Sub-Community Virality Prediction of Fashion on Instagram Using Network Models on Augmented Data Sets

Ramin Ahmari
Stanford University
Department of Computer Science
353 Serra Mall
Stanford, CA 94305

Abstract

With the rise of social media, virality has become a core research interest in different fields given its relation to establishing celebrity and trends. We characterize virality of content across social media if we find cascades of sufficiently large enough scale, speed and reception. Particularly in fast fashion, an industry that picks up on trends rather than setting them, the identification of proliferating trends that are going viral is essential in determining what products to manufacture next. While research in the area of social media cascades and virality has been seeing an increase over the last years, much of it is limited to Facebook and Twitter - platforms not native to supporting and establishing trends. Instagram has become a major hub for fashion trends offering much more granular content-related information than any other social networking platform. This paper establishes a network of authoritative streetwear influencers and hashtags and uses data scraped from these nodes to investigate the characteristics of virality within the streetwear fashion community on Instagram, predicting future virality and cascading. To that end, this paper run network models on a data set augmented with complex and network-based features.

Introduction

Social media has become far more than just a past-time for teenagers or a tool for entertainment and distraction - it has become a core component of the way humans communicate with each other and express themselves in the digital age. From likes, to follows, to reshares and reposts - our interactions on social media leave a digital network that connects us to people across the globe. When information in the form of a post, a tweet or a snap undergoes proliferation through a network at immense speed, we colloquially talk about "going viral". From Beyonce's "Lemonade" to Bella Hadid's leather crop top - examples of virality happen daily, yet the underlying foundations remain elusive and its prediction remains ineffective. Ever since social media has become a ubiquitous tool for marketers all around the globe to push out their products and reach consumers through what is called "influencer marketing", research in the area of social media in relation to influence, virality and trend proliferation has skyrocketed with scientists across the disciplines of social science all the way to computer science trying to make sense out of this new phenomenon that has taken the internet by storm and is governing the intricacies and rise of a new form of distributed celebrity.

In order to understand the current state of social media, its networks, virality, influence and trend proliferation better, we turn to papers across these disciplines. We investigate Mining Social Networks for Viral Marketing (Domingos 2008) first to establish an overview of methodologies for viral marketing and models used. We then turn to Measuring Influence on Instagram: a Network-oblivious Approach (Segev, Avigdor N., Avigdor E. 2018) to stress-test the idea of using a networked approach for this task and deduce that, while they achieved great results, their dataset was heavily augmented and leads us to use a similarly augmented dataset while still maintaining the network-approach. Lastly, we turn to Maximizing the Spread of Influence through a Social Network (Kempe, Kleinberg, Tardos 2003) to investigate different technical cascading models used within this context and on augmented dataset as postulated before.

None of these papers investigate the particularly descriptive social environment of Instagram with a network-based approach that simultaneously relies on heavily augmented data. Additionally, no paper hones in particularly on one specific niche of product (e.g. household items or fashion). It is our believe that focusing on one niche specifically will furthermore allow for much more rigorous investigation and predictive models given the inherent differences relating to a dataset.

Review of Relevant Prior Work

We investigate below different papers relating to virality and cascading across social media. In order to do so, we investigate both high=level and low-level approaches and stress test our hypothesis with research finding alternative ways to predict virality.

Mining Social Networks for Viral Marketing - Domingos 2008 Among these papers, 'Mining Social Networks for Viral Marketing' by Pedro Domingos (Domingos 2005) offers a coherent opening, albeit a sparse one. In it, Domingos

discusses approaches to designing "viral marketing" plans by looking at social networks and their structure and identifies key features that play a role. Domingos explains that users on these platforms should be reclassified based on their network potential, calling that reclassification the "network value of customers" (Domingos 2018). While he explains the shift from the traditional user definition to one of networked potential in terms of a gain in marketing potential, the featurization of a user as a node with a score associated based on "networked potential" seems like an important characterization when analyzing networks in terms of virality - an aspect we keep handy as we dive deeper into these topics and formulate our own approach. Domingo bases his networked value as a function of both the "intrinsic properties of the customer and the product, and the influence of the customer's neighbors in the network" (Domingo 2018). While vague in its approach (leveraging probabilistic inference over the joint model of all customers), this approach generalizes what would be an interesting approach to a viral analysis setup: labeling users / nodes with a function of likeliness to reshare/repost as well as a score of authority / centrality (who is going to see the repost / reshare and how authoritative is the user towards these people?) to draw probabilistic inference over them. Particularly the inclusion of an authority / centrality score seems to be missing in the Domingos' approach - not all networks lead to propagation of information and not all users that are networked also carry enough potential to influence others and fully propagate a call to action / information.

