Horizon 2020 Framework Programme

Grant Agreement: 732328 – FashionBrain

Document Information

Deliverable number: D3.3

Deliverable title: Surveys design and crowdsourcing tasks

Deliverable description: The goal is to handle the so-called unknown-unknown information by continuously probing the crowd for additional insights or data sources concerning new trends, emerging blogs, influencers and subjective opinions. In this task we will develop an additional set of task interfaces and questionnaires that will run periodically on a varying sample of the crowd population. The results of this tasks will be D3.3 and crowd-generated data that will be used in WP5, 6, and, 7.

Due date of deliverable: 30/06/2018

Actual date of deliverable: 30/06/2018

Authors: Alessandro Checco and Inés Arous

Project partners: University of Sheffield, University of Fribourg

Workpackage: WP3

Dissemination Level: Public

Change Log

Version	Date	Status	Author (Partner)	Description/Approval Level
1	30/06/2018	Final	USFD, UNIFR	Public

Table of Contents

1 Introduction

Fashion influencers are a fundamental component that contributes to shaping fashion trends and marketing strategies. FashionBrain has the ambitious goal of using crowdsourcing to detect new fashion trends, by handling the so-called unknown-unknown information by continuously probing the crowd for additional insights or data sources concerning new trends, emerging blogs, influencers and subjective opinions. The expected outcome is to develop a set of task interfaces and questionnaires that will run periodically on a varying sample of the crowd population. However, such an effort requires first to answer some related preliminary questions, that are prerequisites to develop such interfaces:

- Are crowdsourcing platforms reliable and consistent over time? We will explore this question in Section 2, where we study the reliability of two platforms that have been used widely in research and data collection and evaluation. Our findings will help to uncover data reliability problems and to propose changes in crowdsourcing platforms that can mitigate the inconsistencies of human contributions.
- In Section 3, we run a pilot study to assess the conditions necessary to build a crowdsourcing framework to detect fashion influencers.
- We interviewed three leading European companies in the field to understand expert assessment and evaluation of fashion influencers. We report this study in Section 3.

2 Stability and Reliability of Crowdsourcing Output

Estellés-Arolas and González-Ladrón-De-Guevara (2012) define crowdsourcing as: “a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task”.

While in some cases crowdsourced tasks are engaged with on a voluntary basis, there is a significant and growing paid crowd labour market facilitated through platforms such as Amazon Mechanical Turk and Figure Eight. From the task requester’s perspective, the benefits of using crowdsourcing platforms are the possibility of completing a job incredibly fast, with reasonable quality, and with a cost comparable to traditional means. From a worker perspective the key benefit is the ease of accessibility to paid work and flexible hours and commitment.

There are many successful examples on the web of crowdsourcing platforms. However, the features and services provided for the requesters vary from one platform to another, and no single platform meets all the possible requirements that the requesters may have.

We investigate the quality of the output of different platforms when the same task design and dataset is used. To study the reliability and consistency of the output of the platforms and to generalise the findings, we run a continuous evaluation of existing datasets and replicate the task over multiple weeks.

2.1 Related work

Crowdsourcing platforms evaluation

In this context, a study by Crump, McDonnell, and Gureckis attempts to validate Amazon Mechanical Turk (MTurk) as a tool for collecting data in cognitive behavioural research. They designed several types of experiments and compared the results with traditional laboratory ways of collecting data. The findings of this study proved that the quality of the data collected under the experimental conditions in MTurk is highly similar to the quality of the data collected the traditional laboratory way. A similar case study was presented by Bentley, Daskalova, and White, who analysed the results of surveying the workers on their behaviour of using particular technologies. This research compared the results from MTurk and Survey Monkey to those obtained using a traditional survey. They demonstrated that crowdsourcing platforms can provide the same results and do it much faster when compared to the traditional way of collecting survey data (Bentley, Daskalova, and White 2017). Despite some concerns related to the limitations of the technical and visual design of the task and unexpected behaviour such as dropping out of a task before finishing it, collecting data with crowdsourcing saves time and money and reach a wide range of users in a few seconds (Crump, McDonnell, and Gureckis 2013).

A few papers highlighted the differences between crowdsourcing platforms. In one of the recent studies Peer et al. introduced the new platform Prolific Academic (ProA) and compared the result of this platform with CrowdFlower (CF) and MTurk. The findings of this study recorded the highest response rate for participants in CF and the highest data quality for the participants in ProA and comparable to MTurk's (Peer et al. 2016). Another study Mourelatos, Tzagarakis, and Dimara used Rankings website to collect data and compare crowdsourcing platforms over two periods of time and according to a number of criteria: type of service provided, quality and reliability, region, online imprint. The findings of this study discuss the effect of the platforms characteristics of their traffic data and popularity (Mourelatos, Tzagarakis, and Dimara 2016; Mourelatos, Frarakis, and Tzagarakis 2017).

