Prediction of Escalation in E-Cigarette Use Among Twitter Users

BY

Akshay Uppal
B.Tech, Guru Gobind Singh Indraprastha University, Delhi, India, 2014

THESIS
Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Chicago, 2019

Chicago, Illinois

Defense Committee:
Dr. Elena Zheleva (Chair and Advisor)
Dr. Tanya Berger Wolf
Dr. Robin J. Mermelstein (Psychology)

Dedicated to

my family,

for their love and dedication for success in my life.

ACKNOWLEDGMENTS

I would like to thank my mentor Dr. Elena Zheleva for her consistent feedback and guidance throughout the master thesis without which this work would not have been possible. Furthermore, I would like to thank my committee members for their valuable feedback. I would also like to thank all the professors and teachers who I was fortunate enough to interact in my career, who all helped to build a foundation of education and encouraged to dream big.

I would like to like to thank my parents for supporting me to pursue my interests. Special credit goes to my sister *Nandita* for instilling a sense of belief and being a pillar of moral support.

Lastly, I would like to acknowledge the support of all the friends who showed interest in the research and provided valuable feedback.

AU

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

FIGURE **PAGE**

LIST OF ABBREVIATIONS

SVM	Support Vector Machine
LSTM	Long Short Term Memory
Bi-LSTM	Bidirectional LSTM
RF	Random Forest
RNN	Recurrent Neural Network
ML	Machine learning

SUMMARY

Electronic cigarettes have witnessed a growing market in the last few years. While they been advertised as a healthier alternative to combustible cigarettes, they can lead to nicotine addiction and weakening of the immune system. The most popular e-cigarette brand in the United States, Juul, was subject to a lot of public scrutiny due to its marketing practices targeted towards adolescents. There is a growing body of research in the usage of tobacco, cannabis, and illicit drugs, with studies suggesting that the usage of one drug leading to an increase in the likelihood of another drug, commonly referred to as a "gateway drug." However, little is known about whether we can predict the escalation from tobacco use to other drugs and whether e-cigarettes are contributing to such escalation.

In this work, we collect data from Twitter for users who use Juul hashtags and look for temporal patterns that can predict the escalation from tobacco to cannabis. After filtering out the commercial and promoter users out from the data, we look for changes from first e-cigarette to first cannabis use. We reduce the problem to a supervised classification problem that can detect this escalation. We make predictions for different time intervals using several classifiers which consider lexical features from the tweets, together with user features, such as status count. Our findings have implications for making adequate public policies in the area of health and education.

CHAPTER 1

INTRODUCTION

Electronic cigarettes have witnessed a growing market in the last few years. They provide a physical sensation similar to inhaling traditional cigarette through the usage of aerosol to its users [2; 3] with some products also containing nicotine. Some of the studies indicate e-cigarette as a new smoking cessation tool [4; 3]. Although some studies suggest that e-cigarettes reduce the health risk by 95% in comparison to traditional cigarettes, they pose yet another danger. They can increase the likelihood of traditional cigarette consumption either due to nicotine addiction or due to social normalization of smoking behaviour[5].

Among the e-cigarette brands, Juul has become the most popular brand with 75% market share at the end of July 2018 [6]. Juul products contain addictive nicotine. It is most trending in the social network, especially among teens. It became a coveted teen status symbol and a growing problem for high school middle school students [7] (referred to as "Juul Phenomenon" afterward). It was introduced by Pax Labs in June 2015 which split into a separate entity as Juul Labs in 2017. It is claimed to have been targetted towards adults as a healthier alternative to traditional cigarettes but some of the initial marketing strategies were targeted towards the youth [8]. Juul came under a lot of public heat in the middle of 2018 due to its alleged marketing towards teenagers and high school students. This led to an intervention by the FDA and further leading to the suspension of its social media accounts including Facebook and Instagram except Twitter. It also forced Juul labs to tighten its sale of e-cigarette exclusively online [8].

Alcohol, tobacco, and cannabis are the most common drugs affecting adolescents in the U.S. [9; 10]. Substance use related to the above drugs have been studied quite extensively and substance use disorder has been associated with a variety of adverse health, social and economic outcomes [11], specifically affecting adolescents. Various studies have looked at substance use and its correlation with health and socioeconomic factors. Some of the studies have suggested a correlation of drug use with a measure of low self-esteem, aggressiveness, crime, and anti-social behavior [10; 12; 13; 14]. There has been quite a lot of research that looked at the early use of drugs among adolescents and their future dependence. Some of the studies have suggested that early adolescent smoking leads to increased risk of future nicotine dependence substantially [15; 16; 17; 18; 19; 10]. Further there has been growing research suggesting early cannabis use also leading to risk of cannabis use disorder [20; 21; 22; 23; 24]. There also have been studies to look at the substance use of one drug leading to an increase in the likelihood of another drug and different school of thoughts have looked at the problem with different ideas. One idea of thought is relating to acquisition of substance use behaviour, referred to as ”gateway drug” from the field of psychology that suggest that use of psychoactive drug can be coupled to increase probability of other drugs [25; 23; 26; 27; 28; 29; 30; 31]. The second idea in which it is studied is the idea of addiction-prone predisposition associated with behavioral dis-inhibition and conduct deviancy [32; 33; 34; 31], but little is known on how can we predict the escalation of behavior to other drugs among the users. So in this thesis, we utilize the data of e-cigarette use from Twitter to study the shift of user behavior from tobacco smoking to cannabis use. It makes an important area of research to make accurate health policy in the area of education and law.

Social media gives a unique platform for users to describe and explain their emotions. Juul phenomenon has been quite popular in Twitter with over 1.69M tweets from more than 800K users since the inception of Juul from June 2015 to 2018, including but not limited to tweets from commercial users like Juul Labs, teens expressing their sentiment related to Juul e-cigarette on Twitter and various news or research tweet being published against the growing concern of Juul. Younger generation represents a large base on Twitter (Studies conducted in 2014 indicates that 37% of all users on Twitter lie between the age of 18 and 29 [35]) and they also seem to be most active in expressing sentiment related to Juul. With such a fast creation of content by users on a public domain, Twitter gives a great platform to understand their behavior.

In our study, we utilize data from Twitter which contain rich information about various users connected to Juul which could be commercial users marketing Juul product (e.g. Juul Labs), regular users expressing their sentiment or research related tweets being published about the growing concern of Juul. Twitter gives a great platform to understand the Juul phenomenon for mainly two reasons; it has a collection of a lot of tweets relating to Juul from users since the inception of Juul (2015)) and secondly Twitter has been extensively studied by other researchers to either to understand user sentiments [36], social behavior or influence [37].

In this thesis, we intend to look at the escalation in the behavior of users from smoking e-cigarette to cannabis substance. We have collected the data for all users associated with the Juul phenomenon from the year 2015 to 2018. As we require data for cannabis use, we collect the data for the above users who also mention cannabis drug use in their Twitter timeline. The users in the data include commercial users including Juul labs, third-party vendors like vape shops in addition to regular users expressing their

sentiment about Juul. We then filter out those commercial and research-related users from the data. To accurately predict the switch of e-cigarette use to cannabis among the regular users we utilize statistical approach from machine learning. We divide the data within different interval spans by year and month, to make an accurate prediction of a switch in behavior. We organize our thesis by first presenting related work in the field, the second section explains initial data exploration phase of Twitter data, next sections explain the methodology and experimental setup, and we finally conclude with the main takeaways and future scope.

Specifically, the main contributions of the thesis are the following:

- We present new dataset containing all of the tweets of users and their features (e.g. followers count, following count, etc.) related to Juul phenomenon which also contains cannabis tweets of same users from 2014 to 2018 from Twitter.
- We also perform sentiment and retweet cascade analysis on the data.
- We perform classification prediction with different intervals to accurately predict the switch of users from smoking e-cigarettes to cannabis.

CHAPTER 2

RELATED WORK

In our quest to understand the escalation in the behavior of users from smoking e-cigarette to cannabis, we build upon the prior work in the field. Social media gives a unique platform for the user to describe and explain their emotions, which provides an interesting area for researchers to understand their behavior. Some of the prior work in the domain of mining substance use through social media have looked to understand smoking behaviors, and other substance like cannabis, cocaine and alcohol [38; 39; 40]. With some of the studies also related to poly-substance use in teenagers [25; 23; 26; 27; 28; 29; 30; 31]. Studies in relation to specifically e-cigarette have also looked at marketing of e-cigarette on Twitter [41; 42] and sentiment or polarity analysis of e-cigarette on Twitter data [43]. As data is transmitted among users in the social network we can see the development of reshare networks or more popularly studied as "Cascades". Extensive research has been carried out in understating their growth, structural properties and virality [1; 44]. In the following sections, we explain the prior work in the above domains in a bit more detail. We also explain the prior work in utilizing deep neural network employed in context social networks for text classification.

2.1 Data mining for substance use

There have been studies on mining data related to smoking and including other substances (cannabis, cocaine, alcohol) to understand the user behavior [38; 45; 39; 40]. The paper Valente et al. have used social network data to understand and prevent substance abuse [38]. Similar studies have been carried

out to study alcohol, smoking, cocaine, and cannabis substance abuse [46; 47; 48]. Paper by Zhang et al. [46] utilizes a rule-based approach to predict the recovery and relapse of alcohol use disorder in AA (Alcoholics Anonymous group) Twitter data. They combine linguistic and psycho-linguistic features using a structured approach to accurately predict the recovery and relapse of AA attending users. Paper by Hu et al. looked at the correlation between teenage performance and smoking [47]. Jimenez et al. [48] utilized various data mining techniques to understand the usage of alcohol, cannabis, and cocaine among students. They collected drug motives data for each user using survey questionnaire and further utilized classifiers like the artificial neural network, decision tree, nave Bayes, k-nearest and logistic regression [49] to classify the user as a drug or non-drug user based on drug motives and explained the predictive features that separated those users. We also explain the models utilized by us in our work in the following chapters.

There have been studies to understand the pattern of the poly-substance use of alcohol, marijuana and cigarette usage among young adults (Moss et al. [10]) as explained in the previous chapter. Previous work have suggested substance use of one drug affecting the likelihood of usage of another drug among users, specifically in the case of nicotine and cannabis [15; 16; 17; 18; 19; 10], with other studies approaching the idea of ”gateway drug” [25; 23; 26; 27; 28; 29; 30; 31] mentioned in introduction chapter. There has been little work to predict the escalation of behavior to another drug among the users. In this thesis, we utilize the data of e-cigarette use from Twitter to study the escalation of user behavior from e-cigarette smoking to cannabis use. It makes an important area of research for understanding it to make accurate health policy in the area of education and law.

