Influencer Marketing Analytics using SAS® Viya®

Tamás Bosznay, Amadeus Software, UK

ABSTRACT

Marketing nowadays is more and more done via influencers. Influencers are content creators on social media who have a lot of followers and who really influence what people think, but also, what people consider buying. As influencer marketing is costly just as the more traditional marketing channels, it can be crucial for a company to select the best influencers to work with.

In our paper we aim to analyse on examples, who are the real influencers in Social Media, and what is the extent of their influence. The case study focuses on influencing opinions and sentiment about several brands. We analyse the posts and activities of a set of social media users with high number of followers. This requires first downloading posts and their metadata from some social media channels (we will focus on Instagram). Python® libraries and code will be used to complete these steps. We then analyse the network of posts, likes and replies to identify the influencers, the communities they are influential in and the extent of their influence on their follower's opinion and sentiment about the brands being marketed. SAS® Viya® text and network analytics functionality will be used to perform these analyses. The data collection steps (Python®) are using Jupyter as a client, whereas the analysis steps are mainly done using SAS® Visual Text Analytics (Model Studio) and SAS® Visual Analytics.

INTRODUCTION

Marketing is conducted via many channels nowadays. Influencers are a relatively new channel of marketing.

Influencers are basically people that influence other people via their behaviour and the content they produce and provide for their target audience. Nowadays, influencers are active on the different social media platforms, when decades ago influencing was happening via TV, radio, newspapers, magazines and so on. Social media adds the ability of quickly changing influencing behaviour and to reach out to unlimited amounts of people, potential consumers and followers of influencing activity, compared to the old mix of channels.

Marketing via influencers has just emerged in the recent years. Influencer marketing basically means that organisations who would like to promote their products or services are using influencers to get some sponsored content injected into their usual feed. For example, a sportswear company might pick one or more influencers who might be in sports with many followers. These influencers might provide usual content like photos or videos about their matches, coaching, alternative sporting activity like gyms, their meals and so on. The organisation is then asking the influencer to inject some of their products into their usual content, such as wear their sportswear products and give positive remarks about them on their photos, videos or other posts.

As with any marketing channel, an organisation using influencers is naturally interested in its characteristics and effectiveness. In our paper, we aim to show techniques that try to analyse these characteristics and the effectiveness of the channel. With such techniques, an organisation interested in working with influencers, or already working with them, can evaluate the effectiveness of using these channels. If it is worth using these channels for a particular organisation, it is also good to see some insight about the potentially different ways the different influencers work and how their activity is received among their followers. These insights can also drive decisions around which influencer is the best fit to work with a particular organisation.

In the paper we use various modules of the SAS Viya platform, SAS Visual Text Analytics to perform text analytics on social media data via Model Studio and SAS Visual Analytics to create visualisations. Python scripts via Jupyter notebooks are used to collect and clean social media data, as well as some data augmentation.
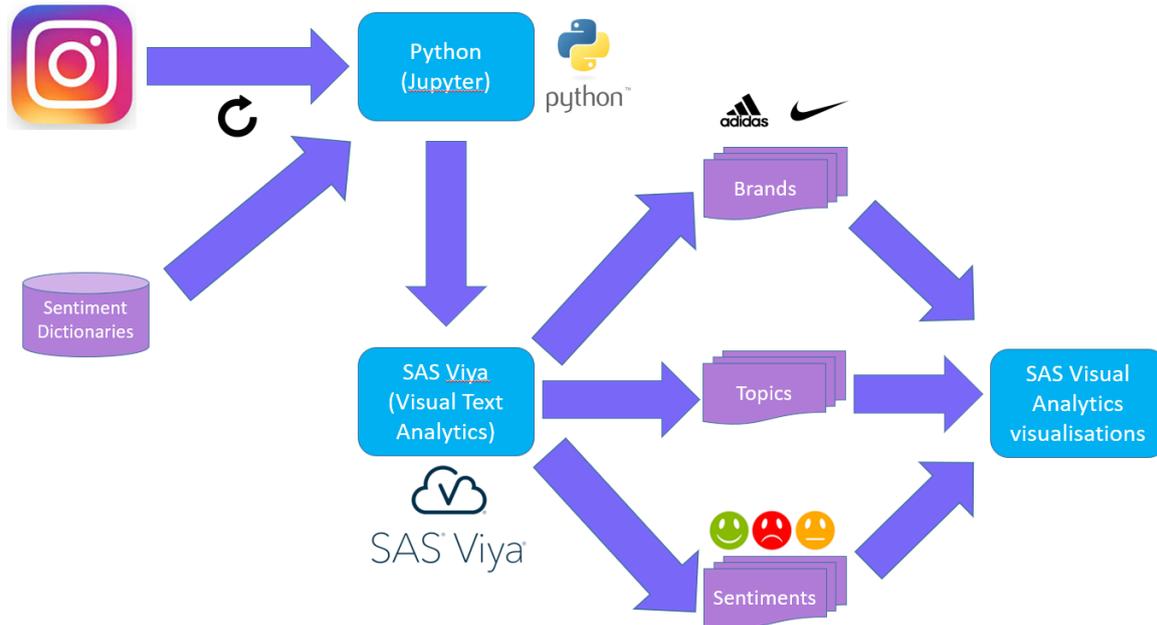
METHODOLOGY

The following figure shows the analysis process followed:

The diagram shows the main steps of the process followed in the paper:

1. Data Collection Phase: first the options available to access social media data are identified and discussed. After downloading the data using the preferred method, some data preparation and cleaning steps are done in a Jupyter notebook using Python.

2. Analytics Phase: The actual analytics steps are aiming to identify brands, extract topics and perform sentiment analysis on the data. External information (sentiment dictionaries) is also used via Python for augmenting the natural language processing capabilities included in SAS® Visual Text Analytics.

3. Visualisations Phase: Finally, results are visualised using SAS Visual Analytics and conclusions are drawn.

Data Collection Phase

EXPLORING RELEVANT CONTENT IN SOCIAL MEDIA

First, we need to decide an approach on what data to use for the analysis – data which is related to the influencer activity and potentially influencer marketing. The data needs to be accessible and be of meaningful size to derive reliable insights. The following social media channels could be considered: Twitter, Facebook, YouTube and Instagram.

From these four, Twitter isn't used much for marketing – although it is used actively for non-marketing related influencing (in politics, for example). Facebook is used very broadly, but when used for marketing, it is rather marketing products and services published directly by their provider organisation. YouTube has loads of users who influence their followers and the public, and it can be used for marketing as well. However, as we plan to do text analytics on the followers' perception on influencer marketing, it is important to have a reasonable volume of comments. A typical YouTube video post has just a couple of tens of comments, which is far less than the average number of comments on an influencer's Instagram post. Therefore, we have decided that we will collect Instagram data for this analysis. We will use Instagram posts created by big impact influencers and comments performed by other users on those posts.

