ASU W. P. CAREY SCHOOL of BUSINESS

ARIZONA STATE UNIVERSITY

**W. P. CAREY SCHOOL OF BUSINESS HONOR CODE MASTER'S PROGRAMS**

We, the students of the W.P. Carey Master's programs are committed to maintaining the highest levels of honesty, integrity, and professional courtesy for ourselves and amongst our classmates. We believe that ethical behavior is not only the cornerstone of a good academic program, but of good business as well. As such, we are steadfast in our commitment to the following principles:

- Academic Integrity – We view the W.P. Carey Academic Integrity Policy as a living, breathing document, not just concepts on a page. As such, we are resolute in ensuring that we abide by the principles contained within, both in letter and in spirit. Furthermore, we will hold our classmates to the same high standards to which we hold ourselves.

- Consideration of Others – As W. P. Carey students, we believe that we should treat others as we wish to be treated. To this end, we will be considerate of classmates, professors, and recruiters in all aspects. Furthermore, we will constructively add to the collaborative environment by helping and encouraging others.

- Professionalism – As W. P. Carey students, we believe that the behaviors we practice today will become habit in the future. To that end, we are committed to act in a professional manner in all school settings, including classes, networking events, and interviews. We will be cognizant of the fact that we are representing the program and our classmates at all times.

Full Name (printed): Jinhang Jiang

Signature Jinhang Jiang Date 4/29/2021

Full Name (printed): Dhruv Tyagi

Signature dhruv tyagi Date 4/29/2021

Full Name (printed): Bhavana Patil

Signature: Bhavana Patil Date :4/29/2021

Full Name (printed): Jinghuan Li

Signature Jinghuan Li Date 4/30/2021

**Onsite MSBA Applied Project Report**
**Spring 2017**
**W.P. Carey, ASU**

| | |
|---|---|
| Topic | *Investigating Influencer Endorsements: A Data-Driven Exploration* |
| Team | *Team A8* |
| Team Members | *Jinhang Jiang, Bhavana Patil, Dhruv Tyagi, Jinghuan Li* |
| Client information | *Adidas, Data Science Team* |

# Executive Summary

Adidas is working on digital branding. In order to increase their brand value, they want us to implement a full social media analytics pipeline (Data Collection, Data Processing, Text Analytics and Data Visualization) and gain deeper insight into consumer reaction towards those marketed with celebrity branding.

In order to capitalize on the fast growing world of social media-market, Adidas needs to select the next influencer who can take the Adidas name to even more masses. In order for Adidas to make a decision on the next influencer, we have come up with a model that will help them make a  good business decision on choosing the next big influencer.

We conducted the whole research with python, and collected the data from Reddit.com and Instagram. We deployed text mining, image quality analysis, and social network analysis in our decision support system. To test the effectiveness and capabilities of our system, we conducted a case study. We were looking to tackle a business question: What if Naeun Son and BlackPink(the Adidas sponsored influencers) are retiring or encounter a controversial issue, who will be the ideal candidate in a given list in the South Korea market. And the result is very promising. You may find the details in the result section.

# Background & Problem Statement

In today's world, there is a rapid rise in the market of social media influencers. Big companies such as Adidas have always in the past endorsed these influencers such as Kanye West, Leo Messi, Sachin Tendulkar to name a few. But their approach in deciding these influencers is largely ad-hoc or relies more on manual effort. Their current approach fails to consider the unstructured data such as text, image audio, video that are present on the social media platform. Thus, while choosing their next influencer that Adidas wants to endorse they are still missing out on a lot of data and thus we decide to create a decision support system that will help them make a good business decision.

Big organisations such as Adidas have a desire to make sure that they are reaching more people every year. In order to do so they like to endorse few influencers that have a vast fan following to increase their reachability across the globe. But so far their approach in this field has been mainly ad-hoc or relies on manual efforts. We have built a computational, data-driven and generalizable system that will help support the business decisions with little human intervention. This approach will not only help Adidas pick the next big influencer but also will help them keep track of the future endorses hence allowing them to expand their brand and make it reach even more people.

