Profiling, Assessing and Matching Personalities Active in Social Media

Ciarán Hennessy[1] and Alan F. Smeaton[2]

[1]School of Computing and [2]Insight Centre for Data Analytics
Dublin City University, Dublin, Ireland
ciaran.hennessy2@mail.dcu.ie

Abstract. The world of social media influencers, bloggers and online "clothes horses" is a relatively new one. New-media personalities, a.k.a. "clothes horses", are now endorsing brands, products and companies in a more subtle way than through traditional advertising. They carefully cultivate and position their personal brands with a view to persuading businesses to use them for relatively inexpensive, often local, online marketing campaigns. In the world of traditional media, companies wishing to advertise use agencies to match their brand and core values to appropriate personalities. In this new media world, businesses must go it alone. In this paper, we present a pipeline for assessing and understanding the online reach of new-media personalities. Using Twitter, our method determines whether the social media followers of a new media personality, as a group, match their perceived brand values. We do this using automatically-determined sentiment and classification of tweets from the new-media personality and his/her Twitter followers. We also look at how businesses might determine which social media personalities would be a good fit for them for a marketing campaign. Finally we look at the evolving nature of the reach and brand of a new-media personality.

1 Introduction

Local businesses are increasingly turning to social media personalities (12) to increase their social media awareness and promote their brands. For most local businesses their choice of social media personality is restricted to people who are actual customers, who know actual customers or who are based in their locality.

More and more of these social media personalities are now earning sufficient income from endorsements to allow them to focus full time on their social media presence and activities. As a result, the issue of "endorsed" posts or output is a somewhat controversial issue at present. For example, a tweet from LeBron James costs in the region of $140,000 (5). At a more mundane level, Irish bloggers can expect to earn €200 to €3000 for a sponsored post (10).

In this context, social media has brought the endorser closer to the endorsed and has fundamentally changed the model. Previously, for this kind of local advertising it would have been necessary to engage with a personality's management and/or an advertising agency to secure their services. Now, businesses can reach out directly to potential endorsers.

While this may seem like progress, or a positive change, it is worth considering the expertise that an advertising agency brings to the engagement (13). An advertising agency will ensure that appropriate personalities are matched to brands and companies. Removing advertising agencies, and their professional knowledge, from the interaction means that companies must now go it alone.

In this paper we considered how to develop a profile for a social media personality based on their online output and the output of their followers. Using Twitter, it is then shown how it is possible for a business to assess the suitability of a social media personality for an engagement. Finally, a case study is presented involving a sponsored post and the effect of said post.

2 Selecting Social Media Personalities and Data Used for Profiles

As a test set for this work, a set of local "new media" personalities were selected. A Google search suggests that some of the the most popular and active social media personalities in Ireland are focused on fashion. For the purposes of this paper a set of 10 of the most popular fashion bloggers (5 male and 5 female) were used, and these are described later. This set of 10 was chosen from two articles published by a popular Irish fashion website (7)(1).

When assessing a social media personality's online presence there are a number of data sources that could be considered, which we now review.

2.1 "Official" data

The first, and probably most obvious data source is based on "official" output from the personality's website. All of the 10 personalities chosen maintain active websites. This source was considered their "official" output; that is thoughtfully composed, proof-read and finally published. We considered this data source to be the one that best reflects a personality's desired profile.

To automate the gathering of the data that was used to generate this profile a combination of Google API Client Library for Python (2) and a custom Google search engine (3) was used. The combination of these tools facilitates gathering a full list of URLs from each website.

2.2 Social data

Social data profiles were generated from the a personality's social media output. Social media output is much more likely to be "off the cuff" and less likely to be pre-authored. While this assumption of spontaneity is likely to be true at the "level" of the personalities included in the study, it may not be true at the upper echelons of social media influencers.

Examples include post of meals, social events, interactions with followers/-fans. All of the selected personalities are very active across a number of social

media platforms e.g. Twitter, Facebook, Instagram, etc. This may have presented additional overhead, i.e. aggregating the output of the various platform. However, due to the use of social media aggregation software the output is typically uniform across all of these platforms. To this end, the platform used to gather the input data for this profile was Twitter. Twitter provides the most mature, developer-friendly and least restrictive APIs (8). The Twitter API allowed access to the most recent 3,200 posts by each personality and to automate the gathering of the data we used the Tweepy Python library (11).

