Monitoring and analysing social media in relation to IP infringement

EUIPO
EUROPEAN UNION
INTELLECTUAL PROPERTY OFFICE

Acknowledgements

Table of Contents

Executive Summary

The growth of e-commerce has been well documented, but how the rise of different technologies and consumer habits has affected intellectual property rights (IPR) infringement on the internet and, in particular, on social media platforms, is not clear. It is within this context, the European Union Intellectual Property Office (EUIPO) decided to conduct a study to better understand the volume and frequency of IPR infringement on social media [1]. This study was organised around three strands of interest to better understand the current activities and trends related to counterfeit physical products and the piracy of digital content.

The first part of the study aims to provide a comprehensive picture of the social media uses related to possible IPR infringement activities or promotion; the second part measures the relative presence of IPR infringement on physical products and digital content on social media compared to genuine products or licit copyright-protected digital content; and the third part intends to identify key indicators in order to better recognise IPR infringement business models on social media.

The scope of the analysis was conducted by data mining four social media platforms (Facebook, Twitter, Instagram, and Reddit) in six European countries: Germany, Spain, France, Italy, Poland and Sweden, although conversation traceability on the social media channels selected for the study was not always possible. In order to mitigate this limitation, language criteria were used to allocate conversations to the targeted countries. Consequently, English was added to the six European Union (EU) languages considered for the study (French, German, Italian, Polish, Spanish, Swedish) as it is widely used on social media, by non-native speakers.

The study applied a methodology based on social intelligence analytics (SIA) and a quantitative and a qualitative analysis. In targeting conversations related to IPR infringement of physical products and digital content, one of the limitations of the study was the right to access certain conversations on social networks. Therefore, the study addressed only public conversations. Another issue highlighted by the study was the

---

[1] As highlighted in the EUIPO 2020 Status Report on IP infringement published in June 2020.

difficulty in distinguishing between licit and illicit content. Finally, the results presented in this report naturally represent a restricted sample and the analyses can only speak for the six countries, seven languages and the four social media channels included. Nonetheless, this variety of sources and the breadth of the data means the conclusions drawn can be considered as holding a broader relevance. Likewise, the significance of this report regarding the behaviours identified on social media can be considered indicative of wider habits that are currently prominent on social media in relation to IP infringement.

Social media uses for IPR infringement activities or promotion

A total of 3.9 million conversations related to the categories and brands chosen to represent physical products were extracted for the study using SIA. Among these categories of physical products, toys, perfume and cosmetics had the highest volumes of conversation. Regarding digital content, films, music and video games had the highest volume of conversations.

An IPR infringement-related conversation was identified as any conversation concerning a breach of an IPR. The study identified 11 % of conversations regarding physical products could be possibly related to counterfeits, and 35 % of conversations on digital content could be possibly related to piracy. Although the quantitative analysis was conducted carefully, the study highlighted the difficulty to identify IPR infringement with certainty.

When examining where there were the highest number of IPR infringement incidences, clothing, footwear and jewellery all featured in the top 3 for physical products. E-books, TV shows and music recorded the highest number of infringements for digital content. The clear conclusion to be drawn from this evidence was that these products collectively had the highest rate of conversations that were identified as being possibly related to IP infringement.

The relative presence of conversations related to counterfeit products and pirated content on social media compared to genuine products or licit digital content

The second part of the study, based on the analysis of three different aspects: platforms, languages and timelines, identified certain trends in IPR infringement activities and efforts to promote them.

In the case of IPR infringement for physical products, the study revealed that Instagram was the social network with the highest total volume of conversations. At product category level, Instagram was also the main platform for conversations about watches, toys, perfume and cosmetics, jewellery and footwear. Twitter was found to have the highest volume of conversations about clothing and toys, and Reddit was used most often for conversations regarding pharma and headgear. In contrast, Facebook [2] showed a lower volume of conversations, which could be explained by an efficient approach from the platform to identify and delete infringing content [3]. This, however, cannot be corroborated by the evidence in this report and, in fact, it may well be that Facebook's low number of recorded conversations was due to IPR infringement-related conversations taking place in private rather than in public. Although all platforms have a direct messaging function, the prominence of private groups appears to be a phenomenon unique to Facebook. As this study only addressed public conversations, this hypothesis could not be confirmed but it should be borne in mind when noting the comparably low numbers of conversations suspected of IPR infringements identified on Facebook by SIA.

In the case of IPR infringement for digital content, a different pattern appeared. Reddit became particularly important for conversations related to films and TV shows, while Twitter was preferred for conversations related to music and e-books.

Regarding the language parameter, the first part of the study highlighted that English was predominant and that some languages, such as Polish and Swedish, were used infrequently. Despite this, it was observed that a high volume of French, German, Italian

---

[2] Private Facebook groups are excluded from this study.

[3] Facebook's measures to face IPR infringement include a 'global notice-and-takedown program, a robust repeat infringer policy, and additional specialised measures going beyond notice-and-takedown.' – Facebook transparency report.

and Spanish conversations were found on Instagram regarding physical products, except for pharma-related mentions, which took place more frequently on Twitter.

As the study's duration was 6 months (April to September 2020) and took place during the COVID-19 pandemic, it was observed that the results obtained were affected in part by the lockdown. In order to contextualise the results, the overall volume of IPR-related conversations was compared in parallel with the overall pattern of IPR conversations recorded over the last 3 years (June 2017 to July 2020). This highlighted that, for physical products, the pharma category showed a noticeable increase of conversations, and the clothing category peaked around the end of the lockdown. In contrast, the digital content trend showed a progressive increase in conversations that could be interpreted as a learning-curve effect as users became more familiar with both licit and illicit digital content consumption.

IPR infringing business models used on social media

The topic modelling and the qualitative analysis conducted to identify possible IPR infringing business models highlighted that it was very difficult to find a pattern regarding IPR infringement on social media. This could be explained by the following observations.

- In the case of a physical product, the analysis revealed that conversations related to IPR infringement concerned mainly promotions and commercial activities. The providers of counterfeit products copied proven and successful business models already used by the legitimate brands.

- In the collected conversations, piracy of digital content was driven by the users' intention to access pirated content. Indeed, all conversations were about finding ways to access content illegally. Consequently, providers of pirated digital contents assumed a passive role, relying on users to disseminate the information and attract new users to their content.

Although tailor-made IPR infringement business models could not be established in this study, the analysis led to the definition of what can be termed a 'bundle-of-clues' approach to help to identify the conversations related to IPR infringement for both physical products and digital content. A more reliable approach to detection would also

be made easier by using trained machine learning models that could provide predictions and recommendations.

The study demonstrates that social media platforms are tools for recurrent IPR infringement for both physical products and digital content. Furthermore, identifying this type of content is complex due to the varied and constantly changing approaches used by providers of counterfeit products or pirated content, which are widespread across multiple platforms, languages and content types.

1   Introduction

This report on monitoring and analysing social media in relation to IPR infringement brings together the findings of research carried out by the European Union Intellectual Property Office (EUIPO) through the European Observatory on the Infringement of Intellectual Property Rights (Observatory).

Social media is understood as 'computer-based technology that facilitates the sharing of ideas, thoughts, and information through the building of virtual networks and communities. By design, social media is internet-based and gives users quick electronic communication of content. Content includes personal information, documents, videos, and photos. Users engage with social media via computer, tablet or smartphone via web-based software or web applications, often utilising it for messaging [4].'

This report is intended to assess and investigate the main characteristics, relevance, trends and impact of the use of social media as a promotion and/or distribution channel of counterfeit products and/or pirated digital content. It investigates and quantifies the role that certain social media platforms play unwillingly in sharing, marketing, selling, promoting and/or advertising IP infringing goods, works, materials and/or services.

This research topic has become of growing interest to IP organisations, rights owners, enforcement agencies, researchers and the media worldwide. According to a study carried out by the EUIPO and the Organisation for Economic Co-operation and Development (OECD) in 2019, estimates on IPR infringement in international trade reached 3.3 % of world trade and 6.8 % of EU imports, equating to EUR 121 billion annually. Both sets of figures are significantly higher than those found in the previous edition published by the two organisations in 2016, indicating that the problem has become even more serious in recent years [5].

This study seeks to establish to what extent this growth can, in part, be attributed to social media, which seems to increasingly supplement the marketplaces that have

---

[4] https://www.investopedia.com/terms/s/social-media.asp.

[5] EUIPO 2020 Status Report on IPR infringement, June 2020. The corresponding figures in the 2016 study were up to 2.5 % of world trade and up to 5 % of EU imports from the rest of the world.

traditionally been misused to sell counterfeit goods [6]. A 2019 study of Instagram showed that more than 1 out of 10 posts contained some kind of content posted by counterfeiters, and that over 15 % of Instagram content related to a specific hashtag or brand was generated by accounts dedicated to illegal activities [7]. Furthermore, the amount of counterfeit content on Instagram grew 341 % between 2016 and 2019 [8]. However, Instagram is not the only social media platform with a growing counterfeiting and pirated content problem. A United Kingdom (UK) Intellectual Property Office (IPO) report suggested that over five times more counterfeit goods were sold in invite-only groups on Facebook and Twitter than in open groups [9].

New promotion and sales strategies use social media to avoid the actions of law enforcement agencies, such as using instant messaging combined with social media promotional messages to enable counterfeit product sales and consumer contact, avoiding the use of e-commerce websites that can be identified and closed more easily when reported. Europol states that 'vendors advertise the counterfeit goods through posts showing the product and price. Then, the details of the transaction are defined through other communication channels. Couriers deliver the packages and payment may be made via prepaid cards, PayPal or other payment methods [10].'

The data underpinning this report was collected throughout 2020, and so the research reflects the impact of the COVID-19 pandemic on IPR infringement on social media, both in terms of the promotion and dissemination of both counterfeit products and pirated content. Europol's COVID-19 report notes that, during the spring of 2020, with millions of EU citizens confined to their homes due to the pandemic, the use of illicit IPTV increased [11]. Furthermore, the number of conversations on platforms such as Twitter about counterfeit goods related to COVID-19 closely coincided with widespread

---

[6] IP crime highlight report 2013/14.

[7] Andrea Stroppa, Davide Gatto, Lev Pasha and Bernardo Parrella, Instagram and counterfeiting in 2019: new features, old problems, April 2019.

[8] Instagram and counterfeiting in 2019: new features, old problems.

[9] https://www.clarionsolicitors.com/articles/intellectual-property-infringement-on-social-media-a-growing-problem.

[10] Social media crime: 20 000 packages of counterfeit medicine, mobile phones, jewellery, sunglasses and watches seized – Europol Press Release, May 2018.

[11] EUIPO 2020 Status Report on IPR infringement, June 2020.

outbreaks and the introduction of restrictive measures to prevent the spread of the virus in different Member States ([12]).

However, despite evidence of an increase in IPR infringing online behaviours related to the impact of COVID-19, consumer sentiment regarding purchasing counterfeit goods compared to consuming pirated content online has also shown a tendency to fluctuate in recent years for other reasons. The 2019 IP and Youth Scoreboard ([13]) shows that, compared to the previous 2016 edition, young people in the EU were less likely to consume pirated digital content, but are slightly more likely to purchase counterfeit goods ([14]).

The scope and methodology applied to conduct the research for this report is described in Chapters 2 and 3 of this report. The results of the quantitative and qualitative analysis in relation to the relative presence of IPR infringement products and illicit digital content on social media compared to genuine products or licit digital content are presented in Chapters 4 and 5. Those chapters discuss the extent to which social media is used for IPR infringement activities or promotion, and possible trends. Finally, the report concludes by identifying some key indicators of IPR infringing business models or behaviours used on social media and the main advantages of social media compared to other communication channels, from the point of view of the infringers.

([12]) Viral marketing – Counterfeits, substandard goods and intellectual property crime in the COVID-19 pandemic, Europol, April 2020.

([13]) https://euipo.europa.eu/ohimportal/en/web/observatory/ip-youth-scoreboard.

([14]) EUIPO 2020 Status Report on IPR infringement, June 2020.

