Prof. Ken Basye
Machine Learning
Final Report

The Role of Machine Learning in Sentiment Analysis

By Daniel Quaidoo

Tuesday, 20th December, 2016

Prof. Ken Basye
Machine Learning
Final Report

In recent years, the quest to interpret data and extrapolate trends has become essential and the backbone of almost every industry. These trends go a long way in affecting how industries are able to solve optimization issues and become more effective than before. A vital technique currently incorporated in realizing this is the use of machine learning as a frontier of data analysis. Thanks in part to current research and development, its role in numerous industries is now apparent as more and more industries adopt some sort of machine learning process to further breakdown the data they collect. The goal of this project is to decipher the role of machine learning in sentiment analysis with a focus on how machine learning can be incorporated in the management of a social media account such as Instagram and to some extent automate the process of maintaining the social media account.

Problem Description

Social media accounts are used excessively to advertise and market the products of various multinational and global companies. Nonetheless, these accounts are more or less primarily managed by humans each day in order to target the right customers and increase the consumer base of these companies via increasing the number of followers of the companies or marketing the right information to the customers. The main problem is that humans primarily view trends in marketing based on trends and already gathered data. Moreover, unless these social media companies were already a big hit or had a large enough social media presence, gaining followers and reaching new market segments is quite a challenging task one which is very time consuming;

sometimes taking forever. So the question now then is how do we consolidate the data already gathered on a social media platform to further improve the lives of social media managers and make it easier for them to effectively grow the companies they respectively represent. With the use of data mining we can address most if not all of these issues. The growth of machine learning has made it extremely possible to solve such problems with great confidence. In this paper I seek to address some of the strategies and mechanisms I've adapted in tackling these issues.

Problem Proposal

For this case study, I'll be implementing some machine learning techniques on a newly created Instagram account. The social media account will be run by a bot named, ***Jarvis*** which seeks to automate some of the tasks performed by a social media manager with the aid of machine learning. These tasks include:

- Finding the right market audience
- Following the appropriate users based on activities of the social media manager
- Attracting the right users based the users' activities (gaining followers)
- 'Interacting' with posts on the social media platform based on previous data of the user.

These are some of the techniques that'll be implored by ***Jarvis***. In the following chapter, I'll address the procedure in attaining these tasks, progress so far and which machine learning techniques were used in achieving any results obtained.

Prof. Ken Basye
Machine Learning
Final Report

Project Implementation

The basis of this project is analyzing the data on the social media account effectively classifying and extrapolating relevant information from it. In our case study for example, our main dataset is photos. The goal here will be to gather as much information from the pictures as we possibly can, classify this information into relevant groups and based on this grouping, perform activities such as following a user or liking their pictures. Moreover, this image classification can be used in future social media interactions such as liking a photo. The figure below outlines a visual description of the basic underlying program flow of the implementation of the ***Jarvis*** bot.