2.3 Follower/fan data

All of the selected personalities have significant follower/fan bases across the multiple social media platforms. By aggregating the data associated with these followers/fans it was possible to create a third profile for a personality.

For the purposes of this paper, followers for each personality were sampled as it simply was not practical to include the entire set of followers for each. The sample size, n, was determined by calculating

$$n = \frac{m}{1 + \frac{m-1}{N}}$$

where

$$m = \frac{z^2_{1-a/2}\hat{p}(1-\hat{p})}{\epsilon^2}$$

where the population size, N, is the total number of Twitter followers for a personality, $\alpha = 0.05$ i.e. 95% confidence level. $\hat{p} = 0.5$ i.e. the sample proportion. $e = 0.05$ i.e. 5% margin of error (4). This meant that sample sizes for our 10 personalities were between 282 and 382 followers.

When considering an account for inclusion in the set of sampled followers, efforts were made to exclude bots, i.e. automated or spam accounts, and dormant accounts. A follower's profile was deemed to be suitable if it posted at least once every 14 days but not more than 100 times per day (14). Using the Twitter APIs, the full list of followers for each personality was downloaded. The appropriate number of samples were drawn, excluding the bots and dormant accounts and for each sampled follower the most recent 100 tweets were downloaded. Thus for each personality there were between 28,200 and 38,200 tweets to use to construct the follower-based profile.

3 Classifying raw data

To build a profile for a personality it was necessary to categorise each data point and to assess its sentiment. Using the IBM Watson^{TM} AlchemyLanguage service (9). which uses "sophisticated natural language processing techniques" and "complex statistics and natural language processing technology" to classify text each tweet or URL in each profile was categorised into a hierarchical taxonomy (6) with 1,092 unique classifications.

At the highest level taxonomical level, examples of categories are "education" and "finance". At the lowest level, examples include "camera bags" and "plasma TVs". An example of a full classification i.e from the highest level to the lowest level is: "technology and computing/consumer electronics/camera and photo equipment/cameras and camcorders/camera batteries". This meant that the classification for each each tweet or URL was very fine-grained, and an individual piece of text or URL may be found to be associated with zero or more classifications, each with a confidence score (see Listing 1).

Listing 1.1. Sample Alchemy API response

```
"text": "Today's Facebook memory highlighting my awful
         fashion choices 3 years ago #Canyounot",
"sentiment": {
        "score": "-0.506493",
        "type": "negative"
},
"taxonomy": [
        {
                "confident": "no",
                "score": "0.313837",
                "label": "/style and fashion"
        },
        {
                "confident": "no",
                "score": "0.238069",
                "label": "/technology and computing/
                  internet technology/ social network"
        },
        {
                "confident": "no",
                "score": "0.137487",
                "label": "/technology and computing"
        }
        ]
```

The example in listing 1 shows that the text has been categorised and a sentiment score returned. For example, if the following text was sent to the Alchemy API: "I love football", a sentiment score of 0.634836 is returned. Similarly, if a negative version of the same text was passed, e.g. "I hate football", a sentiment score of -0.860944 is returned.

Initial processing of gathered data resulted in a vector of length 1,092 representing each profile (16). This involved calculating a score for each individual classification option using

$$\sum_{i=1}^{n} s_i \times w_i$$

where i is an occurrence of a classification, s_i is the score from the taxonomy entry and w_i is a weight applied based on the sentiment. The values in the resulting vector were then normalised using

$$z_i = \frac{x_i - min(x)}{max(x) - min(x)}$$

where $x = (x_1, ..., x_n)$.

In order to assess the effectiveness of this approach we calculated the similarity between any two vectors (profiles) using the cosine similarity k, defined as

$$k(x, y) = \frac{xy^T}{||x||.||y||}$$

4 Taxonomy Depth in Personality Profiles

Before developing profiles for each personality based on the 3 data sources available, it was worth considering whether a profile containing 1,092 data points could be the most effective representation of a personality. as the fine-grained nature of the taxonomy used to generate these data points may not accurately reflect profile similarities. There was also the danger that the classification algorithms may be accurate at recognising the higher-level concepts but less able to recognise the lower-level concepts.

