Data Science for Influencer Marketing : feature processing and quantitative analysis

Anil Narassiguin, Selina Sargent

▶ To cite this version:

Data Science for Influencer Marketing : feature processing and quantitative analysis

Anil Narassiguin, Selina Sargent

Upfluence, anil.narassiguin@upfluence.com, www.upfluence.com

Influencing a target audience through social media content has become a new focus of interest for marketing leaders. While a large amount of heterogeneous data is produced by influencers on a daily basis, professionals in the influencer marketing (IM) field still rely on simple quantitative metrics such as community size or engagement rate to estimate the value of an influencer. As few research papers have proposed a framework to quantify the performance of an influencer by using quantitative influencer data (number of followers, engagement,...), qualitative information (age, country, city, etc...), natural text (profile and post descriptions) and visual information (images and videos), we decided to explore these variables and quantitatively evaluate their dependencies. By analyzing 713,824 influencers on 5 social media platforms over a period of one year, we identified relationships between value proposition data (engagement, reach, audience), demographics, natural text patterns and visual information. The main goal of this paper is to provide IM professionals with a modern methodology to better understand the value of their influencers and to feed machine learning algorithms for clustering, scoring or recommendation.

Key words: Influencer marketing, social media, data science, machine learning, feature processing, natural language processing, image processing, empirical study, methodology

1. Introduction

The last decade has witnessed the proliferation of new paradigms in high-tech fields thank to the emergence of social media platforms. While the big players in IT constantly improve their machine learning algorithms thanks to the huge amount of data generated every day (ref), marketing teams can now rely on social media influencers to better reach their customers. Surprisingly while the two domains have been emerging, almost simultaneously, in the decade of 2010, Influencer Marketing specialists still rely on simple KPIs to find suitable influencer communities for their campaigns.

IM is a sub-domain of marketing where marketing activities are oriented around an influential person on social media [Brown and Hayes (2008)]. The so-called influencers would promote industry products on their blog posts, Instagram feed, tweets, YouTube videos, etc... Before the information era, the concept of influence was already studied by researchers in sociology, marketing and political science [Katz and Lazarsfeld (1966)].

2

Author: *Article Short Title*
Article submitted to ***; manuscript no. (Please, provide the manuscript number!)

Since then, with the emergence of the information age and the social networks, many domain-specific studies have been made : case study of IM campaigns [Glucksman (2017)], relations between IM and brand success [De Veirman et al. (2017)], analysis on sponsoring in IM [Ewers (2017)], etc...

Most of these studies have been performed on small subsets of data (hundreds to thousands of influencers) for specific social media and mostly structured data (number of followers, engagement rates, ROIs, etc...).

Upfluence is an IM company providing services and products for IM professionals. We currently have 2.9 million influencers in our data base and ton better understand how influencers generate enagement, we performed a quantitative analysis on heterogeneous data sources (quantitative data, textual information, visual data). This analysis can also be seen as a methodology for data processing on IM data sets (regular statistics, natural language processing, image recognition, etc...).

2. Influencer's value proposition data

In this section, we will review easily processible value information such as numerical data (community size; engagement) and categorical data (geography; gender).

2.1. Community sizes

On the main social media (Youtube, Facebook, Instagram, Twitter, Pinterest), the potential value of influencers is traditionally estimated by the size of their community, the number of impressions per post (ie the number of times a post has been displayed to a user), the interactions between the fan base and the social media content (likes, comments, views) also called engagement and the return on investment.

The use of certain media has evolved throughout recent years. While Facebook, Twitter and YouTube have been strong references for some time, Instagram is now taking the lead in IM and some platforms such as Pinterest or Twitch are preferred by professionals for specific niche fields (DIY or video games). At Upfluence, the data we gather about different influencers allows us to analyze 572,664 Instagram profiles, 31,526 Facebook profiles, 70,388 Youtube profiles, 24,634 Pinterest profiles and 123,370 Twitter profiles (713,824 unique influencers, as an influencer might have several social networks).

As presented in table 1, the average size of community for each social network is highly biased due to extreme values. Most of our influencers have communities expanding thousands to several tens of thousands of followers, which is usually required by IM professionals for marketing

campaigns. Recently business-oriented categories were proposed to better classify influencers according to their community size [Zietek (2016)] :

- Micro-influencer : less than 15,000 followers

- Regular-influencer : between 15,000 and 50,000 followers

- Rising-influencer : between 50,000 and 100,000 followers

- Mid-influencer : between 100,000 and 500,000 followers

- Macro-influencer : between 500,000 and 1,000,000 followers

- Mega-influencer : more than 1,000,000 followers

Table 1 also shows the distribution of the influencers according to their categories. The business needs are oriented toward Regular and Rising influencers. Indeed large community influencers (Macro and Mega influencers) are usually widely known and therefore have their own connections with quality corporations. Except Pinterest, all the social networks present in the database have a significantly large amount of Regular to Mid influencers which allow us to provide reliable information about the tendencies in IM.

2.2. Engagement

A post's engagement reflects the overall activity and can be defined as the sum of interactions by users on a post. There are various kind of interactions which are dependent on the social platform (Table 2). The total sum can be weighed according to the importance given to each interaction. Indeed, a comment given by a follower on an influencer's content implies a priori more engagement than a simple *like*. Even so, in this study we'll suppose that the total engagement is just the simple sum of all the influencer's interactions.