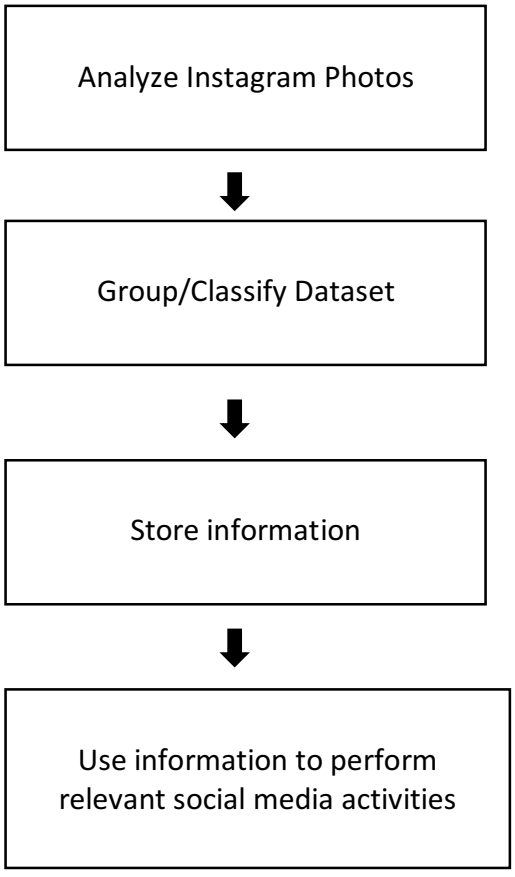


Figure 1.1

Prof. Ken Basye
Machine Learning
Final Report

In analyzing the images off Instagram, the photos were first sifted to using predefined user tags. This procedure is done so as to get a better dataset we could work with. Nonetheless, there was still some false positives after constraining the Instagram photo posts by the predefined photo tag. Consequently, a softmax regression analysis was run on the restricted dataset. With the aid of the softmax regression, I was able to classify the dataset into human readable groups so as to better understand what kind of information was associated with a particular image. For example, valid labeling for an image node could be a car or house. Based on this labelling, the image dataset is classified and the grouping can then be used to perform different interactions on the social media platform.

Why Softmax?

Softmax implementation was adopted due to its use in various multiclass classification methods in machine learning such as multinomial logistic regression. With softmax, the output of the softmax function can be used to represent the categorical distribution of the proposed dataset. That is, the probability distribution over N different possible outcomes can be computed. Converting these probability distribution values into action probabilities via the formula below

$P_t(a) = \frac{\exp\left(\frac{Qt(a)}{t}\right)}{\sum_{i=1}^{n} \exp\left(\frac{Qt(i)}{t}\right)}$ where the action $Q_t(a)$ corresponds to the expected success of finding an related output in the data set (in our case an image probability), given a sample size of t. In our case study, the action probabilities were used in building an undirected edge weighted graph with the values of the action probabilities as the weight of an edge given a particular image. Imagine running our script against an image of a dog. A possible sequence of the soft max is an

output of the action probabilities given the image of the dog. Suppose the output of the action probabilities is given as follows:

- Labels: animal (score = 0.98291)
- Labels: puppy (score = 0.51242)
- Labels: German shepherd (score = 0.04537)

From this data gathered, a graph can be built using these action probabilities where the weight of the edge (***w)*** is equal to 1 – p, where ***p*** is the value of the calculated probability distribution of an image output obtained from running the softmax function. Invariably such a graph will have a structure as seen in the figure below.

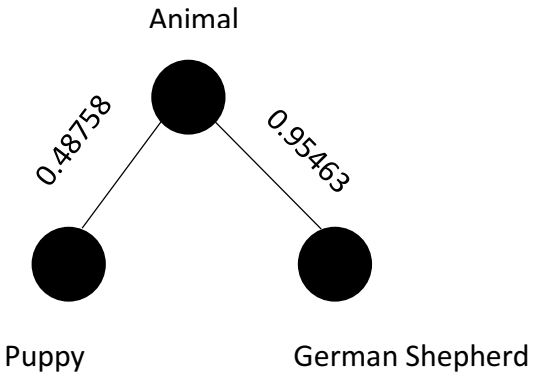


Figure 1.2

Representing the dataset in this manner makes it easier to analyze trends and further predict the 'future' behaviours of the end user. The bot has this three pronged graph system with lateral links to main nodes of the graph. (image shown below)

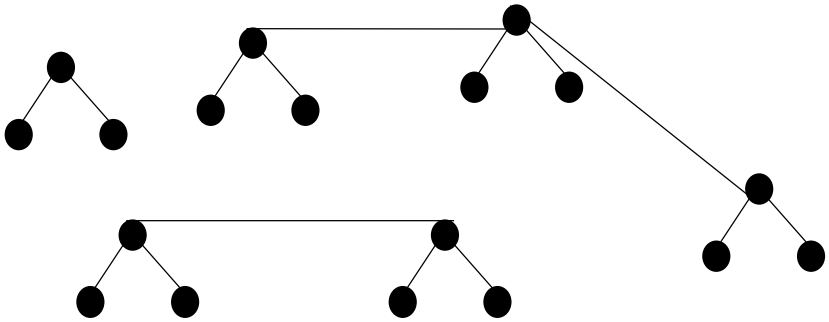


Figure 1.3

Thus in our case study of the Instagram account, to perform an action such as liking a picture related to an animal, the bot only needs to read in the graph input obtained from the trained dataset and find the shortest path from that particular node if it exists. Data used in training the graph is obtained from previous infromation of the end user primarily their already liked Instagram pictures. In the case where the term 'animal' or a particular node doesn't exist in our graph, this implies that the end user hasn't interacted with such a node yet thus situations like this are omited from our prediction system.

