Advertising Brain for Programmatic strategies

Overview:

The world of online advertising has witnessed tectonic shifts in the last decade. The evolution of digital and mobile advertising has disrupted the traditional advertising space confined to print and television and has introduced new execution strategies to target Audiences that are driven by personalization and are more relevant compared to traditional "spray and pray" techniques used for mass targeting.

Programmatic media buying, marketing and advertising is the algorithmic purchase and sale of advertising space in real time. During this process, software and data driven approaches are used to automate the buying, placement, and optimisation of media inventory via a bidding system.

Programmatic play allows the "owner/brand" to tailor a specific message and creative to the right person, at the right time in the right context – using audience insight from the brand (the customers you want to target) around the kind of audience they want to target.

Underlying all the disruption, in Advertising Industry, is the flood of data-- big data-- and superior data processing techniques, *Machine Learning and Deep Learning*, that help organize data into audience segments.

NLP and text processing plays a pivotal role mining the data and automating the datapipe that consists of heterogeneous, structured & unstructured, data sources -- like location data, Point of sales (POS) data, online transaction data, social media data, etc.

The focus of the project would be mine social media data, a rich source of behavioural data, to create programmatic targeting strategies that can be used by the media buyer to bid/buy digital media

Programmatic Targeting Strategies:

Programmatic buyers/traders use following strategies buy/bid for digital media.

- Segment based targeting.
    - *Segmenting a broad target* ***market*** *into subsets of consumers who have common interests and designing strategies to target them.*

- Keyword targeting

    - *Bidding for Keywords that resonates with particular demography, gender or segment of Audience.*

- Retargeting
    - *Target consumers based on their previous Internet actions to help companies reach target audiences who don’t convert right away*

Data Classification & IAB Segments:

The Interactive Advertising Bureau (IAB) has developed and standardized Data Segments & Techniques Lexicon that provides a framework to help understand how all of the data components work together to form the critical audience segments that enhance advertising value.

The first tier is a broad level category defined as a targeting depth of either: category/portal, site section, or page. Tier 2 categories and greater are additional categories nested under Tier 1 categories. Both tier 1 and tier 2 categories are formerly established for the IQG Program so that content classification can be consistent across the industry.

The links provide more details on IAB, Data sources and data classification.

- https://www.iab.com/guidelines/iab-quality-assurance-guidelines-qag-taxonomy/
- https://www.iab.com/news/data-segments-techniques-a-new-lexicon/
- https://www.iab.com/guidelines/social-data-demystification-best-practice-2/

Project Intent:

The intent is to classify user conversations in Twitter,using machine learning algorithms, as sets of IAB segments that can be used as a trading strategy for media buying.

Problem statement:

Behavioural data, *though difficult to mine*, promises rich “information gain” if mined correctly. Social media is a rich source of “Behavioural data”.

Can we use Twitter data stream, *Twitter is public & free*, to mine behavioural traits and classify conversations into various IAB segments to create media buying strategies.

The results of behavioral mining should predict

- Trending level 1 & level 2 IAB segments that a marketer/media buyer can bid.
- Trending Keywords within each segments that can be used by media buyer to bid for right keywords.
- Correlation amongst different IAB segments -- so that media buyer can target audiences in different segments a user is likely to visit in his internet journey.

Datasets and Inputs:

The experiment will use Twitter data stream. Twitter provides, users, access to its historical tweets and Live tweets.
The experiment will create a stub that mines Tweets and stores them locally, in memory, flat files on disk or on Database, that can be analyzed for further analysis by machine learning toolkit -- also developed as the part of the experiment.

Twitter data is quite tricky to handle as it has a 140 character limit. So any user would tend to convey his/her idea in those 140 characters. This may result in shortened word forms and emojis being used in the tweet.