# Methods

### Decision Support System

As we stated above, our design goal is to build a computational, data-driven, and generalizable approach to support business decisions with little human intervention. We aim to use our decision support system to boost the investment decisions in influencer endorsement and help the marketing team improve budget control, exclude unnecessary costs, and achieve the idea of investing precisely and accurately. And here is how it works from a high level of view :

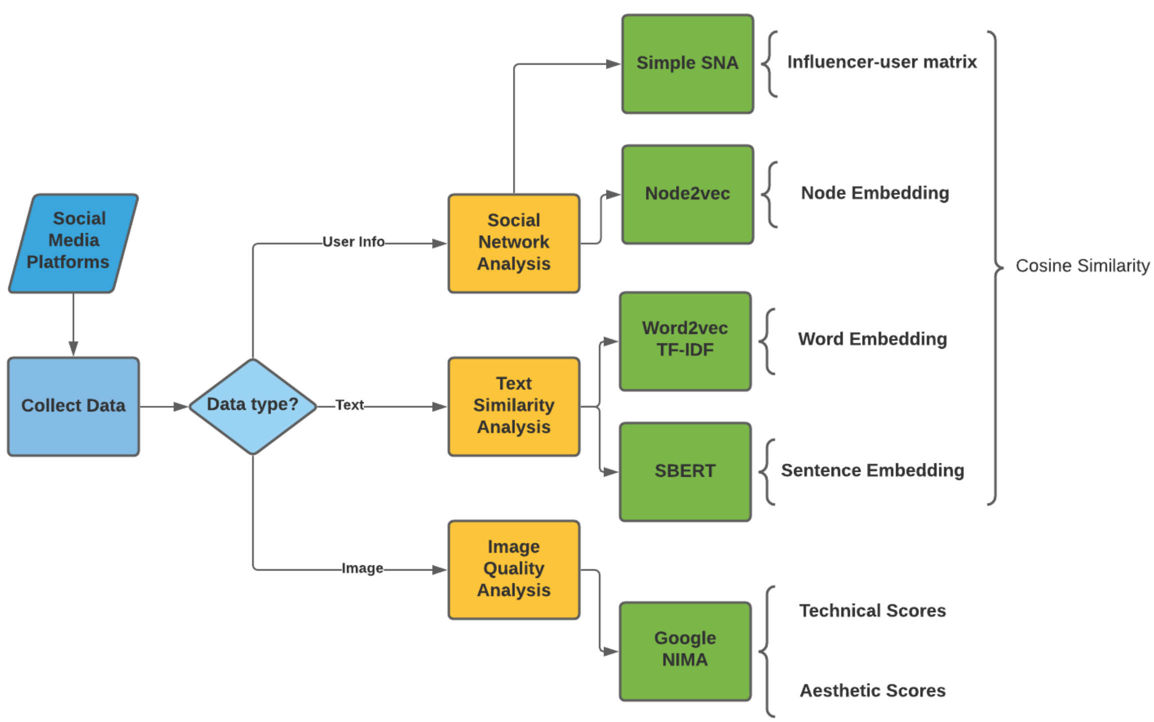



Figure 1: Decision Support System Diagram

Initially, we collected the data from social media platforms like Reddit. And then, we sort the unstructured data by its format. If it contains the user information, we will use it as an input in the social network analysis. We used techniques like simple social network analysis, node2vec. And then, if the data contains raw text, we will do a text similarity analysis. The text-mining models we included in our systems are word2vec+tfidf and SBERT. We can later use the output from the text mining and social network analysis for various analyses, and we mainly used them to make a cosine similarity comparison.

## Data

For the experiments and case studies, we collected two versions of the data. First one contains the data related to 14 Adidas endorsed influencers. It contains over 22 million words and 100 thousand user information. The second is designed for a case study, which composed data for 10 south korean celebrities or girl groups. It contains 15 million words and 62 thousand user information.

The average time to collect text data for 14 celebrities for one round is 7 days 13 hours. And the average time to collect the user information is 3 days 8 hours.

To collect the Instagram stories, we choose to use an open-source API – Instaloader to scrape the media file along with its metadata. The Instagram system is very locked down. We are not able to get any media data from its official APIs.

## Model Designing

### Text Analytics in DSS

The very first thing we brought into our system is text mining. Text analytics is an automated process of translating large volumes of unstructured text into quantitative data to uncover insights, trends, and patterns. Companies will be able to understand the story behind the numbers and make better decisions with this technique.

For our case specifically, it helps to discover the similarities between the language habits of the influencers' fan groups. By saying fans in today's presentation, we mean the people on whom the influencers can impact on the social media platforms.