2 Objectives and Scope of the Study

Online counterfeiting and piracy are challenges that many companies and institutions have tried to address, but despite enforcement measures, they persist as problems and are responsible for considerable losses in sales for legitimate enterprises and IP rights holders as well as public revenue for governments. As stated in the EUIPO 2020 Status Report on IPR infringement, 'while 97 % of Europeans surveyed believe that it is important that inventors, creators and performing artists can protect their rights and be paid for their work, 10 % acknowledged they have intentionally purchased counterfeit goods, and a similar proportion admitted to having intentionally downloaded or streamed content from illegal online sources during the last 12 months ([15]).'

Consequently, the EUIPO launched this study to explore and better understand IPR infringement on social media. More specifically, the study aims to answer three investigatory questions.

1. To what extent is social media used for IPR infringement activities or promotion, and can we observe certain trends?
2. What is the relative presence of IPR infringement products and illicit digital content on social media compared to genuine products or licit digital content?
3. What kind of IPR infringing business models are used on social media, and what are the main advantages compared to other communication channels?

The EUIPO set the study frame according to three aspects: 1) the social media platform, 2) geographical scope, and 3) type of infringement.

([15]) EUIPO 2020 Status report on IPR infringement, June 2020.

Social media platforms

The EUIPO decided to focus the study specifically on four social media channels. The selection of these platforms was made based on the following criteria:

- accessibility to public conversations and authorisation to collect public data to ensure GDPR compliance;
- popularity of the social media platform in the targeted countries;
- geo-localisation of the conversations or emitter account.

With these criteria in mind, it was decided that the study would focus on Twitter, Facebook, Instagram and Reddit. The choice of these social media platforms is explained by the popularity of these four channels and by their communication type. Facebook and Instagram enable a more visual style of communication, whereas Twitter and Reddit use shorter conversational exchanges.

Twitter, Facebook and Instagram all fulfil each of the study criteria. However, there was uncertainty as to the ability of Reddit to fulfil the last criterion, as Reddit does not make conversation locations explicit, and the main language used is English. Having considered all options, the EUIPO decided to include Reddit in the study, and an adjustment to the geo-localisation criterion was made. The official language of the targeted countries was to be considered in addition to the geo-localisation of the conversations when this information was offered by the social media platform. Therefore, in the case of Reddit, the results in this study may originate from broader regions than the country-oriented results from other platforms.

For the implementation of the study, some constraints had to be taken into account when proceeding with the data collection. Each social media channel applies strict GDPR rules. Consequently, the access to data used for this study was strictly limited to public conversations in line with the agreement negotiated between each social media channel and the social listening platform.

More specifically, the data collected from each social media platform in this study corresponds to the elements presented in the following table.

Channels	Data Collected	Metadata	Interactions
	All tweets from tracked keywords, hashtags, and accounts	• # of followers* • # of following • # of updates • # of listed	• # of retweets • # of favourites
	Posts: account-based-crawling Hashtag-based crawling: all public posts from the targeted #hashtag enters the database and is anonymised	(Account-based only) • # of followers • # of following • # of media	• # of likes • # of comments • # of views (videos only)
*Only in public conversations	All posts, comments and time of publication for tracked pages	• likes (for page)	• # of likes and reactions • # of comments • # of shares • # of views (videos only)
	90 % of contents of all major subreddits are accessible for data collection	n/a	n/a

*# of followers defined as 'people following a social media account (Twitter; Instagram; etc.)'
# of following defined as 'people followed by a social media account'

Localisation

In order to better understand IPR infringement on social media, the EUIPO considered it appropriate to address a sample of six EU countries to identify emerging trends. The selection of countries was made on the basis of their population size in order to ensure sufficiently large volumes of public conversations and a balanced representation of the main regions of the European Union, and finally, according to the prevalence of social media use. The countries included in the study were Germany, Spain, France, Italy, Poland, and Sweden. Focusing on these countries implied that the study covered the official languages of these countries as well as English, which was usually the most used language after the national language.

Targeted content

In order to set manageable parameters for the study, the EUIPO chose to address eight product categories considered particularly vulnerable to counterfeiting, as shown in the joint reports of the OECD and the EUIPO's sectorial studies of counterfeits [16], and six content types affected by piracy.

COUNTERFEIT

PRODUCT CATEGORIES

Clothing	Jewellery
Headgear	Perfumery & cosmetics
Footwear	Toys
Watches	Pharma

PIRACY

CONTENT TYPES

TV shows	E-books
Movies	Computer games
Music	Sport events
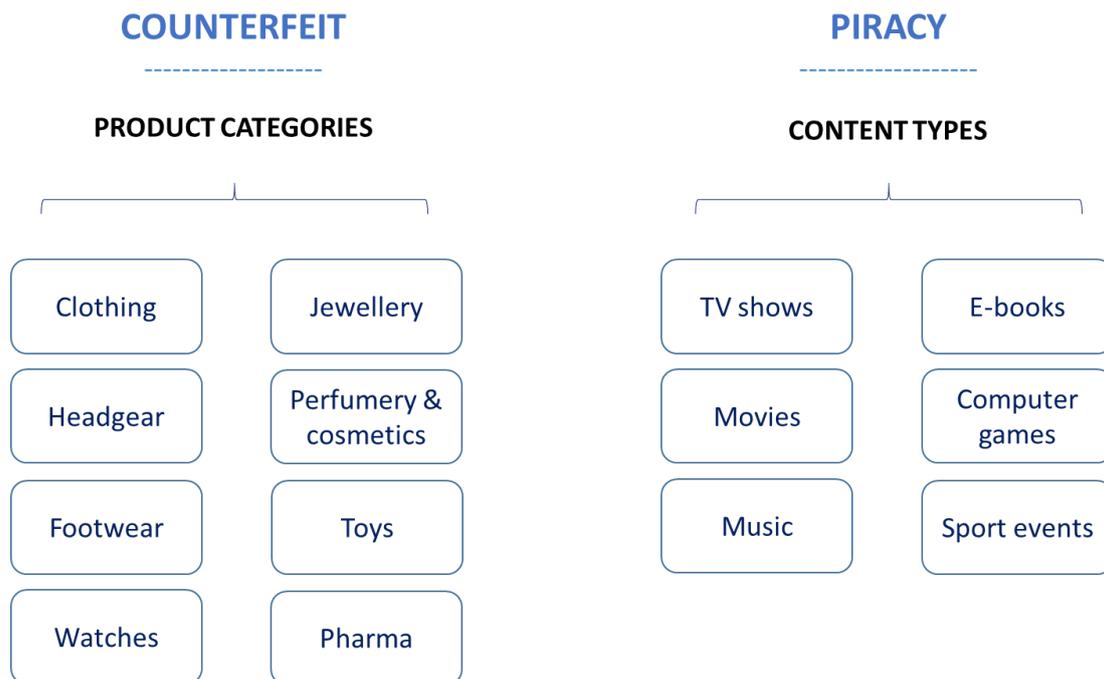
Figure 1 – Product categories and content types addressed in the study

The elements taken into consideration for analysis in the study were limited to text components of the collected public conversations (post, comments, user account, hashtag, etc.). Images and videos were not included in the data-collection process as visual elements require an additional process of codification to treat images (static or animated). However, if images were used in identifying the data corpus, they were taken into account in the qualitative analysis of the data corpus.

---

[16] See: https://euipo.europa.eu/ohimportal/en/web/observatory/trends-in-trade-in-counterfeit-and-pirated-goods for the most recent EUIPO-OECD study. The sectorial studies are summarised in the 2020 Status Report on IPR Infringement.

3   Methodology of the Study

In order to answer the research questions, it was necessary to listen to and analyse public conversations on social media. The accuracy of this activity depended on the volume of conversations listened to. To facilitate this process and to maximise the amount of data that could be gathered and analysed, the use of SIA was required. For this task, the EUIPO used IPSOS SIA services, which relies on Synthesio, a social listening tool that uses AI technology; it collects, aggregates and displays in dashboards social media content from Facebook, Instagram, Twitter, forums, blogs, news sites and thousands of other sources. The first phase of the study therefore focused on data collection and the second phase was dedicated to the analysis of the collected data.

Before the full-scale implementation of the study, a proof-of-concept phase was conducted to ensure the feasibility of the research. This early step consisted of confirming that the data-collection process could be implemented according to the criteria and scope of the study, and that enough material and data would be gathered to proceed with an objective analysis.

The first step of the study focused on extracting online public conversations produced on the four social media channels concerning the selected brand categories and content types. Synthesio helped to identify the public conversations that were of interest and should therefore be 'listened' to and then extracted the relevant conversations according to the set criteria. The technology does not however detect if bots are operating: a bot is an automated account, it publishes, comments, gives likes, etc. automatically with the purpose of acquiring followers, increasing audience reach, improving engagement levels or generating traffic. As a result, some sellers may use bots to gain attention or trust in order to facilitate product sales. Bots are used for promoting counterfeit goods online, although it still remains unclear what the percentage of bots among social network accounts is and, consequently, it is difficult to estimate the impact of bots in this study.

After a 6 month period of collecting online conversations, the study entered into the analysis phase. Using the application of text analytics settings, which consists of tagging relevant posts to inform the SIA tool about the learning algorithm, a statistical modelling approach was applied to the data corpus. It also enabled the EUIPO to identify conversations regarding counterfeit goods and pirated content.

Initial exploration & source selection

Initial exploration to understand the size of social conversations, choose the most relevant sources and provide input into next stages

Query development

Developing key search terms and queries to identify relevant conversations about the brief

Quantitative analysis

Text analytics to understand the volume of the topics and, where appropriate, to execute analytics approaches such as factor analysis, landscape mapping and data modelling vs. survey results

Qualitative analysis

Qualitative deep dive into motivations and occasions relevant to the topics and to the brief

Figure 2 – Four-step methodology

3.1   Step 1 – Initial exploration and source selection

During the proof-of-concept phase an in-depth exploration was conducted to understand the size (volume and distribution) of social conversations according to the study scope. More specifically the feasibility test consisted of compiling a relevant and consistent data corpus. With this objective, the EUIPO, with the cooperation of Europol, compiled lists of **keywords** that were designed to help identify and filter public conversations: one to identify both the targeted physical products and digital content; another outlining potential keywords used in possible scenarios where IPRs were being infringed.

Two examples of these keyword sets are shown below.

perfume	makeup	lotion	cream	fragrance	mascara
gel	serum	eyeshadow	concentrate	concealer	eyeliner
foundation	palette	protection	powder	lipstick	moisturiser
highlighter	nail	cologne	blush	cleaner	mask

Figure 3 – List of keywords to filter conversation about products in the Perfume and Cosmetics category

offers	outlet	copy	mirror image	wholesale	look alike	knock off
cheap	coupon	factory	original	overruns	authentic	DM
rip-off	replica	stock lot	AAA	for sale	PM	best deal
discount	promo	promotion	sales	clearance	dupe	excess stock

Figure 4 – List of keywords to support the identification of counterfeiting conversations

For physical product categories, the lists of keywords chosen related to brands selected to represent each product category. In order to avoid any negative impact on the reputation of the chosen brands, they will remain anonymous.

In contrast, the spread of digital content types was very broad. Therefore, it was decided to attempt to refine them by concentrating the data corpus of the study on the most popular elements of each content type. Desk research was conducted to identify the most popular, defined as the most explicitly liked digital content for each category in 2019 and 2020. This research was conducted via online survey websites and web portals classifying, the most-rated content, among others. However, when setting this list of keywords and applying it across the four social media channels, this approach did not deliver a data corpus consistent enough for a coherent analysis. The most popular contents did not correlate with the most searched for or discussed, and when combined with the keywords that enabled the filtering of possible IPR infringement scenarios, the public conversations were not numerous enough to conduct an objective analysis.

Consequently, the decision was made to keep the content types broad instead of refining them.

With these adjustments made, the feasibility test confirmed that the study could be implemented with a relevant data corpus on the four social media platforms initially selected. This preliminary procedure demonstrated that data collection was possible in the different languages targeted by the study since the list of keywords was translated and enabled the gathering of relevant conversations.