The paper discusses a cursory summary of approaches tested and tried in the section of "Factors that Influencer Network Value" that provide an overview of factors to be considered within a network of virality. High connectivity is stated as one factor (Domingos 2005) alongside whether a customer would "like" the product. As Domingos was working with the unrealistic EachMovie database, it was possible to access whether or not someone liked a certain movie, however in a real setting, particularly when monitoring a live social network, this becomes hard to measure. How can we determine whether or not someone likes a certain content? A guess here is the use of both authority of the user who presented a certain information to another user to influence the other user's relationship to the information. Another approach could be to identify clusters of similar users within a massive network and determine likelihood by proxy via the poster if the poster's audience finds themselves within a similar network. Domingo also briefly mentions the symmetry of influence as a point to consider when determining how the information would flow through a network. The direction of influence could again be approximated via the authority (or an authority score) of one user in regards to another with a large enough difference being necessary for successful propagation of information. The paper also mentions that asymmetric influence is quite widespread in networks and we postulate that it is particularly widespread on platforms such as Instagram in which there are different tiers of "influencers" ("macro" and "micro") that hold varying levels of authority. Lastly the paper mentions the proliferation to a highly dense network through the proliferation of a sparser pathway beforehand / the degree of connections necessary for virality. It is an interesting concept to consider - how do we weigh the different degrees of a network and the varying levels of virality of each when determining whether or not a user / node could proliferate virality? And how do we do so effectively (Depth-first search? Breadth-first search? Some heuristic?)?

Domingos reduces the the main question of the paper in regards to viral marketing to an optimization problem: "choose the set of customers to market to so as to maximize net profits" (Domingos 2005). Similar to Kempe, Kleinberg and Tardos (2003), Domingos was able to approximate this problem with 63% of the optimal by adding each customer only if it improved the overall profit of the set via a hill-climbing algorithm. While probabilistic inference can be quite expensive, the limited size of connectivity and followings/followers of most social media personas amortizes overall. However, as pointed out by Domingo, running this over an entire social media network is unfeasible and constraints need to be set (in the case of fashion, it could be a constraint to a certain area such as streetwear). The paper concludes by suggesting Markov Logic Networks as an improved model to speed up the development of a complex social network model yielding more accurate predictions given its combination of probabilistic modeling of Markov random fields and its expressiveness of first-order logic.

In summary, the paper is a weak paper when considering its technical write-up as most technicalities remain opaque and the models remain high-level - it begs the question of whether it can be replicated in the first place. However, this paper lays out working foundations of social networks and virality analysis that, if true, could be the foundational high-level features and models we could work off of for an analysis for virality in present day social networks.

Measuring Influence on Instagram: a Network-oblivious Approach - Segev, Avigdor N., Avigdor E. 2018 Segev et al. discuss the problem of identifying and measuring high-influence users on Instagram, broadly defined as those whose posts are viewed and shared more than those by their less influential counterparts. Given the inherently social nature of the platform, comprised of follower-following relationships most naturally represented in a graph, the problem often lends itself to network analysis. As the authors mention, past approaches to their question of gauging influence and identifying highly influential users have ranged from simple in/out degree measures to PageRank and variations of such techniques also covered in and relevant to our class material on link analysis. Segev et al. argue that the graph-based approach to measuring influence on other online social networks is inappropriate and an ill fit for Instagram. The main technical content of their paper comprises of comparisons of their "network-oblivious" approaches to graph-based techniques used on other social networks. They extract a handful of intuitive and non-intuitive features available from their Instagram dataset and then attempt to measure influence with a few regression models, including ridge regression, random forest, and multiple-regression. The technical bulk of their

paper, apart from some details on data preprocessing, arises from the comparison of the outcomes of the application of these models.