We manipulated the models and studied the conversations between the influencer’s fans, expecting to capture their language cultures where the users talked about different influencers on Reddit. It will help the business understand better how similar the way fans/users talk about the influencers, the type of customers each influencer can reach to, the similarity and dissimilarity of language habits between the customer segments, and hopefully how the fans view their influencers.

We conducted research and experiments on a list of popular techniques in academic literature and business applications. Here is the complete list of the text mining techniques we have tried during the semester. Then we can apply those techniques all to compute the textual similarity. The number 1, 5, and 6 will require self-training, while others are pre-trained and ready to use.


Figure 2: List of NLP Techniques

After going through the techniques, we finally decided to select the #1 and #2 models as winners since those two seemed to work best. And in the future, if Adidas ever wants to expand this system, we recommend you continue to optimize those two models. The first one is word2vec+tfidf, which produces word embeddings. And it needs training on our collected data. It is somewhat slow compared to the other one, but it is more task-specific and task-oriented. The second one is a natural language processing library called sentence transformers. Instead of producing word embeddings, it will make a sentence embedding or sentence dictionary. It is a substantial pre-trained model, and it is seen as the state-of-the-art model in the current market.

One of the famous examples of word2vec is the concept of analogy. And people tended to illustrate the idea by colorizing the embeddings. So, let's visualize our data with the same idea. We can color code the document embeddings of Naeun Son, BlackPink, and Adriene Mishler. And we can perceive that the users talked about Naeun Son and BlackPink in a more similar way even though the patterns are not identical. So, for the cells painted with the same colors, we can assume they represent gender, music, dancing, new album information, etc., making the two influencers a good match.

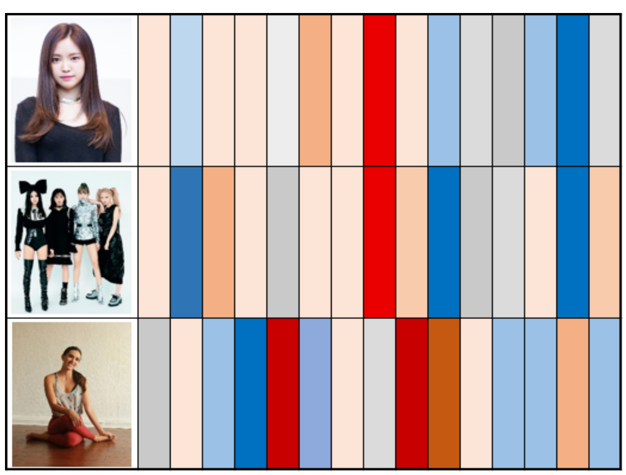


Figure 3: Analogy on Adidas Data

It would be excellent for our marketing team to have this kind of information on hand and be aware of the similarity or dissimilarity between the cultures of influencers' fans, especially when the marketing team tries to launch new marketing campaigns for those influencers.

So far, our model has been able to capture the semantic meaning of the contents. However, a large number of common words across the documents can make the result hard to interpret. For instance, each document will likely have many phrases, such as I think, I love, I don't, etc..So, in each influencer's posts, those common words can make the uniqueness vanished or diminished, leading to a result that similarity scores between the influencers will all go over 99%.

This is why we need to bring in TF-IDF, which is a numerical statistic. It gives us an importance score of a specific word in a particular document.

One of the limitations of the word2vec plus TF-IDF model is that it had a hard time dealing with multi-language text. In this case, we will introduce you to another model called Sentence transformers or SBERT.

Before we get to SBERT, what is BERT? It is a machine learning technique for text mining; it is pre-trained and developed by Google and published in 2018. Sentence BERT or SBERT was built upon the idea of BERT and developed by the UKP lab from Darmstadt university of technology in Germany. It provides an increasing number of state-of-the-art pre-trained models for more than 100 languages. It helps to solve the problems of the language barrier we had with word2vec.

**Social Network Analysis**

Social network analysis characterizes networked structures in terms of nodes and ties. In our case, each celebrity or influencer will be a node, and the Reddit users who have commented about two specific influencers will be the link for the influencers. For instance, user Jay talked about Beyonce yesterday in a post, and this morning he made another post about Pharrell Williams, then user Jay will become the link between Beyonce and Pharrell.

In the private sector, businesses use it to support customer interaction, information system development, and marketing and business intelligence. In our case, we performed SNA to reveal information about the similarity between the influencers' fanbases based on their social media activity patterns.