Time consistency of tasks

studies by Blanco et al. investigate the creation of evaluation campaigns for the semantic search task of keyword-based ad-hoc object retrieval using crowdsourcing task. They used a sample of entity-queries from the Yahoo! log and Microsoft log to evaluate the semantic search result. They prove that the reliability of crowdsourcing workers and the quality of the result was comparable to that of the experts even when repeating the same task over time (Blanco et al. 2011). Following this work, Tonon, Demartini, and Cudré-Mauroux extend the continuous evaluation of information retrieval (IR) systems using stander and crowdsourcing relevance judgments.

2.2 Research Questions

This research will address the following questions:

> **RQ1**: Is there a significant difference in the quality, reliability, and consistency of the results for the same task repeated over a different timescale?
>
> **RQ2**: Is there a significant difference in the quality, reliability, and consistency of the results for the same task performed on different platforms?

Answering RQ1 requires conducting a study where the same experiments will be repeated on a different timescale. We replicated the experiment using the same part of the dataset for the same assumption discussed in (Blanco et al. 2011; Tonon, Demartini, and Cudré-Mauroux 2012) for measuring repeatable and reliable evaluation over crowdsourcing systems. These studies show experimental proofs that a crowdsourcing platform produces a scalable and reliable result over a repetition time of one month. We examined a shorter timescale of one week for the same task design.

RQ2 offers an in-depth analysis and practical comparison of crowdsourcing platforms. We investigated the replication of the same task over multiple crowdsourcing platforms and over different levels of workers' experience and accuracy as provided by each platform. Two of the most popular platforms, that have been used in crowdsourcing business and research studies of data evaluation and acquisitions, that is, Amazon Mechanical Turk (MTurk) and Figure Eight (F8), have been chosen for this study.

For both research questions and for each platform, we ran multiple types of tasks and measured the stability of the performance over the variations of the following factors:

- The quality of the task interface (e.g. instructions, examples, and training question).
- The workers' experience level provided by the platform.

The evaluation of these factors depended on the completion time of the task and accuracy of the result. Moreover, with repeating the same task every week, the overall time of completing the batch on each platform will be recorded.

2.3 Experimental Results

We used the Billion Triples dataset created for the Semantic Web Challenge in 2009 and used in (Blanco et al. 2011; Tonon, Demartini, and Cudré-Mauroux 2012). The task consists of a query and a search result in RDF format and asks the workers to classify each search result into one of the three categories: “Excellent” (describes the query target specifically and exclusively), “Not bad” (mostly about the target), and “Poor” - not about the target, or mentions it only in passing).

Sample Size Estimation

We ran a pilot experiment on both platforms to test the validity of the task design and to calculate the ideal sample size for the main experiment. The task consists of one page showing 20 questions and paying 0.15$ per worker.

To determine the right sample size for the pilot study, (Connelly, 2008) suggests that sample size should be 10% of the population size. However, in our study it is not possible to know the population size of a crowdsourcing platform. (Isaac and Michael, 1995) suggested 10–30 participants for a similar type of studies where the population size is unknown and influenced by many factors. For that reason, we used 30 participants per platform for the pilot experiment.

To calculate the sample size needed for all phases we used the following equation, where the plan is to perform a test of hypothesis comparing the means of a continuous outcome variable in two independent populations:

$$n_i = 2\left(\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{ES}\right)^2$$

where n_i is the sample size required in each group (i = 1 for MTurk and n = 2 for F8), α is the selected level of significance and Z is the value from the standard normal distribution 1-β is the selected power. ES is the effect size:

$$ES = \frac{\mid \mu_1 - \mu_2 \mid}{\sigma}$$

According to the results of the pilot run ES= 0.29, and to have 75% statistical power, the sample size needs to be 150 workers on each platform.

Main Experiment

The experiments in this phase used the plain interface similar to the one presented in (Blanco et al. 2011).

We repeated the same experiment four times (once every week) and it was launched on the same day of the week and at the same time on each platform.

Each task consisted of 20 tweets to be judged by 150 workers. The workers were rewarded with 0.15$ and they could do the task only once since after they finished they were excluded from participating in another batch of the task.