The strengths of this paper include the quality of its dataset and a distinctive first approach to a traditionally graph-oriented problem space. Despite using only 940,439 posts and 115,044 Instagrammers in their dataset (accounting for only roughly 0.58 percent of Instagram users, their distribution of log average views was extremely close to normal, indicating a good selection of sampled accounts and posts. Additionally, their removal of outliers indicating odd behaviors of engagement, including purchased artificial engagements or "Like You, Like Me" behaviors shows a nuanced intuition for the problem space necessary to appropriately clean and preprocess the data for proper analysis of authentic engagements and genuine influence. Additionally, though basic, their limited set of features is a good first approach to the problem without relying on graph analysis. Their choice of features is well-reasoned and justified by either common intuition or an exploration of their data indicating correlation between engagements and certain of the statistics they chose.

The weaknesses of this paper stem from the limited information captured in its features, which while intuitive, suggest only a superficial understanding of influence. The authors briefly acknowledge this, suggesting room for future work with the incorporation of more complex features, including temporal features, user demographic information, structural image components such as the presence of faces, and audience demographic information, such as the audience's location or age. Additionally, their problem formulation and definition of influence as merely the number of views per post is an insufficient measure of true influence. For instance, many people may see the content, but not be effected or moved to any particular behaviors, which would be a significantly stronger indication of influence. While none of their features or the problem definition were unrealistic, they only scratch the surface of understanding the dynamic and subtle nature of influence on Instagram. What they are missing and ought to address include a definition of influence that takes into account stronger evidence of influence, including influence to purchase, comment on, re-post or share content. Additionally, their treatment of traditional graph analysis techniques is also excessively dismissive and they do not substantially justify the superiority of their approach to other techniques, choosing only to run a basic PageRank algorithm on a subset of their dataset as a comparison baseline. What they are missing here is an effective way to capture or account for sub-communities within which highly influential users may emerge exclusively within those local contexts and only a network-based approach might be able to identify.

Promising further research questions to explore would include how augmenting the feature set used in this paper with network-based features affect their measure of influence or other more nuanced definitions of influence. Other interesting avenues to pursue would involve the addition of some of the more complex features they mention, involving computer vision and temporal analysis.

Maximizing the Spread of Influence through a Social Network - Kempe, Kleinberg, Tardos 2003 Kempe et. al consider the problem posed by Domingos (2005) of whether one can trigger cascades of adoption by seeding products or innovations with a particular subset of individuals. They show that with an analysis framework based on submodula functions, a natural greedy strategy obtains a solution provably within 63 percent of the optimal for several classes of models - a preferable alternative to the otherwise NP-hard problem of selecting the most influential nodes. Kempe et. al diverge from Domingos' approach with operational models explicitly representing step-by-step dynamics of the operations, avoiding the vagueness we found as a weakness for Domingos (2005).

The technical content of the paper is mainly reliant on techniques in graph analysis. In their approach, the papers uses a Linear Threshold Model in which a graph is defined with node v that is influenced by each neighbor w according to a weight b(v,w), and every node chooses a uniformly random threshold representing the weighted fraction of v's neighbors that must become active in order for v to become active. Additionally, the paper discusses the Independent Cascade Model starting with an initial active set of nodes and then following a process of discrete steps of activation according to a different randomized rule. The approach in this paper is a rigorous one to defining network dynamics in which they prove worst-case guarantees and explore other heuristics of the performance of the approximation algorithms they discuss. This is evidently quite relevant to the latter material of the CS 224W class, particularly related to cascading behavior, outbreak detection, and influence maximization (cascades being an indicator of influence).

The strength of this paper comes from the rigor of their proofs of approximation guarantees. It ultimately discusses and proves out multiple theorems, filling in the gaps for further explanation and mathematical formalism left by Domingos (2005). A weakness of the papers may be that it was authored in 2003 - the boundaries of approximability may no longer hold for modern online social networks given the massive scale we see today, particularly with Instagram that now boast over one billion users and generates billions of likes every day. For a network at the scale of Instagram, these approaches to seeding nodes to trigger cascades may not appropriately account for the immense diversity of the Instagram graph as well as its sparsity.