At the very first, we started with a simple way to perform social network analysis. We created an influencer-user matrix to compute the cosine similarity scores. In the graph below, the number in each cell will represent the number of times a user has mentioned a given influencer on social media. For example, user one has talked about Jay three times, spoke about influencer 1 two times, and never mentioned influencer 2. And we can use those vectors as inputs to compute the cosine similarity. It perhaps will be one of the most straightforward models for social network analysis.

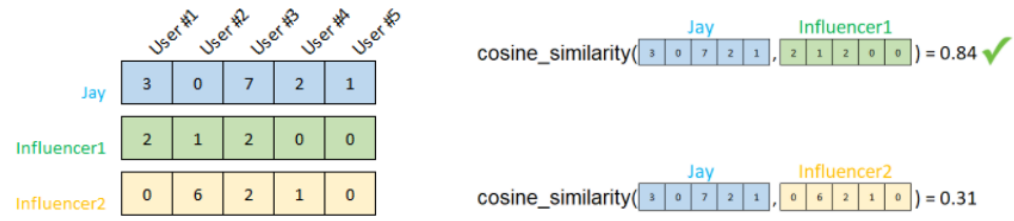



Figure 4: Influencer-User Matrix

To reduce the execution time costs and uncover more deep insights buried in the data, we brought in node2vec. And as you can tell from the name, it is built upon the idea of word2vec. It is an algorithm developed by Stanford to generate node embeddings for social network analysis. There is no good or bad between node2vec and the influencer-user matrix. To optimize the performance of those models, we have to feed them with the proper amount of data. Node2vec seems to work much better with a lot of data, while the influencer-user matrix handles the smaller dataset pretty well.

Here is a diagram to illustrate how the model works. Initially, we collect the data from the social media into a two-column data frame. Then we converted it into a social network analysis graph; on the graph, we can use the distance between the nodes and the color of the ties to interpret the similarities between the nodes. However, it is complicated and hard to read. Right now, we only have 14 nodes in the graph, and it's already a mess. However, what will happen if we want to study 100 celebrities? This is what we call high-dimensional information. Node2vec can help to reduce the dimensionality by converting the graph to a list of embeddings. So, we got a new dictionary for the nodes. Then we can use the embeddings to draw the T-SNE plot or use the built-in feature to calculate similarity or use k-means clustering to find groups or segments.

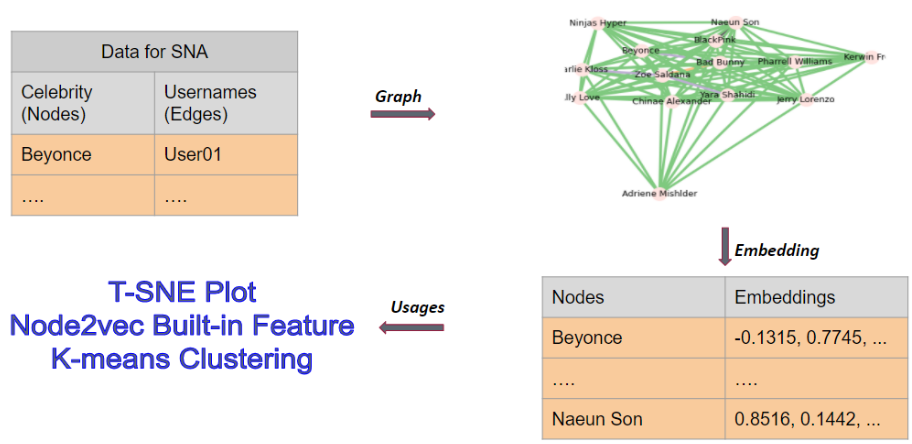



Figure 5: Node2vec Implementation Diagram

**Image Quality Analysis**

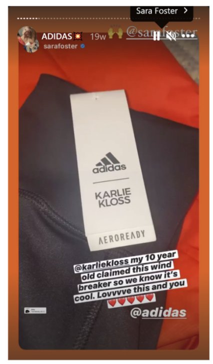
We used Google NIMA (Neural Image Assessment) to give technical and aesthetic scores to images, aiming to minimize human interventions on evaluating the subjective nature of images.