The total engagement of an influencer is always relative to the size of her/his community. To estimate how an influencer performs relatively to others, the average engagement can be compared to the average time the influencer content has been seen (impression). The engagement rate (ER) is formally a quantity that indicates how much an influencer performs relatively to his visibility on social network [Semiz and Berger (2017)]. It can thus be defined as the following :

$$\text{engagement\_rate} = \frac{\#\text{engagements}}{\#\text{impressions}}$$

Even so, the number of impressions for a specific content is hard to estimate, except for the social media companies who have direct access to this data, we usually take the number of followers as an estimate to its potential visibility on social networks :

	Instagram	Facebook	Youtube	Pinterest	Twitter
Count	572,664	31,526	70,388	24,634	123,370
Mean	90,106	244,013	304,447	29,694	93,275
Std	897,271	2,215,912	1,126,288	199,252	1,193,945
Min	1,005	1,000	1,000	1,000	1,000
25%	14,282	2,868	9,031	1,826	2,002
50%	23,427	8,017	43,212	3,841	4,553
75%	51,958	37,611	195,425	11,512	15,200
Max	280,643,232	160,061,728	87,100,920	12,662,073	107,070,300

	Instagram	Facebook	Youtube	Pinterest	Twitter
Micro	174,120	19,628	22,461	19,559	92,271
Regular	266,073	4,978	13,634	3,069	16,552
Rising	63,757	1,862	7,921	842	5,325
Mid	55,473	3,109	15,510	947	6,292
Macro	6,843	734	4,195	127	1,204
Mega	6,398	1,215	4,623	90	1,726

Table 1 **Community sizes quartiles per social networks in Upfluence's data extract. Below : Number of influencers per categories**

	Community denomination	Possible interactions
Instagram	Followers	Likes, comments
Facebook	Fans	Reactions, comments, shares
Youtube	Followers	Likes, dislikes, comments
Twitter	Followers	Retweets, favorites, replies
Blog	Readers	Comments, social media shares
Pinterest	Followers	Pins, saves, comments, likes

Table 2 **Relations between an influencer and his community**

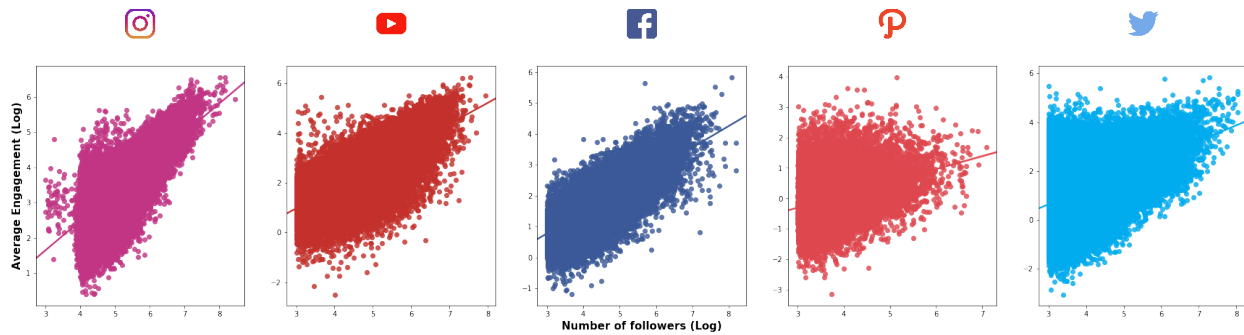$$\text{engagement\_rate} = \frac{\#\text{engagements}}{\#\text{followers}}$$

Figure 1   Evolution of the average engagement vs the number of followers per social media in logarithmic scale

The absolute engagement value has a tendency to disadvantage large community influencers. Indeed, as seen in figure 1, there's a linear evolution between the log of engagement and followers.

Even if this phenomenon is well known by IM specialists, the engagement rate is still used as a reference metric to compare and select potential influencers for marketing campaigns. Practically it may not be bad practice since usually influencers are considered within the same cluster of community size categories (micro, regular, etc...), but this property should orientate the redesign of a new metric to somehow better score influencer performance on specific social media.

Engagement and the engagement rate are actually following log-normal distributions as many human-related behavior metrics (i.e. money distribution across citizens, city population sizes, length of comments on social media, etc...) [Sobkowicz et al. (2013)]. We verify it empirically by plotting the Q-Q plots of the log of engagement rates per social media (Figure 11).

This normality allows for the use of a 2/3-sigma rule in order to identify *usual* values for ER as shown in table 2.2. Depending on the social medium, ER can achieve different ranges of values. Instagram tends to have high engaging influencers while Pinterest tends to have many high community size influencers with low engagement. The later observation has to be deeply studied to compare the potential reach of an influencer on different social media.