	MTurk	F8
Average Time per Assignment	4 m, 16 s	6 m, 09 s
	4 m, 49 s	6 m, 33 s
	4 m, 24 s	6 m, 18 s
	4 m, 25 s	5 m, 30 s
Avg.Accuracy & Standard deviation	0.73 ± 0.20	0.63 ± 0.28
	0.76 ± 017	0.66 ± 0.25
	0.76 ± 0.14	0.67 ± 0.25
	0.74 ± 0.19	0.66 ± 0.27
Completion Time for the Batch	3 d, 00 h, 14 m	05 h, 11 m
	3 d, 01 h, 29 m	04 h, 45 m
	2 d, 08 h, 36 m	07 h, 10 m
	3 d, 13 h, 54 m	04 h, 43 m

Table 1: Results of four runs in MTurk and F8.

Table 1 and Figure 1 present the results of the baseline phase comparison between the two selected platforms. The results show some consistency over the four runs on each platform. Workers were finishing the task faster in MTurk, where the average time per assignment was approximately 4 minutes, while it took approximately 6 minutes in F8. The overall accuracy for each run on MTurk was more than 73% whereas it was in the range of 60% on F8. Although the results from MTurk are significantly better than those from F8, the total completion time for the whole batch took an average of 3 days in MTurk and 4 to 7 hours in F8. The results of the ANOVA tests are shown in Table 2.

	sum_sq	df	F	PR(>)
C(Platform)	2.60	1.0	50.6	0.19e-11
C(Time)	0.17	3.0	1.11	0.34
C(Platform):C(Time)	0.04	3.0	0.26	0.85
Residual	61.24	1192.0	NaN	NaN

Table 2: Results of 2 ways ANOVA test.

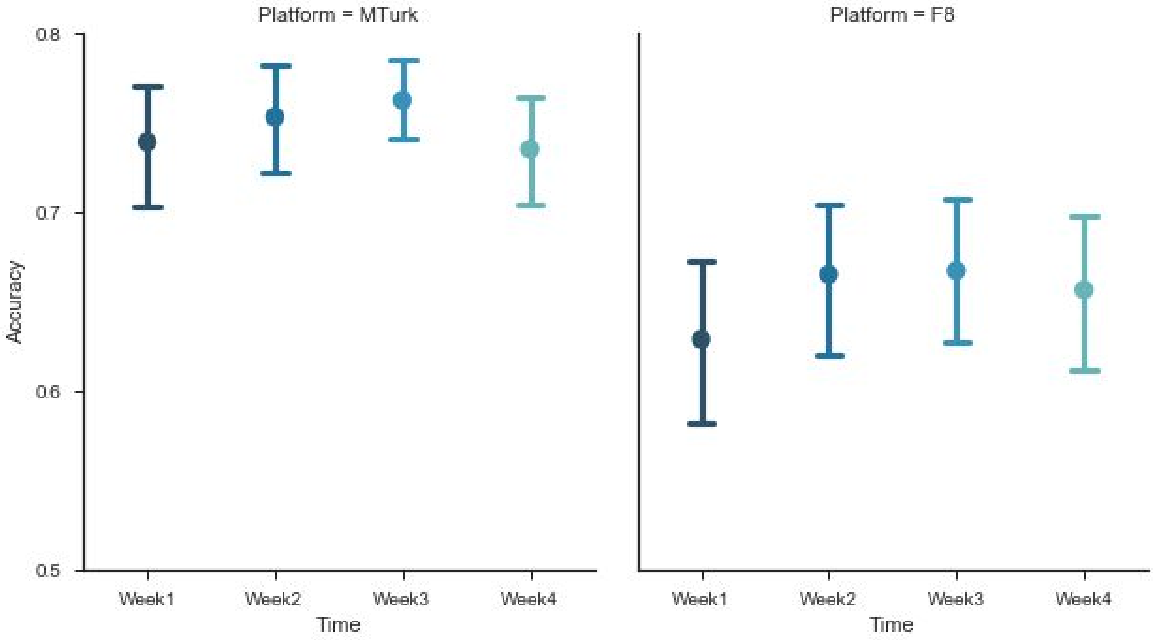


Figure 1: Accuracy distribution over time.

2.4 Future Directions

There are possible unpredictable effects: we are trying to target new workers to complete the task, and with repetition of the same task each week we could lose the attractiveness factor or fail to encourage new workers to participate in the task.

Some developments are unpredictable - e.g. sudden rapid growth of the crowdsourcing platforms. Recently, CrowdFlower became Figure Eight and the changes in the platform policy remain unclear.

For the advanced phases of this study, we will investigate why we had this results in the baseline phase. One of the reasons could be the workers' diversity and their level of experience. Another reason could be the variation of the amount of payment for different channels in F8.

With this study, we hope to reach a reasonable level of understanding what are the best strategies and advise crowdsourcing users on the best way to achieve better service from the system.