Use Case

Data generated from the graph is used in performing actions such as aiding a user like pictures, sampling another followers Instagram information to determine if they're a right fit for the bot or whether that account is one which the bot will be interested in following. A test case run on the graph I trained was on the graph search parameter namely: ***hot dog***. With this graph search

parameter in mind, I went ahead to run the script and obtained the result below.

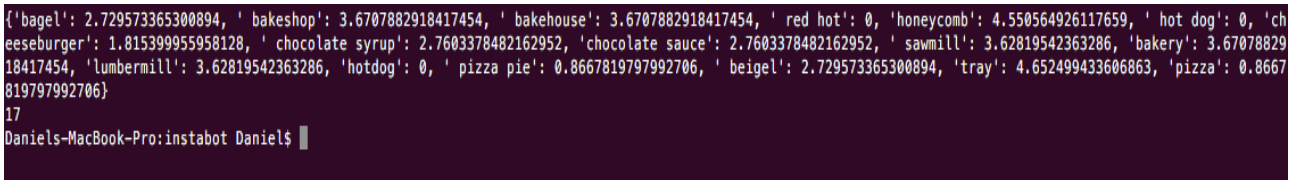


Figure 1.4

This test was run with a threshold in mind so as to further optimize the quality of the results. As depicted in the result set, the outcome obtained was relevant to the search parameter inputted. I was able to obtain similar results given a particular search paramater for other various search parameters. One should not however that the search paramters are defined based on data collected from the users previous interactions with the social media account.

Comments and Improvements

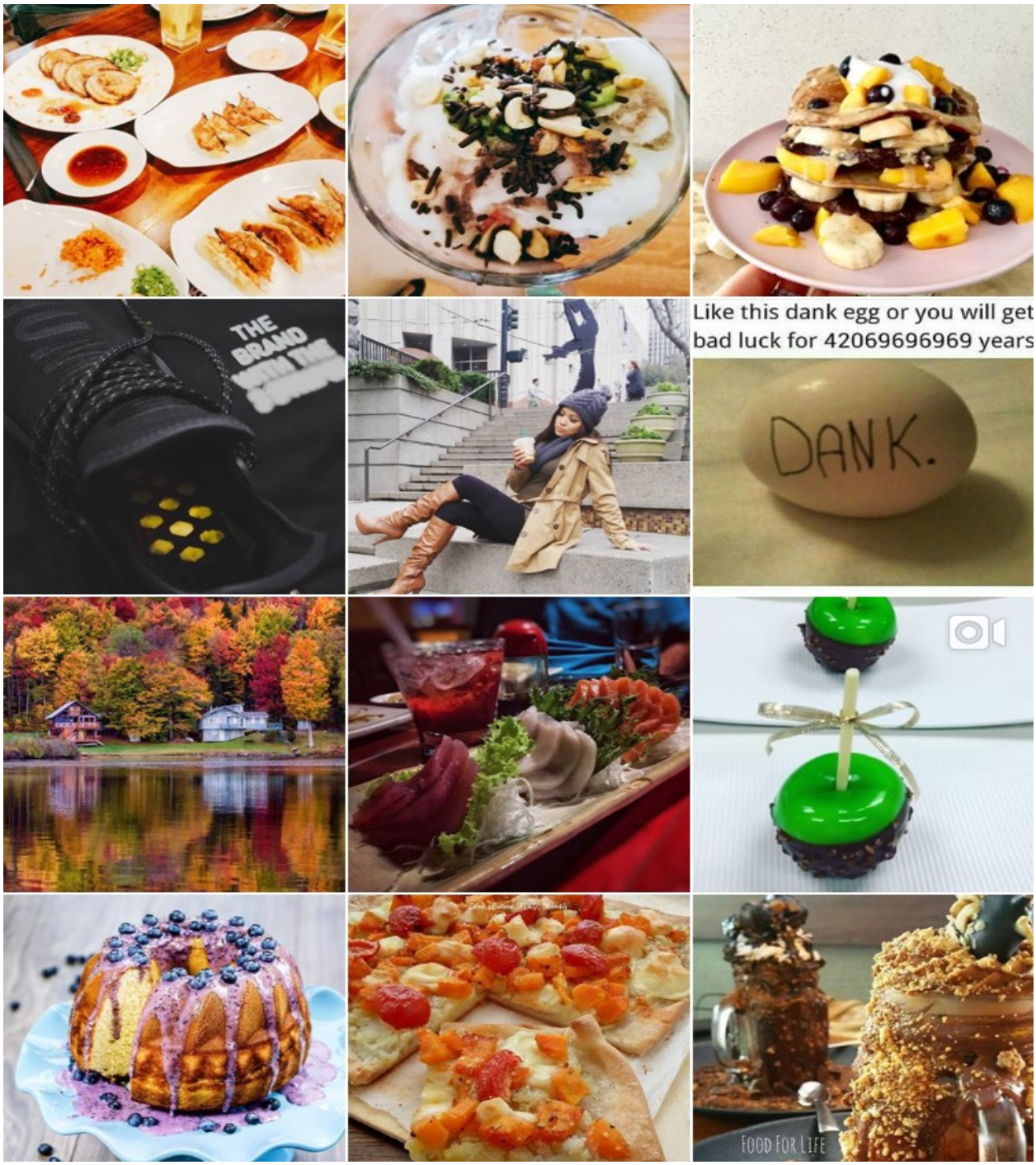
The graph analysis can be improved by inputting more data. Largely the more data obtained in building the graph the higher the chances in improving the accuracy of the script run against it. There was some difficutly in initially determining the source nodes for builidng the graph as the initial data used in drawing out the graph was not enough and created too many disjoint sets. When the input data to the graph was increased, this in turn decreased the number of disjoint sets allowing for a much better result during the graph building process. The script also has the capability of following users on the social media account. From careful analysis, it was observed that per every 100 accounts followed, the account manager could receive 40 new followers. Due to time constraints, the graph wasn't applied to the following mechanism of the social media bot thus a future improvement and addition to this project will be to apply the information gathered

