RightFluencer

https://csil-git1.cs.surrey.sfu.ca/mkselvak/rightfluencer

rightfluencer.ml

By Arin Ghosh, Karanjit Singh Tiwana and Manoj Karthick Selva Kumar

Product

RightFluencer provides a search cum recommendation engine where marketers can enter a product and category to find right influencer for the product by analyzing their Instagram, Facebook, Twitter and YouTube profiles and find their niche/expertise. The marketers can then get more detailed insights about the influencer and visualize the metrics related to the influencer. RightFluencer also allows influencers to gain deeper insights about their online presence and understand their strongholds and weaknesses.

Motivation and Background

Influencer marketing can generate excellent returns on advertising spending if done right. The industry has been growing exponentially for the past few years and recently hit the $2 bn mark [1]. Lots of firms have jumped on the influencer marketing bandwagon without ever thinking about what they had to do, and as a result, these firms have had forgettable, unproductive and expensive experiences [2]. Forbes, in a recently published article, stated that identifying the right influencer is the biggest challenge of all for marketers. [3] This formed the motivation for us to build a product that helps overcome these challenges.

Problem Statement

In spite of the tremendous growth, the influencer marketing industry is still reliant on market watchers and non-scientific methods for finding the best influencers. Other online influencer marketing tools use surface level metrics such as followers and likes and do not exploit the content posted by the influencer to make informed decisions. Tools available online that quantify influence are not tailored to understand the product that a brand wants to market while analysing potential influencers. This provided us with the impetus to develop a product that makes recommendations by understanding the content and expertise of the influencer and the product to be marketed.

The Key differentiators for RightFluencer compared to other products such as Klout or tweetdeck are:

1. Collective analysis of multiple social media platforms using both metrics and content.
2. The influencer score is based on the product and category to be marketed.
3. Analysis of images in posts to understand what products the influencers are experts in.
4. Getting to know the influencers in-depth using their YouTube videos' closed captions.

This is a challenging problem because of the inherently subjective nature of influencer marketing, the difficulty in quantifying the power to sway the decision making of consumers and the lack of any universally accepted forms of measuring influence.

Simplified View of Project Pipeline

Technology Stack

Web Server and Hosting	Flask, mod_wsgi + Apache, Google Compute Engine
Visualization	Plotly, matplotlib, pyLDAvis
UI / UX	HTML5, CSS3, JavaScript, Bootstrap 4, Semantic UI
Data Analysis	KerasR, gensim, nGram, WordNet, PyMongo, R dataframes Watson Personality Insights API, Apache Spark, Pandas
Data cleaning	Spark DataFrames, Pandas, NLTK
Data Storage	MongoDB
Data Collection and Aggregation	RESTful API Clients, BeautifulSoup, Spark, pandas, Mongo-QL
Data Sources	Twitter, Instagram, YouTube, Facebook, Klout

Data Aggregation

An integrated data store using MongoDB was built to combine information from the various data sources. The database was de-normalized for high throughput and low latency. The data sources, the corresponding information collected and the technique used is described below:

Data Source	**Features collected**	**Methodology**
Twitter	Profile information, tweets, followers, likes and hashtags	RESTful API client to extract information from the Tweets API
Instagram	Profile info, images, captions, likes, comments and hashtags	Web scraper using requests and BeautifulSoup for extracting images, profile and posts
YouTube	Profile and video info, likes, views, comments, closed captions	OAuth enabled RESTful API client to extract data from YouTube Data API, youtube-dl
Facebook	Page and post information, description, likes and shares	OAuth enabled API client using the Facebook Graph API

Methodology and Workflow

- Data was collected from Twitter, Facebook, YouTube and Klout through REST API Clients. Data from Instagram was collected by scraping. All data was converted into JSON documents for serializability purposes since the data was semi-structured and lacked a fixed schema.
- The data was then aggregated with the Twitter handle as the primary key into MongoDB using PyMongo. Data from the various sources for every influencer was stored in a single Mongo collection for high throughput and low latency. Data related to each of the other data sources which are large in number such as tweets, posts, images were stored as separate collections.
- The data cleaning step involved removing hashtags, @mentions, URLs, special characters and emoticons. The data was lemmatized to group together words having the same inflected form and then tokenized to remove stop words in the text through an extended stopword list. The cleaned data was then pushed into MongoDB.
- The Data Analysis stage of our project involved four main types of analysis. The data cleaning was done using Apache Spark RDDs and dataframes for scalability and performance.