NIMA implements a deep CNN architecture that is trained to predict which images a typical user would rate as looking good (technically) or attractive (aesthetically). Technical quality assessment deals with measuring pixel-level degradations such as noise, blur, compression artifacts, etc., aesthetic assessment captures semantic level characteristics associated with emotions and beauty in images.

This model is trained with histograms of human ratings for every image, then produces a distribution of ratings for any given image — on a scale of 1 to 10, along with likelihoods to each of the possible scores. Finally, the model calculates a mean score that correlates closely with human perception.

Technical: 5.3163
Aesthetic: 5.4846


Wales Bonner




Karlie Kloss

Technical: 4.3772
Aesthetic: 4.3204

Figure 6: NIMA Scores for Captured Instagram Stories

**Image Style Classification**

The style of an image plays a significant role in how it is viewed. The viewers will receive different visual ideas when encountering images with different styles. The influencers on Instagram present various styles in their postings, while as the company sponsoring them, it is important to diversify the visual styles in marketing one product to expand the customer reach. Unlike the original object detection task, style has received little attention in computer vision research, partly due to the subjective nature embedded under the idea.

We utilized Sergey Karayev's vislab project on visual recognition to complete the task of classifying images based on styles. Our implemented sub model from there is built with a fine-tuned Caffe deep convolutional framework and is pretrained on a dataset containing 80K Flickr images with 20 style labels. The model has been tested on several other datasets like AVA, wikipainting and pinterest, achieving satisfactory performance, which gives us feasibility of transferring the classifier to our own dataset on Instagram.

The default style labels embody several different aspects of visual style, including photographic techniques ("Macro," "HDR"), composition styles ("Minimal," "Geometric"), moods ("Serene," "Melancholy"), genres ("Vintage," "Romantic," "Horror"), and types of scenes ("Hazy," "Sunny"). These styles are not mutually exclusive, and represent different attributes of style. We also have the scalability to add our own training set or define new labels to make a more customized style classifier.

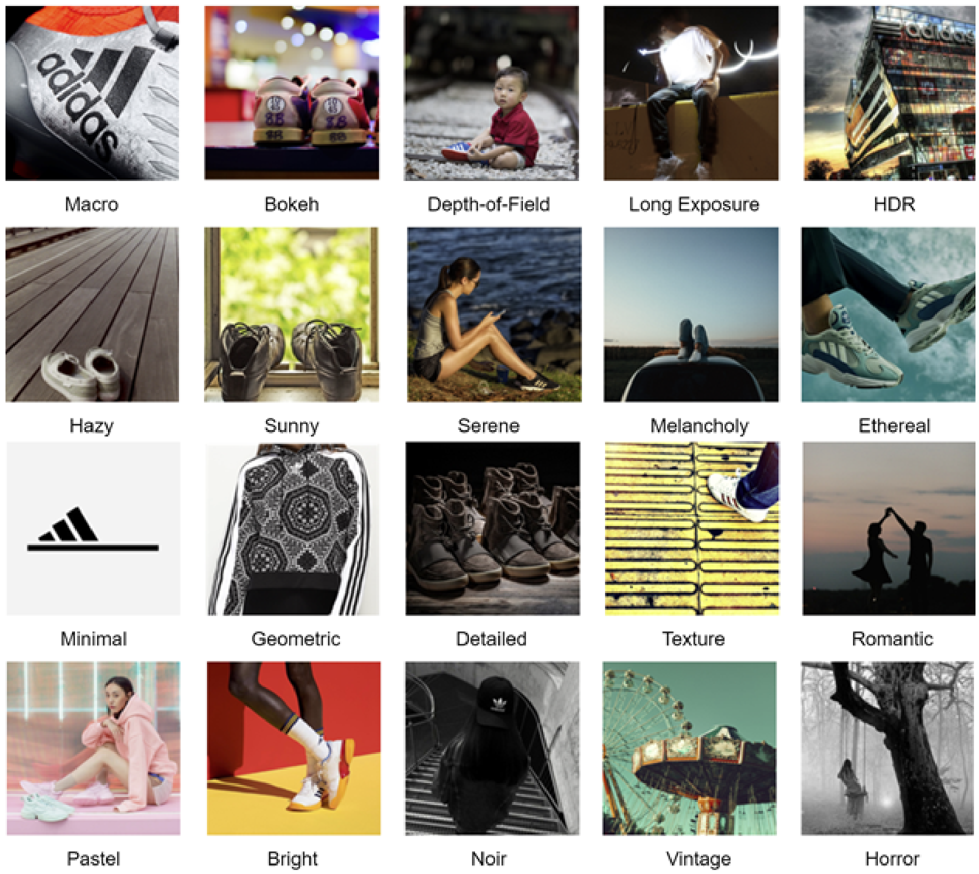
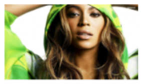