SOCIAL MEDIA CHANNELS	LANGUAGES	PRODUCT CATEGORIES	CONTENT TYPES
• Facebook • Twitter • Instagram • Reddit	• English • French • German • Italian • Polish • Spanish • Swedish	• Clothing (5 brands) • Footwear (4 brands) • Headgear (2 brands) • Jewellery (5 brands) • Perfumes and cosmetics (5 brands) • Pharma (5 brands) • Toys (4 brands) • Watches (4 brands)	• TV shows • E-books • Music • Movie • Sport events • Computer games

Figure 5 – Scope of the study

3.2   Step 2 – Query development

Once the initial exploration and source selection was completed, the AI-powered technology was applied. Firstly, this consisted of preparing accurate queries in English and in the targeted languages to identify the relevant conversations.

A query is a formula combining keywords enabling the filtering of public conversations on the targeted social media channels. Therefore, the list of keywords prepared for the proof-of-concept was adjusted to better fit each brand category, targeted brands of physical product in the case of counterfeiting, and digital content type in the case of piracy. Then, the list of keywords made specifically to identify and filter possible counterfeiting or piracy actions was refined. Once these keywords were optimised, the queries could be designed.

The first query (Q1) was set to filter all conversations and make sure that the data corpus addressed only the previously defined product categories and content types. As all

product categories and content types led to the collection of a certain amount of content
that was not consistent with the study, each query needed to be refined according to the
specificities of the selected brands, and content types identified for this study (Q1+Q2).
However, as explained below, in the final analysis, Q2 was not applied to digital content.

QUERY 1
Identification of the product or content category
+
QUERY 2
Identification of brands and specific content type product

PRODUCT CATEGORY

Keywords to identify Clothing

CONTENT TYPE

Keywords to identify E-books

BRANDS

Keywords to a specific brand of Clothing

PIRACY

Keywords to identify one of the most popular E-books
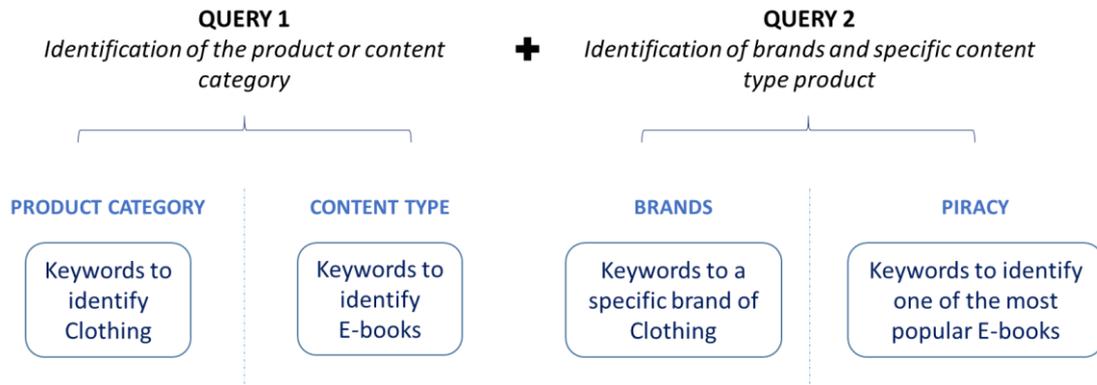
Figure 6 – Illustration of queries combination

A third query was then prepared (Q3) to be combined with Q1 and Q2 to extract
conversations with possible counterfeiting or piracy references.

QUERY 3
Identification of IPR infringement

COUNTERFEITING

Keywords to identify counterfeit

PIRACY

Keywords to identify piracy
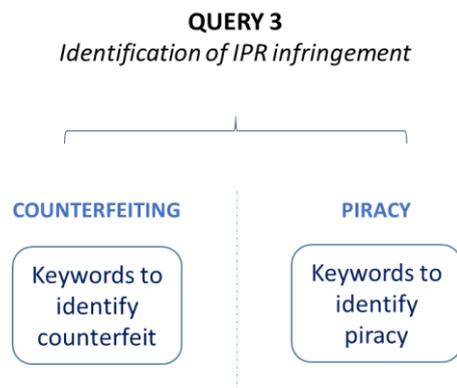
Figure 7 – Query 3 defined to identify IPR infringement

For example, the following combination of queries was set using the SIA tool to proceed
with the data mining for the clothing category. In the first query, the product categories
appeared. This first query was combined through 'AND' with Query 2 on the targeted
brands and the various ways they could be written. The third part further refined the

query by combining the previous elements with keywords helping to identify possible counterfeiting-related conversations. Finally, the last level highlighted by 'NOT' excluded conversations that may bring back 'noise' or unwanted elements that did not fall within the remit of this study.

((███ OR ("█████" OR ████ OR "#█") OR █████ OR ("████████" OR ███████) OR ("██████" OR ███████))) **Q2**

AND

(klänning* OR overall* OR tröj* OR t-tröj* OR underkläder* OR simbyxor* OR boxershorts OR sock* OR blus* OR tröj* OR koft* OR jack* OR täck* OR denim* OR huvtröj* OR collegetröj* OR byxor* OR jeans OR kortbyxor* OR kjol* OR bikini OR kostym* OR underkläder* OR kavaj* OR baddräkt* OR pikétröj* OR badkläder* OR jerseytröj* OR stickat OR byxor* OR kalsong* OR kvinn* OR sportkläder* OR barn*) **Q1**

AND ((oferta* OR descuento* OR autentic* OR promocion* OR venta* OR rebajas OR replica* OR stocklot OR "mejor trato" OR duplicad* OR liquidacion* OR "al por mayor" OR "mirror image" OR copia OR estilo OR barato* OR estafa OR cupon* OR fabrica* OR sobrepasa OR parecid* OR "exceso de stock" OR offres OR "bon marché" OR "pas cher" OR arnaque OR replique OR escompte OR liquidation OR copie OR usine OR dépassements OR authentique OR "surplus d'inventaire" OR ventes OR imbattable OR ressemble OR battre OR "à vendre" OR erbjudande* OR billig* OR plagiat* OR replik* OR kupong* OR utlopp* OR rabatt* OR marknadsföring* OR kopi* OR fabrik* OR lagerparti* OR spegelbild* OR kampanj* OR försäljning* OR grossist* OR överskrida* OR äkta OR utförsäljning* OR "till försäljning" OR "bästa erbjudande*" OR lurad OR kopi* OR överskottslag* OR Angebot OR günstig OR abzocken OR Räumungsverkauf OR Kopie OR Rabatt OR Aktion OR Kopie OR Fabrik OR Spiegelbild OR authentisch OR Verkauf OR Großhandel OR Überschuss OR "gleich aussehen" OR Ausverkauf OR "zu verkaufen" OR "billige Kopie" OR doppelt OR "private Nachricht" OR "direkte Nachricht" OR "bestes Geschäft" OR Überbestand OR okazj* OR tanio OR ofert* OR obniżk* OR autentyczn* OR kupon* OR oryginał* OR replik* OR podobn* OR wyprzedaż* OR likwidacja OR "na sprzedaż" OR duplikat* OR redukcja OR fabryk* OR "wyczerpanie zapasów" OR offerta OR economico OR frode OR repliche OR replica OR sconto OR promozion* OR copie OR fabbrica OR magazzin* OR "immagine speculare" OR "immagini speculari" OR vendit* OR ingross* OR eccedenz* OR simile OR simili OR autentica OR autentico OR liquidazion* OR saldi OR oferte OR sconti OR "immagini riflesse" OR "immagini a specchio" OR "scorte invendute" OR autentiche OR autentici OR truffa OR fabbriche OR "eccesso di scorte" OR "immagine riflessa" OR "eccesso di stock" OR sovrapproduzione OR offers OR discount* OR promo OR promotion OR sales OR ofertas OR descuento* OR promocion* OR rebajas OR descompte* OR promoció* OR rebaixes OR forsale OR € OR SEK OR Kr OR zł OR (% NEAR/1 off) OR DM OR PM OR cheap OR copy OR mirror OR ripoff OR rip-off OR coupon OR factory OR overrun* OR alike OR knockoff OR "knock off" OR "excess stock" OR stocklot OR AAA OR clearance OR dupe OR deal OR wholesale))) **Q3**

NOT (coronavirus OR covid* OR peru OR chile OR donat* OR apple OR nigeria OR $ OR £ OR RM OR ₹)

Figure 8 – Combination of queries to identify public conversations on possible counterfeiting related to the clothing category
The first block of queries covers brands (Query 2)
The second block of queries covers the category (Query 1)
The third block of queries covers IPR infringement (Query 3)

During the proof-of-concept phase, filtering for the most popular digital content created issues with the collection of the data corpus as its size was reduced to such an extent that its validity or relevancy could not be guaranteed. Therefore, it was decided that no

additional filtering through identified products would be done for these content types, leaving the final combination of queries as illustrated in the following schematic.

Figure 9 – Queries combination to filter and identify IPR infringement

During the data-collection phase, adjustments to queries were required to ensure the data corpus would be as relevant as possible. In total, three rounds of query adjustments were conducted.

The first adjustment was made after the presence of some non-relevant conversations were noted that had been collected due to some terms having a shared meaning. At this stage, the queries were configured by brand and category but included all languages in the same query, which led to the creation of some data 'noise'.

The second adjustment was the designing of queries by category and language. To do so, the technical team separated the queries to avoid the terms that had different meanings in other languages. For example, 'hat' is a form of headgear in English but means 'has' (third person singular of the verb 'to have') in German.

Finally, the third adjustment was implemented in the last week of the data-collection period, after the first version of the topic modelling was made. Certain patterns were observed and, based on these findings, the queries needed readjusting to incorporate some new keywords identified during the topic modelling.

3.3   Step 3 – Quantitative analysis

The quantitative analysis was twofold. It took into consideration the main social listening key performance indicators (KPIs), and a topic modelling process was also applied.

Main Social Listening KPIs

Regarding the main social listening KPIs, the quantitative analysis considered four categories of KPIs: total mentions, interactions, impressions and the ratio of interactions to mentions. These are the standard KPIs used to measure social media actions and logically they are part of the quantitative analysis of this study.

1.   The total mentions (volume of conversations) related to the targeted topics of the study are a simple and direct way to measure the number of mentions on a given issue.

2.   The interactions that encompass any kind of engagement such as likes, comments, shares, retweets, etc. are used to quantify and measure the reactions to a conversation and/or topic.

3.   The impressions give an estimation of the number of possible individuals exposed to the conversation.

4.   The ratio of interactions by mentions is an indicator that measures activity across the mainstream social networks by topic, revealing how much an audience responds (likes, retweets, shares, etc.) to a given topical message. It determines the frequency of social 'clicks' related to a given topic.

Topic Modelling

Topic modelling is a technique that groups the whole corpus of conversations into topics with the aim of making a representative selection of comments that are relevant for this research. It relies on text analytics using Natural Language Processing (NLP). NLP is

the technology used to aid computers in their understanding of human natural language and is especially useful when processing a high amount of information. NLP is widely used in our daily life, including tools such as spellchecker, autocomplete, machine translation and voice-to-text messaging. NLP requires the extensive usage of various tools that are highly dependent on language, such as tokenisation (correctly spotting words in sentences), part-of-speech tagging, and lemmatisation (recognising the dictionary form of words). NLP allows the identification of meaningless words and those that are irrelevant for the use case and diminishes their importance.