For example, it is possible to get a classification for one or more of the following, which are very similar, even for a human to assess:

- law/govt and politics/espionage and intelligence/secret service
- law/govt and politics/espionage and intelligence/surveillance
- law/govt and politics/espionage and intelligence/terrorism

Thus it might be better to "roll up" some of the lower-level terms into their parent category.

For example at taxonomy depth 5 the following taxonomy classification represented a data point in a profile: "technology and computing/consumer electronics/tv and video equipment/video players and recorders/dvd players and recorders". At taxonomy depth 4 the value of "dvd players and recorders", if it had a value, was added to it's parent category i.e. "video players and recorders". "dvd players and recorders" was then be dropped from the taxonomy, thus reducing the length of the vector that represents the profile.

As layers are removed from the taxonomy, the number of possible data points for a profile decreases. Table I shows the number of distinct classifications at each taxonomy depth.

To test which taxonomy depth was the most appropriate for representing personality profiles we computed the profiles for each personality at each taxonomy depth and then calculated the mean similarity and variance. Because we have 2 distinct subsets with the group of personalities, i.e. 5 males and 5 females, these

Taxonomy Depth	Number of distinct classifications
5	1092
4	1073
3	894
2	339
1	23

Table 1. Number of distinct classifications at each level in the taxonomy.

two groups were evaluated separately before the evaluation of the super-set i.e. all 10 personalities.

For each of the three profile sets, i.e. male, female and all, a profile was calculated at each taxonomy depth for each data source. Finally, the three data sources were combined and evaluated based on the three sets i.e. male, female and all.

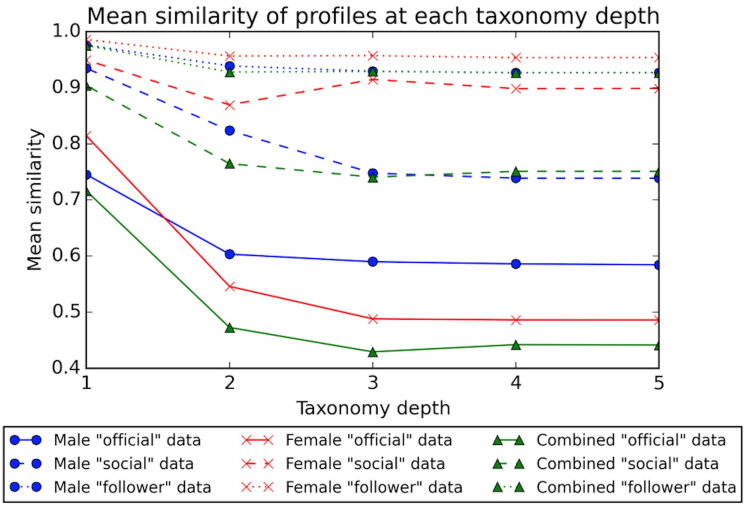

Fig. 1. Mean similarity of profiles at each taxonomy depth

What can be seen from Figures 1 and 2 is that all sets were "most" similar where the taxonomy is at its shallowest. This, intuitively, makes sense. For all 10 personalities, the most prevalent theme is "style and fashion". At this level in the taxonomy there is no difference between "style and fashion/beauty/face and body care/hygiene and toiletries" and "style and fashion/men's fashion" while they are clearly different subjects.