from the graph when deciding which account to follow. Overall, the goal of using previous information from the end user to pursue relevant interactions with posts was achieved as I was able to use data from the account to map out a valid graph and like pictures based on this graph. The project was daunting and interesting at the same time and this is something I will like to build upon.

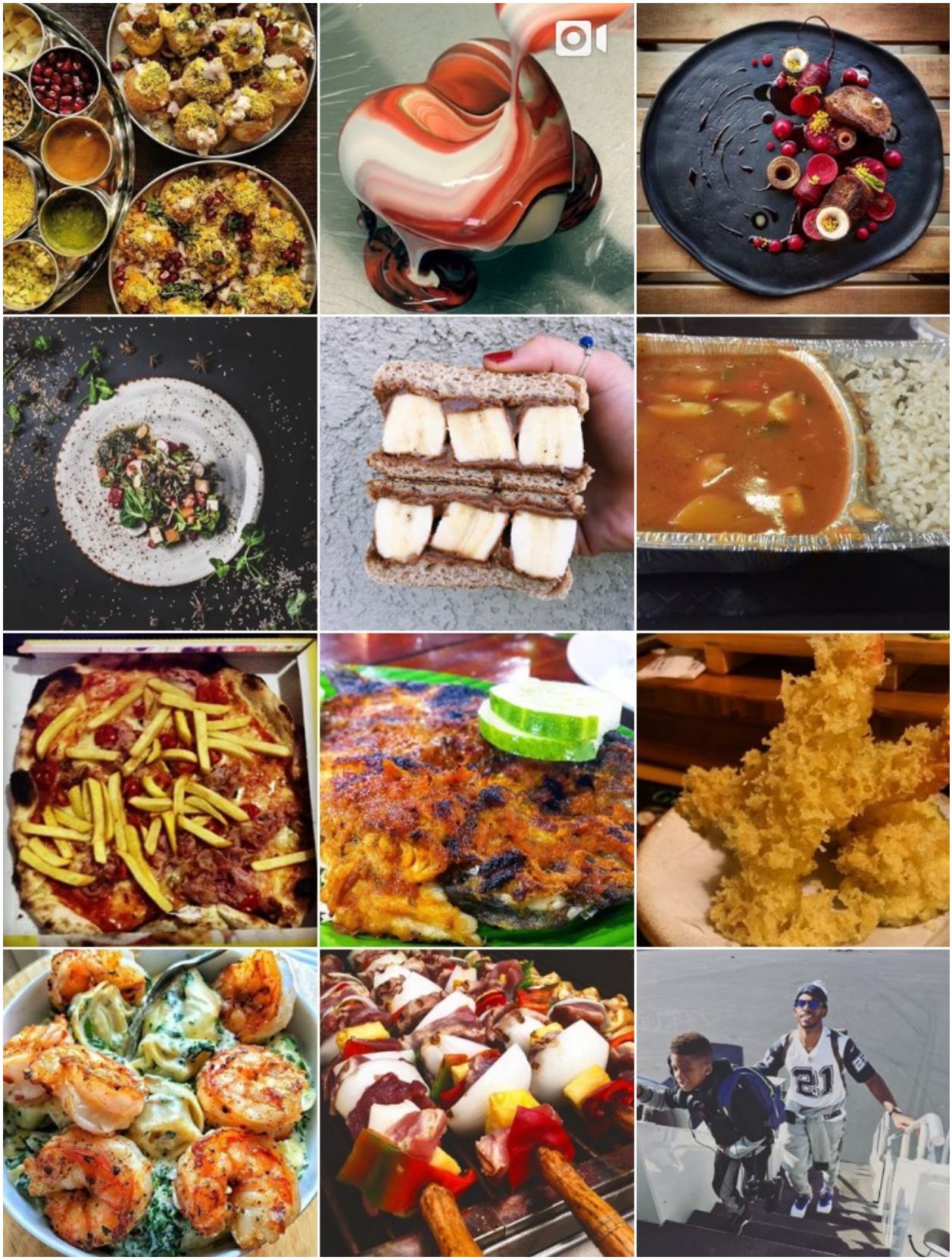
Prof. Ken Basye
Machine Learning
Final Report