The first step in this study was to create a strategy to convert strings to numbers and for this, semantic similarity was used. Semantic similarity is a measure of the degree to which parts of text carry the same meaning. This is useful for obtaining good coverage of the numerous ways that a thought can be expressed in multiple languages. It enables analytics to address sentences having a common meaning. Therefore, semantic similarity enables the most important part of understanding a multilingual corpus of text.
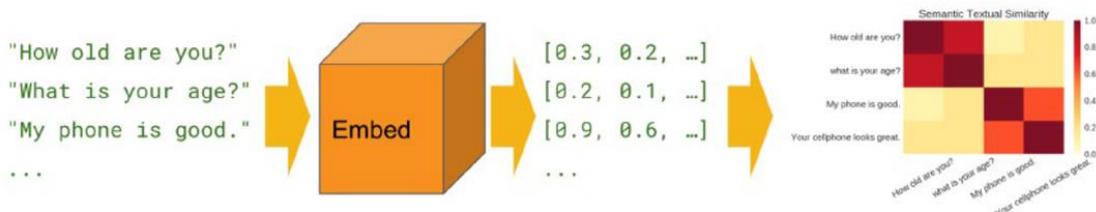
Figure 10 – Semantic similarity embedding

Using the semantic similarity approach, the topic modelling phase could be executed. Topic modelling aims to automatically spot the main subjects of discussion in a group of texts. The machine analyses word usage in each text and then tries to determine to which topics (it can be one or several) each text can be assigned. First, the machine cuts text down into sequences of words, carries out a language-dependent 'cleaning' of the words and, for each word in each text, examines the surrounding words. It will then form the topics depending on the most frequent words and most frequent word associations it has observed. Finally, by analysing the words contained in a text, the machine is able to determine the topic to which it belongs. In other words, topic modelling is a process that will classify posts or texts into different groups with similar features without introducing possible bias caused by human subjectivity. Following this identification of topics,

qualitative analysts refined the identified topics and searched for possible business models.

Once all the comments identified as possible cases of IPR infringement were categorised, the topic modelling process was able to be run. As the study covers two different aspects (actions related to a) counterfeiting, and b) piracy), two different topic modelling processes were run.

In both cases, the topic modelling addressed the whole collection period (1 April to 30 September 2020) and, as the NLP algorithms are most developed and reliable in the English language, all collected conversations were translated into English to ease the process.

3.4   Step 4 – Qualitative analysis

The last phase of the study corresponded to the qualitative analysis. This final step consisted of an in-depth analysis of the conversations in order to explore the comments in more detail on the different topics that had been identified in the topic modelling process. In this part of the study, the qualitative analysis endeavoured to identify nuances within the conclusions already offered by the quantitative analysis.

The SIA was organised in a sequential process, which aimed to deepen the understanding of the learnings from the quantitative stage. Qualitative analysts worked through four steps, enriching the understanding of the topic a little more at each stage. Concretely, the process unfolded as follows:

1.   **setting the scene**: analysts commenced the active reading of the quantitative results, gaining knowledge of the topic, and started to identify key aspects that needed to be explored, studied and analysed;

2.   **crafting**: they analysed all the qualitative inputs from the different social networks, this analysis aiming to identify all the key aspects that would offer the answer to the research questions concerned;

3. **uniting**: having detected all the aspects that were considered to accurately contribute to a complete understanding of the situation, they clustered the different topics to build the fundamentals of the report;

4. **refining**: based on the initial structure that emerged from the uniting stage, they drew logical inferences, confirming patterns that led, step by step, to a logical interpretation of the information.

During the qualitative phase, representative subsamples of comments based on the topic modelling were used in their original languages for English and Spanish (the two most used languages) to ensure the qualitative analysis was not biased by translation. However, conversations in other languages were translated into English before being analysed.

In total, 50 comments for each topic identified by the topic modelling approach were analysed. In the case of IPR infringement for physical products, this led to 1 750 analysed conversations, although, eventually many more were analysed because when qualitative analysts found a pattern, they investigated further into other conversations using the SIA tool to confirm their interpretation. In the case of digital piracy, the qualitative analysis included over 950 public conversations and used the same process as for counterfeit goods.

4 Relative Presence of Conversations Related to Counterfeit Products and Pirated Content on Social Media Compared to Genuine Products or Licit Digital Content

In the first instance, the implementation of quantitative analysis enabled the extraction of conversations concerning possible counterfeiting or piracy topics from the rest of the data corpus. While there cannot be absolute certainty that the data collected always related to counterfeiting or piracy (as conversations addressing these illegal actions are not always explicit), there is a strong indication that these conversations were captured, due to the careful selection of query terms.

4.1 Presence of IPR infringement with regard to physical products

Based on the extracted data corresponding to the eight product categories specifically selected for this study, the following table presents an overview of the total conversations (mentions) gathered by category, and indicates the relative number and percentage of conversations related to possible counterfeiting.

The volume of mentions collected reached 3 913 125 conversations with 11 % of mentions referring to possible counterfeit goods.

	Total mentions Product category and Brand	Total mentions Product category and Brand and Possible counterfeit	% Possible cases of counterfeit
Clothing	165 553	59 751	36 %
Footwear	471 315	98 487	21 %
Headgear	6 333	711	11 %
Jewellery	77 461	15 149	20 %
Perfume and Cosmetics	1 167 572	88 309	8 %
Pharma	40 722	3 837	9 %
Toys	1 868 140	138 398	7 %
Watch	116 029	14 477	12 %
Total	**3 913 125**	**419 119 ([17])**	**11 %**

Table 1 – Overview of the collected conversations according to the product categories

These results also showcase the difference in volume of possible counterfeit-related content between the target categories. It can be seen that the clothing category showed the highest incidence and was followed by the footwear and jewellery categories.

When reviewing these indicators for each category of product, it was observed that the ratio of interactions to mentions was particularly high when it concerned perfume and cosmetics, and toys. This suggests that the original comments related to these categories created a high interest among users who in turn reacted to them with a high number of retweets, likes or comments. By contrast, mentions in the clothing and pharma categories generated a smaller ratio of interactions.

([17]) In the table, the sum of the total counterfeit mentions is calculated on the basis of adding the total number of mentions obtained for each product category. This system of counting the total number of counterfeit mentions may, however, count a small number of mentions more than once if they fit into two or more product categories.

	Mentions	Interactions	Impressions	Interactions/ Mentions
Clothing	59 751	605 153	1 243 835 023	10.13
Footwear	98 487	7 073 575	5 706 067 366	71.82
Headgear	711	37 785	439 589 167	53.14
Jewellery	15 149	1 203 402	3 272 237 112	79.44
Perfume and Cosmetics	88 309	31 496 117	1 372 849 406	356.66
Pharma	3 837	75 441	3 069 832 035	19.66
Toys	138 398	15 899 783	29 096 206 171	114.88
Watch	14 477	715 983	3 492 983 691	49.46
Total	**419 119**	**57 107 239**	**47 693 599 971**	**136.26**

Table 2 – Overall results of social media listening KPIs applied to the mentions presenting possible counterfeit by category

Regarding the breakdown of the possible counterfeiting mentions per social media channel (see chart below), Instagram had 67 % of total mentions, followed by Twitter, whereas Facebook had the lowest incidence (please note that private Facebook groups were excluded and as explained in the introduction, access to data from these private groups would likely increase the proportion of mentions found on Facebook).
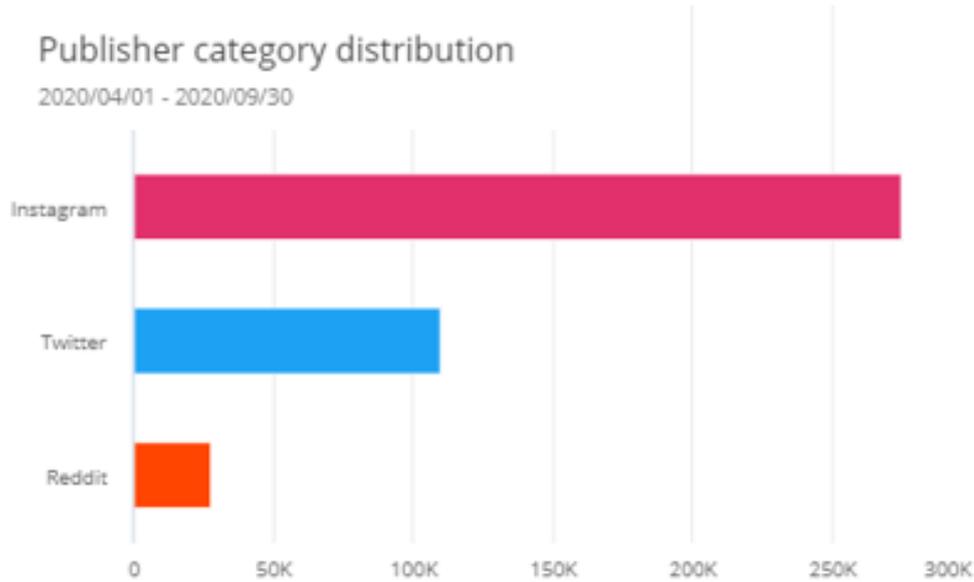
Chart 1 – Distribution of conversations addressing possible counterfeit topic among social media platforms (18)

The study also aimed to analyse and understand the differences and similarities between the different languages used regarding counterfeit. To do so, the main social media listening KPIs were observed.

---

(18) Considering the low amounts of counterfeit-related public conversations collected on Facebook, this data was not included in the visual representation to avoid any misinterpretation regarding the other channels.

In the following table, the mentions of counterfeit were gathered according to the language used, the volume of mentions, and the number of interactions and impressions.

	Mentions	Interactions	Impressions	Interactions/ Mentions
Counterfeit goods – English	355 142	53 554 766	43 743 147 358	150.80
Counterfeit goods – Spanish	39 954	1 613 753	1 827 085 408	40.39
Counterfeit goods – Italian	5 785	244 018	350 516 354	42.18
Counterfeit goods – German	5 256	836 802	577 685 469	159.21
Counterfeit goods – French	5 063	578 865	297 644 685	114.33
Counterfeit goods – Polish	2 052	81 986	185 708 862	39.95
Counterfeit goods – Swedish	274	16 293	4 163 164	59.46
TOTAL	**413 526**	**56 926 483**	**46 985 951 300**	**137.66**

Table 3 – Overall results of social media listening KPIs applied to the mentions presenting possible counterfeit by language

As shown in the table above, conversations in English were predominant and represented 86 % of the total mentions, 94 % of the interactions, and 93 % of the impressions.

Conversations in Spanish were second, representing 10 % of the volume of mentions. Comments in German had the highest interaction rate (number of interactions per mention), followed by those made in English and French.

In contrast, a low volume of mentions in Polish and Swedish was collected, despite the extension of the data-collection period from 13 weeks to 6 months.

4.2 Presence of piracy with regard to digital content

The process of identifying piracy was complicated because both licit and pirated content are frequently offered free of charge to the user, with the platform generating its revenue through advertising or other methods. Therefore, it can be difficult for users to differentiate between licit and illicit content online.

Another difficulty was the method of determining the denominator for the calculation of the incidence of infringement. The high volume of conversations relating to music, e-books, films and other types of content can easily lead to confusion within the collected data. For this reason, it was decided to use content related to **active** online behaviours and not to the digital content itself, for instance, 'watch a film' (and not 'film'), 'play a video game' (and not 'video game'), or 'download music' (and not 'music').

The study attempts to identify piracy in six specific digital content types. In the following table, an overview of the data collected regarding these content types is presented. The total volume of conversations related to each content type is indicated as well as the relative number and percentage of conversations falling under illicit actions.

	Total mentions Content category	Total mentions Content category and possible piracy	% Possible digital piracy
E-books	223 373	136 443	61 %
Sports events	198 221	64 414	32 %
Video games	1 091 561	198 559	18 %
Films	1 621 983	621 141	38 %
Music	1 121 719	478 841	43 %
TV shows	61 255	29 265	48 %
Total	**4 318 112**	**1 528 663**	**35 %**

Table 4 – Overview of the collected conversations according to the content type categories ([19])

The category with the highest volume of possible infringements is 'Films', followed by 'Music'. Meanwhile, 'TV shows' and 'Music' are the two categories that showed the highest percentage of possible infringement.

The distribution of comments relating to digital piracy, presented in the table below, highlights that the category of e-books had the highest interaction rate (interaction per mention), followed by TV shows, sports events and video games. This phenomenon indicates that users of social networks have a high level of interaction or reactions towards each of the different mentions (likes, retweets, comments). It was also noticed that, although the film category had a high volume of impressions, it generated a smaller ratio of interactions per mention compared to other categories.

([19]) In the table, the sum of the total piracy mentions is calculated on the basis of adding the total number of mentions obtained for each piracy category. This system of counting the total number of piracy mentions may, however, count some mentions more than once if they fit into two or more piracy categories.