2.2 Sentiment and polarity analysis

Social media has been extensively researched to either understand the sentiment or opinions of the user. Dai et al. mined social media data to understand public opinion related to e-cigarette [43]. They separated the organic tweets from commercial Tweets to evaluate the general public attitude towards e-cigarette. In the Organic tweets, they consider research study or news that express sentiment towards e-cigarette which is in contrast to the category of users which we choose to assign in our study, in which we consider them as a separate category from organic users. For our case, we consider organic users as individuals who express their sentiment related to Juul and label them as regular users. Dai et al. utilized a Nave Bayes classifier to classify users as support, against, neutral, commercial or irrelevant categories. Their study is motivated to prevent non-smokers youth from consuming e-cigarette. Studies by Martinez et al. [50] also looked at the sentiment of user-related to e-cigarette and indicating e-cigarette as not being a health hazard and a larger collection of positive sentiment as opposed to negative sentiment for the e-cigarette. Further Lazard et al. looked at the public reactions to the regulation of e-cigarette by FDA. Social network data has also been used on understanding user behavior in the context of smoking, Fetta et al. utilized the social network for understanding the cessation of smoking in secondary school adolescents [39]. We provide a brief of change in user sentiments for our data in exploratory data analysis chapter.

2.3 Text classification with machine learning and recurrent models

Various statistical machine learning and neural models have been applied in various domains including the language domain. *Support vector machine*(SVM) is a discriminative classifier that classifies

data based on an optimal hyper-plane [51]. It has been extensively used in regression problems [52; 53], applied in various domains including sentiment analysis [54], spam identification [55].

Combination of ML models to improve generalization has studied as *Ensemble learning*. It is a form of learning in which multiple learners (referred to as base learners) are combined to solve the same problem, hence increasing its ability for generalization. It has proven quite successful in the domain of supervised learning [56; 57; 58; 59]. *Random forest* represent a form of ensemble learning methods that work by training decision trees on sub-sample of data. They represent bagging which is a common ensemble way of learning which reduces the variance in the data. They are better in comparison to vanilla decision trees [60] which are prone to over-fitting. They are applied in various such as spam identification [61], bioinformatics[62], and extensively in language modelling task [63]. *XGBoost* is an optimized, scalable gradient tree boosting [64] implementation of gradient boosting that works on boosting of decision trees [65]. Tree boosting has been a highly effective approach in predictive modeling [64], which has given the state-of-the-art performance in various classification benchmarks [66].

Apart from machine learning models, deep neural networks have been quite successful in different machine learning tasks like image recognition, speech processing, and language processing. RNN models are a type of neural network that can model information in time. RNN uses cells that persist information over time which makes them excellent for text classification task [67]. The problem with RNN is the inability to learn long-dependency in the text due to vanishing gradient problem. LSTM networks [68], (a type of RNN network) avoid this problem by regulating the information in the cell state using the input, forget and output gates, which makes them perform better as opposed to traditional

BOW (Bag of words) models[69] which assumes independent assumption among the words. Unlike other deep neural models, LSTM's share the same weights across all of the steps, which greatly reduces the number of parameters needed by the network to learn. LSTM's has been successfully utilized in various area of bio-informatics, speech recognition and specifically language modelling tasks like entity extraction [70] , document classification [71], sentiment analysis [72], emotional recognition [73] and also in the domain of social media for fake news detection [74]. We utilize LSTM networks for tweet and user classification.

2.4 Information diffusion via cascades in social networks

Online social networks like Twitter, Facebook and LinkedIn have witnessed growth at an unprecedented level in the last decade where millions of people can interact at any given time. In such OSN (online social network) the rate at which the information spreads is quite fast. Social networks have become so pervasive that they can influence the behavior of individuals or groups which has led to research in many areas including information diffusion [75] (referred as "influence maximization" in literature), information cascade (Leskovec et al. [76]) with application in areas including but not limited to viral marketing [76], rumor control [77] and social recommendation [78]. Some of the work has also been concentrated to understand the diffusion trajectories. Understanding the spread of information is crucial for businesses to promote their product and governments to regulate public opinion. As data is transmitted among users in the social network we can see the development of reshare networks or more popularly studied as "Cascades". More generally these Cascades can be represented by a tree, where the root is the source node that initiates the information and each edge would represent the information being transmitted between users. There has been quite extensive research on how these cascades grow,

analyzing their structural properties, virality, and properties contributing to its growth over time [1]. These cascades are more generally tend to be burst information in a short amount of time, often termed as a broadcast network. In some cases, they have also been represented by a branching process infecting only a few users and forming a long chain as shown by Anderson et al. [1] for LinkedIn sign-up cascades. The average path length of a cascade tree also referred to as structural virality (also termed as ”Weiner index”), concerning the structure of tree indicates how deep or shallow a tree can get. In addition to discrimination of cascades based on the structural properties, they have been also classified based on social mechanism specifically social and individual participation cost that govern the growth of these cascades (Cheng et al. [44]). In this paper, we try to look at the properties of retweet cascades on Twitter concerning our data presented in data analysis chapter.

CHAPTER 3

DATASET EXTRACTION

As of 2019, Twitter has about 326M active users [79] with over 1.3 billion accounts (end of second quarter of 2015 [80]) which generates 500 million tweets each day [81]. In our study, we extract data from Twitter which contain rich information about commercial users like Juul Labs, third-party vendors like vape shops and also organic users which are individual users expressing their sentiment about Juul. Twitter gives a great platform to understand the Juul phenomenon for mainly two reasons; there exists a lot of tweets related to user sentiment concerning Juul from its inception (with start of the first tweet by Juul Labs in 21st April 2015)

and secondly, Twitter has been extensively studied by other researchers to understand user sentiments [36], target marketing or modeling social behavior and influence [37]. In this section, we explain how the data was collected for users mentioning tweets related to the e-cigarette *Juul* from Twitter.

Twitter provides public access to its data through Twitter public API [1] but it also poses a limitation on the number of tweets we can gather and also the rate at which we can capture the data. We can only gather 3200 tweets from user timeline and further it is rate-limited by 15 requests per 15 min interval [2], which is the main reason for using Crimson hexagon in conjunction. Crimson Hexagon (now called

[1] twitter.com

[2] https://developer.twitter.com/en/docs/basics/rate-limits

Brand Watch [1]) also has a limit on extracting only $10k$ tweets for any time range. Also using Crimson, we don't get the full range of user features (e.g. following count, followers count, etc.) in comparison to Twitter API [2]. To resolve these limitations we use Crimson Hexagon in combination with Twitter API. We use Crimson Hexagon to gather all tweet Ids (which is a unique identifier for any tweet on Twitter) and all of the user and text features (e.g. tweet-text, followers-count, following count, etc.) are extracted using Twitter API. Further, to resolve the $10k$ count limitation of hexagon, we extract data by each day. We also utilized Tweepy [3] for our extraction process which is a python library for Twitter API.

The meta-data extracted from twitter contains information like tweet text, followers count, following count, when the tweet was created, if it is re-tweeted and geographical information of tweet, etc. In our study, we extract the data for users mentioning Juul on Twitter by crawling the data using the two-step process mentioned above. We first extract the data for all users mentioning Juul on Twitter using Juul related hashtags (#juul, #juulvapor, #juulnation, #doit4juul) from Jan 2015 - Dec 2018. Juul hashtags were based on marketing trend, used by both by commercial (Juul Labs, third party vendors) and individual users [8]. We set the start date as Jan 2015 which is the inception date for E-cigarette Juul (Pax labs introduced Juul e-cigarette on June 1- 2015) [82]. We extract a total of $1.69M$ tweets and $887k$ users and these users form the basis for the second extraction process. We plot the growth of

---

[1] https://www.brandwatch.com/

[2] https://developer.twitter.com/en/docs.html

[3] http://docs.tweepy.org/en/latest/

tweets containing Juul hashtags and users mentioning Juul hashtags by year in Figure 3, illustrates the growth of tweets by year in comparison to the number of users.

To understand the sentiment of users related to consumption of other substances, we look at the tweets of the extracted users mentioning other substances, specifically cannabis. The main reason for using cannabis is that there have been studies suggesting tobacco use leading to consumption of other illicit drugs, in the context of the gateway drug and addiction-prone predisposition among adolescents. Also, cannabis and tobacco are two of the most common form of drugs among teenagers. To predict the escalation in the sentiment from consumption of e-cigarette to cannabis use, we extracted the data for the same users above that also mention cannabis use. We perform a second extraction in which we select the tweets of the selected users from Jan 2014 - Dec 2018 mentioning cannabis using the synonyms for cannabis, the full list is mentioned in (Table I). Some words like *green, grass*, and *joint* which could be ambiguous were excluded from the list. We gather $10.42M$ tweets from $554k$ users mentioning cannabis-related substance. Out of $887k$ users we gather tweets from $554k$ users, the rest of the users we assume are the ones who don't mention cannabis at all in that interval span (termed as *mono* in the following sections). We finally combine both the initial Juul data and cannabis data collected above to form a final dataset (mentioned as dataset afterward). Our final dataset contains almost $12M$ tweets with $887k$ users. Final distribution of users and tweets are shown in Table II, including the count of tweets in Juul dataset, cannabis dataset, and our final dataset.