Even though, extreme values for ER don't correspond to specific profiles, we found that by exploring qualitatively some outliers (the 1% of extreme values), we realised that low ER profiles on YouTube and Facebook usually correspond to brands or news pages (well known organisations, many followers but few real reactions...) while high ER can either be fake profiles or real
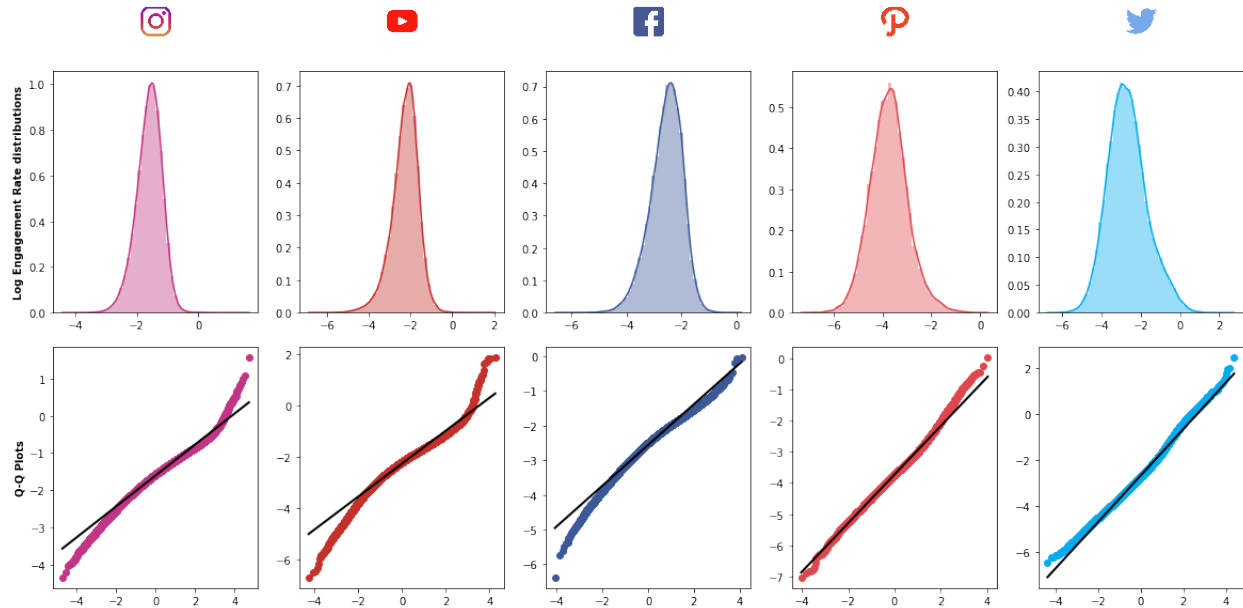
Figure 2   Evolution of the average engagement vs the number of followers per social media in logarithmic scale

	Usual range of values (in %, to be checked with quantile formula)	Low ER outliers	High ER outliers
Instagram	3.62×10^{-1} - 17.14	Niche communities Caption pictures posts	High quality profiles Scam content Fake engagement Small buzzing influencers
Facebook	1.82×10^{-2} - 4.23	News pages Brands pages	Micro influencers
Youtube	2.67×10^{-2} - 10.73	Content aggregation websites Education Media	Buzzing influencers
Pinterest	5.13×10^{-4} - 6.82×10^{-1}	-	Popular fields influencers (fitness, DIY, etc...)
Twitter	2.03×10^{-5} - 2.32×10^{-1}	Automatic profiles (reposts, ...)	Political views Activists Buzzing profiles

Table 3   ER usual ranges

quality niche influencers. This observations should be considered carefully since we didn't perform any statistical analysis on the categories of outliers' profiles but we hope to do it in a near future.

In the rest of the paper, we'll bring to light some interesting correlations between ER and influencer quantitative, qualitative, temporal, textual and visual features. Since the aggregated results are done on large samples of posts and a general selection of influencers, the high differences in engagement rates are usually statistically significant.

This statistical analysis will be an occasion for presenting different data science techniques to extract features from heterogeneous data sources produced by influencers on social media. Indeed, many deep learning techniques allow IM experts to extract meaningful semantics and visual patterns from data that is still ignored by many in the field.
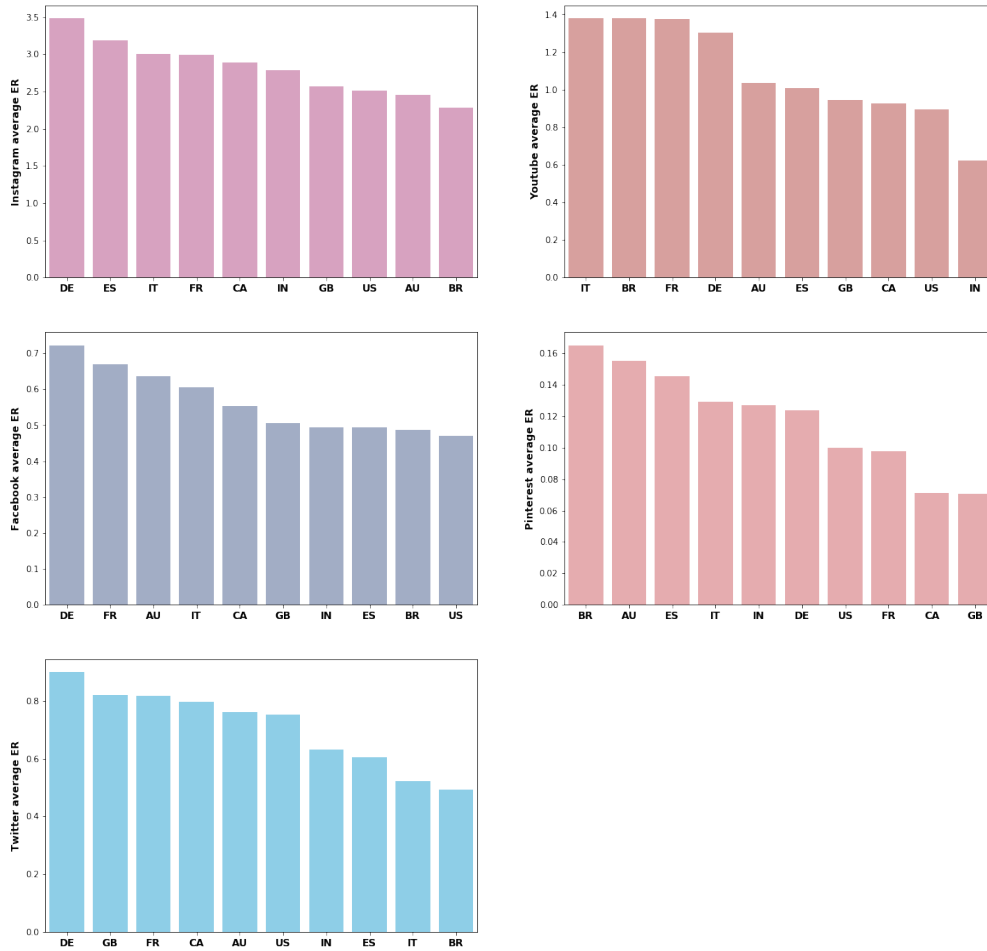
Figure 3   Average engagement rate in percentage per countries and social media