The experiment chooses only IAB categories specific tweets. Hence, the tweets are ring fenced around 10 broad categories in IAB. The training data for the tweets are obtained first by defining IAB category specific keywords dictionary and looking up for those past tweets that mention one of these keywords and then programmatically label the tweets to create the training data set.

The training data distribution for various IAB categories is summarised as follows

IAB Category	**Total Tweets**
Arts & Entertainment	230432
Family & Parenting	235768
Fashion	258808
Food & Drink	283022
Hobbies & Interests	179218
Religion & Sprituality	274056
Science	281992

Society	258812
Sports	340703
Technology & Computing	259477
Travel	288630

The project would employ a convolutional neural network (CNN), explained in Solution statement below, for predicting and classifying Tweets and we would use Twitter Word2Vec for training the model, http://www.fredericgodin.com/software/

IAB-taxonomy details of Level 1 & Level 2 IAB segments against which the Tweets are classified is detailed in the Appendix section .

Solution Statement:

The experiment would attempt to create an “Advertising Brain”, using Machine & Deep learning techniques, that classifies Tweets against the IAB segments.

Tweets would be classified, against IAB segments, based on its content and would be further mined to extract trending keywords.

Any tweet would be classified into one of the 11 broad IAB categories such as

Arts & Entertainment
Family & Parenting
Fashion
Food & Drink
Hobbies & Interests
Religion & Sprituality
Science
Society
Sports
Technology & Computing
Travel

e.g Consider the following Tweets, extracted from the Twitter stream.

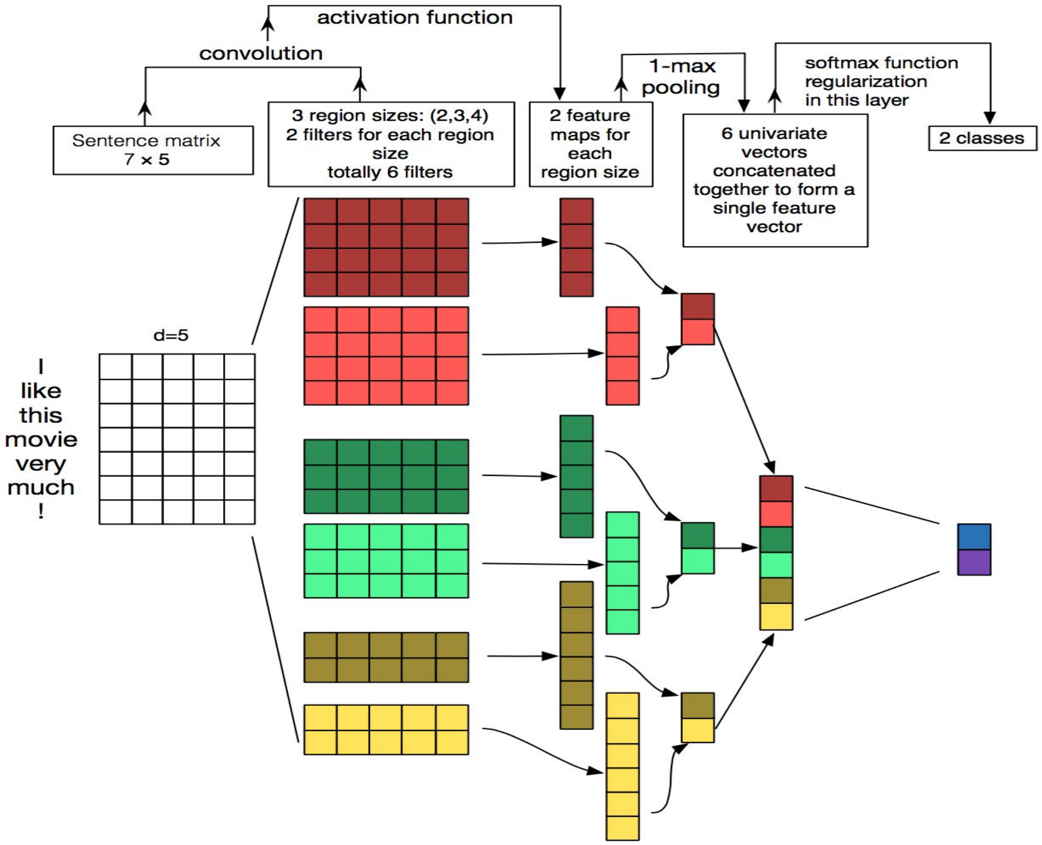
I love to Travel and enjoy Asian cuisines,