3.1 Filtering out commercial users

Our goal is to understand the behavior of individual users who express their sentiment or experience with Juul products. Therefore, we want to exclude commercial users like JUUL labs, and third-party

cannabis words
'weed','ganja','marijuana', 'cannabis', 'mary jane', 'marihuana', 'hash', 'reefer', 'hashish', 'bhang' 'green goddess', 'locoweed', 'maryjane', 'spliff', 'wacky baccy','sinsemilla', 'doobie', 'CBD' 'acapulco gold', 'THC' , 'hemp'

TABLE I: Top Key words for substance tweets. We use the list to Capture poly-type user from mono-type users. Reference : powerthesaurus.org

vendors like vape shops. In the dataset, we find that not all of the tweets are related to sentiments of regular users. As Twitter is extensively used for news, or to promote brands some of the users in the dataset had to be filtered out to isolate the tweets of regular users. By regular user expressed tweet we mean the individual users who mention their experiences or sentiments related to Juul rather than promoters, or tweets related to research about the growing concern for Juul. In our data we identify three main categories of tweets, news or research study related tweets that represent the tweets from news or research channels on Twitter, tweets intended to promote Juul related products like vape shops, etc. and finally some of the tweets that contain information related to individual users which we intend to capture from Twitter. Based on the tweet categories we classify the users into three categories namely

- News or research users
- Promotional users
- Regular users

Dataset	Tweets	Users
Juul data	1,692,201	887,180
cannabis data	10,420,444	554,815
Final data	12,088,584	887,180

TABLE II: Distribution of users and tweets extracted by each extraction process. Final data results combining both of the Juul and cannabis data, which would contain users consuming single subtstance Juul and users who consume cannabis in addition to Juul.

To filter promoters and news users (termed as commercial users afterward) out from the dataset we create a manually annotated dataset which contains 200 samples for each of the category mentioned above. A sample of tweets from each category is shown in Table III. To predict tweets among three categories we used a bag-of-words approach. We use tf-idf approach to reduce it to n * V matrix, where n is the no of words and V is the vocabulary. We further use Truncated SVD to reduce the dimensions of the tf-idf matrix to n* 100 dimensions. We train a linear SVM to train the classifier to predict the three categories. We tested with other classifiers like LSTM [83] which preserves the long term dependency of text and random forest but best results were obtained from the SVM model as shown in Table IV. Note that we predict the category of the tweet for each user instead of predicting the user directly. This decision was based primarily because it would be easier to classify each tweet instead of the user which could be harder to predict if all of the tweets are not consistent by the user.

We preprocess each tweet text by removing the URLs, hashtags (#) and re-tweet symbols ('RT'), removing the stop words(using NLTK [84]), tokenizing and performing stemming to reduce each of the words to their base form. We perform 10-fold cross-validation. We get the accuracy of 71.34% on the test set(Table V). Using the trained classifier we predict each of the tweets in the data. In our case,

Category	Example
News or research	”Current Issues Surrounding the Role of Medical Marijuana in Nursing Homes By #expertwitness Dr. John Fullerton” || ”In 1 minute learn everything you need to know about why we’re investigating JUUL. https://t.co/r2rhJjmznW”
Promotional	”$25 OFF CannabisDelivery promo code: GreenRush420 at #420 #420delivery #weedmaps #weed” || ”The #1 accessory for your Juul Coming soon to https://t.co/Pygk93PXGH RT to win a free case! https://t.co/vKtYC9AbV3”
Regular user	”Some people like to bake pies. I like to bake myself... #stoned #weed #cannabis” || ”You thinned out the liquid in your new pods. It’s a scam now. RIP juul. #juulvapors JUULvapor #juul #juulpods and they’re $25-30 in Cali..”

TABLE III: Tweet type examples for each category (ie. news, promotional and regular) in the annotation data. It contains multiple examples with each tweet separated by ”||”.

we use 200 manually annotated data for each of the three tweet categories but more annotation would provide better accuracy.

The SVM classifier predicts the tweets for each of the user. A distribution of tweet percentage on a subset of 4k is shown in Figure 1. To finally classify each user, we assign the category for the user based on the maximum number of tweets for the category mentioned by the user. e.g. if the user mentions more tweets related to promotion based on the prediction of the classifier, we label it as a promoter. We obtain the distribution explained in Table VI.

model	precision	recall	F1-score	accuracy
LSTM	0.71	0.69	0.69	0.686
SVM	0.72	0.71	0.71	0.7134
random forest	0.60	0.60	0.60	0.60
majority classifier (majority regular user)	1.00	0.28	0.44	0.27

TABLE IV: Classification metrics (on test set) by different classifier to predict tweets in three categories of news, promoter and regular users

category	precision	recall	F1-score	support
news	0.68	0.76	0.72	51
promoter	0.70	0.76	0.73	51
regular	0.78	0.60	0.68	48

TABLE V: Classification report of the SVM classifier (mentioned above) on test for each of the category, (ie. news, promotional and user expressed tweets) with 71.34% accuracy

3.2 Categorization of mono and poly-type users

In this study, we want to differentiate the users who consume only e-cigarette substance as opposed to users who consume cannabis in addition to the e-cigarette. With our final dataset of regular users can be further classified based on their behavior among two broad categories as mono-type and poly-type users. Each of the specific categories is time-sensitive, meaning their label is based on the tweets they mention in a specific time interval.

- Mono User: User who mention tweets about Juul related hashtags at least once but have not to mention Cannabis tweets at all in a specific time interval.

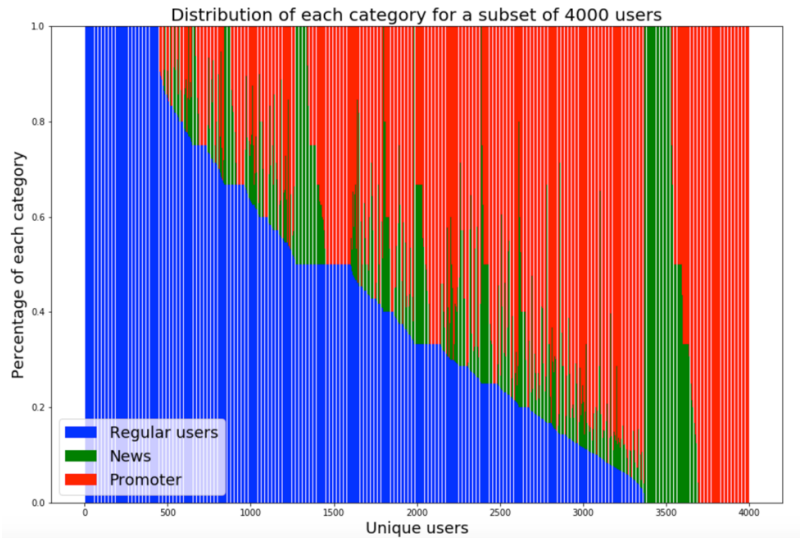


Figure 1: Represent the percentage of tweet categorized in each category for each of the unique user. Blue signify the regular user, green are news users and bars labelled as red are the promoters

- Poly User: User who in addition to Juul also mention Cannabis at least once in a specific time interval on their Twitter timeline

In this study, we are only looking at Cannabis as another drug which users consume in addition to the e-cigarette. We restrain illicit drugs such as cocaine and others from our study. Classification in the above two categories helps us to understand the behavior of the user who consumes only a single substance as opposed to other users who consume more than one substance. Any user who mentions cannabis-related tweet at least once in their Twitter timeline are termed as poly-user and the rest of the users are termed as mono-users (users consuming only Juul e-cigarettes). All of the users in our cannabis data represent the users who consume cannabis in addition to Juul and hence they form poly-

category	# of users
Promoter	91,130
News users	122,326
Regular user	673,724

TABLE VI: Breakdown of the user in each category. We choose the category for each user based on the maximum number of tweets among the three categories mentioned by the user. (e.g. if the user mentions more tweets related to promotion based on the prediction of the classifier, we label it as a promoter). Note: we predict on the category of the tweet for each user and instead of predicting the user directly, as it would be easier to classify each tweet instead of the user if all of the tweets are not consistent by the user.

user category	# of user	total # tweets
Mono-user	241,408 (36%)	374,761
Poly-user	432,316 (64%)	651,2011

TABLE VII: Distribution of tweets in poly and mono category

substance user and the remaining users are termed as mono-users (representing single substance users). (shown in Table VII)

We can see that in terms of count we have a higher number of poly users in comparison to mono users and correspondingly they have a higher number of tweets. In Figure 4 we can see that the poly users almost remain constant with a higher number in the initial years in comparison to mono which are growing exponentially. In the next chapters, we aim to understand if any of the mono users escalate to become poly-user in next year or month and whether can we accurately predict the escalation.

Common words for poly-type user	Common words for mono-type user
['weed', 'i', 'smoke', 'juul', 'like', 'marijuana', 'get', 'smoking', 'say', 'u', 'people', 'make', 'one', 'smell', 'hit', 'go', 'need', 'shit', 'high', 'the', 'bro', 'good', 'pod', 'know', 'want', 'hash', 'time', 'kid', 'it', 'drug']	'juul', 'i', 'pod', 'say', 'hit', 'like', 'it', 'bro', 'dont', 'one', 'get', 'nicotine', 'u', 'class', 'please', 'boy', 'my', 'hitting', 'writing', 'buy', 'me', 'kid', 'give', 'rip', 'smoke', 'the', 'fuck', 'and', 'let', 'life'

Figure 2: Top 30 most frequent words for each category

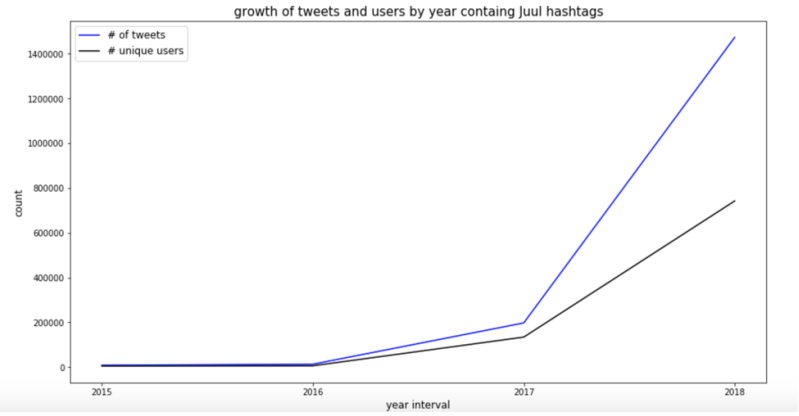


Figure 3: Growth of tweets by year

user category	# of user
Juul-before	52,702 (12.2%)
Juul-after	379,614 (87.8%)

TABLE VIII: Distribution of tweets for Juul-before and Juul-after users

3.3 Further classification of bi-substance users

We are interested in the underlying question whether the poly-user started consuming cannabis before they mentioned Juul or was it after mentioning Juul motivated them to mention cannabis. To understand this question we categorize users based on the occurrence of cannabis or Juul tweet on their Twitter timeline. We sort all of the tweets for each user containing mention of Juul and cannabis on their timeline and check for the first occurrence of their cannabis and Juul tweet. If the Juul tweet was mentioned before the first occurrence of cannabis tweet then we categorize them as Juul-before and on the other hand if cannabis tweet was the first tweet mentioned before any tweet related to Juul they are termed as Juul-after.