3 Crowdworker Fashion Expertise

We decided to run an experiment to assess crowdworkers expertise in fashion, to then successfully devise a framework to probe social networks for new trends. The first question to answer is: are European crowdworkers able to correctly recognise fashion influencers? Are crowdworkers familiar with them and with the most important social networks used for fashion?

To answer these questions, we used the dataset kindly provided by Fashwell. This dataset contains 118 Instagram fashion influencers, with profile picture, biography and 100 posts each, together with some basic metrics as number of comments and likes per post.

We then have built a task asking the crowd to assess each Instagram account, as shown in Figures 3-5. The questions are build so that we could assess crowd worker proficiency in the social network, as well as their ability to recognise famous and less famous influencers. We asked the set of questions to three different workers for each influencer.

From a preliminary analysis we can draw the following observations:

- Workers are able to identify that these influencers work in fashion easily, as shown in Figure 2. We believe that the crowd can be used successfully to recognise fashion influencers from a larger pool of Instagram accounts.
- Only in 47 of 336 cases the worker was able to recognise the influencer.
- Only in 3 of those 47 cases the worker is actually familiar with the influencer and following them on Instagram.

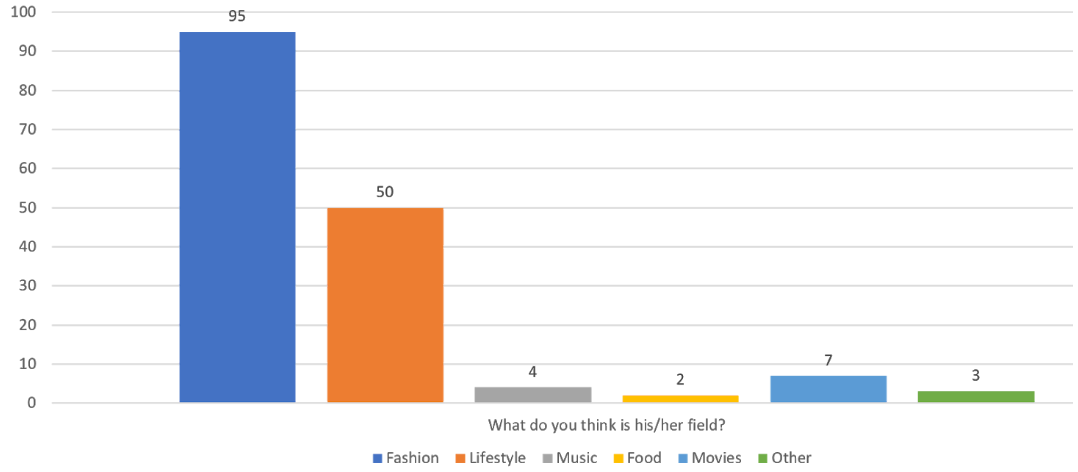
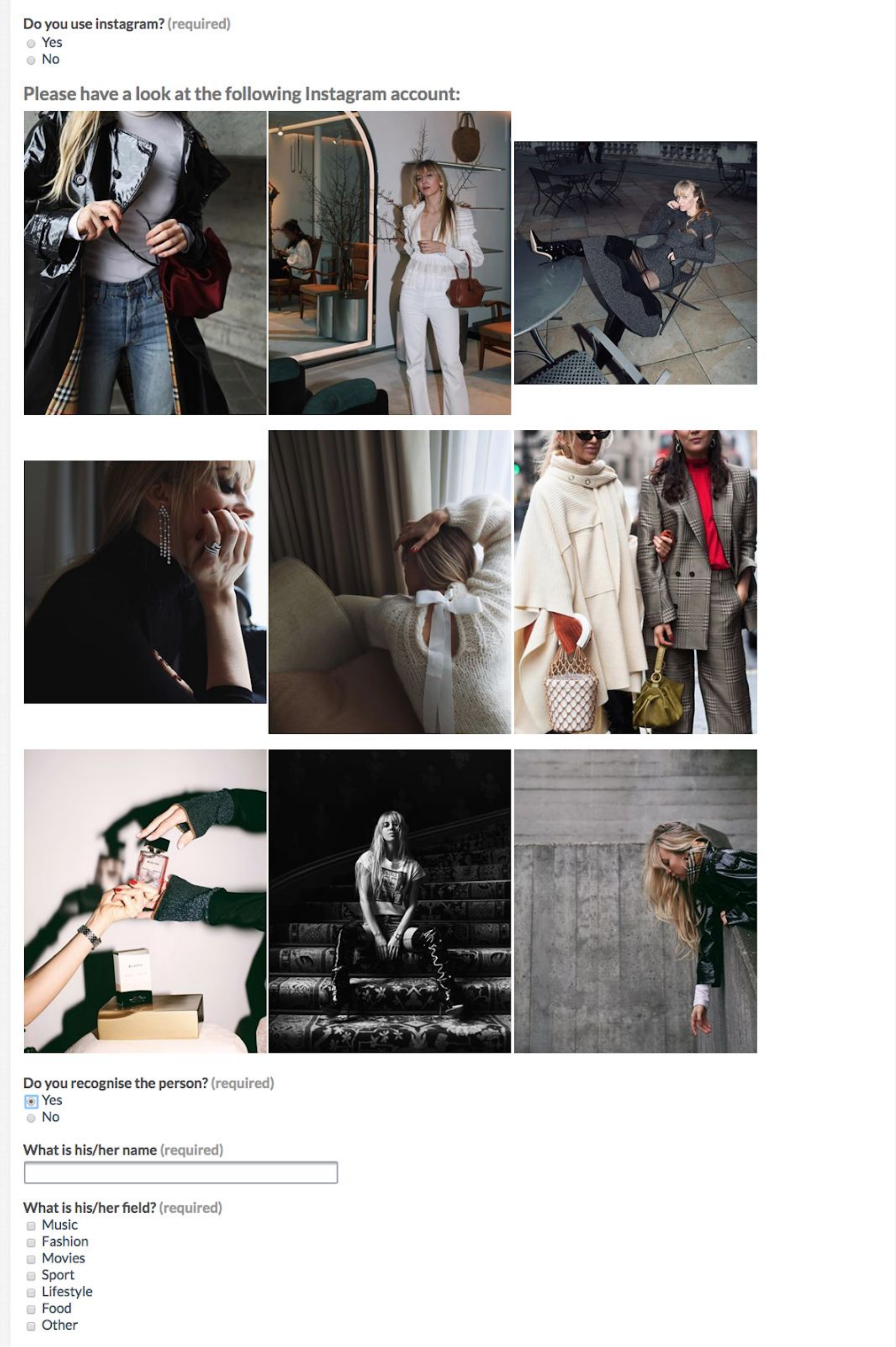