Traveling to Bali was fun. Planning for a Sydney trip in December.
Expecting a thriller from Chelsea & Manchester City match this Saturday
India has dominated England so far in the test series

Our “Advertising brain” would --

- Classify these Tweets belonging to Travel (First two) Sports (last two) and Food (first Tweet) categories.
    - Level 1 IAB segments.
- It will further classify Tweets as Football, Cricket (3rd & 4th Tweet) as level 2 IAB segments.
- It will try to predict the trending IAB segments -- the once generating maximum number of traction in Twitter-- and give an idea to the media buyer the segments worth bidding/buying.
- It will create a word count of the Keywords, for each segments, to predict the keywords worth bidding.
- It will further create the correlation amongst IAB segments.
    - When people are Tweeting about Travel what else are they discussing/Tweeting -- First tweets discusses about Travel and Food.

A Convolutional Neural Network (CNN) would be trained on Level 1 & Level 2 IAB segments using standard gradient descent algorithm with a learning rate that exponentially decays over time.


activation function
convolution
Sentence matrix
7 × 5
3 region sizes: (2,3,4)
2 filters for each region size
totally 6 filters
2 feature maps for each region size
1-max pooling
6 univariate vectors concatenated together to form a single feature vector
softmax function regularization in this layer
2 classes
d=5
I
like
this
movie
very
much
!

Pic courtesy: http://www.wildml.com/

The first layers embeds words into low-dimensional vectors. The next layer performs convolutions over the embedded word vectors using multiple filter sizes. For example, sliding over 3, 4 or 5 words at a time. Next, we max-pool the result of the convolutional layer into a long feature vector, add dropout regularization, and classify the result using a softmax layer.

The experiment would leverage on TensorFlow library for creating/building the Neural network and would employ TensorBoard to visualize and plot quantitative metrics about the execution of TensorFlow graphs.

The training and test sets would be created by applying cross validation on the tweets extracted from Twitter.

The results of the experiment would presented as a dashboard for the end user.

Please refer to the above sections for details on IAB.

Model Benchmark & Evaluation Metrics:

The model would be benchmarked using two approaches

1. Measuring Accuracy & Loss

We measure the accuracy and loss after stipulated number of steps and ensure the loss tends/converges to zero as we increase the number of steps.

We used ***mean cross-entropy*** loss as the ln() function in cross-entropy takes into account the closeness of a prediction and is a more granular way to compute error.

We chose accuracy as a measure because of the problem characteristics as we are more focussed on correct prediction.

The distributions of tweets for prediction in real time is not even as some categories tend to be more popular than others regularly. For instance, Arts & Entertainment, Sports and Fashion may be quite popular among users when compared to Science. However, while training the model we have made sure the model is unbiased by having good distribution of tweets across all the categories.

The sample test run is detailed in the table, below, we started with a loss of 2.21 and accuracy of 56% after completion of 2000 steps. The loss reduced to 0.61 and accuracy improved to 87% for 14400 steps

Steps	Loss	Accuracy
2000	2.21507	0.562134
2600	1.92649	0.623792
3400	1.65499	0.674398
4000	1.49775	0.718656
4600	1.37581	0.73929
5400	1.2461	0.766755
6200	1.14148	0.790692
7000	1.04623	0.803317
8400	0.911524	0.835262

12000	0.69295	0.86411
14400	0.613263	0.878898

2. **Measurement of On-target rate: Practical approach**

On-target percentage is defined as rates of total campaign impressions served to the intended audiences.This is direct measure of the effectiveness of an Advertising campaign.