Summary of Critique & Direction From the papers discussed above, we can see several interesting themes emerge. They touch on questions of how to effectively and formally represent a necessary and nuanced understanding of what influence in an online social network comprises of and how that influence may translate to virality. Together, they suggest an area of exploration around the question of whether or not to accurately capture real-world dynamics of influence and cascades in online social networks, one needs to have features that are heavily descriptive of not only the individual nodes in the network, but also the relationships be-

tween those nodes and their immediate neighbors as well as other more distant sub-graphs. Additionally, they raise the question of the relevance of their findings to a massively scaled network such as modern Instagram, where it may be the case that generalized models cannot possibly holistically describe the underlying social dynamics, inter-connectivity as well as distinctiveness of various contained communities and how those factors may affect measures of influence or probability of cascades. To pursue this further, we extract more complex features involving networked, temporal and language-based features including sentiment analysis, to see how those may affect measures of influence.

Approach

Data Set & Augmentation Using Selenium on the mobile Chrome version of Instagram, this paper scraped high-signaling streetwear accounts on Instagram. We have created an initial, hand-labeled seed set based on influencers mentioned in reputable streetwear publications such as Hypebeast, encompassing 632 profiles. This paper then crawled out from those people, using their "Following" list to identify further high-converting streetwear influencers, determined by a threshold of enough in-edges from other high-profile influencers in our set, bringing our total profile count to 59,035 profiles. In total, there were 128,127 associated posts and 197,094 comments in the dataset. As one can see in Figure 1, the distribution of a PageRank scores expectedly follows one with a longer tail, and tall head. We can also see in Figures 2 and 3 that the largest profiles command the most likes and comments, yet interestingly, there is a large chasm in comments with mid-sized influencers being largely absent. This indicates that as we crawled out the network from our seed set of high-profile influencers, adding new ones if enough in-degrees could be observed, there seems to be a relationship between very small trendsetters and very large ones (in terms of following) that follow each other for inspiration.

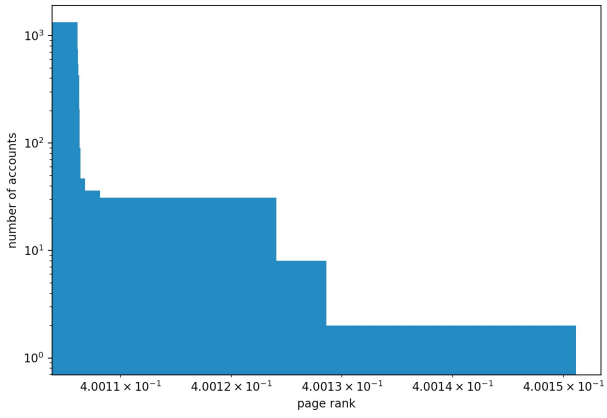

Figure 1: Page Rank Histogram

Given Instagram's complicated rate limitations and web structure, significant effort was used in establishing an infrastructure capable of pulling this information from the internet. This involved building out multi-threaded and parallelized instances that would crawl different profiles simultaneously. Robustness proved to be a key issue here as the infrastructure required building several checking mechanisms to ensure the rate limitations or web structure hadn't changed.

We define influence differently from Segev et al., where it was defined by the number of views. We instead choose to define influence by a post's appearance in the top section of relevant hashtag sub-communities (e.g. #nclgallery #hypebae #cleanfit etc.) on Instagram. Appearing in one of those sections, if looking in the right communities, means the post had been proliferating rapidly and has a very high rate of likes and comments within a short amount of time. We believe this to be a stronger indicator of virality as posts can only be surfaced in that section if they are going viral, as determined by Instagram's own ranking algorithms, whereas numbers of views represents only audience size, but not necessarily influence, and moreover can still be faked.