Figure 2: Crowdworker recognition of field of work of the Instagram account.

From these preliminary results we can conclude that a human-in-the-loop solution is possible, but very arduous if not first preceded by a target recruiting.

For this reason, we will focus in the rest of the work on expert recognition. Expert recognition have been extensively studied in the data mining community (Pal 2011, Pal 2012). We can divide the methods used to identify experts into two main approaches: a graph based approach and a feature based approach. In the graph based approach, the community is represented as a graph and experts are identified with algorithms such as PageRank, HITS and their extensions (Langville, Meyer, 2011). In the second one, the expertise dimensions are learned with supervised machine learning methods and used to identify experts. In this work, we assess workers expertise in fashion. We see this as a first step towards using the crowd network to identify fashion influencers. The plan is to design a task where we optimize the routing between workers to eventually reach those who can reveal emerging fashion influencers.



Do you use instagram? (required)

○ Yes
○ No

Please have a look at the following Instagram account:

Do you recognise the person? (required)

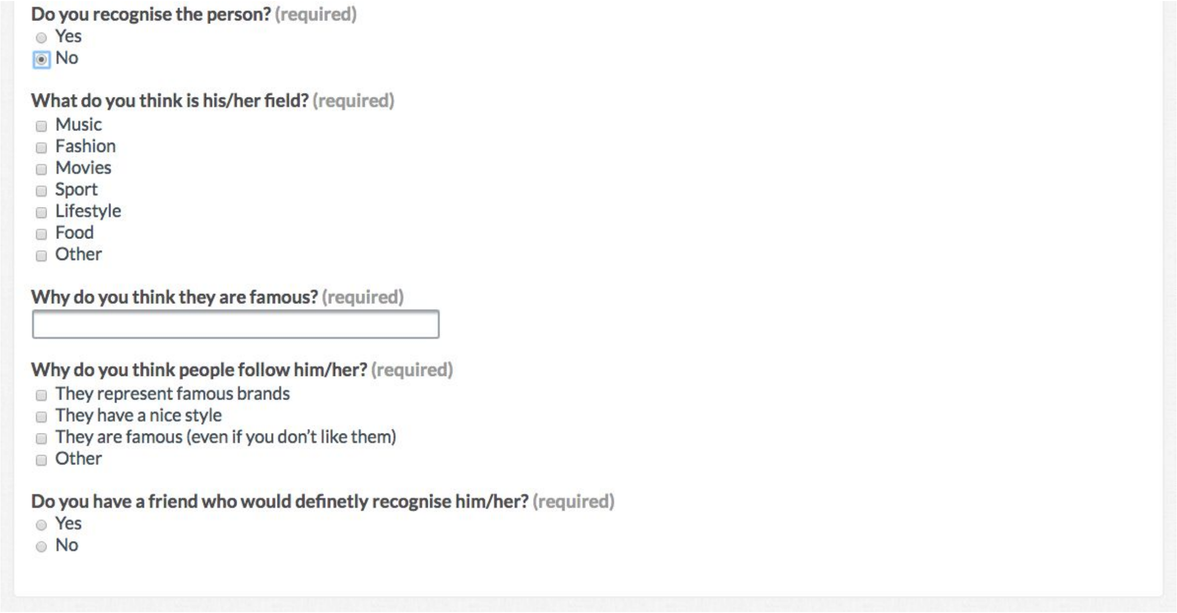
● Yes
○ No

What is his/her name (required)

What is his/her field? (required)

□ Music
□ Fashion
□ Movies
□ Sport
□ Lifestyle
□ Food
□ Other

Figure 3: Crowdsourcing task for Instagram accounts (part 1).



Do you recognise the person? (required)

○ Yes
● No

What do you think is his/her field? (required)

□ Music
□ Fashion
□ Movies
□ Sport
□ Lifestyle
□ Food
□ Other

Why do you think they are famous? (required)

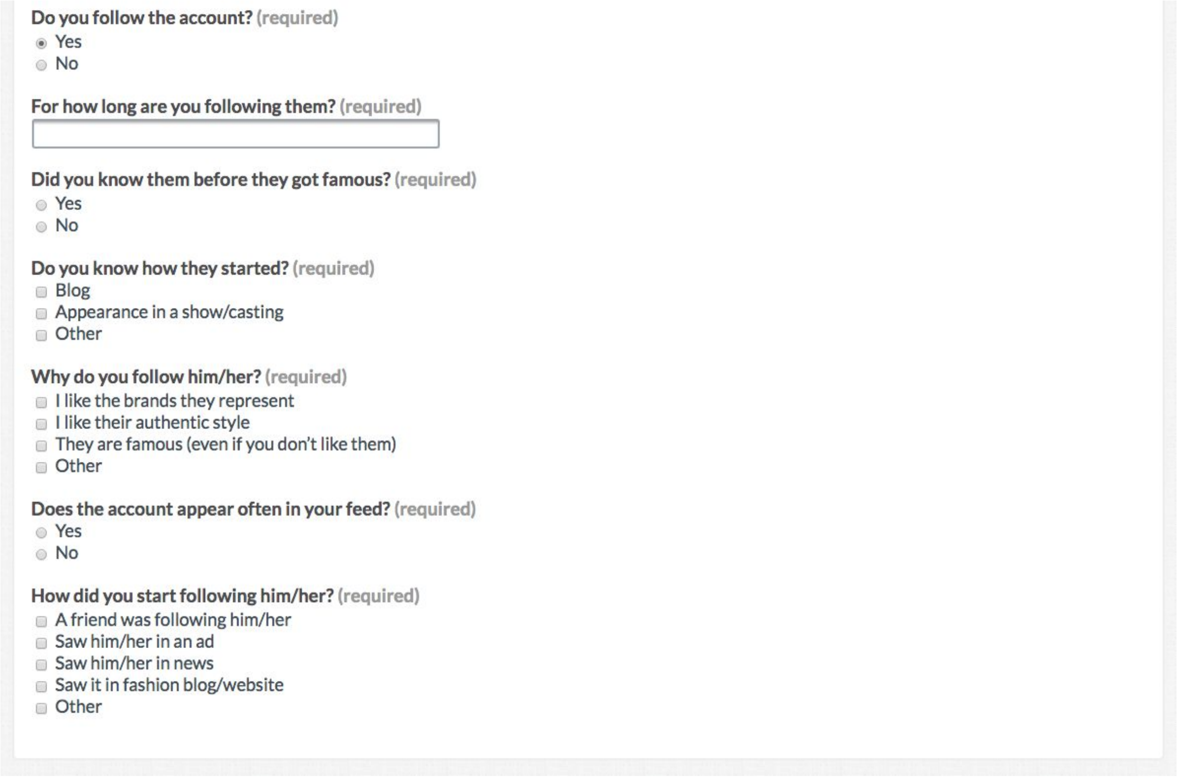
Why do you think people follow him/her? (required)

□ They represent famous brands
□ They have a nice style
□ They are famous (even if you don't like them)
□ Other

Do you have a friend who would definetly recognise him/her? (required)

○ Yes
○ No

Figure 4: Crowdsourcing task for Instagram accounts (part 2).