The trending IAB segments and Keywords, predicted by our Advertising brain, can be tested against the Keywords and segments used for the Media buyer for targeting.

DSPs like Google's DoubleClick Campaign Manager (DBM) provides, upto, past one month summary of segments and keywords that performed best and we can test our segments and Keywords against those.

The final solution would delve on how we can measure the on-target rates but it would more of an academic discussion since integration with a DSP, like DBM is beyond the scope of the experiment

Project Design:

A high level project design is postulated below.

- Twitter stub, *developed as the part of the projec*t, would extract Tweets from Twitter stream and store it in the disk as flat files.

    - We can have a more robust design by storing Tweets in-memory cache but this is not related to the main outcome of what we plan to achieve.

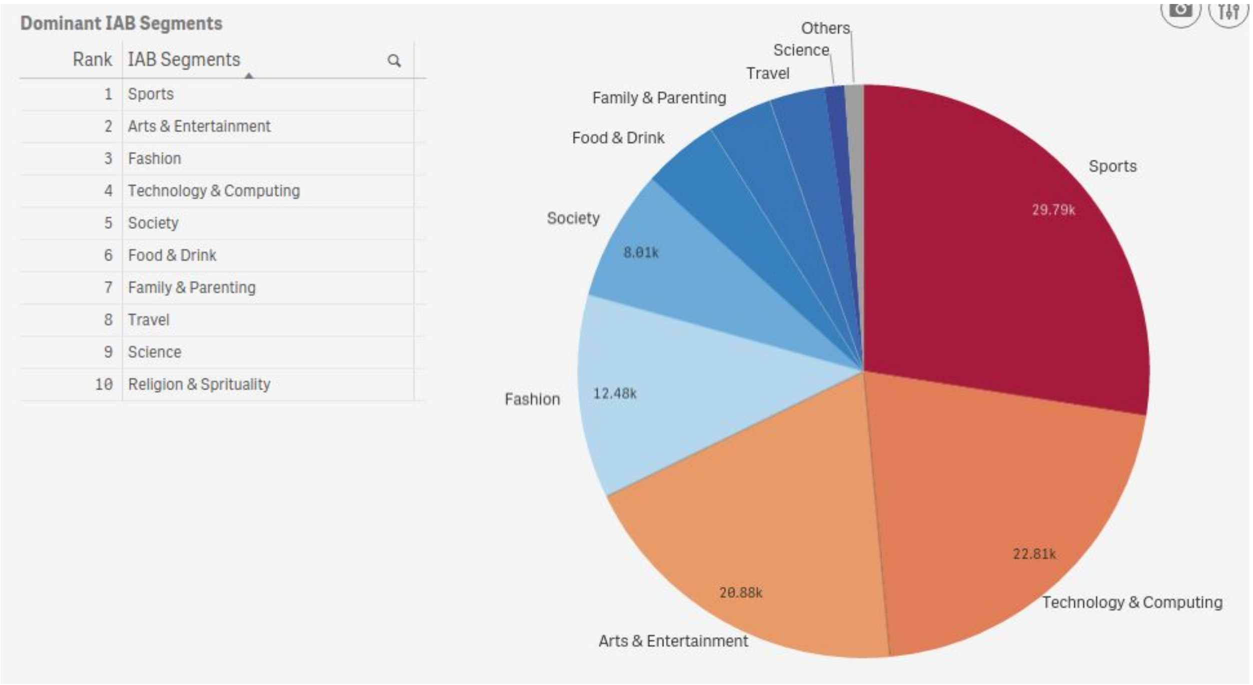
- A Convolutional Neural Network (CNN) would be trained to classify Tweets for IAB Level one and Level 2 categories.
    - The trained model would be stored as a pickle object.
    - Tensor flow would be used for building and training the model.
    - The model file along with the training data would be provided as the part of the project -- sets of python files.