Figure 7: Default Style Labels of the Style Classifier

## Implementation

**Text Analytics:**
In the following chart, we can see two examples of using the embeddings we generated from the previous step. We compute the similarity score between a target influencer and the others and sort the results to find the most similar ones. This similarity is based on the language habits of the influencers' fans on Reddit, not the influencers themselves. It reflects how similarly two groups of fans talked about the given influencers. And it is only valid among the influencers in the given list.



| Top 3 of the most similar influencers to "*Beyonce*" | |
|---|---|
| **Influencers** | **Similarity Score** |
| "Zoe Saldana" | 0.869546415 |
| "Pharrell Williams" | 0.865572384 |
| "BlackPink" | 0.863266794 |

| Top 3 of the most similar influencers to "*Pharrell Williams*" | |
|---|---|
| **Influencers** | **Similarity Score** |
| "Beyonce" | 0.865572384 |
| "Jerry Lorenzo" | 0.856141943 |
| "Bad Bunny" | 0.834991535 |

Figure 8: Word2vec + Tf-Idf Output

The results are not 100% ground truth. Instead, we can use them as a factor to support the final decisions. The model is good at uncovering insights that otherwise will not be easily observed by a human. The output will be slightly different each time when we try to optimize the model. But, during optimizations, we will start to observe patterns that will feed our needs. For example, BlackPink and Zoe Saldana will always stay in the top 3 most similar influencers to Beyonce. And for most of the time, Beyonce and Bad Bunny will show up in the top 3 similar influencers to Pharrell Williams.

**Network Analytics**

To implement the idea of the influencer-user matrix, we wrote a series of functions. Two of the built-in features are best-sub and best-complete. Those two functions will answer two different business questions based on the influencer-user matrix we generated. For instance, if an Adidas endorsed influencer is retiring or has encountered a controversial issue, and you want to replace him or her, who should be the candidate that will minimize the loss or pertain to the existing markets? And best-complete will find the most dissimilar influencers to the target, who will likely help the business expand the marketing and reach out to the customer groups that probably never have been reached before. The function takes two inputs: the target's name, the other one will be the number of candidates you are looking for.

And here is an example: for Beyonce, the best-sub will be Bad Bunny and Pharrell Williams, who are all singers, and all do hip hop. The best-complete will be Adriene Mishler and Jerry Lorenzo; one is a yoga teacher, the other one is a fashion designer.

**Beyonce**

```
model.best_sub("Beyonce",n=2)
['Bad Bunny', 'Pharrell Williams']
----------------------------------
model.best_complete("Beyonce",n=2)
['Adriene Mishler', 'Jerry Lorenzo']
```

Figure 9: Output for Influencer-User Matrix

The following figure shows a table of 4 groups of influencers generated by k-means clustering with k equals to 4. If we look at it group by group, we can find some interesting insights. The first group contains the people who work as content or product creators. The second group is all the musicians and singers. Adriene Mishler stays by herself in group 3, and I believe it is due to the uniqueness of her profession since she works as a yoga teacher. And group four has all the fashion models and actresses. This result is very flexible, if you want to study them in smaller groups, just increase the value of k, and the big groups will be broken down into smaller ones.

| *K-means Clustering with k = 4* | |
|---|---|
| **Groups** | **Influencers** |
| **Group 1** | Ninjas Hyper<br>Jerry Lorenzo<br>Kerwin Frost<br>Ally Love |
| **Group 2** | Bad Bunny<br>Beyonce<br>BlackPink<br>Pharrell Williams<br>Naeun Son |
| **Group 3** | Adriene Mishler |
| **Group 4** | Chinae Alexander<br>Zoe Saldana<br>Karlie Kloss<br>Yara Shahidi |

Figure 10: Node2vec Output

# Results and Conclusions

## Case Study

To better understand the international market, as well as to demonstrate the potential usage of the decision support system we built, we specifically picked 8 more korean groups/individuals (based on popularity, stage style, etc.) along with “Black Pink” and “Naeun Son” to conduct a case study.