After some initial analysis, we decided to collect data on posts from the Instagram users with the highest number of followers, and also the associated comments on those posts. Table 1 shows the users that we collected data from:

Amadeus Software Limited, The Old School Hall, 11 Wesley Walk, Witney, Oxon, OX28 6ZJ
Tel: +44 (0) 1993 848010 email:info@amadeus.co.uk

Page 2 of 16

Influencer Marketing Analytics using SAS® Viya®
Tamás Bosznay, Amadeus Software, UK

	User	Owner	Number of followers (million)
	@cristiano	Cristiano Ronaldo	158
	@arianagrande	Ariana Grande	148
	@selenagomez	Selena Gomez	147
	@therock	Dwayne Johnson	135
	@kimkardashian	Kim Kardashian	131
	@kyliejenner	Kylie Jenner	129
	@beyonce	Beyoncé Knowles	126
	@taylorswift	Taylor Swift	115
	@neymarjr	Neymar Jr.	112
	@leomessi	Lionel Messi	111
	@kendalljenner	Kendall Jenner	106
	@nike	Nike – Agency Account	85
	@victoriassecret	Victoria's Secret – Agency Account	65

The list of top users contains three popular footballers (Cristiano Ronaldo, Messi and Neymar), artists like Ariana Grande, Selena Gomez, Beyoncé or The Rock, and celebrities like "The Kardashians": Kim Kardashian, Kylie Jenner and Kendall Jenner. We also collected data from two agency accounts: Nike and Victoria's Secret. This is to compare characteristics of organisations directly marketing their products on Instagram versus influencers. It is important to mention that data is collected only from the official accounts of these users and not from any fan-based accounts that might exist. Fan-based accounts do not perform any marketing activities and therefore are not considered in the scope of this paper.

ACCESSING DATA FROM SOCIAL MEDIA
How to collect data from Instagram? Various Instagram APIs exist, some of them being originated historically from Instagram itself. Some other Instagram APIs are available via Facebook as Facebook Inc. is the owner of Instagram as well. For example, Facebook's Graph API allows SQL-like queries to be made to the data sources of Instagram.

Social media data APIs can change frequently, but specifically Instagram (and Facebook) APIs have changed dramatically in the last year. This is mainly due to the concerns around data security related to the Cambridge Analytica and similar scandals. As a result of these changes, when consuming Instagram data APIs, the following steps need to be followed:

1. Get an Instagram developer account and the access keys tied with it

2. Study the API documentation

3. Develop an app for Instagram

4. Register the app

5. Get the app approved by Instagram

After the approval, Instagram APIs can be used from the app to collect data, although this process can take some time. However, even when combined with an Instagram business account, data can be collected at quite reduced rates, this means that the amount of data downloaded is limited per time unit. This means for the purpose of volume data analysis, one needs to run their app consuming Instagram data for a long time, probably months or even years long.

Instagram APIs can be consumed from various platforms, including SAS. A combination of PROC HTTP to send API requests and the JSON libname engine to consume requests can be used within SAS.

Influencer Marketing Analytics using SAS® Viya®

Tamás Bosznay, Amadeus Software, UK

To avoid the steps around creating, registering and getting approval for an app, in our paper we followed an alternative method: web scraping. Web scraping is the technique widely used to download and parse potentially any web content for data science or other purposes.

We are interested in basically two forms of data: the posts by the influencers and the comments made on them.

First, have a look at how an Instagram page looks like, Figure 2 shows one.

Each photo (or sometimes video) is basically one post on Instagram, so to get to the actual full-sized photo with the text of the post, one needs to click on the photo on the influencer's landing page. To get more than a handful of posts, one needs to scroll down or hit the page down key loads of times as only that action triggers displaying more posts.

So, the first task in our web scraping technique is to get the URLs (links) to each influencer's posts via scrolling down and save the links. To do this, we have prepared a Python script and ran it from a Jupyter notebook.
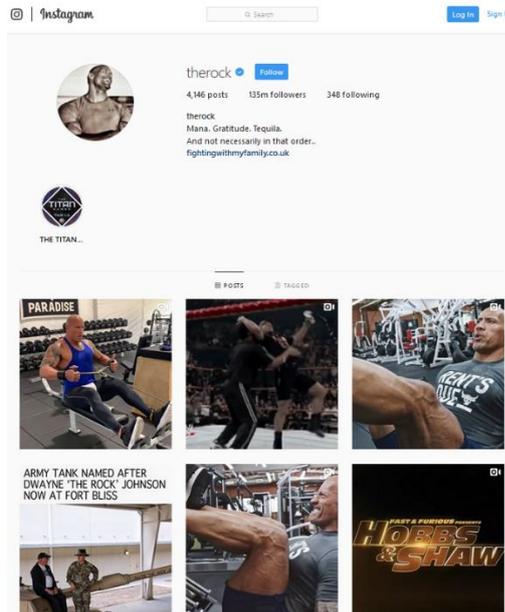

Figure 1: The Rock's Instagram page

Figure 3 shows this Python script:

```python
In [3]:  #opening a headless browser session
         from selenium import webdriver
         url = 'https://www.instagram.com/therock/'
         cap = DesiredCapabilities().FIREFOX
         browser = webdriver.Firefox(capabilities=cap,firefox_binary=r'C:\Program Files\Mozilla Firefox\firefox.exe')
         browser.get(url)
         time.sleep(1)

         #finding the body element
         elem = browser.find_element_by_tag_name("body")
         links=[]

         #scrolling a number of times (650) to get all posts of the Instagram user
         no_of_pagedowns = 650
         while no_of_pagedowns:
             elem.send_keys(Keys.PAGE_DOWN)
             time.sleep(0.35)
             no_of_pagedowns-=1

             #finding all href elements, ie all the links on the page
             elems = browser.find_elements_by_xpath("//a[@href]")
             for eleme in elems:
                 links.append(eleme.get_attribute("href"))
         links=list(dict.fromkeys(links))
         urlpattern=r'instagram.com/p/'

         #saving down all links in the list in case they follow the Instagram post url pattern
         with open(r"c:\mypath\links.txt", 'a') as export:
             for line in links:
                 if line.find(urlpattern)>0:
                     export.write('{}\n'.format(line))
```

Figure 2: Python script for downloading Instagram post links

The script on Figure 3 uses the package Selenium (a web testing framework) and its components, together with a browser driver (Firefox in the example) to perform web scraping. The script performs the following steps:

1. Opens a headless browser driver session and visits the target URL (https://www.instagram.com/therock in the example).

2. Finds the HTML body element of the page.

3. Inside the HTML, scrolls down via "hitting" the page down key. In this example, we hit the page down key 650 times, this proved to be enough to display a couple of thousands of posts. Between each scrolling, the script sleeps 0.35 seconds to allow all non-photo and non-video elements to be downloaded correctly.