	Mentions	Interactions	Impressions	Interactions/ Mentions
E-books	136 443	6 912 555	38 427 917 536	50.66
Events	64 414	1 957 032	41 508 797 633	30.38
Games	198 559	5 961 938	148 646 974 230	30.03
Films	621 141	5 839 271	784 066 432 428	9.40
Music	478 841	11 833 958	140 806 121 099	24.71
TV shows	29 265	1 021 800	45 127 851 284	34.91
TOTAL	**1 528 663**	**33 526 554**	**1 198 584 094 210**	**21.93**

Table 5 – Overall results of social media listening KPIs applied to the mentions presenting possible piracy by category

As regards the breakdown per social media platform, Reddit accounted for 49 % of the total number of mentions, followed by Twitter with 44 %. This distribution was also different from the one observed previously for counterfeit goods, where Reddit had been less important.
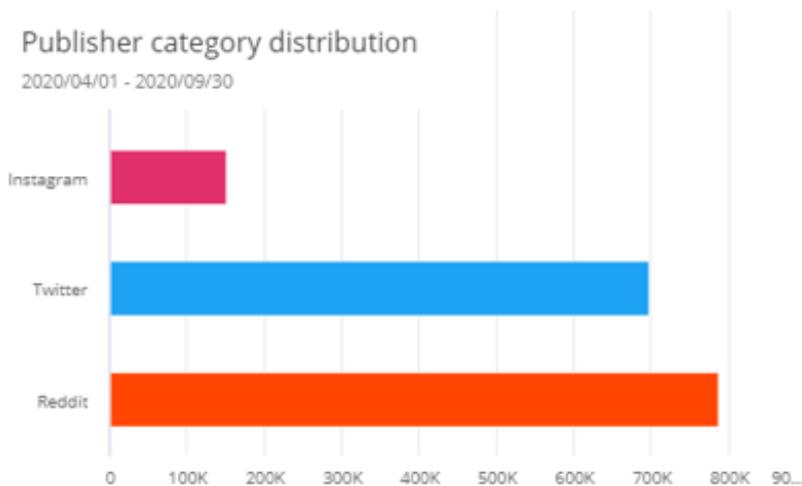
Chart 2 – Distribution of conversations addressing piracy topics among the social media platforms [20]

Regarding social media listening KPIs according to each language, the following results were obtained.

	Mentions	**Interactions**	**Impressions**	**Interactions/ Mentions**
Piracy – English	1 278 485	21 217 262	1 097 444 338 392	16.60
Piracy – Spanish	46 136	1 510 765	1 938 440 729	32.75
Piracy – French	13 310	470 374	1 314 210 807	35.34
Piracy – German	6 503	163 109	2 265 506 791	25.08
Piracy – Italian	5 242	313 635	927 305 158	59.83
Piracy – Polish	690	42 311	790 044 909	61.32
Piracy – Swedish	239	5 911	29 008 105	24.73
TOTAL	**1 350 605**	**23 723 367**	**1 104 708 854 891**	**17.56**

Table 6 – Overall results of social media listening KPIs applied to the mentions presenting possible piracy by language

[20] Considering the low amounts of counterfeit-related public conversations collected on Facebook, this data was not included in the visual representation to avoid any misinterpretation regarding the other channels.

English accounted for most of the conversations recorded with 94 % of the total volume of mentions, as well as the highest number of interactions (89 %) and impressions (99 %). Spanish was the second language in the ranking, accounting for 3 % of the total volume of mentions. These high disparities with the other languages are for two main reasons: English and Spanish are among the three main languages used on the internet [21], and most conversations on Reddit were in English, which created a bias in the overall results when comparing results for other languages across all four platforms.

Comments in French and Italian had the highest interaction rates (number of interactions per mention). The volume of conversations in Polish and Swedish was not considered significant enough to draw meaningful conclusions.

4.3   Conclusions

With regard to the presence of IPR infringement regarding physical products, the analysis highlights the following conclusions.

- More than 419 000 mentions on the social networks analysed in this study could be related to counterfeiting, corresponding to 11 % of all conversations surveyed. The objective of buying a counterfeit product is often to buy a cheaper product, and that is why the queries focused on, inter alia, the price aspect, in an attempt to identify all those conversations in which social network users were offered a cheaper product [22].

- The categories that contained a higher volume of possible references to counterfeits are those that had a higher volume of conversations; toys, footwear, and perfume and cosmetics had the highest volumes of possible IPR infringement conversations.

---

[21] Most common languages used on the internet as of January 2020, Statista.

[22] The main difficulty remained in identifying when a product is being sold under the stated price as a licit offer versus when it is an illicit business. This differentiation was also done through queries, where the search included looking for certain patterns such as the inclusion of a price, a telephone number, or additional words referring to copied products.

- In relative terms, the highest IPR infringement incidences are observed in the clothing, footwear and jewellery categories (respectively 36 %, 21 % and 20 % of mentions).

Regarding the presence of illicit digital content, the quantitative analysis enabled this report to draw the following conclusions.

- 1.5 million mentions of possible copyright infringement were identified from 1 April to 30 September 2020. These mentions generated 33.5 million interactions, an average of 21.9 interactions per mention, which is lower than the interaction rate observed in IPR infringements for the counterfeit products.

- Films, music and video games had the most mentions in absolute terms. In relative terms, e-books, TV shows and sports events featured most prominently.

5 Social Media Use for IPR Infringement Activities or Promotion and Possible Trends

Social networks typically have policies [23] on IPR infringement and some release public transparency reports on a regular basis. Facebook declared in its transparency report at the end of the second quarter of 2020 that 828 789 content pieces had been removed due to counterfeiting following a notification, and for the same period, Instagram removed 480 045 contents. From July to December 2019, Twitter removed 133 920 tweets for the same reason [24], whereas Reddit declared that, in 2019, 10 965 pieces of content had been removed for breaching its content policy for controlled goods and a further 124 257 pieces of content had been removed for copyright infringement [25].

This study was conducted from April to September 2020, a period for which the transparency reports of the four social media networks had not yet been released at the time of going to print. However, having reference numbers from each of the social media channels would help contextualise the results of the study and may help to highlight that possible new counterfeit-related conversations will keep appearing.

This chapter therefore offers an analysis of how each social network is misused to promote counterfeit products and pirated content. It should be noted once more that the results obtained for Facebook were very low and therefore should be considered carefully. As previously stated, availability of data from Facebook private groups would possibly have increased the number of suspicious cases. In fact, the last Facebook transparency report highlighted some infringing conversations, many of which were removed proactively by Facebook. In light of this, the results make it difficult to conclude that there are fewer IPR infringements happening on this channel.

---

[23] Twitter copyright policy.

Instagram and Facebook – counterfeiting fighting actions.

Reddit: content policy.

[24] Twitter transparency report 2019.

[25] Reddit transparency report 2019.

5.1   Analysis of social media use for counterfeit goods

In order to analyse the use of the social media channels, the IPR infringement-related conversations of the data corpus were broken down by category of product and social media channel in the tables below. The breakdown of the collected conversations highlights some social media preference per product category.

	Facebook [26]	Instagram	Twitter	Reddit	TOTAL
Clothing	28	19 851	39 172	699	**59 750**
Footwear	126	82 042	13 883	2 436	**98 487**
Headgear	1	300	124	286	**711**
Jewellery	166	13 478	917	588	**15 149**
Perfume and Cosmetics	59	84 538	3 182	529	**88 308**
Pharma	22	185	1 438	2 192	**3 837**
Toys	511	68 837	49 731	19 319	**138 398**
Watch	129	10 513	1 502	2 333	**14 477**
Total	**1 042**	**279 744**	**109 949**	**28 382**	**419 117**

Table 7 – Overview of the conversations collected according to the product type categories and the social media channel where the mention was published (in absolute numbers)

As illustrated in the table above, Instagram showed the highest number of counterfeit-related conversations across all categories of product, with the exceptions of the clothing category, which was the highest on Twitter, and the pharma category, which was highest on Reddit.

More specifically, as shown in the table below, Instagram was the social network that had the highest number of conversations of possible counterfeit goods (67 %), in particular footwear, jewellery, and perfume and cosmetics.

Twitter was the second highest in terms of the volume of conversations of possible counterfeit goods (26 %), and clothing, pharma and toys were the categories with highest volume of mentions.

---

[26] Private Facebook groups are excluded.

Reddit accounted for 7 % of the total conversations, but had especially high results for the pharma and headgear categories, with 57 % and 40 % of the total conversations, respectively.

Facebook had a very low volume. However, as mentioned above, this may have been because private groups on Facebook could not be included (or be due to proactive removal of posts discussing infringing activities), suggesting that conversations related to possible IPR infringement primarily happen in private groups, not accessible to social listening tools. For reference, a UK IPO report suggested that over five times more counterfeit goods were sold in invite-only groups on Facebook and Twitter than in open groups [27].

	Facebook [28]	Instagram	Twitter	Reddit	TOTAL
Clothing	<1 %	33 %	66 %	1 %	100 %
Footwear	<1 %	83 %	14 %	2 %	100 %
Headgear	<1 %	42 %	17 %	40 %	100 %
Jewellery	1 %	89 %	6 %	4 %	100 %
Perfume and Cosmetics	<1 %	96 %	4 %	1 %	100 %
Pharma	1 %	5 %	37 %	57 %	100 %
Toys	<1 %	50 %	36 %	14 %	100 %
Watch	1 %	73 %	10 %	16 %	100 %
TOTAL	**<1 %**	**67 %**	**26 %**	**7 %**	**100 %**

Table 8 – Overview of the breakdown of IPR infringement conversations by product category

Nevertheless, the analysis by social media channel showed that most of the IPR-related conversations on Facebook related to the product categories toys (49 %), jewellery (16 %), and footwear and watch (12 %).

---

[27] https://www.clarionsolicitors.com/articles/intellectual-property-infringement-on-social-media-a-growing-problem.

[28] Private Facebook groups are excluded.

IPR infringing conversations on perfume and cosmetics (30 %), footwear (29 %) and toys (25 %) constitute most conversations on Instagram.

On Twitter, the highest percentage of conversations related to possible counterfeits was about toys (45 %) and clothing (36 %); and the most discussed topic on Reddit was toys, with 68 % of conversations on counterfeits referring to counterfeit toys.

	Facebook ([29])	Instagram	Twitter	Reddit
Clothing	3 %	7 %	36 %	2 %
Footwear	12 %	29 %	13 %	9 %
Headgear	<1 %	<1 %	<1 %	1 %
Jewellery	16 %	5 %	1 %	2 %
Perfume and Cosmetics	6 %	30 %	3 %	2 %
Pharma	2 %	<1 %	1 %	8 %
Toys	49 %	25 %	45 %	68 %
Watch	12 %	4 %	1 %	8 %
Total	100 %	100 %	100 %	100 %

Table 9 – Overview in % of the breakdown of IPR infringement conversations by social media channel

5.1.1   Impact of the COVID-19 pandemic on conversations related to counterfeits

When examining the conversational trends in these different categories, the above results were compared with the results obtained over a longer duration (i.e. 3 years: from July 2017 to July 2020). This enabled the observation of fluctuations and trends in the volumes of conversations. It revealed that the context of the pandemic only affected a few of these categories when compared with the data on conversations from previous years.