2.3.   Influencer's value proposition data

In this section, we'll see the correlation between the influencers' personal demographic data such as age, sex, place of residence and language spoken, and the influencers' engagement rate.

2.3.1.   Geographical and language features In Figure 2.3 and 4, these graphs display the rate of engagement with correlation to country and language. It's worth noting that community size will have, statistically a significant impact on ER due to country size.

2.3.2.   Gender In Figure 5, for these graphs, we analyzed data collected using image recognition API on social profile pictures. This approach represents an innovative alternative to manual data, however it is supported by gender identification entered manually by the user. This is enough to get a sufficiently accurate result and allows us to categorise gender identification into two classes.
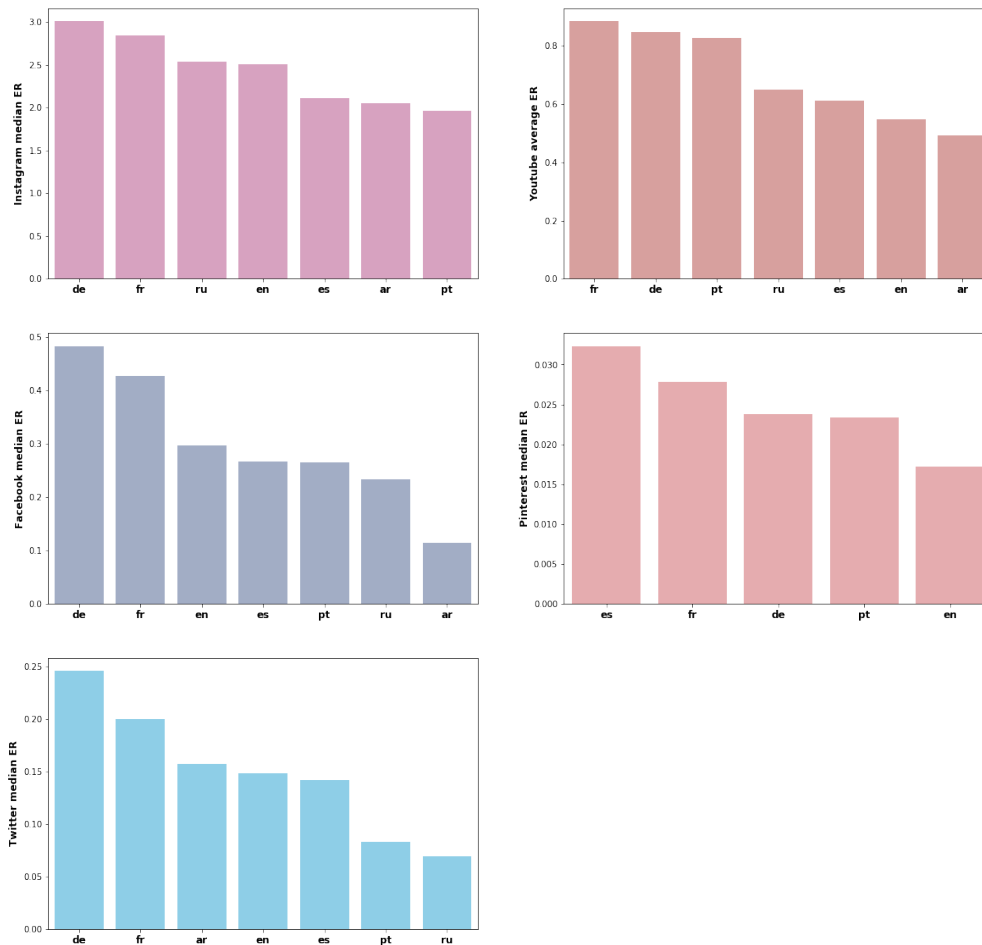
8

Author: *Article Short Title*
Article submitted to ***; manuscript no. (Please, provide the manuscript number!)

Figure 4   Median engagement rate in percentage per languages and social media

The nature of the social platform may indicate a gender preference. However, the overall gender split remains similar across all platforms expect Pinterest, where female users are more prevalent (1 man for every 10 women).

2.4.   Temporal behaviors

Figure 6 shows that there are trends in our data that suggest the influencer's activity likely correlates to the most active times for their audience across all platforms. Similarities in activity due to social platform algorithms can not be ruled out.

Evidence of ER vs Time show no significant difference, other than on Youtube. It can be reasonably assumed these differences are due to the nature of and use of the platform.