- Juul-before: User who first mention Juul before any cannabis tweet
- Juul-after: Users who mention first Juul tweet after their tweets related to cannabis

We obtain the final count for each category as shown in Table VIII.

We see that there are considerable (12%) of the mono users who escalate their sentiment after consuming Juul. These are the users who after using Juul started consuming cannabis substance which form the main users we are trying to predict. In the next chapters we explore when do they escalate from mono to poly-substance and how can we accurately predict the escalation for those users.

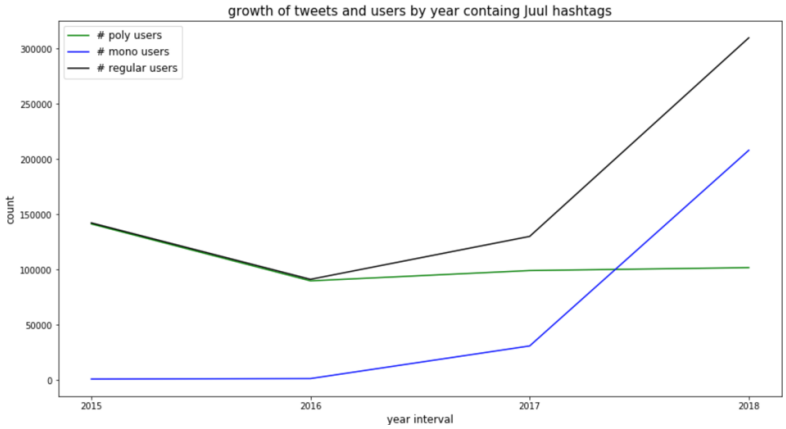


Figure 4: Illustrates the growth of poly and mono users

CHAPTER 4

EXPLORATORY DATA ANALYSIS

In this section we explore the data to form a preliminary data analysis. We present the sentiment change for users for Juul form 2015 (inception) - 2018. We also look at the re-tweet cascades of the users to understand the structural virality and diffusion rate.

4.1 Sentiment analysis

From the start of Juul product the sentiment towards it has changed a lot based on the events that happened in the press like FDA issuing warning to Juul labs regarding its alleged marketing towards youth and suspension of its social account from Facebook and Instagram due to growing concerns. We try to understand the change of sentiment over the period of 2018 which had lot of press events and larger amount of data. We have utilized LICW (Linguistic Inquiry and word count) which have been the popular among traditionally psychology based approach [85]. It is a dictionary based approach, which is based on frequency count for 64 manually different categories constructed in psychology theory. These include categories related to part-of-speech and emotions. In this study we only look at the change of positive and negative emotion over a period of a year.

Furthermore, we utilize NLTK Vader analyzer [86] which is capable to detect positive and negative emotions. The analyzer contains three categories of positive, negative and compound. Compound is the resultant emotion which can be positive and negative. In the Figure 6 we plot the change positive and negative emotion in the JUUL data (data containing Juul hashtags) for the 2018 year. We look at the

data of regular users filtered from promoters and news. We finally map the change of sentiments to the actual events in the 2018 year. We can see that April 2018 mentions about the FDA warning and we see the negative emotion spike in May 2018. We have a sort of a delayed effect of sentiments after the event.

4.2 Cascade analysis

In this section, we look at the reshare network of users mentioning Juul related Hash-tags(#juul, #juulvapor, #juulnation, #doit4juul). So as we discussed in [1] paper we try to create a cascade retweet network for out dataset. As the Twitter API does not return us the cascade or network of users involved the retweet in any chronological order, we had to build the cascades based on the following network of users. Initially, we sort all of the tweets from their creation date and we filter the tweets that have the re-tweet count of less than 1. We take the source node as the user which created the tweet initially. Then the remaining users re-tweeting the same tweet are considered as followers in the cascade. We utilize the twitter API to get the users who follow the source node. Users following the source user directly are considered to be at level-1. Then we keep building the cascade recursively at each level, for the next level we look at the users who are the followers of the node at the previous level. We keep building the cascade recursively until we exhaust all of the remaining users or no user can be assigned to the next level in which case we terminate building our cascade. We look only at the users following the source user directly or indirectly through a series of users. An example of the cascade is shown in Figure 7, each node represents users belonging to one retweet cascade, with the source node at the center of the graph highlighted in yellow, green nodes (which are at level-1) represent the users following the source node and the blue nodes are nodes at level-2 following the green nodes.

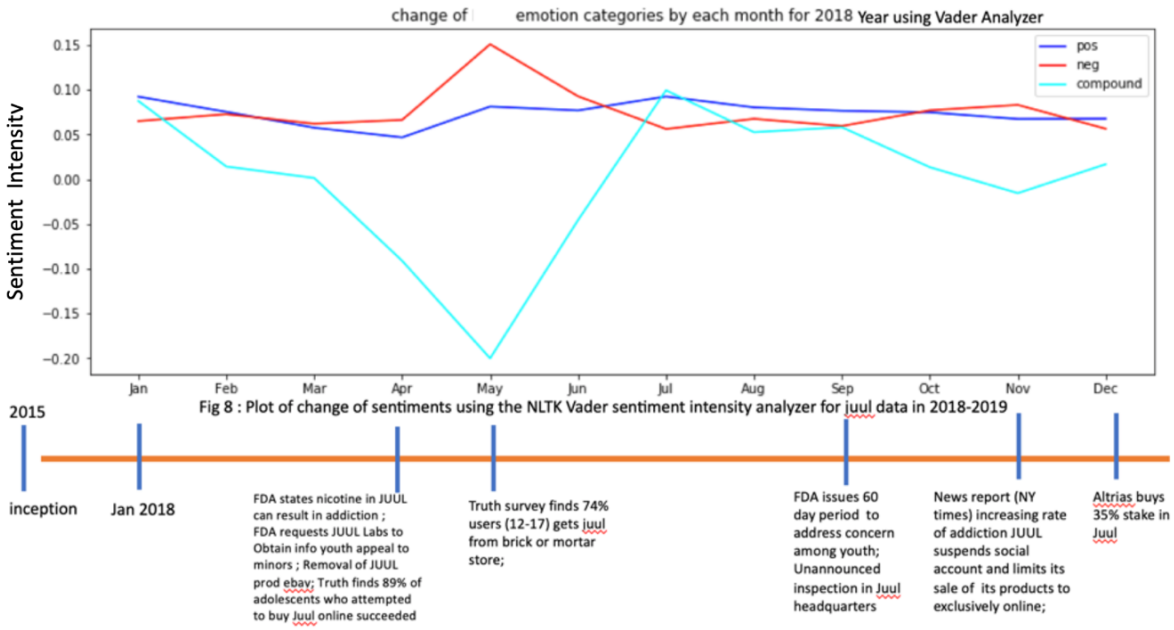


Figure 5: Emotion categories for each month for the year 2018 using NLTK vader analyzer

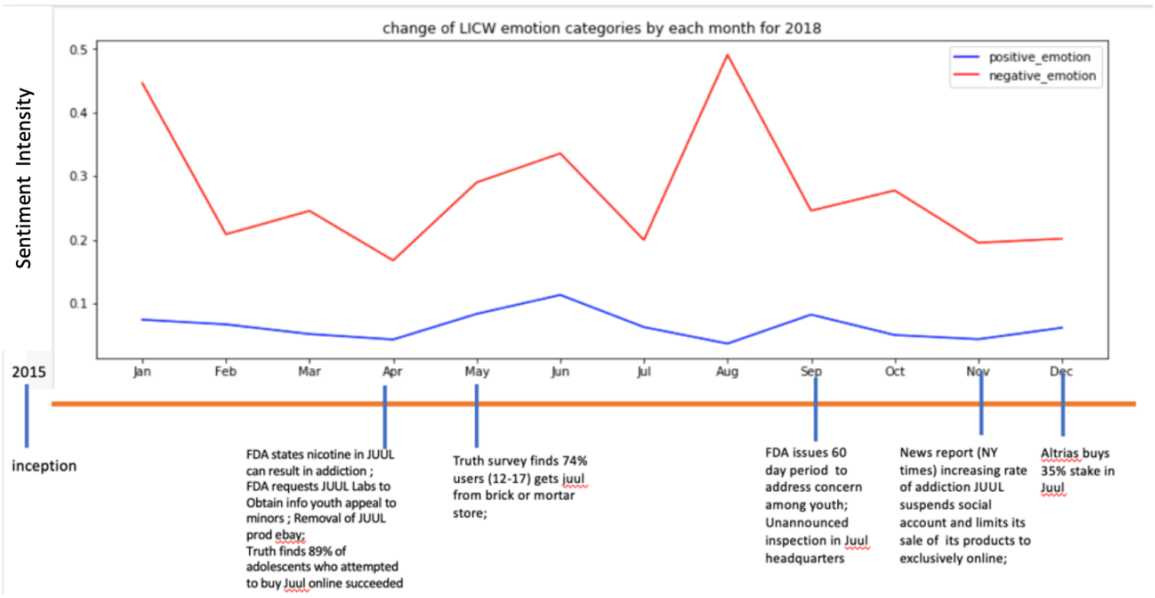


Figure 6: LICW (using Stanford empath) categories of positive and negative emotion categories for each month (based on psychological word count for the specified categories for the year 2018

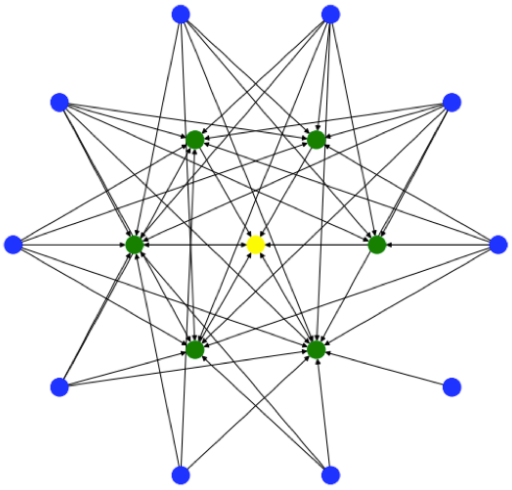


Figure 7: An example of retweet cascade. The yellow signifies the source node, blue nodes are the follower node of the source node. Cascade is created recursively at every depth with each of the node being the followers of the preceding depth

We look at the dataset containing Juul hashtags as we are trying to understand the Juul phenomenon. In the data, we have 41,391 cascades with retweet count > 1. In this analysis, we look at 828 of the subset of cascades (termed as cascades afterward) which is a random sample of the initial set. In the next section, we look at the properties of these cascades. The average number of nodes in a cascade are 6, with 372 and 2 as the maximum and the minimum number of nodes respectively. 99th percentile of cascades have 43 nodes, signifying way less no of nodes as compared to LinkedIn cascades presented by Anderson et al [1].