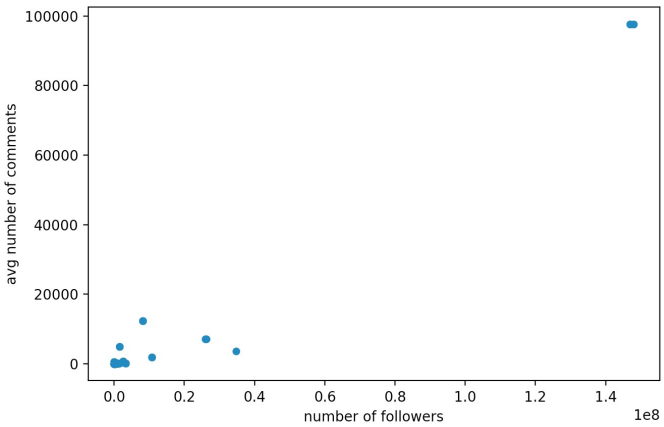

Figure 2: Number of Comments Against Number of Followers

We identified the right hashtag-communities by modeling them as nodes and crawling out via the recommendation section at the top as well as via the mention on high-profile user that we had identified earlier. Again, a threshold of in-edges had to be reached for a hashtag-community to be considered high-signal - in this case, in-edges stemmed from both profiles as well as other communities. It was possible to discover relatively small but highly engaged communities such as #nclgallery or #cleanfit that while only having about several tens of thousand of posts were boasting extremely engaged content. Additionally, posts in one community often became trending in an adjacent one. While our label was a binary of whether or not a post had been appeared in any of the sections, one could further granularize the experiment

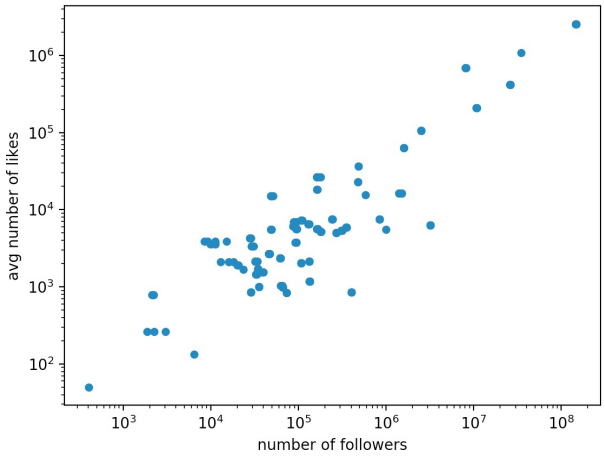

Figure 3: Average Number of Likes Against Number of Followers

by establishing the amount of virality as how widespread its virality was across all different sub-communities, a higher frequency indicating higher virality.

All information collected was stored in an anonymized format with usernames hashed out and any further individually identifying information (including pictures, etc.) was removed to keep the privacy of users fully intact.

Methodology After having established the hand-labeled seed set of high-performing influencers and using their network of people they followed and a threshold of in-degrees from the established network to find further high-performing influencers, we utilize further network analysis to establish network features that aid in our prediction. In particular, features can be found within the nodes and communities in the form of authority PageRank scoring, of in-degrees, of out-degrees, and further information on the sub-communities the node finds itself in such as modularity connectivity.

- **PageRank**
  From our dataset of following relationships we are able to establish a social graph. Given the limitation of accessing this data from Instagram, we were able to collect only out-edges from nodes directly, and thus in-edges indirectly. Nonetheless, we use this representation to run a PageRank to determine the importance of each profile in our network. We use the PageRank score of a post's user as a feature value for that post.
- **Number of Neighbors**
  The number of direct neighbors (followers and followings) are taken into account when evaluating the positioning of the node within its network.
- **In-Degree**
  As mentioned above, we indirectly collect in-degree data from following outbound edges from nodes in our network. A higher in-degree suggests stronger influence within the network, as it means that profile is highly sought after to be connected with.
- **Out-Degree**
  As mentioned above, we directly collect out-degree data from following outbound edges from nodes in our seed and 'top posts' network of profiles. A higher out-degree might suggest weaker influence as highly influential individuals tend to be more often followed by others than the ones to follow others.
- **Mean Neighbor Degrees**
  The mean of the neighbor's degrees was added a a further feature, indicating how well-connected the neighbors are.
- **Sub-Community Modularity**
  Modularity of a network measures the strength of division of it into modules as high modularity would indicate many connections between nodes in the modules but only few connections between nodes in different modules. The modularity of the sub-networks the profile was found in was added as a network feature.