The observed differences can be explained in part by the varying requirements in content creation for Youtube in comparison to the other platforms and in how it is consumed.
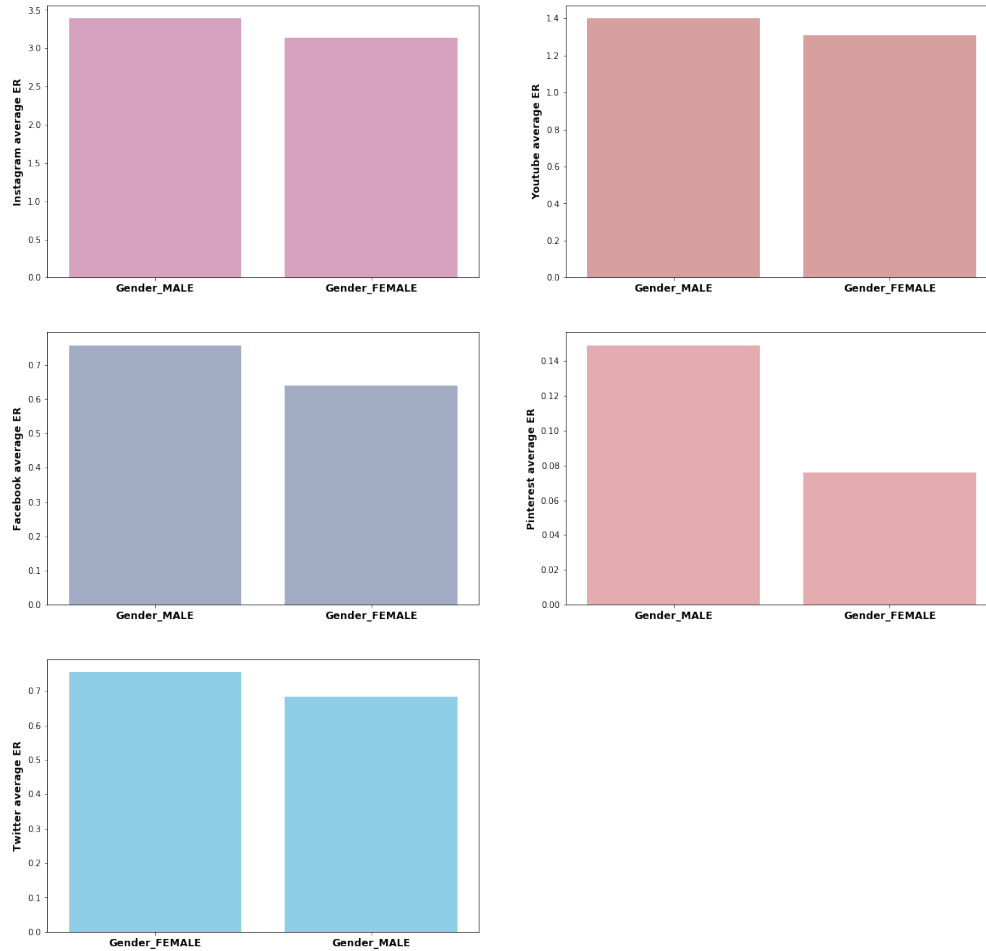
Figure 5   **Median engagement rate in percentage per gender and social media**

The most prevelant use of textual and visual information is available on Instagram, which, outweighs all other platforms when it comes to Influencer preference. Thus, the remaining sections will focus on data collected from Instagram.

3. Textual information

Textual information has been for decades the first kind of semi structured data available online. Natural language comes with inevitable perceived context which by its nature makes it a difficult medium to infer intent when mined computationally. Recently by using the last neural technologies (word embeddings) for NLP it has been possible to capture strong semantic meaning. Before the emergence of these technologies, tags for the Web and then hashtags for social media, have been a good way for users to specify the content of their posts.
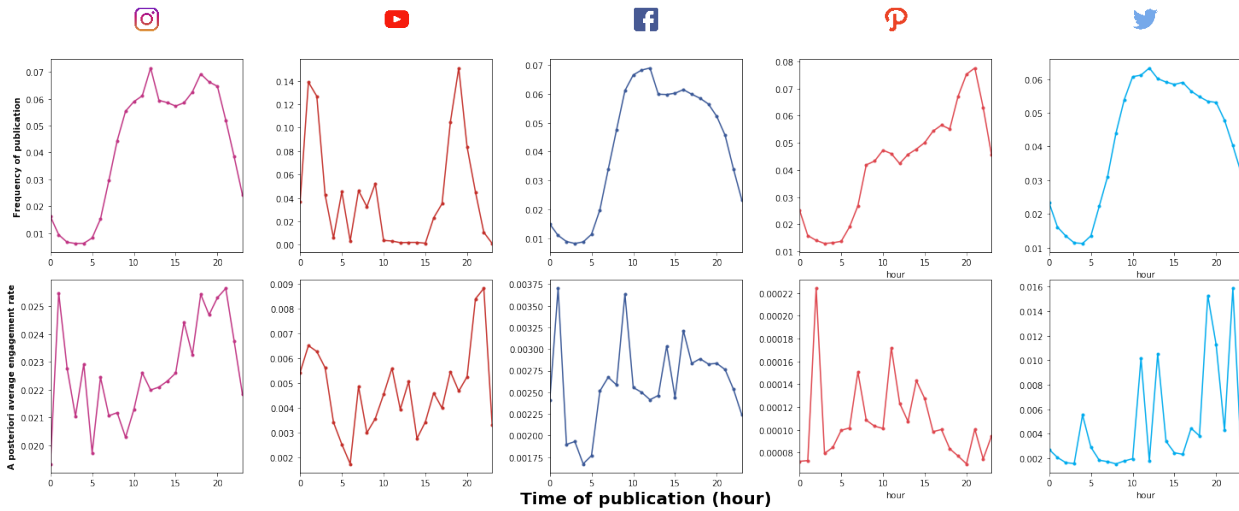
10

Author: *Article Short Title*
Article submitted to ***; manuscript no. (Please, provide the manuscript number!)