Amadeus Software Limited, The Old School Hall, 11 Wesley Walk, Witney, Oxon, OX28 6ZJ
Tel: +44 (0) 1993 848010 email:info@amadeus.co.uk

Page 4 of 16

4. After the scrolling, it locates each link (a href) inside the HTML body and appends each link to a Python list.

5. Opens a .txt file in append mode and appends each link from the list which follows the pattern for an Instagram post link (where the link contains the pattern "instagram.com/p/").

6. This process needs to be repeated for each influencer's landing page (possible to further automate).

Now we have our list of links for all the influencer posts that we are interested in. The next step is to visit each post URL and download the comments and their attributes. With web scraping, just like when finding links in the HTML body, we need to locate the actual HTML elements that contain the information we need. To locate these, browser developer tools can help a lot. Let's open up a post in our browser and open the developer tools. Figure 4 shows an example post with its comments in the developer view (inspector) in Firefox as the browser.
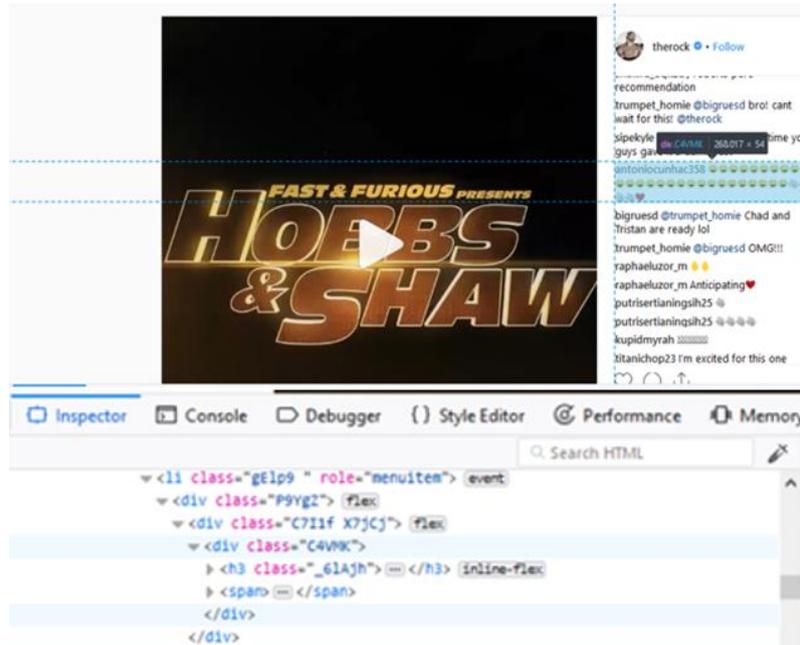
Figure 3: An Instagram post and its comments

In the inspector view (available under Tools->Developer tools->Inspector in Firefox) hovering over the desired content shows the actual HTML code. As Figure 4 shows, we have highlighted one user comment, and at the bottom, the inspector view tells us that we need to focus on the "div" element in the HTML which in this case is the class C4VMK (div class="C4VMK").

Amadeus Software Limited, The Old School Hall, 11 Wesley Walk, Witney, Oxon, OX28 6ZJ
Tel: +44 (0) 1993 848010 email:info@amadeus.co.uk

Page 5 of 16

We can also see what structure this highlighted comment has: it contains the user name who made the actual comment and the comment content. The comment is a collection of emojis and no plain text, but other comments can have plain text or plain text mixed with emojis. To display more than a handful of comments, the user needs to click on the link "Load more comments" a couple of times. Figure 5 shows a Python script implementing the above:

```python
In [27]: def openpost(browser,file,csvexport):

    #getting links from file
    with open(file) as f:
        for line in f:
            url = line.strip()  # to remove the trailing \n
            #go to link
            browser.get(url)
            no_of_clicks = 30
            while no_of_clicks:
                try:

                    #find the button to load more comments and click on it
                    browser.find_element_by_xpath("//button[contains(text(), 'Load more comments')]").click()
                    time.sleep(0.25)
                    no_of_clicks-=1
                except:
                    no_of_clicks-=1

            #find the user and the content of the user comments
            usernames=browser.find_elements_by_xpath("//div[@class='C4VMK']//a")
            elems = browser.find_elements_by_xpath("//div[@class='C4VMK']//span")

            #append it to a csv file
            for eleme, username in zip(elems,usernames):
                with open(csvexport, 'a') as csvfile:
                    writer = csv.writer(csvfile)
                    writer.writerow([url,username.get_attribute('title').encode("utf-8"),eleme.get_attribute('textContent').encode

openpost(browser,r"c:\mypath\links.txt",r"c:\mypath\allposts.csv")
```

Figure 4: Python script to open Instagram posts and parse and save their comments

The script on Figure 5 uses Selenium and the Firefox headless browser driver to perform the following steps:

1. Open the file containing the posts' links.

2. Visit the link in the headless browser session.

3. Find the button with the caption "Load more comments" and click on it 30 times. 30 proved to be the optimal number of clicks.

4. Between each click, sleep for 0.25 seconds to allow all non-photo and non-video contents to be loaded. This is crucial so that all the relevant HTML content is downloaded.

5. In the result HTML, find all of the div elements with class="C4VMK".

6. Locate the "/a" component inside it. This will contain the user making the comment.

7. Locate the "/span" component inside it. This will contain the comment contents.

8. Parse the text content from both HTML elements and save it using utf-8 encoding.

9. Append the results to a new csv file allposts.csv to keep them recorded.

PREPARING SOCIAL MEDIA DATA FOR ANALYTICS

The results of collecting the post comments can be further enriched. The following actions have also been done:

*   Searching for hashtags – as they occur quite frequently both in posts and in comments (in this paper this is done in SAS Visual Text Analytics)

*   Distinguishing between emojis and textual content in comments (in this paper this was done with some extra Python data manipulation steps)

*   Searching for other entities such as links or mentioned user names (in this paper this is done in SAS Visual Text Analytics)

As a result of running the data collection steps, altogether 2,059,070 English language comments were collected

Amadeus Software Limited, The Old School Hall, 11 Wesley Walk, Witney, Oxon, OX28 6ZJ
Tel: +44 (0) 1993 848010 email:info@amadeus.co.uk

Page 6 of 16

on the posts from the 13 analysed Instagram influencer accounts. Non-English comments have been filtered out, however, there are a reasonable amount of mixed language comments remaining in the data, as we will see later.

The next step is to upload our data to SAS Viya, namely an appropriate CASLIB.

For this, first we need to connect to a CAS session, then upload our data, and promote it so that other CAS sessions can access it. As a result, we will have a table named **allposts** in our PUBLIC CASLIB. From this point onwards, the power of the SAS Viya platform will be used for the Analytics Phase on the uploaded data. The actual tools being used include SAS Visual Text Analytics via Model Studio and SAS Visual Analytics.