If this was abstracted higher it could be said that everyone is interested in "something" therefore all people's interests are identical. The trade-off now becomes between the level of similarity versus the richness/quality of the cat-

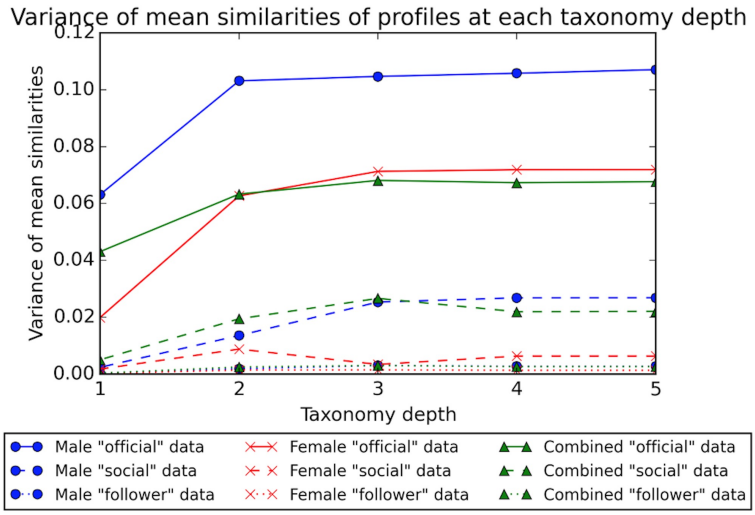

Fig. 2. Variance of mean similarities of profiles at each taxonomy depth

egorisation. For the purposes of this study we chose taxonomy depth 3 as the preferred option.

There is very little difference in similarity or variance between depths 3, 4 and 5 for all profiles. While depth 2 provided an increase in similarity for most profiles the trade off, i.e. decrease in data quality, was considered too expensive. For example, "style and fashion/clothing/wedding dresses" vs. "style and fashion/footwear/sneakers" is likely to be an important distinction (in the context of the fashion domain at least).

5 Comparing personality profiles

We can now develop three profiles for each personality and make some interesting comparisons, namely the official output compared to the social media output, the social media output compared to the followers and the followers compared to the official output. These types of comparison might be interesting to businesses considering engaging with a personality as it would allow them to see whether a personality's follower base are a good match for a company/brand/product(s). This is obviously a more important insight than simply knowing the number of followers, or the expected reach of a sponsored post.

Again, the similarity is measured by computing the cosine similarity between profiles represented as vectors (15) and Table 2 shows the results and has a number of interesting results. What is quite clear is that the profile developed from the personality's "official" data is, on average, most different from their social and follower profiles.

Personality	Official vs. Social	Official vs. Fan	Social vs. Fan
Darren Kennedy	0.5158	0.3586	0.9154
Damien Broderick	0.3058	0.3165	0.8741
James Patrice	0.4205	0.4856	**0.9446**
Conor Merriman	0.3598	0.4157	0.9061
Rob Kenny	0.4007	0.5050	0.9259
Suzanne Jackson	**0.8618**	0.3827	0.6202
Pippa O'Connor	0.7085	0.2764	0.6467
Louise O'Reilly	0.7770	0.4452	0.6558
Anouska Proetta Brandon	0.7027	**0.6584**	0.8550
Nuala Gorham	0.0858	0.1427	0.7975

Table 2. Comparisons within personality profiles

6 Case study

6.1 Introduction

An in-depth case study of the matching between new media personality and a business was performed using one of the personalities, Suzanne Jackson, and a business who agreed to "sponsor" a tweet. The business is a popular restaurant with 2 locations in Dublin. This personality and business have a standing agreement for this type of engagement. The business that agreed to be part of the case study provided some details to allow us to consider the impact of the engagement.

Before the impact is considered, it is worthwhile developing a profile for the business so that we might speculate as to the impact of the sponsored tweet and the type of results that could be expected. By doing this it will be possible to attempt to measure the level of similarity between the personality and the business. The business had 1500 followers on Twitter so in keeping with the data gathered for each of the personalities, the followers were sampled, as described above. The most recent 100 posts by each sampled follower were categorised and included in a profile that represented the business. The business also maintains an active Twitter account and a dedicated website. This allowed for the three types of profiles that have been used so far to be compared.

6.2 Profiles

The three profiles were calculated using the same process as developing the three profiles for each of the 9 other personalities. The results of the comparison can be seen in Table 3 below.

Comparison	Similarity score
Official restaurant v Official personality	0.1171
Social restaurant v Social personality	0.2768
Follower restaurant v Follower personality	0.7196

Table 3. Comparison of personality & business profiles

6.3 Sponsored post and interpretation of results

The sponsored post was photograph of Suzanne Jackson, holding a cocktail with the following accompanying text; "Dinner with my < 3 @ Siam Thai restaurants".