Figure 6 Activity and Engagement rate vs Time of the day

3.1. Hashtag vs ER

Primarily used to identify a keyword; topic of interest or to facilitate search, however, increasingly used as a blanket approach in an attempt to increase reach, and can be perceived as a greedy tactic for attention. Therefore posts with zero or demonstrably fewer than excessive use, tend to have higher ER (Figure 7).

In figure 8, we employed a data collected method that included the identification of a predetermined list of specific sponsored hashtags in five languages (French, English, Portuguese, Spanish, German). The correlation showed that ER on posts containing sponsored hashtags was higher in contrast to non-sponsored posts.

Anecdotal (#natgeo; #earthpix) references, as well as mainstream (#ig_italy; #greece) references to travel have significant ER performance. On the other hand, generic beauty (#hairstylist; #nailart) and lifestyle (#decor; #chic) related hashtags have a low ER.

4. Image processing

Since the beginning of the 2010s, image processing techniques have improved consequently thanks to deep learning methods helped by the performance of powerful GPU clusters and the massive amount of unstructured data generated on the Internet. Thanks to image recognition it's possible to detect objects in pictures or video with comparatively high accuracy.

Using a conventional neural network for object detection based on the *YOLO* method [Redmon et al. (2016)] and trained on the Microsoft *COCO* data set [Lin et al. (2014)] it has been possible to automatically tag tens of thousands of images to estimate the average ER for usual object categories (person, dog, cake, etc...).
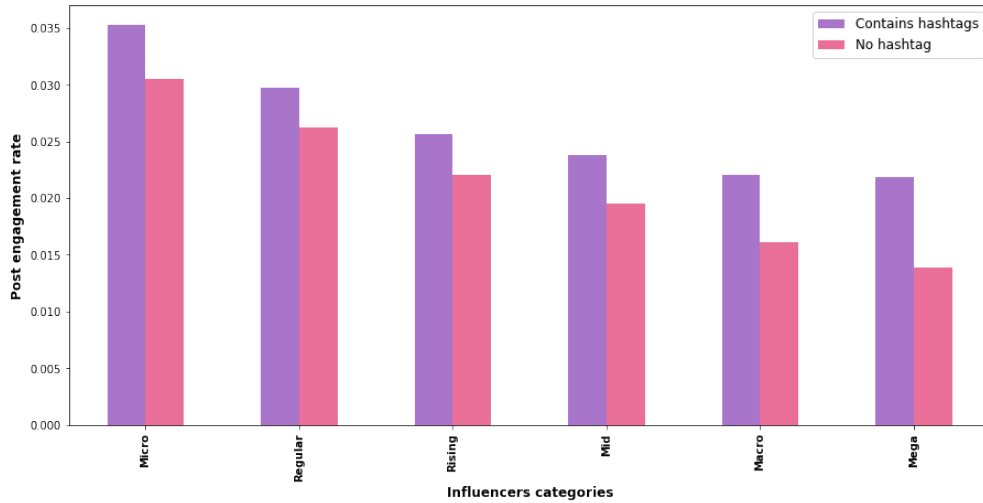
Figure 7 Average engagement rate for posts containing hashtags vs posts containing no hashtag
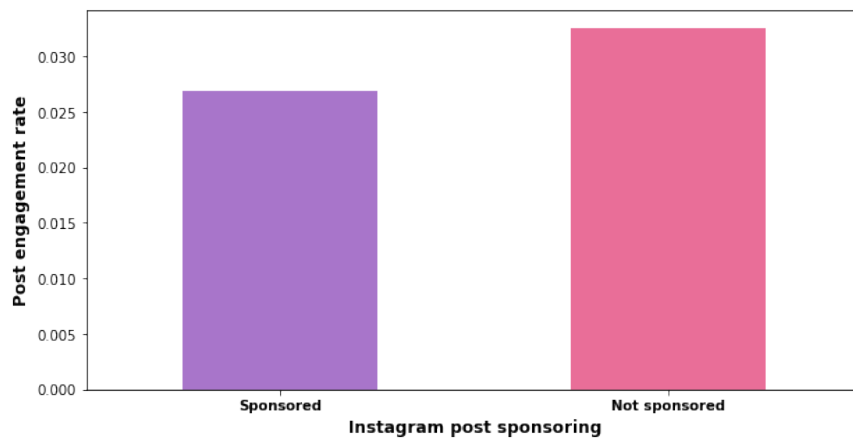

Figure 8 Average engagement rate : sponsored posts vs non-sponsored posts

Our data revealed that profiles with cats, dogs or people had significantly higher ER.