Analytics Phase

TEXT ANALYTICS STEPS USING MODEL STUDIO

We will now use Model Studio to explore our Instagram data, identify and extract specific information. The following figure shows the pipeline used:
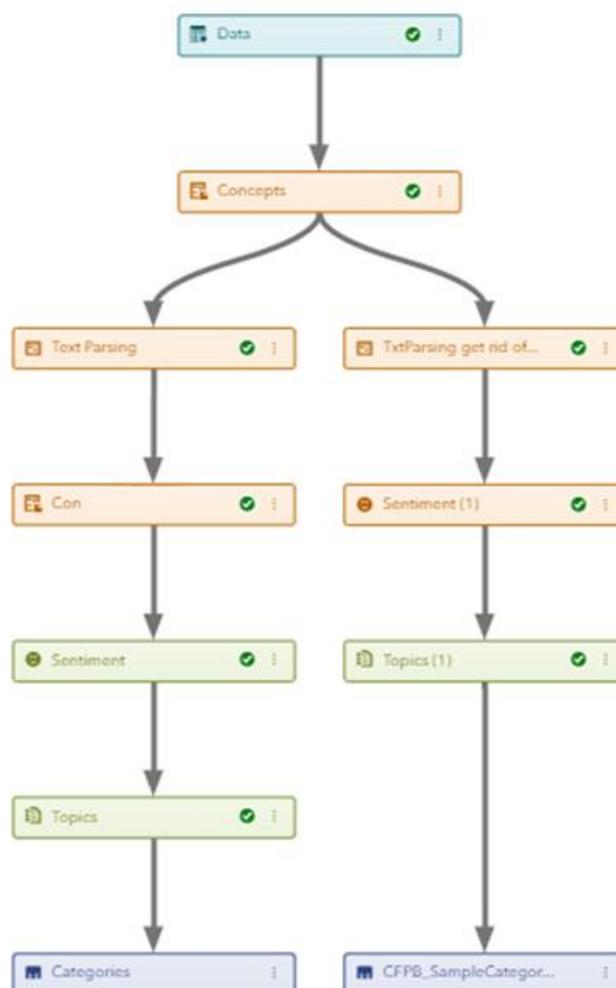

Figure 5: SAS Visual Text Analytics pipeline in Model Studio

As Figure 7 shows we have added several nodes to the default SAS Visual Text Analytics pipeline in Model Studio. The Model Studio pipeline on Figure 7 performs the following steps:

1. Data. The Data node points to the **allposts** table in the PUBLIC CASLIB library. We have to assign the variable containing the posts text as the variable with the Text role.

2. Concepts. The Concepts node after the data node defines some concept rules. We will use nlpOrganization from the 9 predefined concepts to mark organisations and potential brands for us, but we also define some custom concepts.

Tamás Bosznay, Amadeus Software, UK

The first custom concept rule is for hashtags. Figure 8 shows how the custom concept rule is defined for hashtags using a REGEX type custom concept rule (note that the hashmark character (#) is escaped using a backslash (\), this is because the hashmark character marks comments in these regular expressions and therefore it has to be escaped):
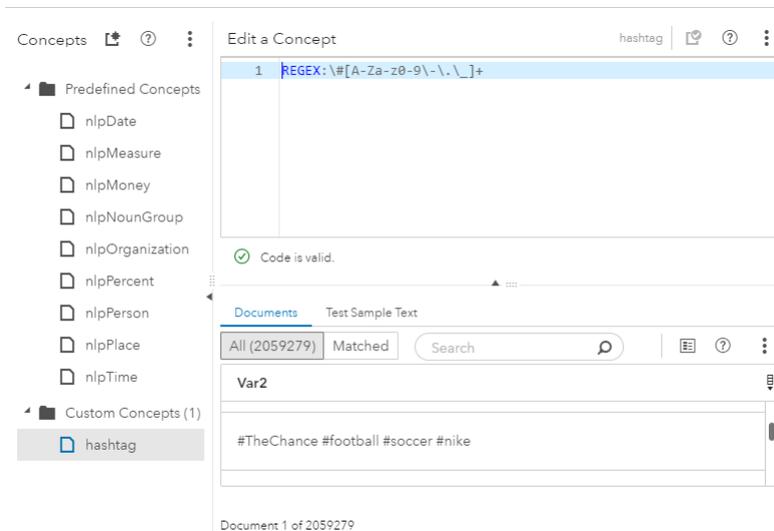
Figure 6: Defining a REGEX type custom concept rule to identify hashtags

3. After the Concepts node, the pipeline branches into two. The left branch is for raw text data analysis and the right branch is for cleaned text analysis (getting rid of hashtags, links and mentioned user names).

4. Left branch: Text Parsing. The Text Parsing node on the left branch is used to identify terms and to drop unneeded terms (the analysis on the left branch is using the out-of-the-box English stop list), where default settings and findings have been used. This means that those terms (words and expressions) are kept.

5. Left branch: Con. The node named Con (which is another concepts node) is used to analyse the terms marked as Organisations by the nlpOrganization predefined concept rule. We have augmented the list of organisations to identify the brands being mentioned in the posts and the comments. This augmentation was done searching for terms which are either marked as an organisation and are a product or service brand. We have also manually searched for some further brand names to make sure everything relevant is captured.

    Once we have identified the initial terms that are brands, we can use the Similarity scores in the Con concepts node to look for similar terms. This technique is needed to capture "spinoff" brands of main brands. For example, Nike might not always use the brand Nike for its posts, but might also use Nikefootball and Justdoit as hashtags or brands. After entering the term in the Textual Elements tile, we need to click on "Generate Similarity Scores" and then select further relevant items. Once we have selected all the relevant items, we can click on "Generate Concept Rule" to actually generate a concept rule that defines the augmented Nike brand.

6. Left branch: Sentiment. The Sentiment node on the left branch generates sentiment scores for us for the whole textual part of the comments.

7. Left branch: Topic. The Topics node on the left branch is used to generate topics in the comments using Singular Value Decomposition. The user here can play around with a couple of settings, including whether to use the different term roles (part-of-speech tags and concepts) in the topics or not, and also it is possible to determine the maximum number of topics generated. We will use our results from the topics generation later in SAS Visual Analytics visualisations.

8. Left branch: Categories. In the Categories node at the bottom of the left branch we can export our data via right-clicking on the node and selecting "Save Data table". This allows further analysis in potentially other Model Studio pipelines, visualising results using SAS Visual Analytics or even via calling CAS actions on the result table from Python via a Jupyter notebook.

9. Right branch: Txt Parsing get rid of….. This text parsing node excludes the terms from the analysis that are flagged by custom concept rules in the first Concepts node. Namely, we get rid of hashtags, and also mentioned user names in comments (which follow the @username pattern and therefore can be

Amadeus Software Limited, The Old School Hall, 11 Wesley Walk, Witney, Oxon, OX28 6ZJ
Tel: +44 (0) 1993 848010 email:info@amadeus.co.uk

Page 8 of 16

easily captured using a REGEX concept rule). The reason for getting rid of these is that they can affect both the topics generated and the sentiment calculated on the comments. We use two branches in our analysis to have the ability to explore and analyse comments both raw and cleaned.