| Groups | Ranking | Reddit Subscribers |
|---|---|---|
| Black Pink | #2 | 122k |
| IZ*ONE | #5 | 27.4k |
| BTS (boy group) | #1 | 214k |
| NCT (boy group) | #6 | 29.4k |
| GFriend | #9 | 23.8k |

| Individuals | Groups | Ranking | Reddit Subscribers |
|---|---|---|---|
| Naeun Son | Apink | #14 | 7.5k |
| Seolhyun | AOA | (Nike) | 4k |
| Sana | TWICE | #2 | 102K |
| Solar | MAMAMOO | #7 | 18.5k |
| Miyeon | GI-DLE | #4 | 15k |

Figure 11: Table of Influencers for Case Study

Business Question: Due to a controversy, if we had to substitute Naeun Son and BlackPink with someone else, who would be the ideal candidate?

Hypothesis made before digging into the data:

- Solar or Seolhyun could be the most similar individuals to Naeun Son.
  Naeun and Seolhyun look very similar to each other and Naeun and Solar are both prominent members of a girl group.

- IZ*ONE could be the most similar girl group to BlackPink based on ranking (#5 & #2) and origin (Seoul, South Korea)
- Sana and Miyeon could be very unique in these datasets since their stage positionings and artistic style are very different from the other groups or individuals.

```
most_similar("BlackPink",
             Pairwise_similarities,
             'Cosine Similarity',
             topn=2)
--------------------------------------------------
Name: BlackPink
Similar Influencers:
Celebrity: iZone : 0.7479952518415461
Celebrity: GFriend : 0.7131607796176859
```

```
most_similar("Naeun Son",
             Pairwise_similarities,
             'Cosine Similarity',
             topn=2)
--------------------------------------------------
Name: Naeun Son
Similar Influencers:
Celebrity: Seolhyun : 0.6698078801647593
Celebrity: Solar : 0.6178915677265608
```

Figure 12: Text mining - Word2vec + Tfidf

```
model.best_sub("Naeun Son",n=2)
['Seolhyun', 'GFriend']
--------------------------------------
model.best_complete("Naeun Son",n=2)
['Sana', 'Miyeon']
```

```
Model.best_sub ("BlackPink",n=2)
['GFriend', 'NCT']
--------------------------------------
model.best_complete("BlackPink",n=2)
['Sana', 'Miyeon']
```

Figure 13: Social Network Analysis - basic decision modeling

```
model.wv.most_similar("Naeun Son",topn=4)
(the groups were filtered out)

[('Seolhyun', 0.9898267984390259),
 ('Solar', 0.9860391616821289),
 ('Miyeon', 0.9639727878570557),
 ('Sana', 0.9561481475830078)]

--------------------------------------------
model.wv.most_similar("BlackPink",topn=4)
(the individuals were filtered out)

[('iZone', 0.996224582195282),
 ('GFriend', 0.9955752491950989),
 ('BTS', 0.9739093589782715),
 ('NCT', 0.9657029318809509),]
```

Figure 14: Social Network Analysis - Node2vec

Here is the summary of the case study:

| Techniques | Naeun Son | BlackPink |
|---|---|---|
| Text Mining - Word2vec+TFIDF | Seolhyun | IZ*ONE |
| Social Network Analysis - Basic decision modeling | Seolhyun | GFriend,NCT |
| Social Network Analysis - Node2vec | Seolhyun | IZ*ONE |

Figure 15: Summary of the Case

With these results we can conclude that Naeun can be replaced by Seolhyun and BlackPink can be replaced by IZ*ONE.

## Limitations and future work

### Data

- Different backgrounds.
- Not all of them on Reddit .
- 22% of users aged from 18 to 29 (as of February 2019)[6]
- We observed that the functionality of the Reddit API is not constant when we try to extract large amounts of data at once.

### Text mining

Limitations:

- No ground truth that can help measure the performance of the models
- Better performing models need labeled data

Potential Future Work:

- Additional preprocessing of data & vetting of the processed data
- Move on to supervised models

### Social Network Analysis:

Limitations:

- A small number of nodes will limit the performance (e.g. we only had 14 influencers in the dataset)
- The data is unbalanced, so the model still could be biased.

