May 2020.

Federal Trade Commission (2019). Disclosures 101 for social media influencers. `https://www.ftc.gov/tips-advice/business-center/guidance/disclosures-101-social-media-influencers`. Online; March 2019.

Ferrara, E., Interdonato, R., and Tagarelli, A. (2014). Online Popularity and Topical Interests Through the Lens of Instagram. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14, pages 24--34, New York, NY, USA. ACM.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1):3--42.

Giannoulakis, S. and Tsapatsoulis, N. (2019). Filtering Instagram Hashtags through crowdtagging and the HITS algorithm. *IEEE Transactions on Computational Social Systems*.

Jang, J. Y., Han, K., Shih, P. C., and Lee, D. (2015). Generation Like: Comparative Characteristics in Instagram. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 4039--4042, New York, NY, USA. ACM.

Klear Influencer Marketing (2019). The State of Influencer Marketing 2019. `https://klear.com/TheStateOfInfluencerMarketing2019.pdf`. Online; January 2019.

Koutsopoulos, I. and Spentzouris, P. (2016). Native advertisement selection and allocation in social media post feeds. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 588--603. Springer.

Lithium Technologies (2016). 74 percent of digital natives tired of brands shouting at them. https://www.prnewswire.com/news-releases/74-percent-of-digital-natives-tired-of-brands-shouting-at-them-300263367.html. Online; May 2020.

Loude, E. (2017). *#Sponsored?: Recognition of Influencer Marketing on Instagram and Effects of Unethical Disclosure Practices*. PhD dissertation, University of Minnesota.

Mathisen, A. and Stangeby, M. F. (2017). *Factors influencing advertising effectiveness and purchase intention on Instagram*. PhD dissertation, BI Norwegian Business School.

Pal, A., Herdagdelen, A., Chatterji, S., Taank, S., and Chakrabarti, D. (2016). Discovery of Topical Authorities in Instagram. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 1203--1213, Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Pinder, C. (2017). The Anti-Influence Engine. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '17*, pages 770--781, New York, New York, USA. ACM Press.

Rokach, L. and Maimon, O. (2005). Clustering methods. In *Data mining and knowledge discovery handbook*, pages 321--352. Springer.

Segev, N., Avigdor, N., and Avigdor, E. (2018). Measuring influence on instagram: A network-oblivious approach. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 1009--1012, New York, NY, USA. ACM.

Shatnawi, M. and Mohamed, N. (2012). Statistical techniques for online personalized advertising: a survey. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pages 680--687. ACM Press.

Song, J., Han, K., Lee, D., and Kim, S.-W. (2020). Understanding emotions in sns images from posters' perspectives. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 450--457. ACM Press.

Speicher, T., Ali, M., Venkatadri, G., Ribeiro, F. N., Arvanitakis, G., Benevenuto, F., Gummadi, K. P., Loiseau, P., and Mislove, A. (2018). On the Potential for Discrimination in Online Targeted Advertising. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*'18)*.

Wang, L., Liu, H., Chen, G., Ye, G., Meng, X., Hu, Y., and Wang, H. (2019). Learning theory and algorithms for revenue maximization in sponsored search. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 434--440. IEEE.

Ward, J. and Joe, H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236--244.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. (2011). Comparing twitter and traditional media using topic models. In *European conference on information retrieval*, pages 338--349. Springer.

Zhao, X., Yang, J., Xie, T., and Wang, Z. (2017). Examining advertising intrusiveness on Instagram: Hedonic and utilitarian attributes of brand and sponsored content. In *Proceedings of the Conference of the American Academy of Advertising*, pages 243--255.

Zhou, K., Redi, M., Haines, A., and Lalmas, M. (2016). Predicting Pre-click Quality for Native Advertisements. In *Proceedings of the 25th International Conference on World Wide Web - WWW '16*, pages 299--310, New York, New York, USA. ACM Press.