Do you follow the account? (required)

● Yes
○ No

For how long are you following them? (required)

Did you know them before they got famous? (required)

○ Yes
○ No

Do you know how they started? (required)

□ Blog
□ Appearance in a show/casting
□ Other

Why do you follow him/her? (required)

□ I like the brands they represent
□ I like their authentic style
□ They are famous (even if you don't like them)
□ Other

Does the account appear often in your feed? (required)

○ Yes
○ No

How did you start following him/her? (required)

□ A friend was following him/her
□ Saw him/her in an ad
□ Saw him/her in news
□ Saw it in fashion blog/website
□ Other

Figure 4: Crowdsourcing task for Instagram accounts (part 3).

4 Analysis of Influencer Market Ecosystem

In order to understand how to reliably and efficiently detect new trends in fashion, it is necessary to first examine the state-of-the-art from the industrial side. We reached out and interviewed three influencers experts working for three European companies: *Collabary*, *Influencer-Check.ch* and *Reachbird.io*, asking specific questions on how the influencer market works and what are the techniques used to detect fashion influencers.

4.1 Fashion Influencer Ecosystem

The fashion influencer ecosystem is made essentially of three parties: The fashion brands, the influencers and a platform connecting the two. Every party benefit from one another.

Brands and retailers have adopted fashion influencers to promote new trends and reach a wider audience. In fact, fashion influencers help brands to present their new products to consumers in an authentic way. Therefore, their impact on spreading awareness of these new products is more relevant than paid advertisement. In fact, 65% of fashion and retail brands launched campaigns with influencers over the past year and 74% of those experts found that influencer marketing was effective at driving sales in 2016 (Launchmetrics, 2017).

Fashion influencers need to continuously create content in their social media accounts (Instagram, Youtube, etc.) and have an engaged relation with their audience. By definition, influence is the "act or power of producing an effect without apparent exertion of force or direct exercise of command"[1]. In fact, the fashion influencers followers look for an authentic content therefore influencers need to "stay true to their style" and have the right collaborations.

This is where it comes the role of the platforms that connect brands with influencers. For example Collabary offers "a marketplace that give access to all relevant players and hence enables the brand to reach their audience in an authentic way […] Collabary covers the campaign creation, the discovery of influencers and the management of their participation in the campaign." These platforms also play a role in creating new campaigns and then provide "extensive reporting on the campaign performance".

4.2 Profile of Fashion Influencers

Understanding whether a person is a fashion influencer can be rather difficult in the fashion marketing world: celebrities that have other professions, like models, actors, and athletes, can sometimes be considered influencers, because celebrities also start to boost their own social media channels and some Influencers that started off with blogging as a hobby, now are considered celebrities, e.g. Chiara Ferragni (Roussina, 2018). They both get either paid by a brand or genuinely like it and tell the world with either themselves or their social personality (Faltl and Hauser, 2018).

However, we can identify three key differences between influencers and other fashion actors:

1. Celebrities core profession is related to an industry – singers, actors, professional sportsmen, politicians – they can of course be brand ambassadors, but their main professional activity is not being a full-time influencer. On the other hand, fashion influencers have this as profession (full time) – they focus heavily on creating and curating content for their Social Media

[1] As defined by the Merriam-Webster dictionary.

accounts, that is in line with their persona/brand and cater to their community (Roussina, 2018).
2. Influencers are closer – one could say they have a personal relationship – to their followers, while celebrities might have a greater and worldwide following, but not as close relationship to them (Roussina, 2018).
3. There is a stronger proclivity among influencer to actively engage in the creative process, as in contrast to many celebrities they draw their credibility directly from the content. Content creation is at the heart of their business model and not just one way of monetization as for e.g. an athlete (Faltl and Hauser, 2018).

Influencers are usually tied to specific brands, and their "specialisation skills" (such as lifestyle, fashion, beauty, food, etc.), their geographic location and the ones of their followers are important factors taken in consideration when looking for a fit between an influencer and a brand for a collaboration (Faltl and Hauser, 2018; Roussina, 2018).

The companies we contacted consider five main characteristics an influencers should have:

- Authenticity: the ability stay true to their style/brand and community when communicating and deciding on collaborations.
- Communication: the ability to engage with one's audience and the relevant influencer community (get to know other Influencers IRL and support colleagues, even planning co-creation sessions), as well as being professional (responsive) in the communication with brands during collaborations.
- Dedication: the ability to manage their account as a full-time job, meaning, continuously creating and curating content (postings/video/stories) for their accounts as well as being active on social media.
- Branding: the ability to treat and work on their social media account as a brand, meaning, the ability to find and keep a consistent and unique style, imagery (feed) and tone of voice.
- Mission: the ability to generate value either for society in general or their community.