4.2.1 Properties of cascade

The cascades mentioned by Anderson et al. [1] comprised of 332 million LinkedIn signup cascades with large depth size which follows a tree-like structure. On the other hand, the cascades in our study

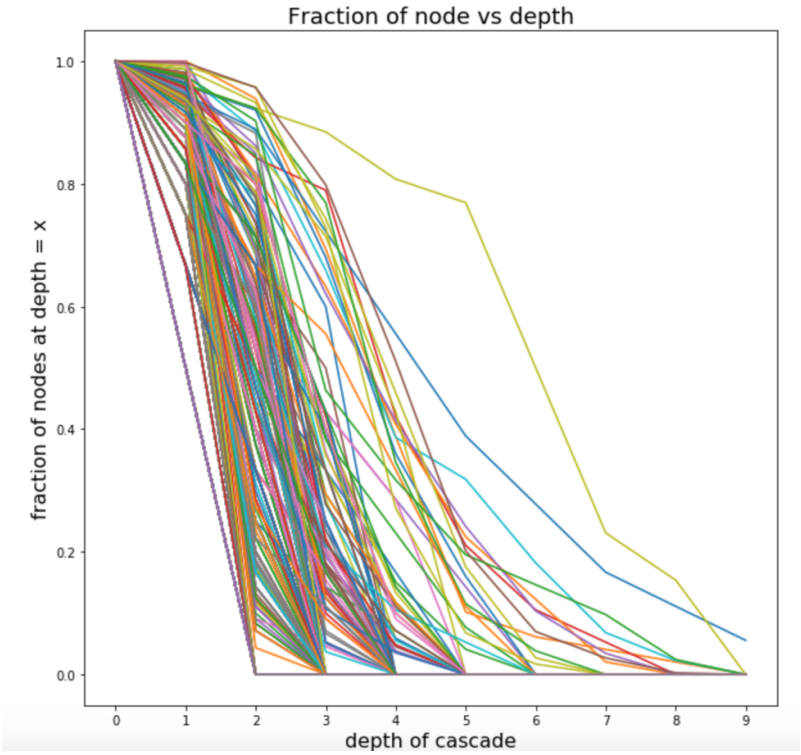


Figure 8: Plot of depth vs fraction of nodes at that depth. In our retweet cascades of Juul users have a smaller depth of with 1.77 as the average depth size. (max depth of 9 and a minimum depth of 1) We can see that most of the cascades have a larger fraction of nodes at earlier depth, signifying a broadcast-based network.

have a lower number of nodes with a shallow network. We look at the fraction of cascade as a function of its depth size, Figure 8 represents the fraction of nodes at a particular depth. Most of the cascades have fraction of nodes at depth 2 or 3 and then very few nodes remain in the next depths. It highlights the point that Twitter is a transient network as mentioned by Cheng et al. [44], where retweets grow rapidly in the early stages and then die suddenly. Broadcast diffusion networks are shallow networks, where information flows in a broadcast fashion. We see that our cascades are broadcast networks with the shallow network. The Cascades in our data have an average depth size of 1.7 and a max depth of 9. Most of the cascades in the dataset dont form deep networks and representing a shallow network. Wiener index is a metric used to highlight the virality of a network [1], it refers to the average shortest path in graphs. Similarly, we plot depth of the cascades vs Weiner index (shown in Figure 9). As most of the cascade size were quite small and few were outliers with more than 100 nodes so the cascade size is normalized (log base 10 scales). All of the cascades in our datasets have Weiner index < 1 which highlights low structural virality, a trend common in broadcast diffusion-based network [1].

We look at the growth of cascades over a period of time. As highlighted by Anderson et al. [1] with LinkedIn cascades, the growth of large cascades is somewhat linear in time with medium nodes (nodes < 500) but in our dataset, we have a smaller number of nodes ($<< 500$). We look at the pattern of cascades over a period of time. We plot the growth of cascade vs. fraction of node depth as shown in Figure 10. Figure 10 represent growth of cascade since the inception of the first tweet of the cascade by the source node to the tweet by the last node in the network which represents the lifetime of the cascade. We see somewhat approximately linear growth with some dispersion as shown in Figure 10.

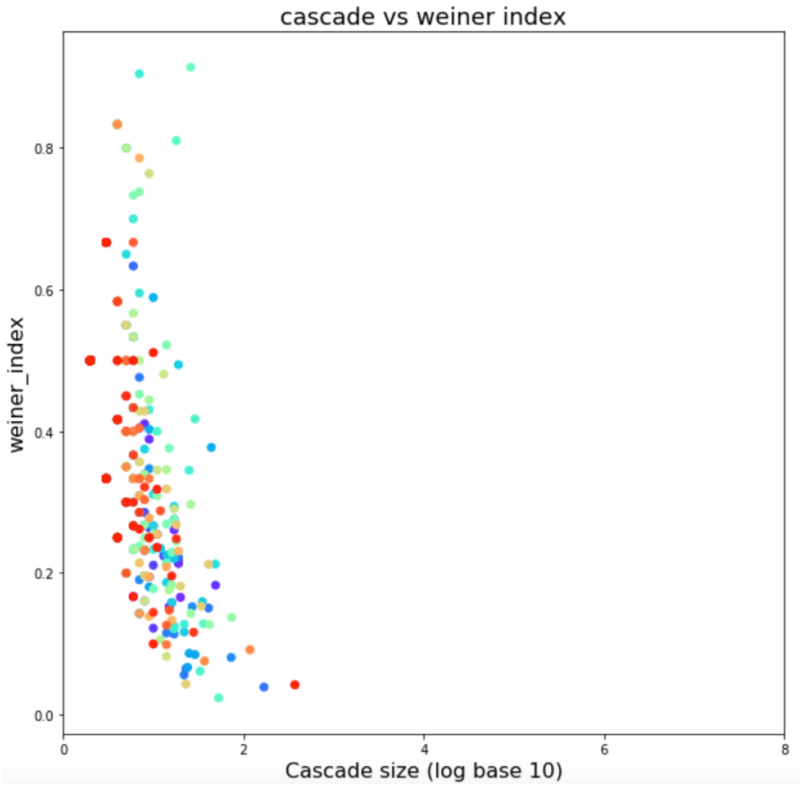


Figure 9: Structural virality(weiner index) as a function of cascade size(log base 10). We have shallow based networks leading to a low structural virality, it highlights a broadcast based network (Anderson et al. [1])

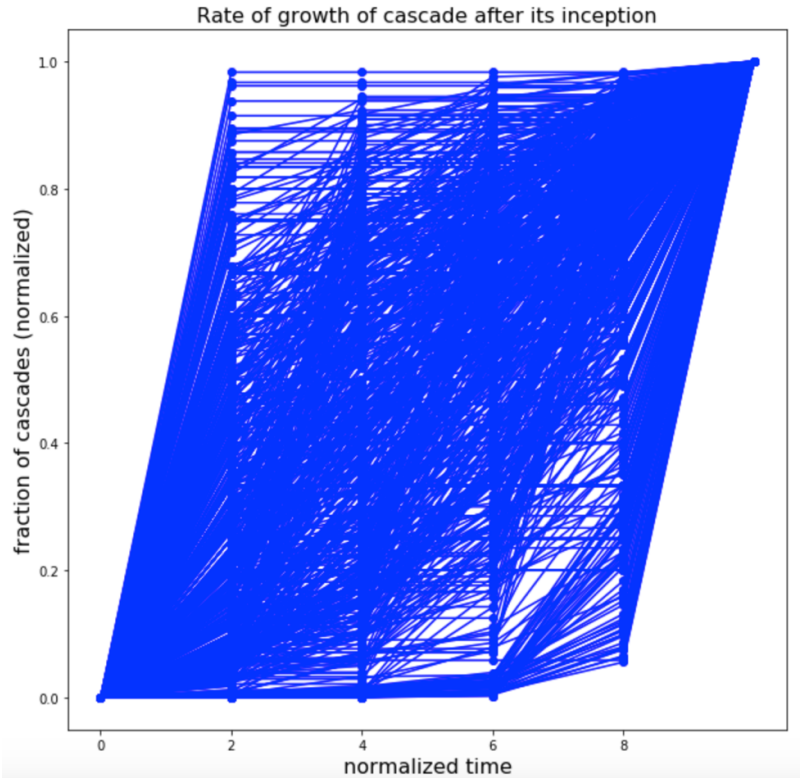


Figure 10: Pattern of cascade growth over time for Juul cascades. It signifies the growth of cascades over time. As the data has small number of cascades most of the nodes ($<< 500$), they seem more dispersed than LinkedIn signup cascades mentioned by Anderson et al. [1]. Most of the cascade growth is approximately linear.

CHAPTER 5

METHODOLOGY

In this work, we are interested in predicting the escalation from mono to a poly user and we cast the problem as a binary classification problem. In this section, we present the methodology used to solve the problem. We discuss the organization of data, feature engineering, and machine learning models.

5.1 Organizing user timelines

We saw in the data exploration chapter that there is a considerable amount of poly-users (12.2%) who have started consuming cannabis after they expressed sentiment towards Juul e-cigarettes. This signifies an escalation of mono-user to poly-user. Escalation here refers when a user is mentioning a single substance i.e. Juul also starts conveying information about Cannabis. We aim to predict the users who escalate when they first started sharing information about Juul to when they first expressed tweet about cannabis on their timeline, which we term as Juul-before users in the previous chapter. To capture first cannabis and first Juul tweet we utilize the same hashtags for Juul and cannabis that we used to extract user data. Figure 11 illustrates the delay for each of the Juul-first users from when they first mention Juul to the time when they mention cannabis on their Twitter timeline, the delay in the figure is divided in 30 days interval for each user. We can see that majority of them escalate within a 30day period, few of them escalate after a year($>$366 days) and a very tiny fraction of them take more than two years to escalate to cannabis.