Potential Future Work:

- Collect data periodically (e.g. every 3 month) to study seasonal patterns and capture the fandom changes over time
- Node2vec can be applied for online product recommendations

### Exploration on Instagram Stories:

Limitations:

- Need to create an Instagram account and follow all target influencers
- The script needs to be run every 24 hrs to catch all data
- Instagram blocks high-volume abnormal data flow
- Style classifier may be biased by the 80K images from Flickr
- Pre-trained style classifier with fixed labels may not fit to individual business

Potential Future Work:

- Feed more generic labeled image data from other social media platform like Pinterest
- Define business-specific style labels and manually label the training data, then retrain the classifier.

# References


- Alammar, J. (2019, March 27). The Illustrated Word2vec. Jay Alammar's Blog. https://jalammar.github.io/illustrated-word2vec/.

- Cohen, E. (2018, April 23). node2vec: Embeddings for Graph Data. Medium. https://towardsdatascience.com/node2vec-embeddings-for-graph-data-32a866340fef.

- Graf, A., & Koch-Kramer, A. (n.d.). Instaloader. https://instaloader.github.io/.

- Grover, A., & Leskovec, J. (2016, July 3). node2vec: Scalable Feature Learning for Networks. arXiv.org. https://arxiv.org/abs/1607.00653.

- Jiang, J. (2021, March 23). Use R to Calculate Boilerplate for Accounting Analysis. Medium. https://towardsdatascience.com/use-r-to-calculate-boilerplate-for-accounting-analysis-f4a5b64e9b0d.

- Jiang, J. (n.d.). ASU_Adidas_Applied_Project_2021. GitHub. https://github.com/jinhangjiang/ASU_Adidas_CapstoneProject.

- Karayev, S.(n.d.) vislab. Github. https://github.com/sergeyk/vislab

- Karayev, S., Hertzmann, A., Trentacoste, M., Han, H., Winnemoeller, H., Agarwala, A., & Darrell, T. (2014). Recognizing Image Style. Proceedings of the British Machine Vision Conference 2014. https://doi.org/10.5244/c.28.122

- Lennan, C. (2018, July 9). Using Deep Learning to automatically rank millions of hotel images. Medium. https://medium.com/idealo-tech-blog/using-deep-learning-to-automatically-rank-millions-of-hotel-images-c7e2d2e5cae2

- Lennan, C., Tran, D., et al. (n.d.) image-quality-assessment. Github. https://github.com/idealo/image-quality-assessment/

- Reddit Usage and Growth Statistics: How Many People Use Reddit in 2021? Backlinko. (2021, February 25). https://backlinko.com/reddit-users.

- Reimers, N., & Gurevych, I. (2019, August 27). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv.org. https://arxiv.org/abs/1908.10084/.

- sdaylorsdaylor. (2019, March 1). What does embedding mean in machine learning? Data Science Stack Exchange. https://datascience.stackexchange.com/questions/53995/what-does-embedding-mean-in-machine-learning.

- Sieg, A. (2019, November 13). Text Similarities : Estimate the degree of similarity between two texts. https://medium.com/@adriensieg/text-similarities-da019229c894.

- Talebi, H., & Milanfar, P. (2018, April 26). NIMA: Neural Image Assessment. IEEE Transactions on Image Processing, 27(8), 3998–4011. https://doi.org/10.1109/tip.2018.2831899

- TensorFlow Team, G. (2017, November 20). Introducing TensorFlow Feature Columns. Google Developers Blog. https://developers.googleblog.com/2017/11/introducing-tensorflow-feature-columns.html.

- [UPDATE] January 2021 Top 50 KPOP Popularity Ranking. KPOP OFFICIAL. (2021, February 1). https://kpopofficial.com/top-50-kpop-popularity-reputation-ranking-january-2021/#3_Kpop_Idol_Group_Popularity_Brand_Reputation_Ranking_All_Kpop_Groups.

- What Is Text Analytics? MonkeyLearn Blog. (2019, November 20). https://monkeylearn.com/blog/what-is-text-analytics/.

# Appendix – Reproduction of Results

*Location or Links to Data Sources:*
*https://github.com/jinhangjiang/ASU_Adidas_CapstoneProject/tree/main/Data*

*Tools/Platform: Python, docker, bash*

*List of scripts :*
*https://github.com/jinhangjiang/ASU_Adidas_CapstoneProject/tree/main/Scripts*

*Guidance on how to run the scripts:*
*All the scripts are ready for "run-file". All the necessary packages are listed in the beginning of every script.*