Additionally, social networks offer a vast amount of additional information in the form of likes, follows and, in particular, a wealth of textual information - on Instagram, this can be found in the form of comments underneath pictures that come with its own set of likes as well. In Segev, Avigdor N., Avigdor E. 2018, we saw how important and telling non-network features are in the determination of influence and virality and, consequently, we will embed additional feature information into the network we are establishing, augmenting the nodes and thus the data inherent to the social network with the aim to improve our predictive capabilities. Language data comes with its own set of particularities and requires pre-processing and vectorization. The following steps were followed to clean exploit the textual data into a format that could augment our network features.

- **Stemming & Lemmatization**
  Real-word textual data is inflected, conjugated and manipulated. However, no matter its manipulation, words such as 'wants' and 'wanted' carry the same essential meaning. In order to count their instances more easily and remove the grammatical manipulation, we utilize stemming and lemmatization that groups words of derivationally related words together by removing inflection and reducing them to common base forms. Stemming is usually a rather crude heuristic process in which suffixes and affixes are removed. Lemmatization is a more serious version of stemming that takes further account of morphology, returning the very base of the word, otherwise known as the lemma. This also required POS Tagging explained further below. Lemmatization consistently outperformed stemming and, although computationally more expensive, was incorporated into the final results of this paper.
- **POS Tagging**
  POS Tagging or Part-of-Speech Tagging is a natural language understanding technique in which words within

a document are annotated based on their types (e.g. verb, noun, adjectives, etc.). This is not only required for lemmatization to function, but is furthermore helpful for any linguistic models that go past a simple frequency counter or bag-of-words model. It can be particularly helpful for bigrams and trigrams.

- **Stop Word Punctuation Removal**

  Empty words such as articles ('the') and filler words are empty in meaning and thus would only obfuscate as a feature (or contributing to a feature) when analyzing textual information. As such, they need to be removed in the preprocessing step along with punctuation.

- **Emoji Retrieval**

  Given that we are investigating a social media corpus, our textual data is highly dominated by linguistic idiosyncrasies stemming from generation Z and millennial. Much of the textual information includes references to emojis. As particular character sequences, these are separately extracted and treated as their own words.

- **Unigrams, Bigrams & Trigrams**

  Beyond a simple word histogram in which instances of singular words (or unigrams) are counted, bigrams and trigrams add additional semantic complexity. Bigrams and trigrams are tuples of words that follow each other. When utilized, this allows for more complex and sequential information to be captures (e.g. a trigram of '(I, want, this)' is a strong indicator of purchasing intent and requires sequential information as the words by themselves are rather meaningless. The higher order the n-gram, the exponentially more computationally expensive the processing of the textual data becomes. As such, this paper used both bigrams and unigrams when evaluating its textual data. Word features were limited to the top 10,000 to limit the amount of irrelevant information but still allow for enough breadth to cover the diverse vocabulary across different posts on social media.

- **TFIDF**

  TFIDF, or term frequency-inverse document frequency, is a statistical vectorizer for our textual data. Rather than raw counts frequencies, TFIDF statistically established the importance of a document within a corpus- often used on in text mining. As a measure of uniqueness of the word to a certain document within a corpus, TFIDF rewards words that are specific to a certain document (or in this case Instagram post) - a measure that is highly applicable here.

- **Sentiment Analysis**

  Textual data harbors a plethora of sentiment, particularly within a social context. While sentiment analysis is a large field within natural language understanding itself, simple methodologies can be used to enrich a dataset. Most simply, we compiled a dictionary of psychological words that could indicate strong or weak desire (e.g. 'need', 'want', 'love', 'stan' as strong words and 'like', 'vibe' as weaker words) and counted the occurrences within a posts' comments. This information was then normalized and appended to our feature vector.

We were required to sample a subset of our full dataset in order to create a balanced representation of examples for training, as posts that appear in the top section of hashtag sub-communities are significantly more sparse than those that do not (approximately less than 1 percent of posts). Additionally, as suggested by Domingos (2005), this paper will limit the size of the community we are looking at drastically by solely focusing on the streetwear community on Instagram. This paper also applies some of the regression models used in Segev, Avigdor N., Avigdor E. 2018 with the addition of networked and other more complex features. In order to efficiently identify the users / nodes necessary to trigger a cascade, Kempe, Kleinberg, Tardos 2003 suggest heuristics such as the Linear Threshold Model or the Independent Cascade Model, and Domingos takes this further by suggesting a hill-climbing model. While these heuristics offer a great starting point for addressing the computational efficiency of this problem, we found in the course of our investigation that such techniques would have been excessive for our particular dataset, as all processing, training, and predictions were computable within minutes on the current data set.