- Pharma conversations suspected of referring to counterfeit medicines peaked twice, during the lockdown period (Spring 2020) and, to a lesser extent, around the

---

([29]) Private Facebook groups are excluded.

time people were allowed to go out again (Summer 2020, depending on the country).
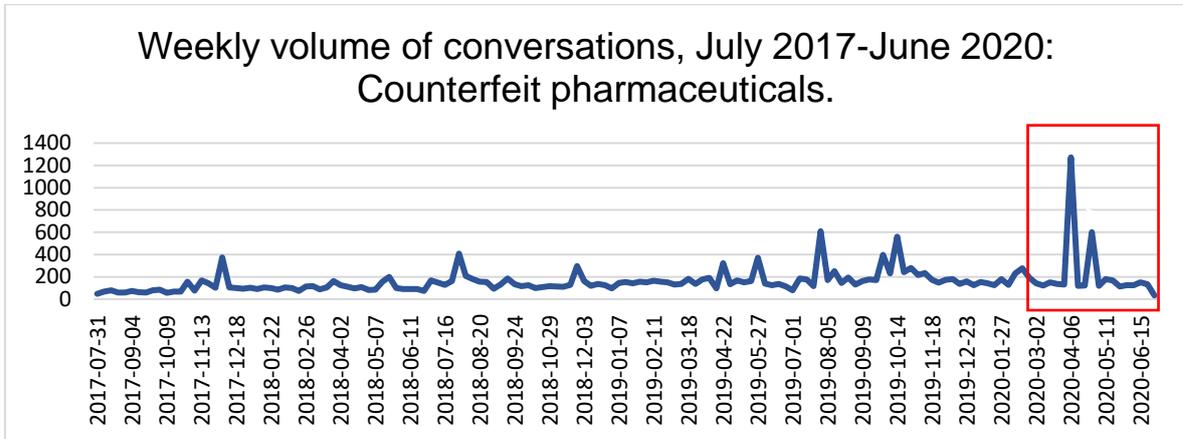

Chart 3 – Volume of conversations related to IPR infringing conversation in pharma from July 2017 to June 2020

- Conversations about clothing increased at the end of the lockdown period in the summer.
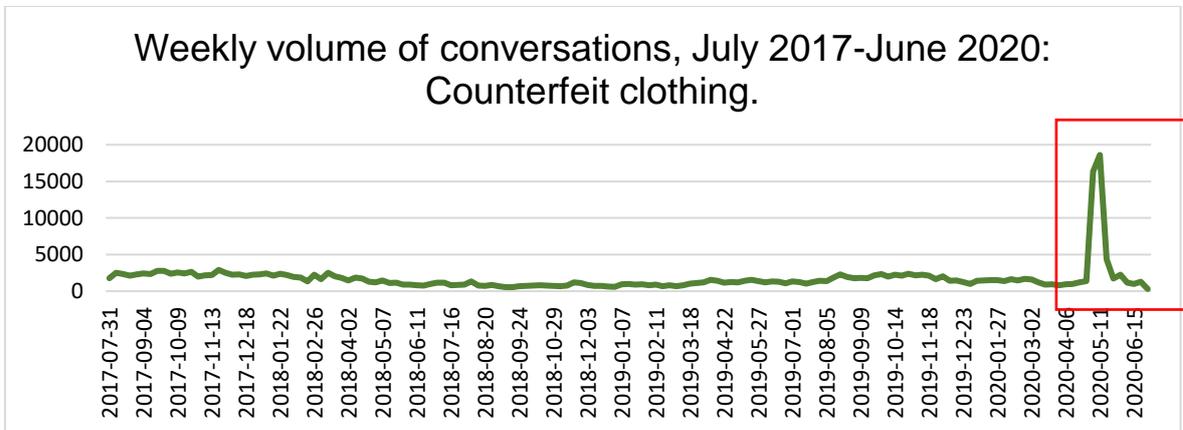

Chart 4 – Volume of conversations related to IPR infringing conversation in Clothing from July 2017 to June 2020

- In contrast, it can be seen that in 2020, the headgear category did not experience the winter and summer peaks of 2019 and previous years.
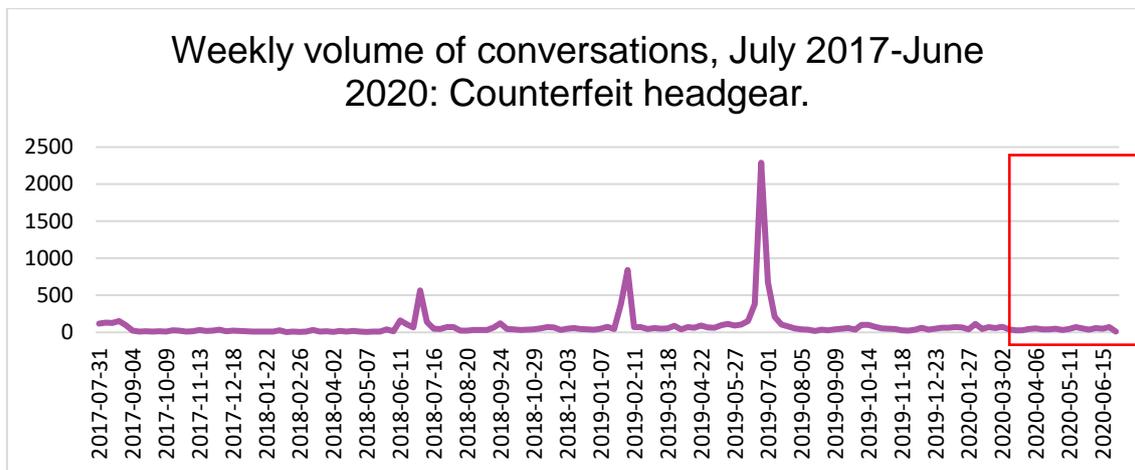

Chart 5 – Volume of conversations related to IPR infringing conversation in Headgear from July 2017 to June 2020

As far as can be interpreted in the conversation trend analysis, no other category showed anomalies during the pandemic.

5.1.2 Language trends for conversations related to counterfeit

As mentioned above, the English language dominated most of the conversations (86 %), meaning the analysis of the total number of conversations was significantly influenced by conversations in English. However, some differences between languages were observed.

	English	Spanish	French	German	Italian	Polish	Swedish	TOTAL
Total mentions	359 635	40 963	5 223	5 211	5 752	2 056	277	419 117
Share of mentions	86 %	10 %	1 %	1 %	1 %	<1 %	<1 %	100 %

Table 10 – Overview of the collected conversations by language in which the mention was published

When considering the use of different languages in relation to the categories, it can be seen that certain languages were used more in some categories than others. More

concretely, English and Spanish featured heavily in the most discussed categories (perfume and cosmetics, footwear and toys), which is logical as they are the most frequently used languages on social networks.

All languages generated high volumes of conversations about toys (in the top 3 in all analysed languages), while a second category could be added according to the language:

- French, Italian, and Polish – perfume and cosmetics;
- German, Polish, and Swedish – footwear.

	English	Spanish	French	German	Italian	Polish	Swedish
Clothing	15 %	11 %	18 %	8 %	16 %	10 %	12 %
Perfume and Cosmetics	21 %	21 %	38 %	15 %	27 %	40 %	9 %
Footwear	23 %	29 %	5 %	23 %	14 %	29 %	17 %
Pharma	1 %	1 %	4 %	2 %	1 %	<1 %	3 %
Headgear	<1 %	<1 %	<1 %	<1 %	<1 %	<1 %	<1 %
Toys	33 %	32 %	24 %	46 %	31 %	16 %	42 %
Jewellery	4 %	2 %	5 %	2 %	4 %	2 %	3 %
Watch	3 %	3 %	7 %	3 %	5 %	3 %	14 %
TOTAL	100 %	100 %	100 %	100 %	100 %	100 %	100 %

Table 11 – Overview in % of the breakdown of counterfeit goods conversations by language

5.2   Analysis of social media use for pirated digital content

The conversations related to pirated digital content – represented in this study by six content types (TV shows, films, music, sports events, e-books and video games) – were observed on the four target social media platforms. An analysis of the distribution of the conversations is presented in the following tables.

As previously shown, even though films and music showed the highest volume of conversations related to IPR infringement, the content types showing the highest percentage of potential digital piracy were e-books (61 %) and TV shows (48 %).

	Facebook [30]	Instagram	Twitter	Reddit	TOTAL
E-books	1 172	33 509	78 318	23 444	**136 443**
Sports events	407	3 883	32 554	27 570	**64 414**
Video games	571	28 484	71 920	97 584	**198 559**
Films	510	16 589	50 805	553 237	**621 141**
Music	838	59 253	325 362	93 388	**478 841**
TV shows	3	54	488	28 720	**29 265**
Total	**3 501**	**141 772**	**559 447**	**823 943**	**1 528 663**

**Table 12 – Overview of the collected conversations according to the content type categories by
social media channel where the mention was published (in absolute numbers)**

Regarding the overall use of social media channels in terms of the volume of digital
piracy-related conversations, it was observed that Reddit was the most used channel,
followed by Twitter.

More specifically, most conversation on e-books, sports events and music occurred on
Twitter, whereas Reddit generated the highest percentage of conversations in the video
game, film and TV show categories. Facebook and Instagram showed a less significant
number of conversations as the overall quantity of digital piracy-related conversations
for these two channels was under 10 % of the total conversations, as illustrated in the
table below. Instagram was the second most popular channel regarding e-books, when
all the other categories of digital products featured more predominantly on Twitter and
Reddit.

[30] Private Facebook groups are excluded.

	Facebook ([31])	Instagram	Twitter	Reddit	TOTAL
E-books	1 %	25 %	57 %	17 %	100 %
Sports event	1 %	6 %	51 %	43 %	100 %
Video games	<1 %	14 %	36 %	49 %	100 %
Films	<1 %	3 %	8 %	89 %	100 %
Music	<1 %	12 %	68 %	20 %	100 %
TV shows	<1 %	<1 %	2 %	98 %	100 %
TOTAL	**<1 %**	**9 %**	**37 %**	**54 %**	**100 %**

Table 13 – Overview of the breakdown of IPR infringement conversation per product category

5.2.1 Impact of the COVID-19 pandemic on conversations related to piracy

Piracy-related conversations were analysed in the context of the pandemic and were found to increase immediately lockdowns were implemented, reaching a peak just before the summer. However, they did not spike like some of the physical products categories. Levels of piracy-related conversations returned to a more normal level in June 2020.
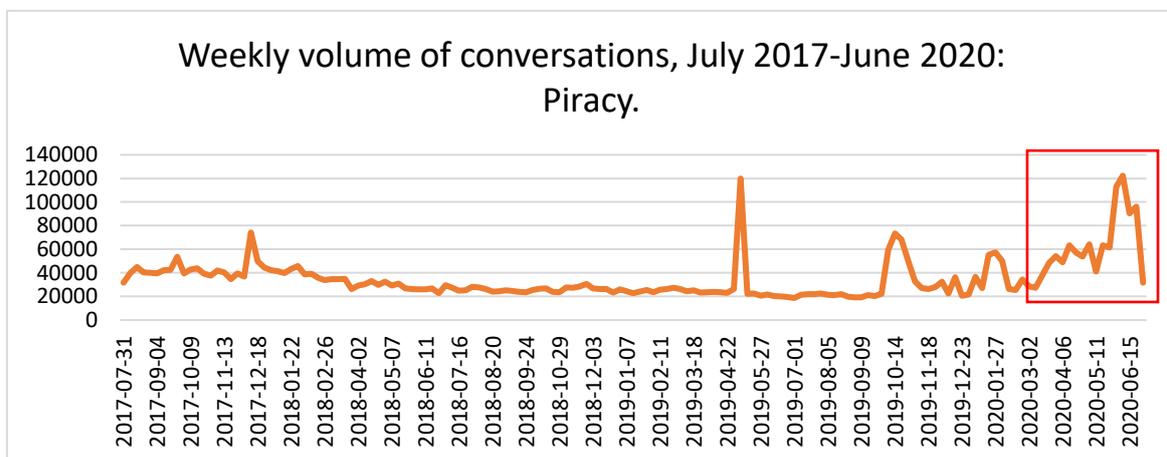

Chart 6 – Volume of conversations related to piracy from July 2017 to June 2020

([31]) Private Facebook groups are excluded.

5.2.2   Language trends in conversations related to piracy

The analysis by language shows that English had the highest volume of conversations
(94 %).

	English	Spanish	French	German	Italian	Polish	Swedish	TOTAL
Total mentions	1 437 503	59 774	17 001	7 693	5 720	713	259	1 528 663
Share of mentions	94 %	4 %	1 %	1 %	<1 %	<1 %	<1 %	100 %

**Table 14 – Overview of the conversations collected according to the content type categories
crossed with each social media channel where the mention was published (in absolute number
and %)**

When analysing results by language, the music category was the most discussed in all
languages except for English (where it was the second content type after films). Video
games was the second category in both Spanish and French and e-books had higher
relative importance in Italian, French and Swedish.