A number of observations immediately stand out from the table presented:

- the official outputs were vastly different;
- the social media output, although more similar than the official output, still differed greatly;
- despite these two observations, the profiles derived from the output of each set of followers was quite similar.

What is interesting is that the follower output is quite similar. The Cosine similarity between the profile generated from the followers of the business and the followers of the personality score 0.7196. This suggests that while the personality and business exist in different domains, i.e. "style and fashion" v "food and drink", the people who follow both are similar in their output.

The business saw some interesting results from the engagement. Firstly, the post attracted only two new followers and two "likes" on Twitter. Instagram attracted more interest where 1084 people "liked" and 11 followers commented on the post. There were 22 new followers for the business' Instagram account. Secondly, there was no increase in food sales in the immediate aftermath of the sponsored post. Most interestingly however is that they noticed an increase in sales of the particular cocktail that was featured in the sponsored post. For the business this was a welcome, if unintended, result.

By inspecting the aggregated profile generated from the sampled followers of the personality, it is might have been possible to predict this increase in alcohol sales. The top 10 most popular taxonomical categories among the sampled *followers* of the personality are presented in Table 4. This result suggests that it is possible to predict the effect or consequences of such an engagement. At the very least, it is possible to identify the interests of the people who the sponsored post is being targeted at.

6.4 A better choice ?

As mentioned earlier, the business in question has a "standing" arrangement with this personality for this type of engagement. If the business wanted to choose a personality based on the information available as part of this study,

Category	Normalised score
travel/tourist facilities/hotel	1.0
art and entertainment/music	0.8534
art and entertainment/movies	0.8015
style and fashion/beauty/cosmetics	0.7787
shopping/gifts	0.7595
food and drink/beverages/ alcoholic beverages	0.7241
style and fashion	0.7170
art and entertainment/shows and events/ festival	0.7069
business and industrial	0.6857
education/school	0.6369

Table 4. Top 10 most popular taxonomical categories among followers of Suzanne Jackson

would an alternative personality have been more appropriate ? Table 5 presents the profile similarities for all 10 of the personalities included in the study, with the business in question.

Personality	Official vs. Official	Social vs. Social	Fan vs. Fan
Darren Kennedy	0.0376	**0.4253**	0.8073
Damien Broderick	0.0443	0.4234	**0.8475**
James Patrice	0.1207	0.4062	0.8187
Conor Merriman	**0.1642**	0.2493	0.7846
Rob Kenny	0.0847	0.3624	0.7857
Suzanne Jackson	0.1172	0.2768	0.7196
Pippa O'Connor	0.0624	0.3308	0.8026
Louise O'Reilly	0.0493	0.2695	0.8026
Anouska Proetta Brandon	0.1428	0.3962	0.7774
Nuala Gorham	0.0056	0.3876	0.8304

Table 5. Comparison of 10 personalities with the restaurant business profile.

The results show that Damien Broderick is the best match for the business, His follower profile is comfortably the best match the with the restaurant's follower profiles. His social profile is the second best match with the restaurant's social profile by a margin of only 0.0019. It is worth considering the most popular taxonomical categories that appear in his followers profile, shown in Table 6 to explain why this might be so.

The list shows why Damien Broderick would indeed by a better match for a promoted post by the restaurant. "Food and drink" and "alcoholic beverages" both appear, which for a restaurant, would be considered important. Also, the

Category	Normalised score
travel/tourist facilities/hotel	1.0
style and fashion	0.9705
hobbies & interests/arts & crafts/ photography	0.8653
business and industrial	0.8284
food and drink/beverages/ alcoholic beverages	0.7891
health and fitness	0.7225
food and drink	0.7097
art and entertainment/movies	0.6807
art and entertainment/music	0.6639
style and fashion/beauty/cosmetics	0.6451

Table 6. Personality Damien Broderick, top 10 categories in his user profile

appearance of "music" and "movies" (which, in the interest of fairness, also appear in the Top 10 taxonomical categories for Suzanne Jackson) would be welcomed. The restaurant has a bar area which hosts a DJ at weekends and there is a large multiplex cinema within the same complex.