- The results of the classification would be visualized as a dashboard using Google Fusion tables and would be the part of the project

Sample visualizations are detailed below for reference.

- First visualization depicts reach.
    - Number of Audience that can be reached for given IAB segment.
- Second visualization depicts rank order of IAB segments along with the trending keywords per segment.
- Third visualization depicts correlated IAB segments.
    - IAB segments that discussed in conjunction.

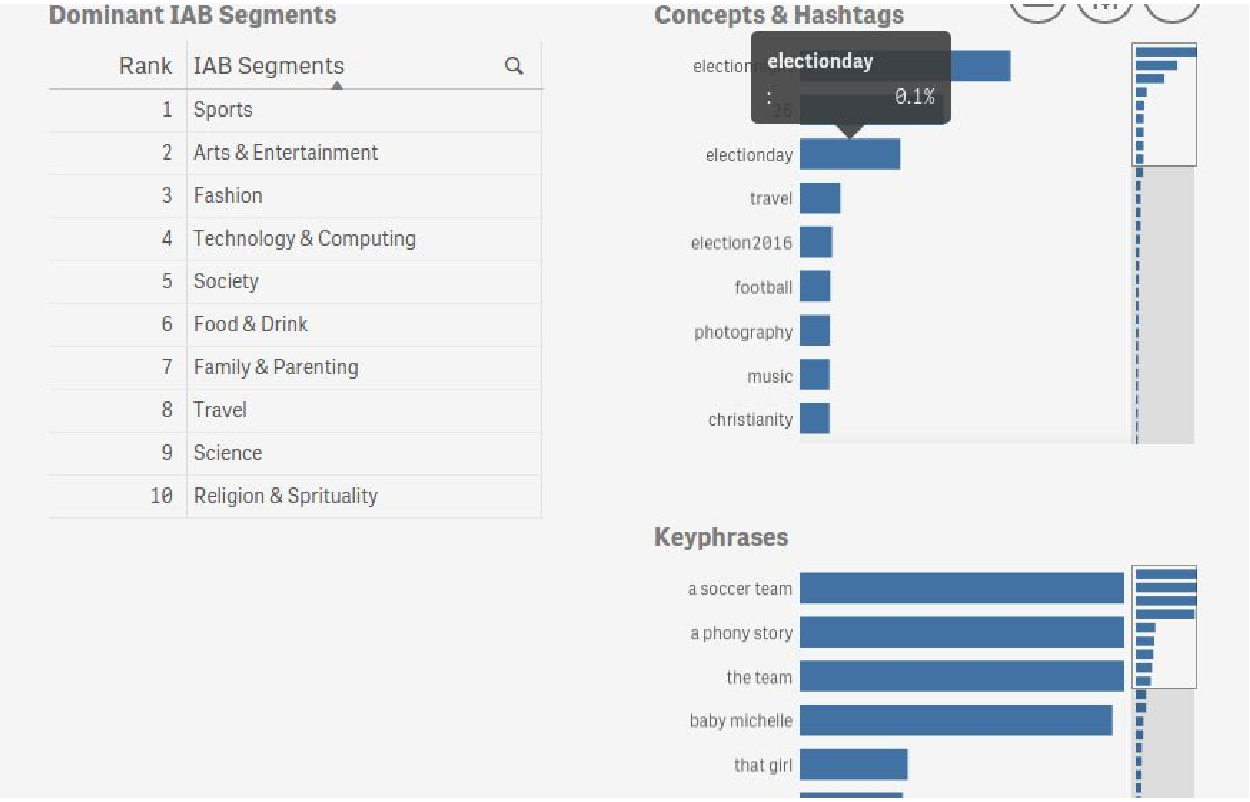
Audience reach in Twitter, per IAB segment