10. Right branch: Sentiment (1). This node on the right branch is calculating sentiment for cleaned comments.

11. Right branch: Topics (1). This node on the right branch is used to identify topics in the comments. We will use our results from the topics generation later in SAS Visual Analytics visualisations.

12. Right branch: CFPB_SampleCategories. Using this node at the bottom of the right branch we can export our data via right-clicking on the node and selecting "Save Data table". As earlier, this allows further analysis in potentially other Model Studio pipelines, visualising results using SAS Visual Analytics or even via calling CAS actions on the result table from Python via a Jupyter notebook.

EXPLORING USER POSTS AND COMMENTS

As a couple of cleaning and enrichment steps have been performed, it is a good time now to explore our comments.
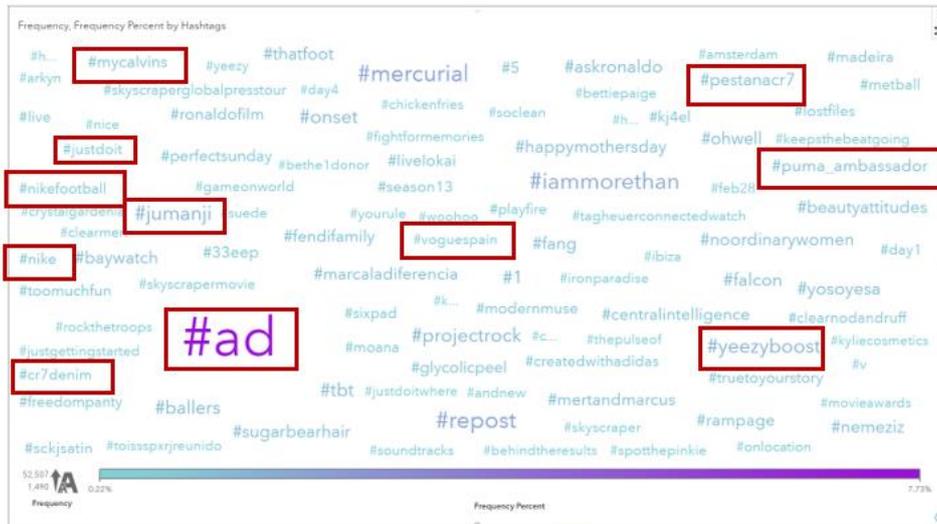
A good way to explore how different words and terms appear together is using the text parsing node in SAS Visual Text Analytics (interfaced in Model Studio). The text parsing node offers some interactive analysis tools, one of which is the term map. We just need to highlight a term and click on the "Term Map" icon for the map to be displayed. Figure 10 shows an example term map for the term beautiful:
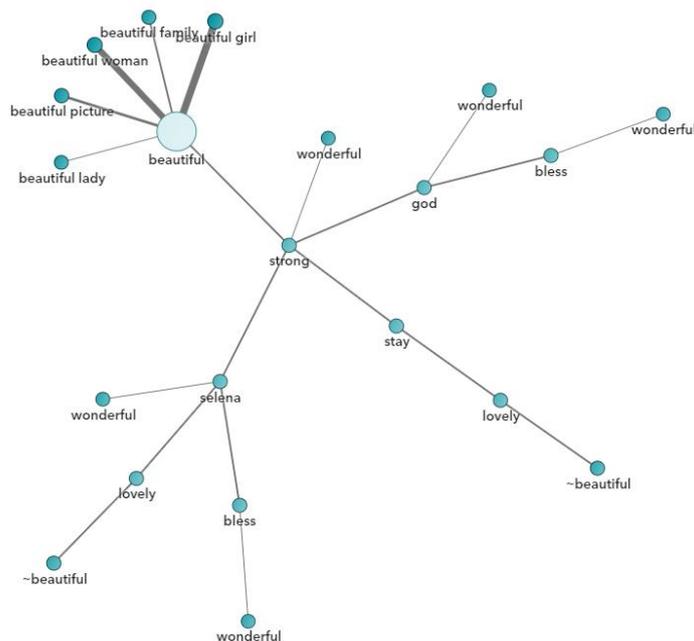
Figure 7: Term map of the term beautiful

The term map displays what other terms beautiful is frequently used with in comments. So here in our example, the term beautiful appears frequently together with the terms picture, lady, woman, girl and family the most. A bit less frequent occurrence happens with the term strong. The term strong is then also frequently occurring with terms selena, stay, wonderful, lovely, god and bless. The term map is a good technique to explore frequent phrases or motions in comments.

Another good technique to visually explore posts and comments is the Word Cloud visualisation available in SAS Visual Analytics. Figure 11 shows a word cloud:

Amadeus Software Limited, The Old School Hall, 11 Wesley Walk, Witney, Oxon, OX28 6ZJ
Tel: +44 (0) 1993 848010 email:info@amadeus.co.uk

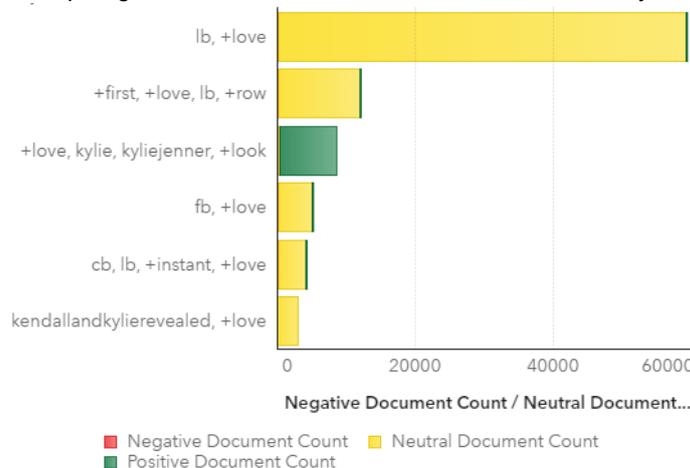Page 9 of 16

Figure 8: Word cloud of hashtags

The word cloud on Figure 11 is generated using the hashtags from the posts and comments collected. We first identified the hashtags with a custom concept rule in Model Studio, and then used only the hashtags for this visualisation. The most frequent hashtag is #ad – the one usually used if an influencer would like to show that their post is sponsored. However, in many other circumstances they do not use this hashtag, just directly the hashtag or hashtags related to the brand or brands they are posting content about. A few brand hashtags are highlighted on Figure 11. Other hashtags identified are not marketing or brand related, they are just general hashtags, adding extra information of the posts or comments in a concise way (or repeating information already provided in the actual photo or video and the textual part of the post or comment).

Using text similarity scores and automatic generation of concept rules and word clouds, we came up with a list of relevant and frequently appearing brands, which we used for identifying which original post by influencer was marketing related and which not. Later in this paper, we used this flag (marketing VS non-marketing) in further analyses and visualisations.