Appendix

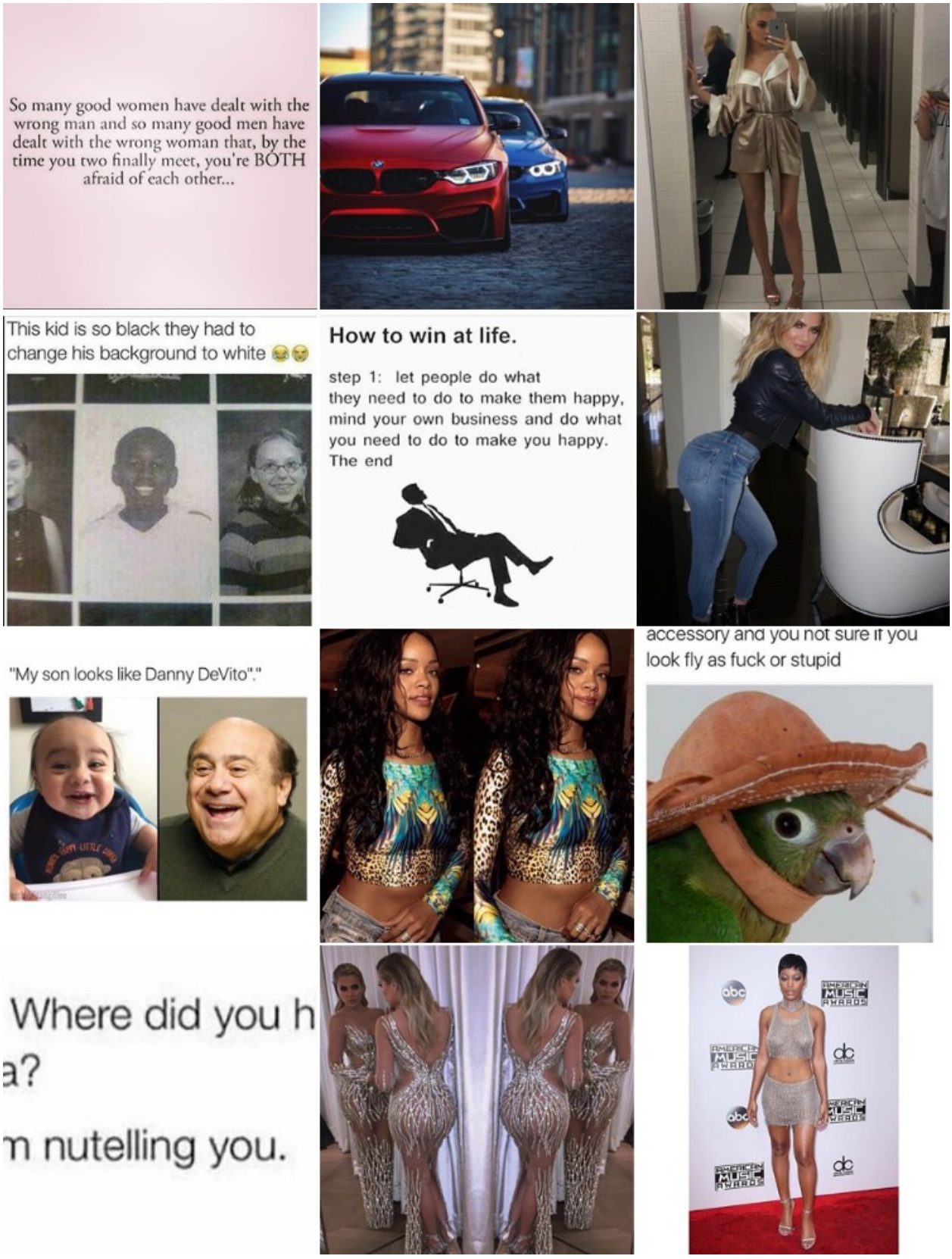
Liked photos on Instagram account:



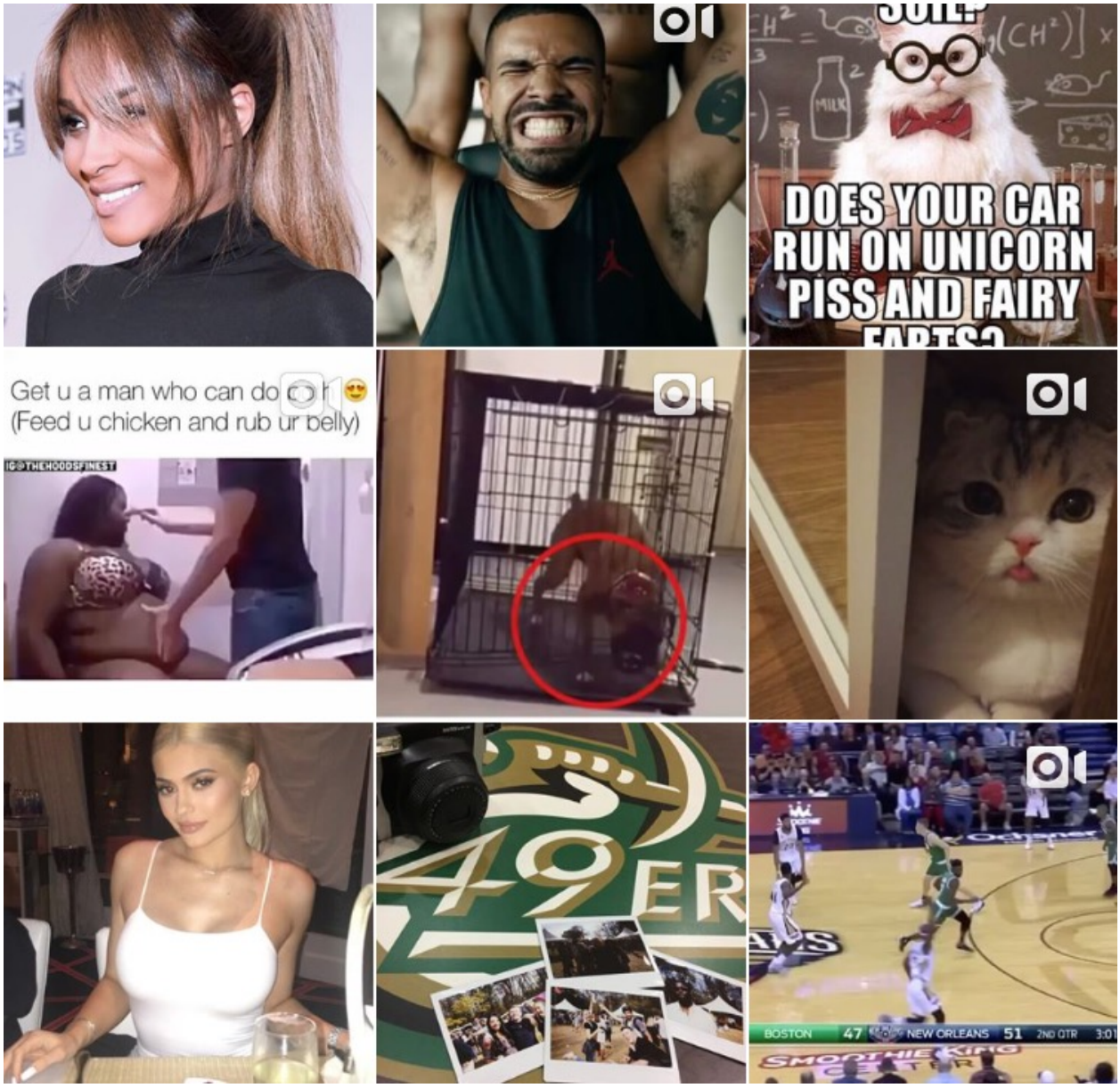
Prof. Ken Basye
Machine Learning
Final Report