    - **Metrics**: KPI's and absolute metrics were directly obtained from the data sources. Derived metrics like audience interaction, popularity, audience growth and post frequency were obtained by aggregating on the fly using MongoDB aggregation operators. Additional aggregations for which Mongo's aggregations were insufficient was done using pandas.
    - **Topic Modelling**: The cleaned data was used as a training corpus to build models using LDA and LSI. LDA was chosen due to its flexibility and it's probabilistic nature that allowed us to use the values during influencer score calculation. Choosing two approaches also helped us cross-verify the results. We found that LDA consistently outperformed LSI.
    - **Image recognition:** Pretrained VGG-16 and VGG-19 models were further trained using the instagram images of the influencers. We used a combined weighted score of the two CNNs' results to get the most popular items in influencers' image by frequency and confidence.
    - **Personality Insights:** The personality of the influencers such as Extraversion, Openness, emotional range, etc were analysed to help understand what traits the influencers had since certain traits are more valuable than others for certain categories. This helps marketers understand the influencers persona at a glance. Post data from the various data sources were combined using Spark and analyzed using Watson Personality Insights API.
- The Flask web server was used to create the search engine and dashboards for the influencers. The server dynamically fetches data from MongoDB related to the search and influencers on-the -fly. The aggregations and metrics mentioned in the data analysis stage is calculated at run time.
- The aggregated data is then displayed on the web pages using Plotly and matplotlib. An important advantage of this approach is that the plots are dynamically generated, ie. if the underlying data related to the influencers changes, there will be no need to re-render the plots. The results of topic modelling were visualized using pyLDAvis in the form of a inter-topic distance map.
- The web pages were designed using bootstrap and Semantic UI and created and rendered using the Jinja2 framework. The completed application was then hosted on the Google Compute Engine. The application uses flask + mod_wsgi to display the web pages using the Apache web server.

Recommendations and Scoring

- Posts, tweets and videos of from all the influencers are analyzed for the search term using an nGram model that uses unigrams and bigrams. The results of the n-gram model are laplace smoothed to obtain fail-safe probabilities. n-gram being a probabilistic language model allows us to calculate probability of occurrence of the search which is used in the weighted scoring model.
- **Scoring**: Every influencer is given a category rank and product rank using a weight scoring model. The audience growth, popularity, #posts, followers and interaction are combined together to calculate a static category score and this category score decides the ranking of the influencers within the category. The product score for the influencers is calculated on the fly using the search query. Their product score is calculated using the frequency and relevance of the search terms within the data collected. The results from the topic modelling and image processing to recognize what the influencer's expertise are also taken into account during score calculation. All these scores are weighted into a combined product score for the influencer and ranked accordingly.

***Visualization*: Influencer's expertise [Topic Modelling]** **Target:** Marketers

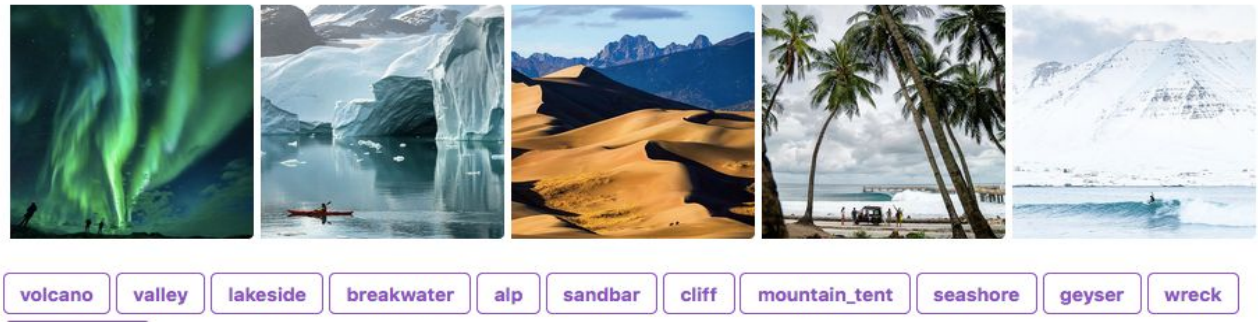
Problem: A famous food brand has introduced a new pastry kit that makes it easier for DIY bakers. Analyzing the posts of the influencer across platforms can help marketers understand the what the influencer's niche or expertise is and if he/she is a good choice for the marketing campaign.