TOPIC IDENTIFICATION AND ANALYSIS

The Topics node in SAS Visual Text Analytics is a great tool to create topics from textual using the LDA (Latent Dirichlet Allocation) technique, which is based on Singular Value Decomposition. In our paper, we used the Topics node in Model Studio as it gives more interactivity. When analysing different influencers' behaviour, it makes sense to create and display topics separately for each influencer. So, in this paper we generated topics for each influencer to make comparisons. We visualised the final top topics using the Text Topics visualisation in SAS Visual Analytics.

Figure 13 and 14 show the topics generated and word cloud of the comments on Kylie Jenner's posts:

Figure 9: Topics of Kylie Jenner's comments

Amadeus Software Limited, The Old School Hall, 11 Wesley Walk, Witney, Oxon, OX28 6ZJ
Tel: +44 (0) 1993 848010 email:info@amadeus.co.uk

Page 10 of 16

Figure 10: Word cloud of Kylie Jenner's comments

The first thing we can see on the word cloud is that the most frequent word is "lb". This term is a social media specific abbreviation and stands for "like back". Loads of comments solely contain lb on Kylie Jenner's comments. Users who only comment in this manner tend to aspire for the influencer to like (and follow) them back. We can also see that "first" is also a relatively frequent word as it is usually a comment by the user who manages to write the first comment on an influencer's post.

When we look at the topics, we see that most of them contain these terms in one, or another format, but also contain positive words like love. The bar for each topic is coloured in the ratio of how many negative (red), neutral (yellow) and positive (green) sentiment polarity comments are there. In this case most of the topics have mainly neutral comments, and possibly this is due to the posts being casual photos, with no particular content, and also most of the commenters just send emojis along with a couple of words. However, one topic has rather positive sentiment comments, mainly the one that contains the term look.

Further explorations show that these posts (for which the comments have been made) tend to be using more marketing related hashtags and also the post photos are being more professional. We therefore can infer that in the case of Kylie Jenner these posts were rather the marketing ones, where the content (photo or video) also tends to be more professional (and maybe attractive) than the other posts of Kylie.

To deal with social media specific abbreviations, it is possible to use an augmented stop list of words in Model Studio.

Figure 15 shows the topics generated from the comments on the posts of the agency account Victoria's Secret:

Figure 11: Topics of Victoria's Secret's comments

Amadeus Software Limited, The Old School Hall, 11 Wesley Walk, Witney, Oxon, OX28 6ZJ
Tel: +44 (0) 1993 848010 email:info@amadeus.co.uk

Page 11 of 16

We can see from Victoria's Secret's comments that they are quite different from Kylie Jenner's. They contain quite positive words like love, cute, beautiful and nice. Quite a few topics have overwhelmingly positive comments with some neutral though. As Victoria's Secret is an agency account, all the posts they make can be considered marketing, no matter which actual model or product their photo or video displays. Also, all photos and videos are very professional and of very high quality. This difference between a celebrity's comments and an agency account shows how marketing posts can invoke more positive engagement from the followers.

Now let's see some other influencers' topics. We will compare two footballers: Cristiano Ronaldo and Messi. Figures 16 and 17 show their top topics.
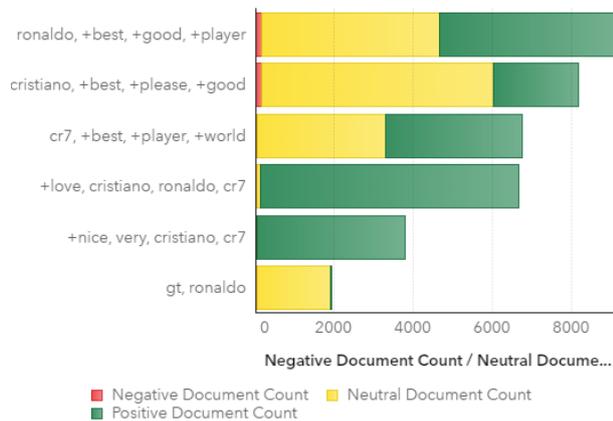
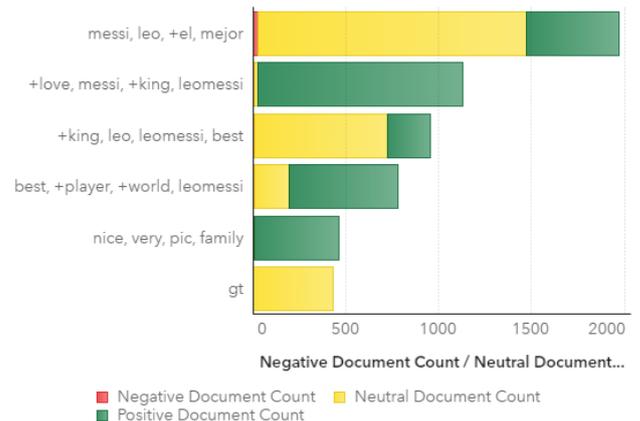**Figure 12: Topics on Cristiano Ronaldo's comments**   **Figure 17: Topics on Lionel Messi's comments**

When we look at the topics, similar patterns can be seen for both footballers. The top topics tend to be about how great the players are (king, mejor, player, best, player, world), but also how nice their images are. Now both Ronaldo and Messi tend to post photos (and videos) about quite a few areas. They post loads of pictures about coaching, matches, successes, when they score some goals or provide assists, then also marketing content about some of the brands they work together with. They also post pictures together with their kids and family, and also general casual photos like visiting the gym or their meals.

We see from the topics and their sentiment polarity that they have about the same ratio of neutral and positive comments, but they also have some negative ones as well. This is not a surprise as both footballers attract many followers, but they also have some users who they irritate. Altogether, their marketing related posts tend to generate engagement, either positive or sometimes negative. This shows that they are effective in their influencing behaviour and that organisation can definitely benefit from working them as influencers.

AUGMENTED SENTIMENT ANALYSIS

Now we have seen the topics among some of our influencers, let's now aggregate the general sentiment in the comments about each influencer.

Until now, we have dealt with the textual part of comments and their sentiment. We have used the Sentiment node in SAS Visual Text Analytics to generate sentiment scores and polarities. In the Sentiment node, it is possible to customise the sentiment analysis with uploading a custom sentiment model. This is useful when working in a domain where specifics are very important and therefore general sentiment models are not appropriate.

However, either with the default or a customised sentiment model, SAS only deals with sentiment of the textual part of comments. There are loads of smileys, emojis and other symbols in social media comments though, which hold some specific sentiment polarity. This type of reaction on posts cannot be ignored when analysing influencer activity. So we need to come up with a way to estimate sentiment polarity for emojis.