7 Conclusions

What has been presented in the study is an effective way to assess the suitability of an active social media personality for an engagement with particular business, most likely through some form of endorsement such as a sponsored tweet. This suitability was evaluated by incorporating three distinct sources of online information for a personality. Automated natural language processing and taxonomical classification tools were then applied to the data, the output of which was used to generate profiles to which standard similarity measures to compute similarity were applied.

A case study was presented which documented an engagement between a social media personality and a restaurant. The effect of this engagement for the business was outlined. The documented effect appeared to validate the approach that was taken in the study.

In the future it would be interesting to gather metrics across different industries e.g. celebrity chefs or sports stars. Of course, as a personality transitions through the normal set of life events e.g. marriage, children, becoming a "former" personality, divorce, death, etc., it is likely that their follower/fan base will also evolve. Studying this evolution could allow personalities themselves to make predictions about their future earning potential and/or their own "shelf life".

Acknowledgement

The contribution to this work by AFS was supported by Science Foundation Ireland under grant number SFI/12/RC/2289.

Bibliography

[1] 5 Irish Guy Bloggers You Should Totally Be Following — Stellar. http://stellar.ie/klaxon/5-irish-guy-bloggers-you-should-totally-be-following, accessed: 2016-07-12

[2] API Client Library for Python — Google Developers. https://developers.google.com/api-client-library/python/, accessed: 2016-06-23

[3] Custom Search — Google Developers. https://developers.google.com/custom-search/, accessed: 2016-06-23

[4] Estimating a Proportion for a Small, Finite Population — STAT 414 / 415. https://onlinecourses.science.psu.edu/stat414/node/264, accessed: 2016-07-10

[5] LeBron James-sponsored tweets valued at $140K, or $1K per character. http://www.espn.com/nba/story/_/id/13470682/lebron-james-sponsored-tweets-232-million-followers-cost-140k, accessed: 2016-06-23

[6] List of the possible taxonomy categories. http://www.alchemyapi.com/sites/default/files/taxonomyCategories.zip, accessed: 2016-07-12

[7] Power Players: The Irish Bloggers You Need To Be Following — Stellar. http://stellar.ie/fashion/power-players-the-irish-bloggers-you-need-to-be-following, accessed: 2016-07-12

[8] REST APIs — Twitter Developers. https://dev.twitter.com/rest/public, accessed: 2016-06-23

[9] Text Analysis Features — AlchemyAPI. http://www.alchemyapi.com/products/alchemylanguage, accessed: 2016-07-12

[10] They can make thousands from an Instagram post - but how transparent are Irish bloggers? http://www.thejournal.ie/irish-bloggers-money-transparency-2722012-May2016/, accessed: 2016-06-23

[11] Tweepy. http://www.tweepy.org/, accessed: 2016-06-23

[12] Booth, N., Matic, J.A.: Mapping and leveraging influencers in social media to shape corporate brand perceptions. Corporate Communications: An International Journal 16(3), 184–191 (2011), http://dx.doi.org/10.1108/13563281111156853

[13] Freberg, K., Graham, K., McGaughey, K., Freberg, L.A.: Who are the social media influencers? a study of public perceptions of personality. Public Relations Review 37(1), 90 – 92 (2011), http://www.sciencedirect.com/science/article/pii/S0363811110001207

[14] Jull, A., Bermingham, A., Adeosun, A., Ni Mhurchu, C., Smeaton, A.: Using twitter for public health infoveillance a feasibility study. In: Conference on Twitter For Research (2016), http://doras.dcu.ie/21188/

[15] Wang, J., de Vries, A.P., Reinders, M.J.T.: Unifying user-based and item-based collaborative filtering approaches by similarity fusion. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 501–508. SIGIR '06, ACM, New York, NY, USA (2006), http://doi.acm.org/10.1145/1148170.1148257

[16] Ziegler, C.N.: Social Web Artifacts for Boosting Recommenders: Theory and Implementation. Springer Publishing Company, Incorporated (2015)