Insight: This image shows the objects most frequently found in the images posted by Izy Hossack, a popular food influencer. The result shows that Izy Hossack is more interested in Baking through the terms "baked",

"bagels", "rye", "porridge" - all items that are baked or involve baking. This goes to show her expertise is baking and that is a perfect fit for the baking brand.

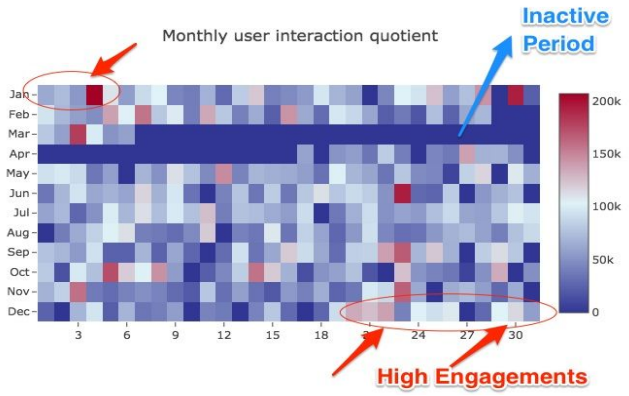
***Visualization*: Influencer's expertise [Image Recognition]** **Target:** Marketers

Problem: Thomas Cook has created a new travel package exploring Canada's northern lights, coasts, mountains and the arctic and wants to find social media influencers to promote their package and brand.
Insight: The image below shows that travel influencer Chris Burkard is an adventurous guy whose images show the northern lights and icy mountains. The popular objects in his pictures are "alps", "seashore", "cliff", "mountain_tent". Thus a marketer can use the images and the objects in the images to ascertain if a particular influencer is a good choice for the marketing campaign.



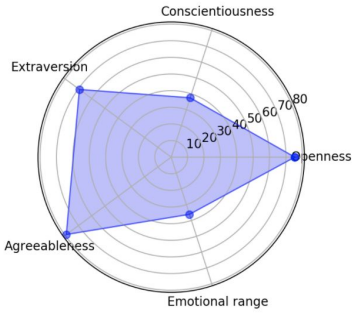
***Visualization*: Audience Engagement [Heatmap]** **Target**: Marketers, Influencers

Problem: Some marketing campaigns are seasonal. For example for a fitness club would like to hire a fitness influencer to promote word of mouth around the New Year, when people take up new resolutions The audience engagement plot can help a brand identify if a social influencer who they plan to hire can be effective during that time of year.
Insight: Take a look at Rachel Brathen, #9 in Fitness. By looking at her Monthly User Interaction Quotient we see that she stirs up conversation during the turn of the year amassing 100k+ audience engagements during the time but inactive from March to mid April. If a marketer is looking for a fitness influencer, she seems promising for resolution targeted lifestyle marketing campaigns around the holiday season.



***Visualization*: Personality Insights [Radar Plot]** **Target:** Marketers

Problem: It is very important to hire influencers who reflects the same values as the brand he/she represents. Often times, marketers wonder if the influencer shows the right attitude that fits with the ethics and values of the brand and the campaign.
Insight: This plot shows the personality analysis of Lilly Singh [superwoman], a popular entertainment influencer. The radar plot shows that Lilly is intellectually curious, emotionally-aware, and willing to try new things. Also, she is spontaneous person. If a particular marketing



team looking for a energetic and curious social influencer with high throughput engagement, she would be a very good fit. *[other visualizations and insights added in the appendix]*

Evaluation and Improvement

The scores obtained through our weighted scoring model shows positive correlation (*spearmanr*=0.4 *and pearsonr*=0.3) with Klout score, a popular influencer marketing and score tool available online. We have also integrated a rating and review system for as a feedback mechanism to analyze the quality of the results. Since there is a lack of product based influencer recommendation system in the market, and the subjective nature of influencer marketing, there is no ground truth to explicitly evaluate the results against. Based on our domain knowledge and discussions with few marketing industry professionals we were able to confirm that the rankings and recommendations produced by the system tends to be accurate more often than not. In the 10th Annual Shorty Web creator awards on 15th April, MKBHD won creator of the decade and Rossana Pansino won Food creator of the year award. These influencers are ranked #1 in the Tech and Food category respectively according to RightFluencer's algorithm and this is real proof of the accuracy of our product.