The nltk package in Python has a component VADER, which does calculate sentiment of emojis. Figure 18 shows a Python code snippet to do this:

```python
import nltk
from nltk.sentiment.vader import SentimentIntensityAnalyzer
def compsent(sentence):
    score = analyser.polarity_scores(sentence)
    return score['compound']
```

```
In [7]: compsent('😄')
Out[7]: 0.3612
```

```
In [8]: compsent('😄 😄')
Out[8]: 0.6124
```

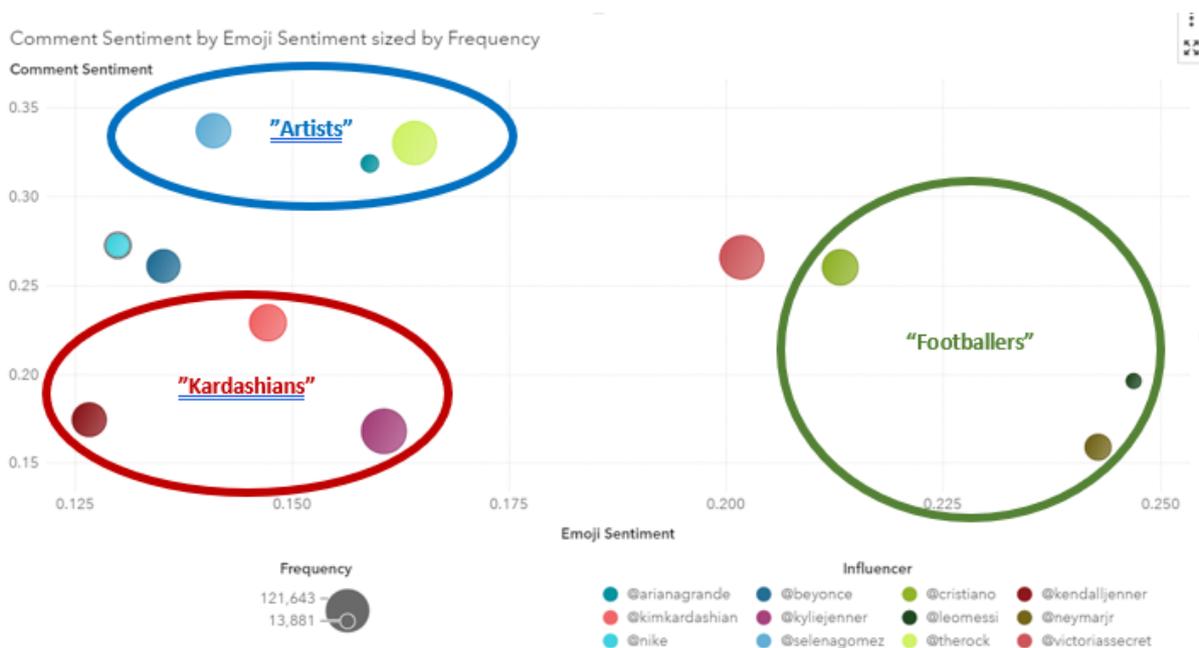**Figure 18: Python snippet to calculate sentiment score of emojis**

In the Python example above, we first import the required packages and then define a small user function called compsent which returns the compound sentiment score for a given string. The two examples shown above display the sentiment score of one and two smiley emojis. From here we can see that repetitions of the same emoji mean different sentiment scores, as it is expected.

The sentiment scores generated from Model Studio are measured on a scale from 0 to 1, with 1 being the maximum positive polarity and 0.5 being neutral. The VADER sentiment scores on emojis are however measured from -1 to 1, with 1 being the maximum positive polarity and 0 being neutral. To be able to use sentiment scores combined from Model Studio and Python, we need to transform their scale in order to be the same.

Creating visualisations

Now we are going to create some visualisations using SAS Visual Analytics to display the results of our analysis.

We first have a look at how the average textual sentiment and emoji sentiment differs on the comments of the different influencers. Figure 19 shows a bubble plot displaying these. On the figure we can see that there are differences among the different influencers. The chart shows values for all the comments, not just the marketing related ones. Emoji sentiment is displayed on the x axis and comment sentiment on the y axis. On our chart, the "Kardashians" show both relatively low average comment sentiment and emoji sentiment. We have seen that they have loads of neutral comments from followers that would like to be followed by them. The "Footballers" (Ronaldo, Neymar, Messi) show much higher emoji sentiment in average – this probably means more engagement even from a casual post by them. The "Artists" (Selena Gomez, Ariana Grande, The Rock) show much higher comment sentiment, which could be seen as a way of the followers showing true respect or even true love about them.

Figure 19: Sentiment of comments and emojis on different influencers

Amadeus Software Limited, The Old School Hall, 11 Wesley Walk, Witney, Oxon, OX28 6ZJ
Tel: +44 (0) 1993 848010 email:info@amadeus.co.uk

Page 13 of 16

Tamás Bosznay, Amadeus Software, UK

These differences in the average sentiment do show how the different influencers work and influence followers and this gives insight about their effectiveness.

Finally, we aim to visualise the influencers and their networks. For this, we consider our dataset with all the comments as a network (edge-type) dataset, where edges are the comments the followers have made on the influencers' posts. Visualising the network of influencers can also help evaluating their activity and their effectiveness.

To do this, we use the Network Graph visualisation in SAS Visual Analytics. First, we have to split our comments dataset into a "marketing" and a "non-marketing" network dataset, based on the flag we created earlier to show whether a comment is on a post which is marketing-related or not.

The Network Graph won't be able to directly visualise our networks with hundreds of thousands of nodes and millions of edges as it is limited to display 25,000 data points. Therefore, we have created a sample of the datasets using stratification. This way we can ensure that the network graph is showing the network structure in a realistic way. We have also coloured our network graph using the sentiment values, red showing negative, brown showing neutral, green showing positive sentiment on comment emojis. Figure 20 shows a non-marketing related network of the "Kardashians" and the three footballers (Ronaldo, Neymar, Messi). On this network chart, each edge is a comment, displayed as pointing from the influencer to the commenter. A thicker arrow means more comments by the same user on the same influencer. Each influencer has his or her own community: a circle containing followers that only commented in their posts. Further connected nodes are followers who commented on more than one influencers' posts and therefore they "stand between" their influencers. As the chart is displaying only a sample, the number of followers commenting on more than one influencer is in fact much higher than single digits. These common commenters among different influencers can be used generally to find new people to influence in a quick way. If someone follows both influencer A and B, influencer A can try to grow its network using the common followers, as they are likely to convey the content of influencer A to the network of influencer B via shares, likes and comments. The closer two communities are, the more common commenters they have on the chart.
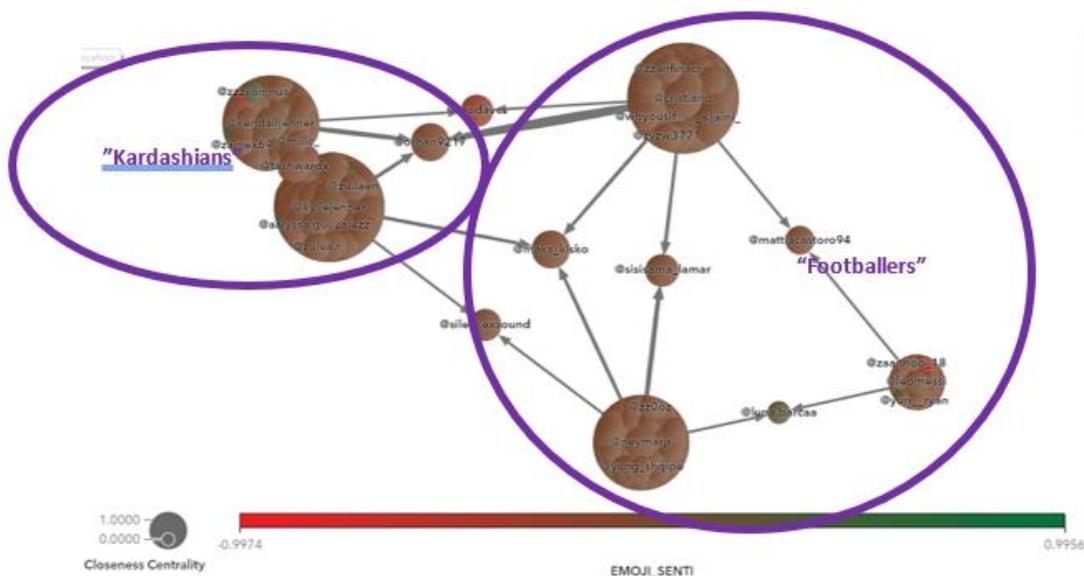
Figure 20: Non-marketing network by emoji sentiment