The data isn't significant enough to give a complete conclusion, however, it appears that food related images have lower engagement rates.
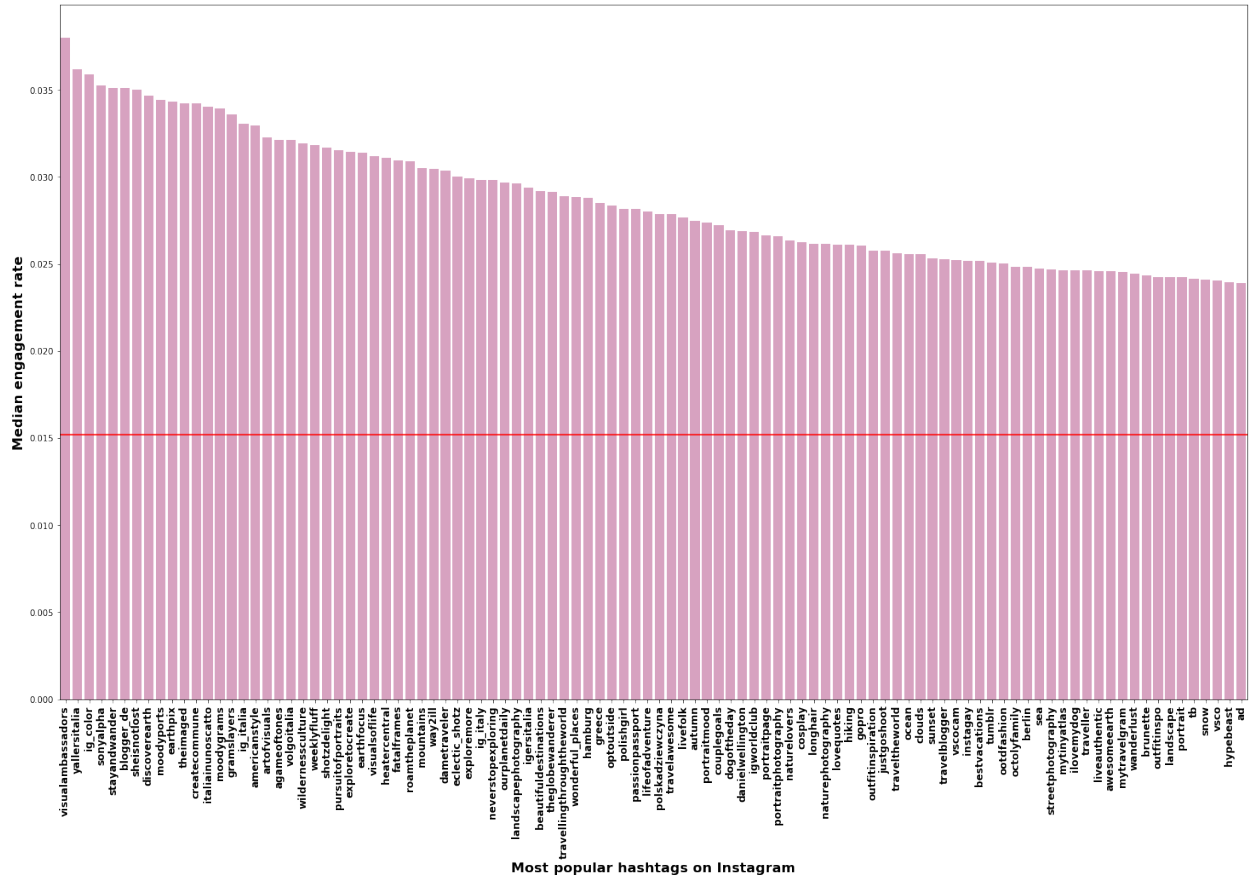
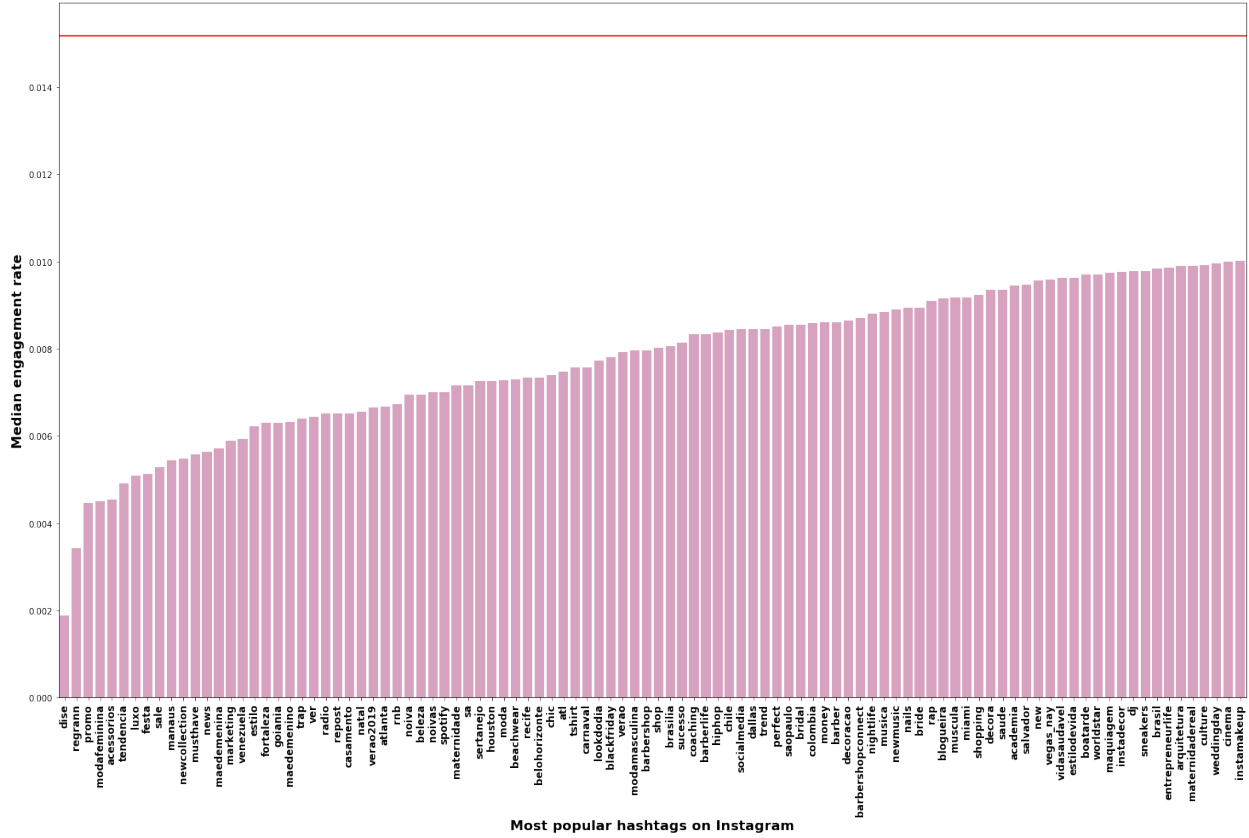**Figure 9** Best performing hashtags in terms of ER (red line = overall median ER)