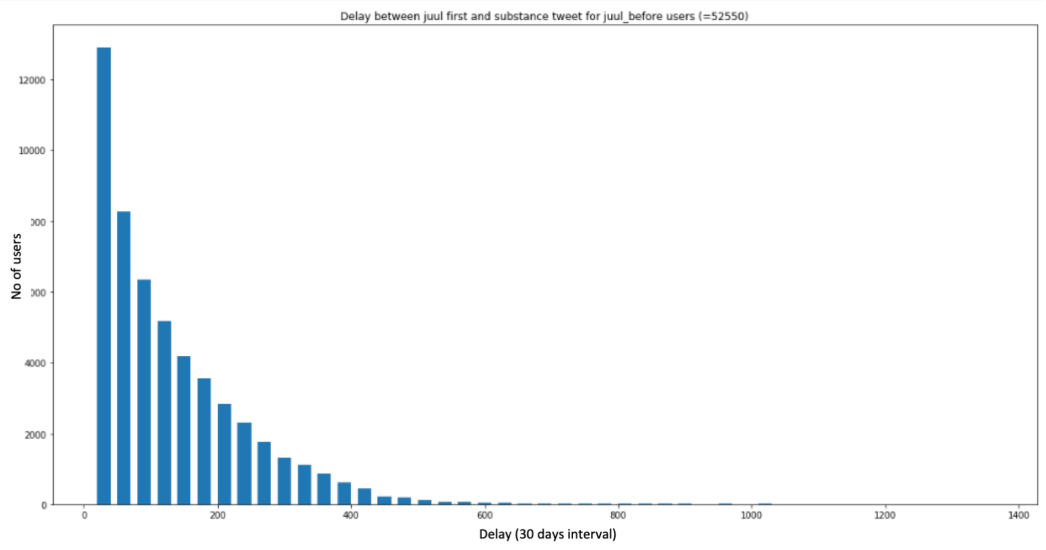


Figure 11: Delay distribution between the mention of the first Juul to first Cannabis tweet on the Twitter timeline. A large proportion of user mention Cannabis after Juul within a delay period of 30 days.

5.2 Feature engineering

In our study, we have utilized two sets of features to predict the escalation of users. One set of features represent the features related to the user, and another set of features is the text features that come from utilizing the tweet text. We describe these two types of features next.

5.2.1 User features

User features refer to features associated with the user on Twitter, which is unique to each user. Some of the features we gather from Twitter API help to understand the discriminating features of the user. We explain those features below:

- Followers count: The number of followers associated with this account

- Following count: The number of accounts the user is following
- Status count: The number of Tweets (including retweets) issued by the user
- Unigrams: Number of unique words in the Tweet
- Favourite count: The number of Tweets this user has liked in the accounts lifetime
- Listed Count: Number of lists the public user is a member of

We don't use the above features directly. To have a common scale and prevent bias due to uneven ranges while training the machine learning models, we utilize a log version of them. Inspired by Pietro et al. [87] paper,(in which they deploy a range of user features to for mental illness detection for CL-Psych 2015 task), we use the similar set of user features highlighted in Table IX. To predict the escalation of mono to a poly user, user features convey the information about the user behavior which forms discriminating features to distinguish mono and poly users. For the prediction task we describe below we consider the users in a specific interval, therefore we obtain the features in that specific time interval. Unigrams and status count mentioned in Table IX are time-sensitive but rest of the features due to the constraint of data are have the most recent values of the dataset (i.e. Dec 2018).

5.2.2 Text features

Text forms an important part of expressing user sentiment, which can help us discern users. Our data contains rich information of tweet text for each of the user, which is the text posted by the user on Twitter. The text represents a continuous string of words in which the context is determined by the neighboring words. Although traditional models have looked at the text as an independent collection of words (termed as bag-of-words approach [88]), in our study we have analyzed the tweet text using

both as bag-of-words approach [88] using the statistical machine learning models, and a recurrent neural network that preserves context-dependency in-text [89]. For the machine learning models except for the recurrent model mentioned below, specific feature engineering part includes TF-IDF vectorization. This changes the text sentences into TF-IDF vectors and all the tweets reduce to TF-IDF matrix. TF-IDF score is calculated as follows:

$$tfidf_{\text{t,d}} = tf_{\text{t,d}} * idf_{\text{t}} \tag{5.1}$$

where tf stands for term-frequency and idf stands for inverse document frequency. For further details, one can refer to the book by [90]. Using TF-IDF we obtain (n * V) matrix where n is the number of words and V is the vocabulary of the text. As the vocabulary size can be quite huge for our collection of tweets, we also utilize truncated SVD (single value decomposition) to reduce TF-IDF matrix to lower dimension of 100 features, a process termed as latent semantic analysis. It is a common technique used in conjunction with TF-IDF, in natural language processing and information retrieval domain to reduce the dimensions, presented by Deerwester et al [91]. Due to the large vocabulary size the the TF-IDF matrix returns quite high dimension features, therefore we reduce each of the text features to 100 dimension. The above features obtained by from above are used for all of the experimental setup using text features. The text features obtained from above are used in all of the machine learning models except in the case of recurrent model. For the recurrent model, we utilize the embedding layer which maps the word into a specific vector space. The embeddings are based on the weights of Glove [92] trained on a corpus of 2B tweets. The machine learning models are explained in the next sections.

f1	log number of followers
f2	log number of following
f3	followers/following ratio
f4	log number of listed count
f5	log number of favourite count
f6	log number of unigramns
f7	log number of status count

TABLE IX: User features for predicting escalation of mono to poly user, inspired by Pietro et al. [87]. Except number of unigrams(f6) and number of status count(f7), rest of the features are not-time sensitive due to the constraints of the data.

5.3 Combining user and text features

To improve our prediction task we utilize the user and text features in combination. We combine the above 7 user feature with the 100 features of text obtained by TF-IDF and SVD. In this approach, we utilize all of the machine learning models except the recurrent based model as RNN model are more suited for sequence-based tasks.

5.4 Predicting escalations

Some of the users mentioning Juul related tweets have also talked about cannabis-related substances as we saw in the previous sections. We aim to predict those users who escalate from e-cigarette to Cannabis. To accurately capture the escalation, we look at the range of year intervals spanning from 2015 to 2018, to compare if any of the users have escalated as compared to the previous year. We divide the data into the different time range intervals of year and month span. We are only looking at the mono users who will escalate to poly in the next year. In the following sections, we look at

predicting escalations by year and month interval. We utilize the input data of the previous year to make an accurate prediction for the future year. By dividing the data for each year range, we aim to understand whether the previous year data can be predictive to understand the escalation of the user from mono to a poly user. We look at the data in a different range of buckets and the output labels are based on next year information for the user. For example, the 2014-15 interval we take the input data containing user and text features and the user data of next year (2016) to serve as the output labels for machine learning models. The features we consider in this problem are user and text features. Each of the features uniquely determine the user in that interval. Some of the user features in our dataset are not time-sensitive due to the constraint in the data. Number of unigrams and status count Table IX mentioned in the table are taken based on the interval span mentioned, however, rest of the features are not time-stamped so we have taken those users value based on the end date of the dataset (i.e. Dec 2018). Text features are time-sensitive hence they are filtered based on the interval span selected. The user can only be among two classes either representing a mono user or poly user, if a user mentions any tweet related to cannabis in next year, we treat it as poly user otherwise we term it as a mono user. We have reduced the problem domain to a supervised learning task in which we are predicting if the user will escalate its sentiment from a mono to a poly user. As we are predicting the escalation of users, we have combined all of the tweets for each of the user in that specific input time range. In the next section, we explain the machine learning models used to learn the binary classification task.

5.5 Machine learning models

We employ a range of classifiers from classical machine learning models like SVM, random forest to more recent machine learning models like XGboost [65]. For understanding text, we also utilize LSTM

[93] (RNN based classifier) which have been quite successful in the language domain as it captures the context of the text as opposed to the bag-of-words approach.

Except for XGBoost which is an open-source package, all of the models are part of Scikit-Learn library [94].

Decision tree learning is a predictive modeling approach that constructs decision rules to classify the data into predefined classes [95]. It is a supervised learning approach used widely in the statistics, data mining and machine learning domain. We use Random forest model in our task that is based on decision trees.

Random forest represent ensemble learning methods that work by training decision trees on sub-sample of data. We utilize this method for our classification as they reduce the problem of over-fitting encountered with vanilla decision trees [60].

Support vector machine(SVM) is a discriminative classifier that classifies data based on an optimal hyper-plane [51]. It has been extensively used in regression problems [52; 53], applied in various domains including sentiment analysis [54], spam identification [55] and news classification in a social network, which makes it great to be used in our domain for tweet classification.

XGBoost is an optimized, scalable gradient tree boosting [64] implementation of gradient boosting that works on boosting of decision trees [65]. Tree boosting has been a highly effective approach in predictive modeling [64], which has given the state-of-the-art performance in various classification benchmarks [66], which forms the main motivation for using it.

Long short term memory(LSTM) (Recurrent Neural network (RNN) based classifier) have been quite successful in the language domain as it captures the context-dependency of the text as opposed

to bag-of-words approach [93] which assumes independence among the words. We exploit the Bi-LSTM network which is a special case of LSTM, which allows information from past and future states simultaneously. It has proven the state of the art in the language domain [70]. Although the version mentioned in [70] is for entity extraction, it can also be extended for our case of tweet classification as well.

We further utilize a *embedding layer* as well in our network, whose purpose is to map each word in the discrete library to lower-dimensional continuous vector space. This helps in the extraction of semantic feature detection without manual feature detection.

We remove the top Cross Random Field layer mentioned in Huang et al. [70]) which is used to make a structured prediction for each entity. As it is more suitable for entity extraction task, we remove from our model, keeping a simple bi-LSTM model. We also don't utilize character-based embeddings as mention by Huang et al. We utilize an embedding layer, followed by a Bi-LSTM layer and have a fully connected layer at the top, followed by element-wise sigmoid activation function as we have binary classes (basic outer layer hierarchy is; Embedding Layer $->$ Bi-LSTM $->$ FC). Figure 12 shows the architecture layers for the Bi-LSTM model. We use sparse cross-entropy as the loss function for training the network.