Lessons

- ❏ Since open APIs are constantly evolving and introducing restrictions, enterprise access and/or partnerships might be required to maintain system functionality up to date.
- ❏ During data collection process, it is imperative to affix an uniform timestamp format, to make time series calculations easier and consistent.
- ❏ Many influencers' followers may be bots. Methods to identify bots need to be developed in order to get a more accurate picture of the influence exerted by a social media influencer.

SUMMARY

Data sources, Getting the data	Twitter, Instagram, YouTube and Facebook were the important data sources. Data was obtained from these sources using REST API Clients and Web scraping.
ETL	The data was cleaned by removing stopwords, emojis, URLs, etc, lemmatized and transformed into JSON for serializability and pushed into MongoDB.
Problem	Current marketing tools do not take into account the content, images and videos and the product to be marketed while scoring and suggesting influencers.
Algorithmic Work	Analyze influencers images using CNNs and find topics in their posts using LDA/LSI and to find their area of expertise, find personality of influencers to understand them better, used n-grams to estimate relevance of product searched for, weighted scoring model integrating metrics, image and topics relevancy, Find related products through WordNet for higher quality results and suggested search.
Scalability	Scalable data transformations and operations through Apache Spark and storage using MongoDB; High emphasis on scalability by careful usage technologies that scale well with larger data; search engine has high efficiency and low latency
UI	Interactive web dashboard built using Bootstrap and Semantic UI, Hosted on Google Cloud platform, web pages served using Flask web server.
Visualization	Dynamic plot generation through Plotly, data aggregation for viz on-the-fly, wide range of visualizations such as bar/bubble plots, timeline plots, waterfall plot, heatmaps, image and topic pills used to help understand the influencers
Technologies	Wide range of technologies such as Spark, Flask, MongoDB, Plotly, gensim, NLTK,

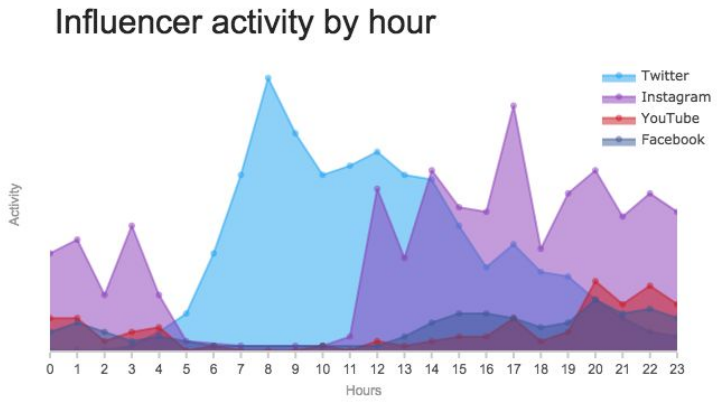
	KerasR were used throughout the lifecycle of the app. Hosted on Google Cloud.

Appendix

This appendix contains more visualizations available online at rightfluencer.ml to understand the influencer activity and expertise and help marketers make the right choice.

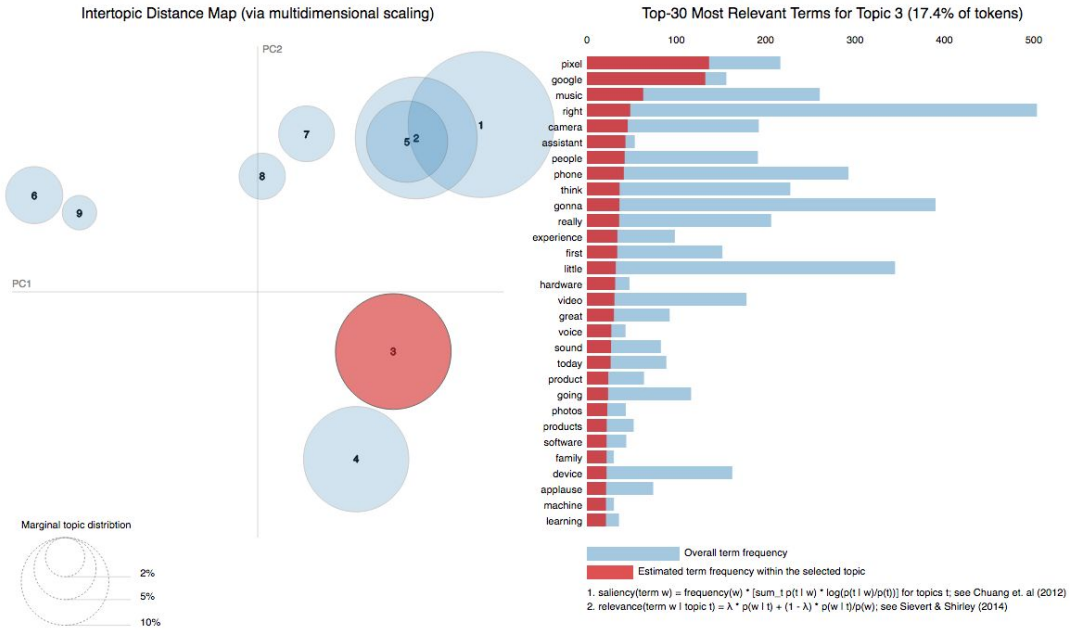
***Visualization*: Influencer Activity [Timeline plot]** **Target**: Marketers, Influencers