Figure 10 Worst performing hashtags in terms of ER (red line = overall median ER)
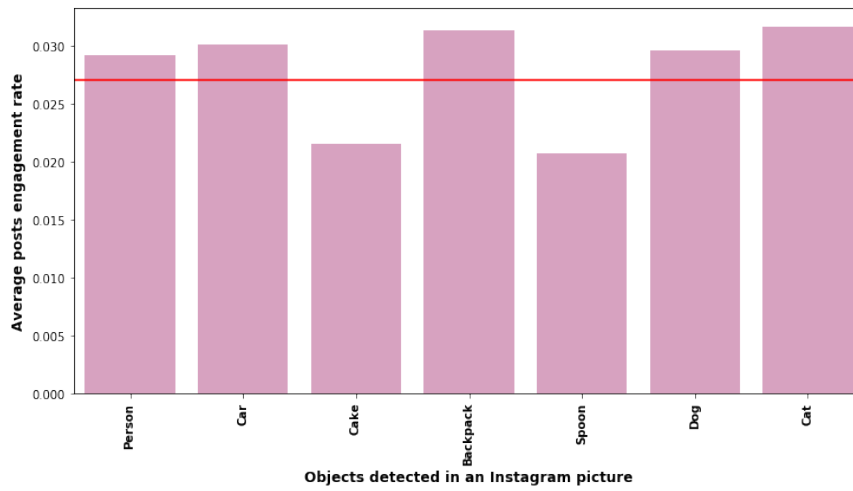

Figure 11 Average engagement rate per tags detected in images (red line = overall average ER)

14

Author: *Article Short Title*
Article submitted to ***; manuscript no. (Please, provide the manuscript number!)

5. Conclusion

Overall, the results support the idea that content is consistent with high engagement, as is demonstrable in certain posts that display a combination of certain visual and textual patterns. When non-structured data is present, this variation of content can be captured by hashtags, high level deep learning features such as word embedding or convolutional neural networks. Combined with more readily available influencer information (followers; engagement; country; gender), it gives a more contextual view on the piece of content produced by the influencer.

These findings have significant implications for the understanding of AI and social post data mining. We presented feature engineering techniques that help IM specialist and data scientist to better represent their data and feed it to machine learning algorithms for recommendations; outlier influencer detection; brand recognition, etc.

However, more research on this topic needs to be undertaken before the association between machine learning techniques and influencer presence on social media is more clearly understood.

References

Brown D, Hayes N (2008) *Influencer Marketing: Who Really Influences Your Customers?* (Elsevier/Butterworth-Heinemann), ISBN 9780750686006, URL https://books.google.fr/books?id=250VSJ5JwjIC.

De Veirman M, Cauberghe V, Hudders L (2017) Marketing through instagram influencers: the impact of number of followers and product divergence on brand attitude. *International Journal of Advertising* 36(5):798–828.

Ewers N (2017) *# sponsored–Influencer Marketing on Instagram: An Analysis of the Effects of Sponsorship Disclosure, Product Placement, Type of Influencer and their Interplay on Consumer Responses.* Master's thesis, University of Twente.

Glucksman M (2017) The rise of social media influencer marketing on lifestyle branding: A case study of lucie fink. *Elon Journal of Undergraduate Research in Communications* 8(2):77–87.

Katz E, Lazarsfeld P (1966) *Personal Influence, the Part Played by People in the Flow of Mass Communications.* A Report of the bureau of applied social research Columbia university (Free Press), ISBN 9781412830706, URL https://books.google.fr/books?id=rElW8D0D8gYC.

Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. *European conference on computer vision*, 740–755 (Springer).

Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, real-time object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

Semiz G, Berger PD (2017) Determining the factors that drive twitter engagement-rates. *Archives of Business Research* 5(2).

Sobkowicz P, Thelwall M, Buckley K, Paltoglou G, Sobkowicz A (2013) Lognormal distributions of user post lengths in internet discussions-a consequence of the weber-fechner law? *EPJ Data Science* 2(1):2.

Zietek N (2016) Influencer marketing: the characteristics and components of fashion influencer marketing.