Ensemble learning is a form of learning in which multiple learners (referred to as base learners) are combined to solve the same problem, hence increasing its ability for generalization. It has proven quite successful in the domain of supervised learning [56; 57; 58; 59]. Base learners can be decision trees, neural networks or other kinds of machine learning models. Most of the ensemble learning algorithms use single-type base learners leading to homogeneous base learners but there are methods which use a

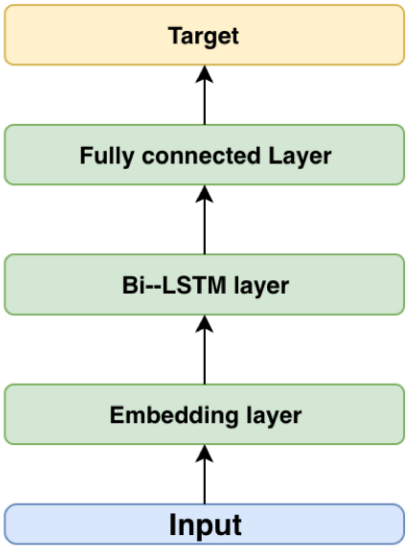


Figure 12: Network layer architecture of Bi-LSTM model

heterogeneous form of learners [96]. Construction of ensemble can either be in sequential or parallel style. A most common strategy for the combination of base learners is to predict the output from base learners which uses either weighted average for regression or majority voting for classification. Also, the ensemble learning can be employed is 2 or more stages, which combines the prediction from the previous stage. In our case, we intend to combine the user features and text features to improve the tweet classification. We use a 2 stage parallel (commonly referred to as stacking (inspired by Wolpert et al [97].) ensemble-based to solve our problem. As we have two distinct types of features (i.e. text and user features), we use a parallel based approach which exploits independence between the base learners. We use base models such as SVM and RF trained on user and text features for the first stage of the

ensemble. For the second stage, we use another classifier such as SVM to make the final prediction which uses the predictions of previous models to make the final predictions.

CHAPTER 6

EXPERIMENTS

To accurately predict the escalation in the behavior of users from Juul to cannabis substance use, we look at two different interval ranges: i) year spans from 2015 to 2017 interval ii) two-month interval spans for the 2018 year. In these experiments, we are focused only on the mono users who will escalate to poly.

6.1 Evaluation metric

Evaluation measures play a key role in assessing the performance of classification models. Traditionally accuracy was commonly used measure for classification but it doesn't work in the case of imbalanced data. In the case of an imbalanced dataset, the rare class will have the least impact on predictive performance using accuracy, e.g. if the rare class represent 1% of the full data then the classifier that predicts all prevalent class will have 99% accuracy without any training required. In our case as well as we have imbalance data, so we use a different metric called F1-score.

Precision and recall are commonly used metrics used in the information retrieval domain and forms the basis for F1-score.

Recall : In information retrieval task *recall* refers to fraction of relevant documents that are successfully retrieved. It is also referred to as True positive rate.

$$R = \frac{TP}{TP + FN} \tag{6.1}$$

Precision: *Precision* refers to the percentage of documents that are relevant to the query, also referred to as positive predictive value.

$$P = \frac{TP}{TP + FP} \tag{6.2}$$

F1-score refers to the harmonic mean of both precision and recall.

$$F1 = \frac{2}{\frac{1}{R} + \frac{1}{P}} \tag{6.3}$$

For further reference for F1-score metrics one can refer to [98]. For our results as well, we utilized F1-score to report the classification-performance.

6.2 Predicting escalation by year interval

We are interested to capture the escalation of the mono user to poly. Our prediction task looks at whether the user who was mono (never mentioned about Cannabis) in a period (input period) remain mono or escalate to poly in a second period (output period). We represent the input and output period as (period 1, period 2). We select the users and their features who remain mono for the input period and label them as mono or poly based on their activity in the output period. In the first experiment, our input period comprises of 2 years and the output period consists of next year after the input span. We divide the intervals of input and output into the following buckets:

1. (2014-15, 2016)
2. (2015-16, 2017)
3. (2016-17, 2018)

For each interval, we first select the mono user and their features for the input period and then label the user based on the activity in the output period. e.g. in 2014-15 we first select the users and their features who are mono and have not mentioned about Cannabis in the 2014-15 period and then label them as poly if they mention Cannabis at least once in 2016 otherwise term them as mono. For each of the experiments, we train machine learning models using the binary labels of mono and poly to classify them under two classes (i.e. binary classification). As we can see in Table X that there are fewer users who will turn to a poly user in the next year, to handle the dis-balance in the dataset we have tried 2 strategies including downsampling the majority class (mono class) and oversampling the minority class (poly class). Oversampling works best for our dataset so we report only the scores calculated with oversampling the training data. We report average F1 scores of 5-fold cross-validation in each input interval for each of the machine learning model. We plot the F1 score for each of the mono and poly class, to clearly understand the prediction performance of both classes. We will compare the scores of the poly class for all of the experiments as this is the class we are more interested in predicting. We follow the above approach in all of the experiments. For the baseline, we plot the performance of the majority classifier which predicts the positive class in each plot (i.e. for mono class it will predict all as mono and similarly for poly class).

We also represent the relevance of features used for each of the experiments. In our data we have two sets of features namely the user features such as status count and text features which is tweet text. We utilize user and text features individually and also with different combinations. Following are the different combinations we experiment in our case:

- User only features

(a) prediction performance on mono users

(b) prediction performance on user who turn to poly in next year

Figure 13: We plot the 5-fold cross validated F1 performance for each of the classifiers in each year interval by utilizing user features and illustrate for each of the class mono and poly (i.e. who will escalate to poly in next year). Status count and unigram count are calculated based on input time interval of data but rest of the features are fixed based on the Dec 2018 value for each user. SVM perform better as compared to other classifiers, (with average F1 scores of 21.6%, 29.8% and 44.23% in 2014-15,2015-16, 2016-17 intervals respectively, and mean score of 31.9% for all of the intervals).

- Text only features
- Simple-combination of both the user and text features
- Combining the classifiers trained on the user and text features in an ensemble
- Comparison analysis of each approach.

For relevance calculation, we utilize the weights of the classifiers to explain the most predictive feature. We split data into 80% - 20% as train and test data. We use all of the training data to train the ML model for feature importance. The motivation for this is to understand the predictive feature by looking at most of the data.

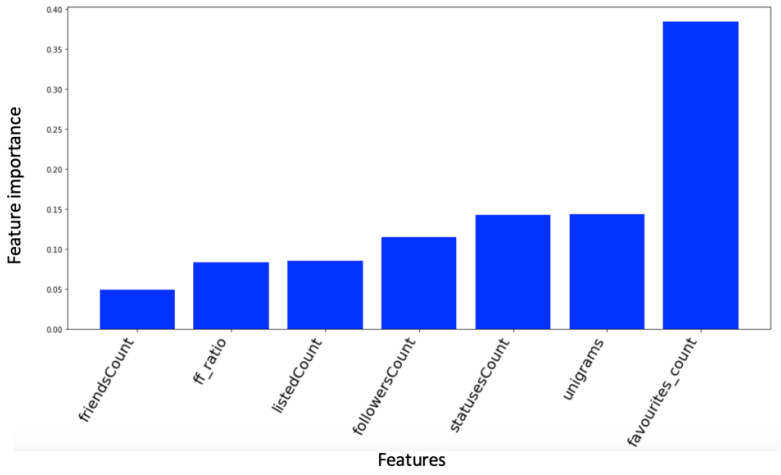


Figure 14: Top most predictive user features using the weights of XGBoost classifier trained on user features for 2016-17 data.

6.2.1 Predicting with user-only features

We have performed our results using the previously mentioned machine learning models. We have not utilized Bi-lstm for user features as it is more suitable for sequence-based input. For the parameters for ML models, we use Linear SVM model with C = 1, with L2 regularization and squared hinge loss function; we use 100 base estimators for Random Forest with ”Gini” criteria for split; and finally set the max depth = 3, learning rate = 0.1 with 100 base estimators for XGBoost classifier.

The plot Figure 13 shows the change of classification performance for each of the classifiers in the different year range buckets. Based on the comparison of two plot, we can see that F1-score for predicting mono user is higher, which is explained by the disbalance of users in the test set (i.e. fewer users who turn to poly in next year). It is harder to predict the users who will escalate to the poly user in comparison to predicting mono user. In the user features we can see that SVM seem to perform better as

Year	Mono user	Turn to poly in next year (% of total user)
2014-15	1043	97 (8.5%)
2015-16	2548	313 (10.94%)
2015-17	32405	9497 (22.66%)

TABLE X: Showing distribution of mono and poly users in different input intervals. This table shows the number of users who will turn to poly and the user who will remain mono in the next year after the input period. We can see there are fewer number of users who will turn to poly in the next year in comparison of user who remain mono.

compared to other classifiers, (with average F1 scores of 21.6%, 29.8% and 44.23% in 2014-15,2015-16, 2016-17 intervals respectively, and mean score of 31.9% for all of the intervals). We see that in Fig 13a that the performance of each of the classifiers seems to decline in the last interval for the mono case. As the amount of data in the last interval is highest (Table X) it could add more noise in the data, making it harder to predict. One would also notice that the performance of majority classifier is too high in mono class and it goes to low in poly class, it is attributed due to the same reason of imbalance in the data discussed above.

We also plot the most predictive features among the user features using weights from SVM classifier, which indicates favourites count (i.e number of tweets liked by the user in accounts lifetime) as the most distinguishing feature, a similar view is captured in other classifiers as well (shown in Figure 14).

6.2.2 Predicting with text-only features

Similar to the above approach we conduct a similar experiment by utilizing only the text features. In this case, we employ all of the machine learning models mentioned in the previous chapter, including bi-LSTM. Each of the tweets is clean and preprocessed to remove URLs, author mentions (Eg @john),

the hashtag symbol ('#'), retweet keyword ("RT"), punctuation or symbols (","). Furthermore, we also remove stop words from the text. (using NLTK [99] stop word dictionary). This ensures to remove the words that don't convey any additional information in the data. All of the tweet text of each user has been concatenated, an example of each tweet for each class is shown in Table XI. In Table XI each of the tweets have been trimmed for easy reference. We use 5-fold cross-validation to report the scores similar to above.