The same is measured by closeness centrality, which is the size of the discs. Probably it comes as no surprise, that the "Kardashians" have communities standing close to each other, as a result of a reasonably big group of common commenters. Similarly, the footballers have a reasonably big group of mutual commenters. The chart on Figure 20 (the non-marketing network) is fairly brown in terms of emoji sentiment, which in this case means neutral.

Amadeus Software Limited, The Old School Hall, 11 Wesley Walk, Witney, Oxon, OX28 6ZJ
Tel: +44 (0) 1993 848010 email:info@amadeus.co.uk

Page 14 of 16

Tamás Bosznay, Amadeus Software, UK

Figure 21 shows a marketing network of the same influencers (the "Kardashians" and the footballers):
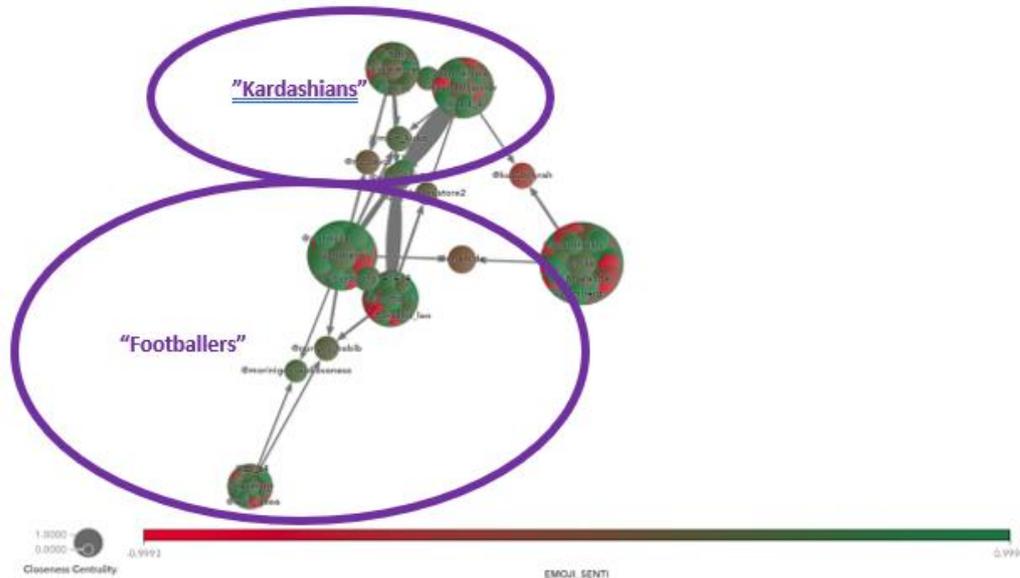
Figure 21: Marketing network by emoji sentiment

In contrary to the non-marketing network, the marketing network on Figure 21 does show emoji sentiment relatively polarised: some negative comments, but mostly positive. This is true for both the "Kardashians" and the footballers. This altogether shows that marketing related posts can generate more engagement, especially more positive engagement. The reason for this can be the higher quality of the marketing posts by influencers, compared to normal casual non-marketing posts. Top influencers usually work with professional photographers for marketing posts, whereas for casual ones, they might just create content on their own (they might even take selfies with their smartphones). This difference in quality can be one reason for marketing related posts generating more positive engagement.

On the other hand, this actually means that there is definitely space for marketing via influencers, as they can convey messages about products or services very effectively and this does generate positive reactions among their followers. The more positive the reactions of the followers are, the more likely it is that they will consider buying those products or services.

CONCLUSION

Analysing social media data related to influencers reveals the characteristics of their posts. When analysing comments on the influencers' posts, topics and sentiment patterns directly show not only how their network of followers perceive their activity and content provided, but also gives insights about the effectiveness of marketing in those networks. With such insights, organisations can assess and decide who is the best influencer they should work with to maximise return on influencer marketing.

Instagram data can be collected via web scraping in Python, however, Instagram APIs can be directly consumed from SAS. This data can then easily be used within SAS Viya applications such as Visual Text Analytics (either via Model Studio or from Python via SWAT) to generate topics, identify brands, hashtags and other important information. Analytics results generated through the SAS Viya applications are best visualised with SAS Visual Analytics. SAS Visual Analytics can easily present the data analysed and help by using the relevant visualisations to derive the best influencers to work with.

REFERENCES

[1] SAS (2017) SAS® Visual Text Analytics 8.3: Programming Guide. Python Programming Language [Online] Available from:
https://documentation.sas.com/?docsetId=casvtapg&docsetTarget=p0vouc3o8s7gq0n1p1b68jydb8id.htm&docsetVersion=8.3&locale=en [Accessed 21/03/19].

[2] SAS (2017) SAS® Visual Text Analytics 8.3: User's Guide [Online] Available from:
https://documentation.sas.com/api/docsets/ctxtug/8.3/content/ctxtug.pdf?locale=en

Tamás Bosznay, Amadeus Software, UK

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author:

Author Names	Tamás Bosznay
Company:	Amadeus Software Limited
Address:	The Old School Hall, 11 Wesley Walk, Witney, Oxon OX28 6ZJ
Phone:	+44 (0) 1993 848010
Email:	Tamas.bosznay@amadeus.co.uk
Web:	www.amadeus.co.uk

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

Amadeus Software Limited, The Old School Hall, 11 Wesley Walk, Witney, Oxon, OX28 6ZJ
Tel: +44 (0) 1993 848010 email:info@amadeus.co.uk

Page 16 of 16