4.3 Influencer Detection - State-of-the-art

We now investigates what are the techniques used to detect fashion influencers. Usually, a weighted average of the following indicators is used:

- Average engagement rate: the ratio of number of comments/likes to number of followers.
- Comments/like ratio: the ratio of number of comments to number of likes.
- Followers/followed ratio: the ratio of number of followers to the number of the following accounts.
- Mentions: the number of mentions of the influencer.
- Ad/No-Ad ratio: the ratio of ads to the number of posts without ads.
- Follower growth rate: change on the number of followers within one month.
- Sentiment of the comments: analysis of the mood in the comments.
- Klout score (Rao et al. 2015).

However, often these metrics are not enough to properly detect new influencers, because of the plague of bought followers, and for the difficulty or properly predict authenticity and engagement rates. Because of that, multiple solutions are taken into considerations:

Manual screening: the use of experts that will manually screen the influencers posts, taking attention to imagery quality, feed consistency etc. Experts can decide to onboard accounts based on exceptional results even when the metrics are below the established thresholds, for example because of exceptional engagement rate/imagery or a recent fast-growing follower community (Roussina, 2018).

The shift towards micro influencers, where screening is easier and potential manipulations are easier to spot (Faltl and Hauser, 2018).

5 Conclusions

The observations at the end of Section 3 make clear that influencer detection can be a candidate problem to be solved in a human-in-the-loop fashion: some features are easily detected automatically (number of followers or metrics related to understand activity level like number posts over time), while others require more manual annotations (like authenticity or quality of content). This is why a hybrid crowdsourcing approach using automatic metrics together will manual screening is a solution that FashionBrain will pursue. However, we discovered that a preliminary targeted recruiting phase is necessary to select the most competent crowdworkers in this niche field.

6 Acknowledgments

We want to thank Rehab Qarut for her contribution on Section 2.

We thank Collabary (https://www.collabary.com/), Influencer Check (www.influencer-check.ch), and Reachbird (www.reachbird.io) for the valuable insights.

References

Bentley, F. R.; Daskalova, N.; and White, B. 2017. Comparing the Reliability of Amazon Mechanical Turk and Survey Monkey to Traditional Market Research Surveys. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* - CHI EA '17, 1092– 1099.

Blanco, R.; Halpin, H.; Herzig, D. M.; Mika, P.; Pound, J.; and Thompson, H. S. 2011. Repeatable and Reliable Search System Evaluation using Crowdsourcing. *Journal of Web Semantics* 21:923–932.

Connelly, L. M. *Pilot studies*. Medsurg Nurs, 17(6):411–412, 2008.

Crump, M. J. C.; McDonnell, J. V.; and Gureckis, T. M. 2013. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE* 8(3).

Estellés-Arolas, E. and González-Ladrón-De-Guevara, F., 2012. Towards an integrated crowdsourcing definition. *Journal of Information science*, 38(2), pp.189-200.

Faltl, M.; Hauser, C. Influencer Check (influencer-check.ch), *personal communication*, 2018.

Launchmetrics, *The state of IM 2017*, Jan 2017.

Mourelatos, E.; Frarakis, N.; and Tzagarakis, M. 2017. A Study on the Evolution of Crowdsourcing Websites. ISS) *European Journal of Social Sciences Education and Research* 11(1):2411–9563.

Mourelatos, E.; Tzagarakis, M.; and Dimara, E. 2016. A Review of Online Crowdsourcing Platforms. *South-Eastern Europe Journal of Economics* 14(1):59–74.

Peer, E.; Samat, S.; Brandimarte, L.; and Acquisti, A. 2016. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70(January):153–163.

Rao, Adithya, et al. "Klout score: Measuring influence across multiple social networks." *Big Data, 2015 IEEE International Conference on.* IEEE, 2015.

Roussina, V. Collabary (https://www.collabary.com/), *personal communication*, 2018.

Tonon, A.; Demartini, G.; and Cudré-Mauroux, P. 2012. Combining inverted indices and structured search for ad-hoc object retrieval. In *SIGIR*, 125.

Tonon, A.; Demartini, G.; and Cudré-Mauroux, P. 2015. Pooling-based continuous evaluation of information retrieval systems. *Information Retrieval* 18(5):445–472.

Pal, A., Farzan, R., Konstan, J. A., & Kraut, R. E. (2011, July). Early detection of potential experts in question answering communities. In International Conference on User Modeling, Adaptation, and Personalization (pp. 231-242). Springer, Berlin, Heidelberg

Pal, A., Chang, S., & Konstan, J. A. (2012, June). Evolution of experts in question answering communities. In ICWSM.

Langville, A. N., & Meyer, C. D. (2011). Google's PageRank and beyond: The science of search engine rankings. Princeton University Press.