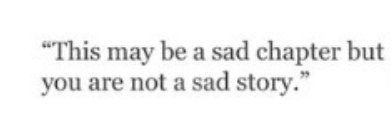
Prof. Ken Basye
Machine Learning
Final Report

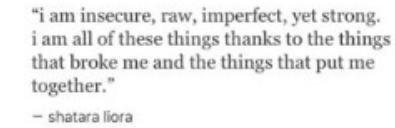


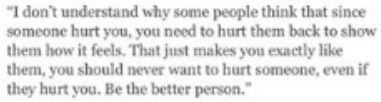
Prof. Ken Basye
Machine Learning
Final Report

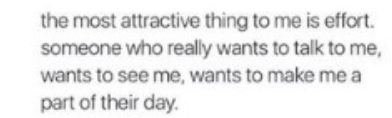


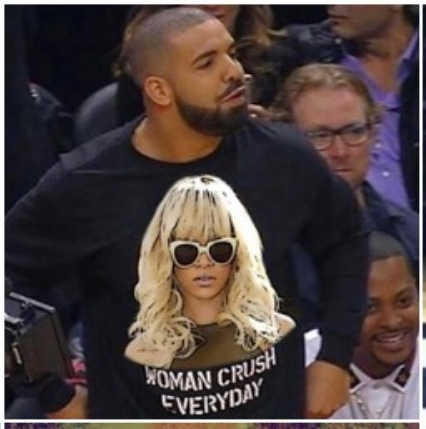
Prof. Ken Basye
Machine Learning
Final Report

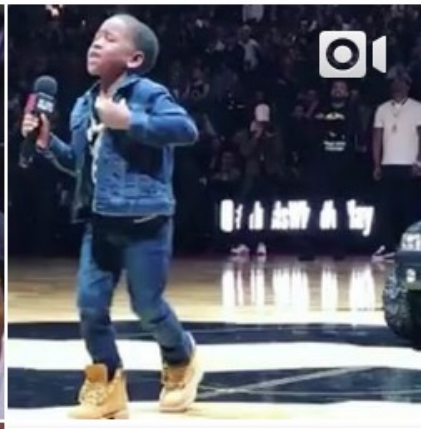


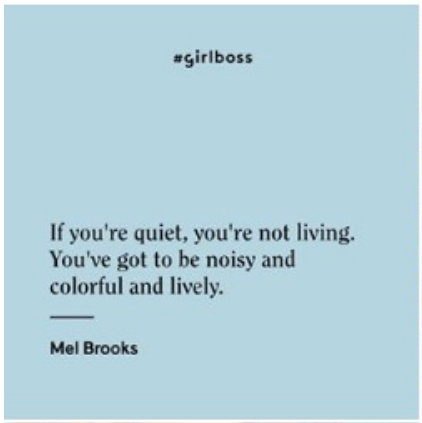


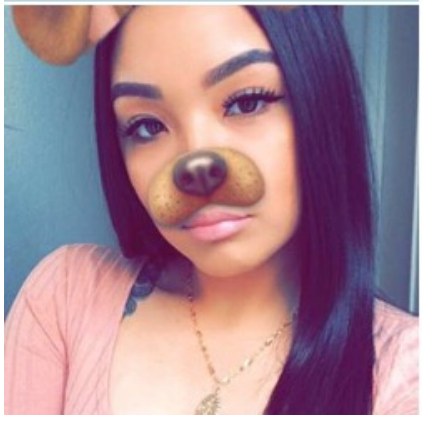












Prof. Ken Basye
Machine Learning
Final Report