	English	Spanish	French	German	Italian	Polish	Swedish	TOTAL
E-books	9 %	4 %	11 %	9 %	14 %	7 %	12 %	**9 %**
Sport event	4 %	4 %	15 %	3 %	13 %	0 %	5 %	**4 %**
Video game	13 %	10 %	15 %	6 %	10 %	0 %	7 %	**13 %**
Movie	**43 %**	4 %	7 %	15 %	12 %	**93 %**	7 %	**41 %**
Music	**29 %**	**77 %**	**52 %**	**67 %**	**51 %**	0 %	**69 %**	**31 %**
TV shows	2 %	0 %	0 %	0 %	0 %	0 %	0 %	**2 %**
Total	100 %	100 %	100 %	100 %	100 %	100 %	100 %	**100 %**

**Table 15 – Overview in % of the breakdown of IPR infringement conversation per language
(% vertical)**

5.3   Conclusions

Just as social media platforms offer a lot of possibilities for legitimate market promotion,
they also offer diverse means of promoting counterfeit products and pirated content. As

demonstrated in this chapter, this abuse not only differs between each platform and category, but also between languages.

Counterfeiting of physical products

- The footwear, perfume and cosmetics and toy categories accounted for the highest volumes of conversations on Instagram. Clothing and toys showed the highest volumes of conversations with regard to the use of Twitter, and the toy category also showed the highest volume of conversations on Reddit.

- Overall, Instagram accounted for the highest share of conversations about potential counterfeits.

- Analysing the relative incidence of each platform within each category, it can be observed that Reddit had a special importance for the pharma and headgear categories. Instagram was the main platform for watches, toys, perfume and cosmetics, jewellery and footwear, while Twitter had a high relative presence for the toys, pharma and clothing categories.

Piracy of digital content

In the case of piracy, a different pattern can be observed.

- Reddit played a very important role, especially in conversations related to the film category for which it had the highest volume of conversations, while Twitter had a high volume of conversations related to music.

- In relative terms, Twitter had a high number of conversations about e-books and music.

- Reddit was the main social network for conversations on TV shows and for the film category.

- When considering conversations per language, English was again the most frequently used language across categories.

Impact of the COVID-19 pandemic on counterfeiting and piracy conversations

When analysing the possible effects of COVID-19 on the different categories of physical and digital content, increases in the volume of conversations about potential counterfeits could be noticed.

The pharma and clothing categories were the main categories showing an increase in conversations during the lockdown period of the pandemic.

Regarding piracy, a notable increase in conversations occurred during the lockdown, but this seemed to return to normal levels in the summer of 2020. However, given that the data used for this study ends in July 2020, it is difficult to get a full picture of the pattern of these conversations during the pandemic.

6 IPR Infringing Business Models Used on Social Media and Their Main Advantages Compared to Other Communication Channels

The approaches to the sale of counterfeit products and the offer of pirated content have some important differences. When dealing with physical products, there are two distinct scenarios, a) where buyers make their purchase knowingly, and b) where buyers are the victims of deception. In many cases in either scenario, the marketing approach relies on a push strategy where messages are intended to generate appeal among potential buyers. However, as regards piracy, this intent is a key component of the conversations encountered on social media; it seems that no 'marketing' strategy is needed as users proactively search for tips, recommendations and sources to download content, regardless of whether it is pirated or licit.

In this chapter, the different approaches are explained, highlighting how social networks contribute to the different business models, and identifying patterns that would help to detect IPR infringement situations. A practical approach is taken, building on the analysis in the previous chapters.

6.1 Business models for IPR infringement with regard to physical products

As has been established throughout this report, it is very difficult to find a pattern that can identify when a product is a counterfeit or not – regardless of whether the processing is done by a human or AI. There are, however, a series of clues that may indicate sales of counterfeits, and when various indicators combine, the likelihood of the product being a counterfeit increases.

In order to analyse the large number of comments on possible IPR infringement, a semantic analysis was conducted, using a topic modelling approach as described in

Chapter 3. NLP is an automatic process, used to identify groups of topics of conversations for marketing and research processes.

NLP identified seven big 'groups' of conversations in relation to counterfeiting. Six of these coincided with the categories that generated the greatest volume of conversations (including toys, perfume and cosmetics, footwear, clothing, watches and jewellery) – the seventh group being promotions. These seven groups could be broken down into 35 different subtopics.

The qualitative analysis, in which an expert reviewed the original comments to analyse the content in depth, provided more information on the process and corroborated the first indications that had been identified in the queries phase.

Focusing on business models used to market counterfeited goods through social networks

As identified in the study, 'Research on Online Business Models Infringing Intellectual Property Rights' [32], 'many of these business models are based on generally applicable business models, i.e. business models that can be used and are being used for commercial activities that are entirely legitimate.' This was corroborated in this study as various business models emerged, each having specific objectives in the context of marketing counterfeit goods. The following six business models were identified.

- **Business model A – Orders through direct messaging or a messaging app**
  Objective: rapid sale, quick result.
  The social network showcases the product, and the option is offered as available to order directly via a messaging app or direct message.

- **Business model B – Building a customer database for future direct sales**
  Objective: sustained growth.
  This model seeks to expand the base of users and followers and to establish loyalty with the potential customer. To achieve this objective, sweepstakes and contests are often organised in which, for an opportunity to win, users have to follow the

---

(32) EUIPO – Research on Online Business Models Infringing Intellectual Property Rights, June 2016.

account, post about it and share it with their contacts. This ultimately contributes to enhanced visibility and to increasing the number of potential clients.

- **Business model C – High promotions and discounts on product price**

  Objective: draw traffic, rapid sale.

  Discounts play a role in advertising to lure consumers and are usually in the 20 % to 30 % range but can be higher than 50 %. It is conceivable that the higher the discount, the more likely it is that the product offered for sale can be suspected of being a counterfeit.

- **Business model D – Pre-order**

  Objective: generate income with no cost.

  This is a sales process where the buyer reserves and orders a product that is not available yet. The reservation implies full or partial payment. A typical example detected in some senders' accounts on Instagram (which have been removed) seem to have rather prejudicial conditions, that is: pre-order plan with a delay of between 2 to 3 weeks and a 30 % advance payment. When applying this model, there is also a risk of scam – the buyer never receives the counterfeit item.

- **Business model E – Reinforced customer service through personalised advice through messaging**

  Objective: long-term relationship and income.

  The sales model can be based on different mechanisms, processes and procedures, but its key feature is its focus on customer service. The main objective is to create a trusted sales context and determine a more direct and personal relationship with the potential buyer. The added trust contributes to the commercial success of the illicit business, luring unaware consumers.

- **Business model F – Selling through tutorials**

  Objective: build credibility.

  This business model is especially relevant in the case of perfume and cosmetics. It consists of videos (YouTube) or direct messaging (Instagram) to disseminate make-up tutorials during which products are promoted.

These different business models are not necessarily indicators of IPR infringing activities, and specific indictors have to be considered to assess the probability of a possible infringement.

In the context of the application of these business models, two approaches were detected.

- Serving consumer's intentional need for a counterfeit good: throughout the entire process, from searching to the point of purchase, the consumer is aware of what they are doing. In this context, users take the initiative in the conversations, asking for tips and recommendations. Vendors rely on positive word-of-mouth communications to generate interest in their products; to that extent, various business models help them achieve this level of reputation – for example, effective customer service or low prices.

- Scamming consumers by presenting the products as licit ones to sell at a higher price: consumers become aware they have acquired a fake upon receiving it. Vendors post their offer to which potential buyers react. Many of these posts convey a sense of urgency regarding an offer that will shortly expire to ensure a quick sale, while reassuring consumers about the authenticity of the product – for example, mentioning 'original'; offering an exclusive pre-order; insisting on a quick contact via direct message or WhatsApp.

The following table summarises the attributes of a post that can indicate counterfeit goods.

Indications that may suggest IPR Infringement ([33])

	Clues to look for …
Post content	• Very direct, even rude vocabulary • Redundancy and repetition of messages • Poorly translated posts, mistakes hindering understanding, text with untranslated parts (especially slang and abbreviations) • Lack (even absence) of information about the product • WhatsApp number • Intensive use of emoticons and hashtags
Price and discounts	• Very low prices • Excessive discounts above 30 % or 40 % suggest a complementary motivation with respect to a purchase made in normal circumstances, and as this discount becomes larger, the advertising claim becomes a hook to precipitate the sale/purchase. Many tracked publications worked on this basis, offering products with discounts above 50 %. High discounts can be an indication of a possible counterfeit. • Offering payment methods that limit traceability – e.g. Bizum (Spain), cash or PayPal
Product's picture	• Visible image modification with superimposed elements, alterations, etc. • Overemphasis on a brand or logo • Poor picture quality although the vendor appears to be a professional
Redirection	• To marketplaces such as Taobao, AliExpress and eBay • To new and/or unknown online stores
Lack of personal information	• Difficulty to identify an actual person linked to the account (no name or surname or other contact information) • Reduced quantity of information provided regarding the vendor
Refund policy and delivery	• No exchange, return or refund policies • In-person delivery

---

([33]) Some of these indications work in conjunction. For example, the fact that there is a WhatsApp number in the text does not in itself point to a counterfeit but must be seen in combination with other indicators. Other aspects of a post can by themselves be an indication, for example, very high discount or poor picture quality although the vendor appears to be a professional.

A few examples of how these clues can coexist in a post.

Figure 11 – Publication from Instagram

Inconsistent branding and lack of personal information with multiple paying options

**Figure 12 – Publication from Twitter, with product picture, alluding to be an original product and
redirection to a sales web page**

Figure 13 – Publication from Instagram
High discount (60 %), multiple payment platforms, using WhatsApp

6.2   Business models for pirated digital content

A study carried out in Spain in 2019 found that approximately 60 % of consumers who access pirated content use search engines (Google, Bing, Yahoo, etc.) to find that content. Other preferred access points are via direct downloads on the internet and social networks, but these methods of access are much less popular than search engines, with 29 % and 23 % respectively ([34]).

Social networks are an access channel to pirated content. The average internet user is likely to use the most popular social media channels. These platforms therefore also function as a search engine and as a mediating channel between the user and digital content, whether legal or illegal.

([34]) GFK – Digital Piracy and Consumption Habit Observatory, 2019.

As indicated earlier in this report, in the context of conversations on social media, users often proactively ask for advice to gain access to pirated content, which demonstrates a clear act of intent on the user's side. In this case, two types of communication emerge.

- The provider of the pirated contents does not apply a specific model to attract users and the success of the provider's business merely depends on the positive word-of-mouth advertising they are able to generate; users become message emitters, working in the interest of those offering the pirated contents.

- Information regarding the availability of pirated contents is shared actively through a process akin to advertising of legitimate content, often with promises of free access to contents of interest to potential users.

The ways pirated content is marketed through social networks can be illustrated by the following indicators.

	Clues to look for …
Wording	Certain words appear repeatedly in posts that promote piracy: • download • watch • free • streaming • online • link.
Structure	• As with sales of counterfeit products, the posts can be confusing and difficult to understand. • Many posts try and give tips/recommendations, some are even written in the form of 'tutorials', often presenting steps (1, 2, 3, 4 …).
Links	• Links are fundamental as the piracy action usually does not take place on the social network itself. • They can lead to websites or cloud storage service where contents can be streamed or downloaded.
Free	• The use of the word 'Free' in many cases • When considering piracy of legitimate video streaming services, there is often an intention to sell an illegal access to the service. The process therefore becomes similar to what was identified for counterfeit products (direct messaging, appealing prices, etc.).