We will run four main experiments, varying our feature sets for each. As a baseline, we begin by using the basic features as alluded to by Segev (2018) as well as the ridge regression model. We then incorporate networked features, which examine the surrounding graph structure of nodes in our dataset, and finally incorporate NLP features, which are intended to account for the sentiment towards the posts in our dataset. In addition to applying the ridge regression model in Segev (2018) as a baseline, we also leverage 7 other models to identify the best approach.

Models

In order to identify the best results, we employ different models with varying strengths on our data set.

- **Logistic Regression**

  This models serves as our starting model as a more simple approach to the problem. A simple model that is easily interpretable and predicts the probability of a variable. It is a simple and efficient algorithm with wide flexibility.

- **ADA Boost Classifier**

  Creating a strong classifier from many weak classifiers, the ADA Boost algorithm is particularly good at precision. It is, however, senstitive to data imbalances which could become problematic in our case working with social data.

- **Ridge Classifier**

  A more common classifier, the Ridge Classifier avoids collinearity between features which could come handy given the tight relation between some of our network features.

- **Stochastic Gradient Descent (SGD)**

  Stochastic Gradient Descent is a model that is particularly efficient on large datasets as it can converge quickly. As

we collect more and more data, this model can be particularly helpful, albeit lacking in complexity.

- **Passive Aggressive Classifier**

  This model continually adjusts to bad inputs, and overcorrects. It has been successfully used on Twitter data is efficient for the classification of large, social data sets.

- **K-Means Classifier**

  Examines the distances between cluster centroids and new data points for classification. This models performs well with low-dimensional, hand-chosen features so its performance should be limited in our case.

- **Decision Tree Classifier**

  A Decision Tree or Decision Tree Classifier is a supervised learning method that can be used both for classification and regression. In a tree-like structure, decision trees embed complexity and rules within their leafs, decision nodes and branches. Given the amount of different features and particularly textual data, a decision tree is a good model choice as it allows for the establishment of rather complex yet versatile models. While prone to overfitting, an ensemble model resulting in a random forest could be used to counter that if promising.

Results and Findings

Basic Features

The models were run on four different feature scenarios: standard social features (likes, followers, number of comments, etc.), standard social features + network features, standard social features + language features, and standards social features + network features + language features. We ran all seven different models across all different scenarios and the results can be observed in Table 1.

Network features only caused a slight increase in performance across all models while the inclusion of language features caused a significant spike in prediction accuracies. Best results were obtained when network features and language features were combined, however, the difference to language features only is rarely picked up. This could be caused by the overwhelming amount of language features compared to more classic features and more complex models and feature engineering might be needed to optimize the modeling. Decision Trees and ADA Boost were most attuned to network features. Given their complexity and rule-based nature, Decision Trees would model networks well.

That being said, the accuracy with network only features maxed out at 22%, which is quite the low accuracy. This outcome differs greatly from the findings in Segev (2018), which indicated that even simplistic summary statistics could by highly indicative of influence. This may be because their measure of influence (views) is much more strongly correlated with the features they selected, such as the follower or like count, whereas our measure of influence (appearance in the 'top section') is much more weakly correlated with those simplistic measures. Given the rarity of appearing in the top posts section across all posts, it would make sense that cursory features derived only from a particular profile as in Segev (2018) would not sufficiently predict virality across the entire Instagram ecosystem or fashion sub-communities. Given the unbalanced nature of the data set as well as the very small subset that we were working with, the accuracy could also be severely limited by the circumstances and further work would need to be done with a much larger data set.

Network features might require

Decision Trees are also well attuned to language and thus, the Decision Tree Classifier achieved the highest overall accuracy at 67.2% as well as the highest accuracies for each scenario.