Dominant IAB Segments

Rank	IAB Segments
1	Sports
2	Arts & Entertainment
3	Fashion
4	Technology & Computing
5	Society
6	Food & Drink
7	Family & Parenting
8	Travel
9	Science
10	Religion & Sprituality

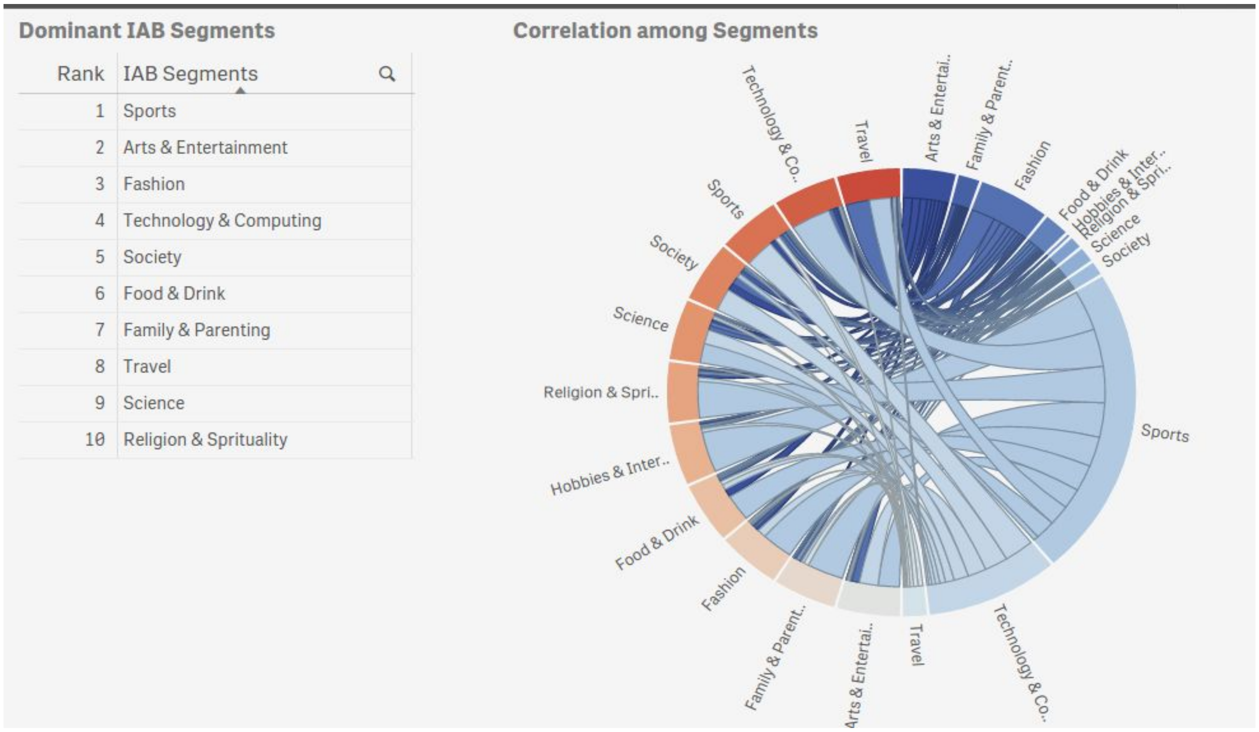
Trending IAB Segments & Keywords



Dominant IAB Segments

Rank	IAB Segments
1	Sports
2	Arts & Entertainment
3	Fashion
4	Technology & Computing
5	Society
6	Food & Drink
7	Family & Parenting
8	Travel
9	Science
10	Religion & Sprituality

Correlated IAB segments



Dominant IAB Segments

Rank	IAB Segments
1	Sports
2	Arts & Entertainment
3	Fashion
4	Technology & Computing
5	Society
6	Food & Drink
7	Family & Parenting
8	Travel
9	Science
10	Religion & Sprituality

Correlation among Segments