This plot allows marketers understand which time of day the influencers are active in which platform to time their marketing posts. A common scenario is that influencers tend to be active on Twitter during the day and Instagram during the after-hours. (Plot shown here from: MKBHD)



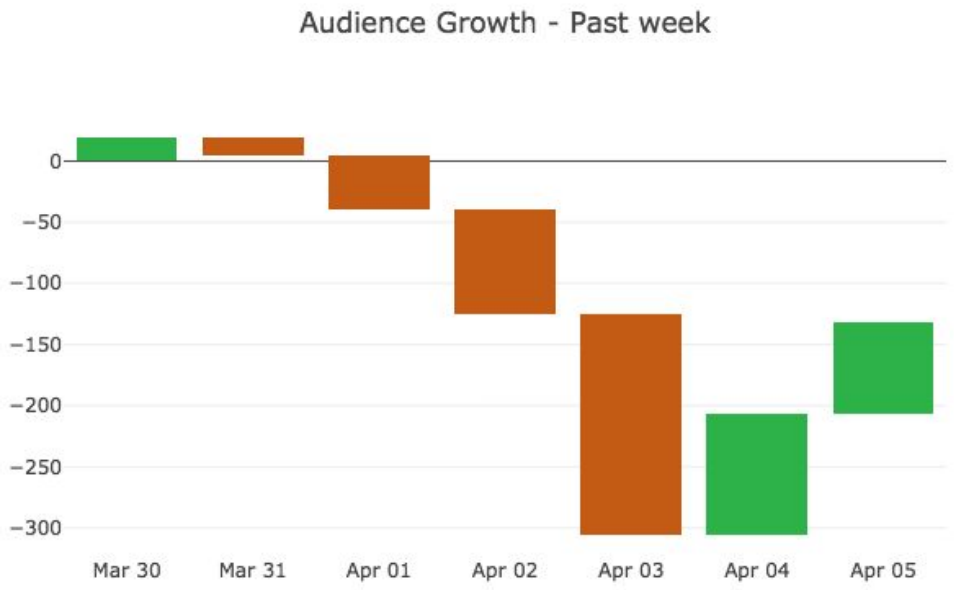
***Visualization*: Influencer Expertise [Topic modelling - LDA]** **Target**: Influencers, Marketers

This plot shows the results of topic modelling on YouTube video data from the popular tech influencer Unbox Therapy. This result is interesting because Clusters 3 and 4 are separate but closer to each other.
Cluster 3 refers to products and items from the ***Google*** ecosystem - "*google*", "*pixel*", "*assistant*", "*camera*"
Cluster 4 refers to products and items from the ***Apple*** ecosystem - "*iphone*", "*apple*", "*watch*", "*iphone* 8, 10"



***Visualization*: Audience Growth [Waterfall plot]** **Target**: Influencers, Marketers

This plot shows the audience growth in the past week for travel influencer Murad Osman. Murad has had a steady drop in audience from Mar 31 and reached a peak low on Apr 03 before rebounding the next day. This can help influencers understand why they are losing followers (maybe due to post or opinion) and help marketers analyze the recent momentum of the influencers they plan to hire.



References

[1] Influencer Marketing in 2018: Becoming an Efficient Marketplace - Adweek (Link)
[2] Influencer Marketing challenges - InfluencerMarketingHub (Link)
[3] The Three Biggest Influencer Marketing Challenges And How To Overcome Them - Forbes (Link)