Networked Features

Models using this subset of features, which incorporates the structure of the social graph as input, performed the second worst. We looked at both the PageRank scores of the profile nodes associated with posts as well as their in-degrees. One explanation for the relatively poor performance here is in the sparsity of the network data we were able to collect within Instagram's limitations. As we were only able to crawl through 'Following' lists and not 'Followed by' lists, we were only able to approximate a true representation of the network as we could only follow out-edges. In particular, the in-degrees of the nodes was significantly underrepresented in our dataset. We see this as the maximum in-degree was only 42, which seems low as only less than 1% of the network in our dataset given that this is a highly focused and curated subset of Instagram profiles. The findings here somewhat confirm Segev et. al's conclusion that networked features may not add significantly to predictions of influence. However, from our later findings, we see that that while they may not be sole predictors, they can at least enhance the predictive power of other features.

Interestingly, both the networked feature set and the basic feature set perform extremely poorly with the K-Means classifier (accuracies of only 5.7% and 5.7% respectively). This reflects the great variance in basic and networked features across all of the top posts.

Language Features

We see a significant jump in accuracy with models leveraging language features extracted from the comments under posts. Apart from the standard text processing of stemming and lemmatizing, we also made sure to specially handle and coalesce emojis, which carry strong sentiment and are pervasive across social media. Emojis have to be handled particularly as, for example, an emoji repeated consecutively within a string should not be treated distinctly from that single emoji appearing only once in its own string, as the meaning of both are not the same. We used the TFIDF vectorizer to capture the relative importance of our unigrams, bigrams, and trigrams to the rest of the corpus. We used bigrams to capture the strong intent in phrases such as "need this" or "love your", which tend to reflect strong sentiment specific to the content of the post. Not surprisingly, the most complex

Prediction Results				
Model	Standard	Network	Language	Network & Language
Decision Tree	20.3%	22.0%	76.2%	**76.2**%
SGD	14.6%	15.0%	76.2%	76.2%
Logistic	14.8%	15.0%	75.8%	75.8%
Passive Aggressive	14.6%	15.4%	75.8%	75.8%
K-Means	05.7%	05.0%	75.0%	75.0%
ADA Boost	19.9%	21.0%	73.6%	73.8%
Ridge	14.6%	15.0%	70.1%	73.8%

Table 1: Table of Prediction Results and Accuracies

model here, the Decision Tree, performed the best given its abilities to form complex rules.

Networked + Language Features

This feature set yielded the highest accuracies amongst the models overall, but did not perform much better beyond the feature set of NLP features except for with a few of the models. This can be attributed to the likelihood that the comments carry the strongest signals for predicting virality as opposed to the network structure. Our explanation for this is that because of the rarity of a post appearing in the 'top posts' section, even highly connected nodes rarely have posts that appear there, so it must be that features apart from the social connectivity of a post's owner differentiate the post from others enough to go viral. As the networked features remain relatively constant across posts for a particular profile in a limited time window, only spikes in likes and comments and the actual content of comments can account for the remaining variability that could possibly separate posts likely to go viral from those that are not.

Discussion & Future Work

A large amount of effort for this paper has been spent on establishing the right pipelines to extract data from Instagram as well as interesting features to allow for a complex and augmented data set. The rate limitations inherent to Instagram were more aggressive and creative than initially anticipated and required creative workarounds on our end that consumed more time than expected. However, with the pipelines set, more data is incoming and with a growing data set, further and more substantiated conclusions can be drawn from the data set. Most notably, we expect network features to play a larger role once the sparsity of network connections is alleviated through the addition of more profiles.

Nevertheless, our findings suggest that the most predictive features of virality within subcommunities are language-related. Further enhancements could be made to the language feature set by examining trigrams in addition to bigrams as well as granular sentiment analysis work.

Further, while our label was a binary of whether or not a post had been appeared in any of the top sections of selected hashtag communities, one could further granularize the experiment by establishing the degree of virality as how widespread its appearance was across all sub-communities, a higher frequency indicating higher virality.

References

Pedro Domingos. *Mining Social Networks for Viral Marketing*. 2005.

Segev, Avigdor N., Avigdor E. *Measuring Influence on Instagram: a Network-oblivious Approach*. 2018.

Kempe, Kleinberg, Tardos. *Maximizing the Spread of Influence through a Social Network*. 2003.