A few examples of how these clues can coexist in a post:

**Figure 14 – Twitter post extracted by the SIA tool with the step-by-step structure**
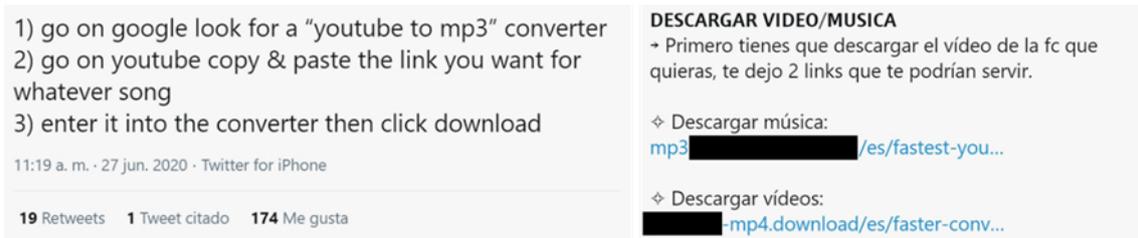
**Figure 15 – Twitter posts extracted by the SIA tool showing the use of 'mp3', step-by-step format,
use of the word 'descargar' (download) …**
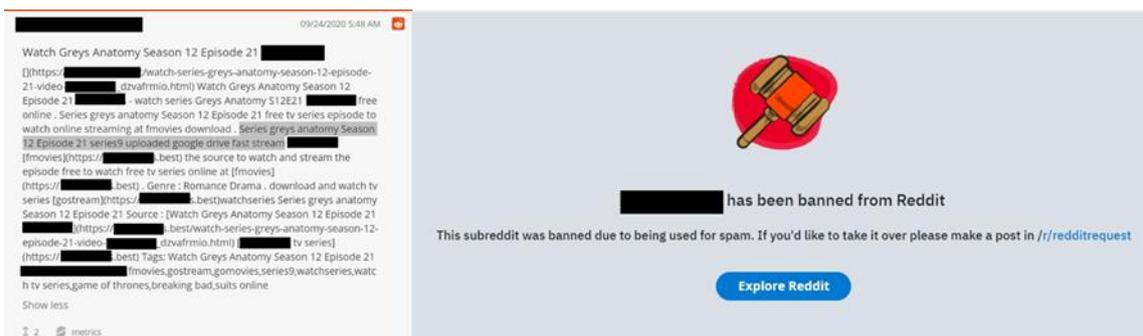
Figure 16 – Reddit post extracted by the SIA tool and deleted message from Reddit

6.3   Advantages of social media for IPR infringement

Both the sales of counterfeit products and the provision of pirated content can rely on social networks to achieve their respective objectives. These platforms unwittingly offer an environment conducive for such activities, as they offer:

- free access;
- massive and global 'market' in which to sell to or from which to get information and tips about how to access pirated content;
- high flexibility to test new approaches and to avoid detection, including the capacity to set up accounts and encourage transactions quickly before deleting the account;
- direct and fast communication with buyers and other users.

Just as many types of brands and businesses are using social networks to promote themselves through a wide range of business models, IPR infringing businesses have also invested in their activities on social media platforms as they grant them access to a large volume of potential consumers, and they provide an environment in which they can easily mirror legitimate brands' practices to better serve or lure buyers. The fact that they are present on these platforms contributes to the success of IPR infringing business operations.

7 General Conclusions

This study demonstrates the presence of IPR infringement-related conversations on social media, as shown by the examples collected from Facebook, Twitter, Instagram, and Reddit. The large amount of content identified as related to possible counterfeiting or piracy over the 6 month data-collection period shows that conversations concern both physical products from the categories of clothing, headgear, footwear, pharma, toys, perfume and cosmetics, jewellery, and watches; and also digital content in the categories selected for this study: films, music, e-books, video games, sports events and TV shows.

Some social media platforms are preferred by infringers due to their characteristics. According to the results, Instagram is the preferred channel for IPR infringement of physical products, which can be explained by the very visual nature of the platform, therefore, making it more appealing as a virtual showroom and more effective in promoting physical products. In contrast, conversations about digital content mainly take place on Twitter and Reddit, as they offer short and direct messages accompanied by links that seem to better fit IPR infringement-related conversations for these content types. It was also noted that public conversations on Facebook are used less than other social media networks to share IPR infringement-related conversations (as noted above, this could also be a result of successful proactive removal of such conversations). In the case of Facebook, it is important to reiterate that social listening platforms only have access to public profiles, and so the extent of IPR infringement activity in private conversations and groups was beyond the scope of this study. Drawing conclusions about Facebook's seemingly low numbers would therefore be unadvisable at this stage.

As well as highlighting which social media channels are preferred depending on the type of product to be advertised, the study also shows disparities within the product categories regarding IPR infringement-related conversations. Among the physical products, toys represent the most important cluster of conversations, while films represent the highest number of IPR infringement-related conversations for digital content.

The quantitative and qualitative approaches used in this study enabled the establishment of a list of cross-cutting patterns when it comes to business models of IPR infringement on social media. One of the detected trends is that sales of counterfeits are very often linked to the promotion of a product using common marketing strategies, whereas digital

product piracy more often occurs through the exchange of 'good tips' or 'opportunities' to access pirated content. This limits the ability to detect these IPR infringement situations, however, a 'bundle-of-clues' approach could be utilised to identify probable IPR infringements. These clues relate to the way posts are written, for example, in a very emotional tone or with the overuse of emoticons; their wording, with certain key words, such as 'discount' or 'opportunity' for physical products and 'free' or 'download link' for digital content; and their structure or lack of.

Brand owners sometimes claim that sellers of counterfeits are switching to social media as the internet sales platforms are improving enforcement [35]. Examining this interplay between internet sales platforms and social media in the context of IPR infringement is beyond the scope of this study, but it is certainly a relevant research topic for possible future studies.

---

[35] For example, the Share and share alike report, published by the UK Intellectual Property Office in 2017, highlights the role played by social media in marketing counterfeits.

References

Clarion (2020). *Intellectual Property Infringement on Social Media – A Growing Problem*.
Available at: https://www.clarionsolicitors.com/articles/intellectual-property-infringement-
on-social-media-a-growing-problem. [Accessed 1 March 2021].

European Union Intellectual Property Office (EUIPO) (2016). *Research on Online
Business Models Infringing Intellectual Property Rights – Phase 1*. Available at:
https://euipo.europa.eu/tunnel-web/secure/webdav/guest/document_library/observatory
/resources/Research_on_Online_Business_Models_IBM/Research_on_Online_Busine
ss_Models_IBM_ex_sum_en.pdf.

European Union Intellectual Property Office (EUIPO) (2019). *Intellectual Property and
Youth Scoreboard 2019*. Available at: https://euipo.europa.eu/ohimportal/en/web/
observatory/ip-youth-scoreboard.

European Union Intellectual Property Office (EUIPO) (2020). *2020 Status Report on IPR
Infringement*. Available at: https://euipo.europa.eu/ohimportal/en/web/observatory/
status-reports-on-ip-infringement.

Europol (2018). *Social media crime: 20 000 packages of counterfeit medicine, mobile
phones, jewellery, sunglasses and watches seized*. [Press Release]. Available at:
https://www.europol.europa.eu/newsroom/news/social-media-crime-20-000-packages-
of-counterfeit-medicine-mobile-phones-jewellery-sunglasses-and-watches-seized.

Europol (2020). ***Viral Marketing** Counterfeits, Substandard goods and Intellectual
Property Crime in the COVID-19 pandemic*. Available at: https://www.europol.europa.eu/
sites/default/files/documents/report_covid_19_-_viral_marketing_counterfeits.pdf.

Facebook for Business (2019). *Good Questions, Real Answers: Protecting Brands
Against Counterfeits*. Available at: https://www.facebook.com/business/news/good-
questions-real-answers-how-facebook-helps-brands-protect-against-counterfeits.
[Accessed 1 March 2021].

Facebook Transparency Report (2020). Available at: https://transparency.facebook.com/intellectual-property.

Growth from Knowledge (GFK) (2019). *Piracy Observatory and Digital Content Consumption Habits 2019*. Available at: lacoalicion.es/wp-content/uploads/executive-obs.piracy_en_2019.pdf.

Investopedia (2020). *Social Media Definition*. Available at: https://www.investopedia.com/terms/s/social-media.asp. [Accessed 1 March 2021].

Reddit (2019). *Transparency Report 2019*. Available at: https://www.redditinc.com/policies/transparency-report-2019.

Reddit Content Policy. Available at: https://www.redditinc.com/policies/content-policy. [Accessed 1 March 2021].

Statista (2021). *Most common languages used on the internet as of January 2020, by share of internet users*. Available at: https://www.statista.com/statistics/262946/share-of-the-most-common-languages-on-the-internet/. [Accessed 1 March 2021].

Stroppa, A., Gatto, D., Pasha, L., Parrella, B. (2019). *Instagram and counterfeiting in 2019: new features, old problems*. Available at: https://ghostdata.io/report/Instagram_Counterfeiting_GD.pdf.

Twitter Transparency Report (2019). Available at: https://transparency.twitter.com/en/reports/copyright-notices.html#2019-jul-dec.

Twitter Counterfeit Policy. Available at: https://help.twitter.com/en/rules-and-policies/counterfeit-goods-policy. [Accessed 1 March 2021].

UK Intellectual Property Office (2014). *IP Crime Highlight Report 2013/14*. Available at: https://www.gov.uk/government/publications/ip-crime-highlight-report-2013-to-2014.

UK Intellectual Property Office (2017). *Share and share alike*. Available at: https://www.gov.uk/government/publications/share-and-share-alike.

Glossary

AI	The simulation of human intelligence in machines that are programmed to perform tasks normally requiring human intelligence and mimic human actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving.
Copyright	A type of intellectual property that gives rights holders exclusive rights to authorise or prohibit the use (e.g. reproduction, distribution, adaptation, translation) of their content (e.g. films, programmes). Works covered by copyright range from books, music, paintings, sculpture and films to computer programs, databases, advertisements, maps and technical drawings.
Counterfeiting	Counterfeit goods, as used in this report, are physical products that infringe a trade mark, design rights or a patent.
Data corpus	A collection of linguistic data, either compiled as written texts or as a transcription of recorded speech. The main purpose of a corpus is to verify a hypothesis about language – for example, to determine how the usage of a particular sound, word, or syntactic construction varies.
Engagement	The shares, likes and comments received by a post on social media.
EUIPO	The European Union Intellectual Property Office (EUIPO) is responsible for managing EU trade marks and registered Community designs. It also hosts the European Observatory on Infringements of Intellectual Property Rights.
GDPR	The General Data Protection Regulation (Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC)

	is a regulation in EU law on data protection and privacy in the European Union and the European Economic Area. It also addresses the transfer of personal data outside these areas.
Golden file	The expected output of some tests (usually automated) stored as a separate file rather than as a string literal within the test code. When the test is executed, it reads the file data and compares it to the output produced by the system being tested.
Impressions	Metric used to quantify the total number of times **social media** users have been shown a piece of content, usually an advertisement, digital post, or a web page. Impressions are not action-based and are merely defined by a user potentially seeing the content.
Incidence	The occurrence, rate, or frequency of something.
Interactions	Communication or direct involvement between the audience and a brand, platform, individual or any such entity responsible for the creation and dissemination of content. Interactions may vary between platforms such as mentions or shares on Twitter or likes and comments on Facebook.
IP	Intellectual property
IPR	Intellectual property rights
IPR infringement	In this report: counterfeiting or piracy.
IPTV	Internet Protocol television
KPI	Key performance indicator
Keyword	A word that acts as the key within a (data) query; indicator of significance relative to the study objectives.
Lemmatisation	The process of grouping together the inflected forms of a word so they can be analysed as a single item, identified by the word's lemma or dictionary form.
Mentions	A measure of how many times a brand name has been mentioned on social media channels.
Natural language processing (NLP)	A subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between

	computers and human language, in particular the programming of computers to process and analyse large amounts of natural language data.
Piracy	The act of making copyright-protected content available to others without the permission of the owner of the copyright.
Query	A formula combining keywords enabling the filtering of public conversations on the targeted social media channels.
Semantic similarity	A metric defined over a set of documents or terms, where the idea of distance between items is based on the likeness of their meaning or semantic content as opposed to lexicographical similarity.
Social listening	The process of monitoring social media channels for mentions of a brand, competitors, product, etc.
Speech tagging	The process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context.
Synthesio	Social listening tool provided by Ipsos Group S.A.
Topic modelling	A machine learning technique that is capable of scanning a set of documents, detecting word and phrase patterns within them, and automatically clustering word groups and similar expressions that best characterise a set of documents.