For the machine learning models except for bi-LSTM model, specific feature engineering part includes TF-IDF vectorization. This changes the text sentences into TF-IDF vectors and all the tweets reduce to TF-IDF matrix. TF-IDF score is calculated as follows:

$$tfidf_{\text{t,d}} = tf_{\text{t,d}} * idf_{\text{t}} \tag{6.4}$$

where tf stands for term-frequency and idf stands for inverse document frequency. For further details, one can refer to the book by Manning et al. [90]. Using tf-idf we obtain (n * V) matrix where n is the number of words and V is the vocabulary of the text. As the vocabulary size can be quite huge for our collection of tweets, we also utilize truncated SVD (single value decomposition) to reduce TF-IDF matrix to lower dimension of 100 features, termed as latent semantic analysis. It is a common technique used in conjunction with TF-IDF, in natural language processing and information retrieval domain to reduce the dimensions, presented by Deerwester et al [91]. As the vocabulary size is high in the dataset we reduce all of the features to 100 dimension and keep the experimental setup same for all of the

Example	Tweet Text	label
a)	...; Why this Juul taste like cantaloupe ; First date ideas : - Smoking a juul thru a bong	1
b)	..; Smoked two juul pods in 24 hours thats sad ;imagine having a nicotine addiction from a juul u bought just to fit in	1
c)	..;You know what Im just gonna say it I like the tabaco juul pods ;Took my 5 juul Rips and now Im ready to sleepn	1
d)	..; This girl in my glass just bought 200 dollars worth of juul pods and said this will last me about 2 weeks	0
e)	..;a possible refund, they gave me the refund on my card and said I could keep the item I purchased. They let me finesse em	0
f)	..;There are two types of fiends on a Greyhound bus: the ones who get off at every stop to smoke a stoge and those that secretly hit their Juuls inside the whole ride	0

TABLE XI: Example tweet text for each of mono and poly user. Examples have been trimmed for easy visualization. Each of the tweets for a user in the dataset has been concatenated and separated by ";"

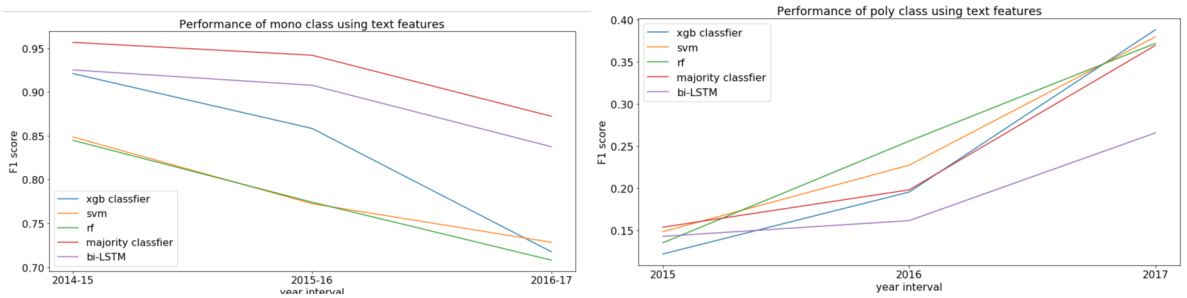
experiments. These 100 text features are used by all of the machine learning models mentioned except for the bi-LSTM model.

We experimented with various parameters for the bi-LSTM model using 60%-20%-20% split into train, validation and test data respectively. We selected the ones which gave good evaluation results with the validation data. For the embedding layer of the bi-LSTM model, we have utilized Glove [100] 100-Dimensional twitter embeddings which are trained on 2B twitter tweets. The deep learning models like bi-LSTM don't work with different length of the text so we apply post zero paddings to the embedding

vectors to pad the input to a fixed length. Also for the fixed input length of text in case of the neural model, we take the 95% percentile length of all tweets as the fixed length of the text, with a limit of 60. We take whichever is smaller (i.e. cap length $=$60 or 95 percentile length of text). We truncate text if it is greater than this max length and pad in case of the shorter text. We take 0.02 L2 regularization, with 50% dropout for the bi-lstm layer of the bi-lstm model, which is done to avoid over-fitting the bi-LSTM model to the training data. We have a sigmoid activation function for the output layer. We train the neural model for 20 epochs by utilizing cross-entropy loss, which works great in case of binary classes. Similar to above we plot the average F1-score performance (based on 5-fold cross-validation) for each of the models in 3 buckets. Figure 15 explains the prediction performance of each of the classifier using an average 5-fold cross-validation F1-score for each of the machine learning models in the 3 buckets. We plot the performance of each binary class performance separately as shown in 15a and 15b.

We can see in Figure 15 (a) that prediction performance for classifier fares a little better at predicting the mono class (shown in Fig 15a) as compared to poly class (shown in Fig 15b), leading to convey that it is easier to predict mono class than poly, similar to above. Another thing that we notice in 15a similar to above is that the performance of classifiers declines in the last interval span. In 15b we can see that most of the classifier perform with almost similar scores with random forest performing the best among them. We get 15.8%,25.4%,37.2% respectively for each interval, with a mean average F1-score of 25.4% for the RF model in all interval.

We train the all of the ML models except bi-LSTM using the TF-IDF matrix,(i.e (n*V) matrix where n is the number of features and V is vocabulary size) to understand the relevance of words in the data. We utilize 80% of data to train the ML models to calculate the relevance. In the Figure 16 which plots the



(a) prediction performance on mono users

(b) prediction performance on user who turn to poly in next year

Figure 15: We plot the 5-fold cross-validation F1 score for each of the classifiers in each year interval utilizing the text features. We plot a separate figure for each of the binary class for clarity. RF seems to fair better in comparison to other classifiers, we obtain 15.8%,25.4%,37.2% respectively for each interval, with mean F1-score of 25.4% for all intervals.

feature importance of words using the trained XGBoost classifier on 2015-16 interval data we can see that words like ”pod”, ”cigarette” and ”vape” seem to be the most predictive feature for discriminating mono and poly user. Surprisingly the bi-LSTM model which preserved the context-dependency of text seem to perform poorly in the dataset. We tried various approaches including L2 regularization and 50% dropout for LSTM layer but none of them gave better results on the dataset. As the data is imbalanced so it could be harder for the bi-LSTM model to tune the weights to learn the discriminating features.

6.2.3 Simple-combination of the user and text features

In this section, we combine the 7 user features and 100 text features obtained from TF-IDF in conjunction with SVD. In this approach, we utilize all of the above-mentioned machine learning models except the recurrent model which is more suited for the sequence-based task. We compare the performance of each of the classifiers using the same experimental setup described above. We plot the

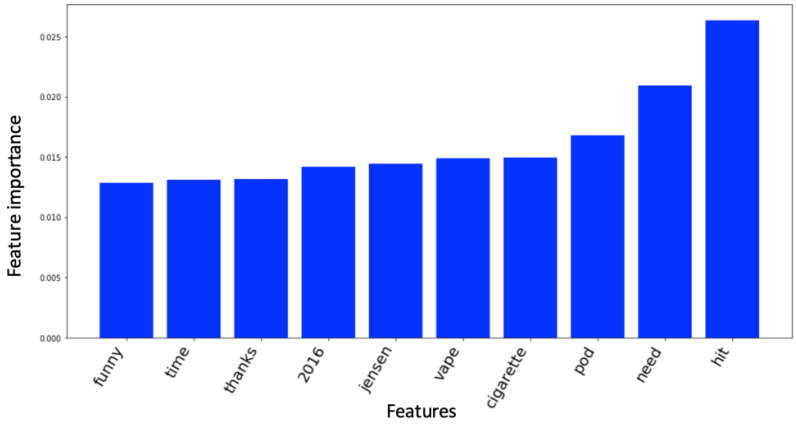
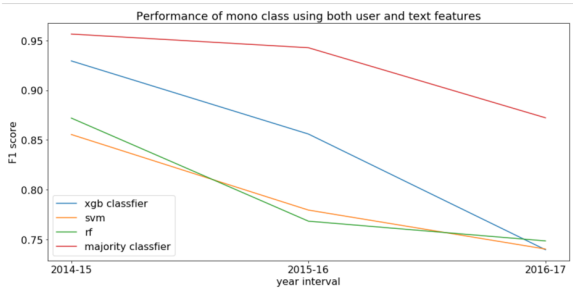


Figure 16: Feature importance of words in tweet text using the XGBoost classifier trained on text features for 2015-16 interval data. We can see that words like ”pod”, ”cigarette” and ”vape” seem to be the most predictive feature for discriminating mono and poly user.

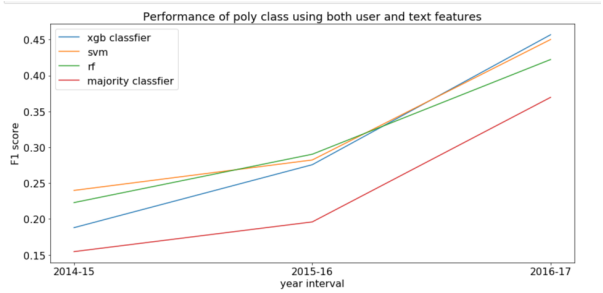
performance of each of the binary class as shown in Figure 17. We can see in this case SVM classifier seem to perform the best with a mean F1-score of 32.4% in all of the intervals. We can see that the combination of user and text features seem to perform better in comparison of individual text (25.4% F1-score) and individual user features (31.9 % F1-score). We also look at the feature importance among the user and text features using weights of XGBoost classifier trained on 2016-17 data as shown in Figure 18. We can see that user features like favourites count, unigrams have higher relevance as compared to text features represented numerically.

6.2.4 Ensemble learning using both the user and textual features

To utilize both the user and text features, we combine both the set of features to analyze the escalation of mono to the poly user in an ensemble manner. We used a heterogeneous stacking model (inspired by Wolpert et al [97]) to combine 3 user feature classifiers (including SVM, rf, and XGBoost trained



(a) prediction performance on mono users



(b) prediction performance on user who turn to poly in next year

Figure 17: We plot the 5-fold cross-validation F1 score for each of the classifiers in each input year interval utilizing both of the user and text features. We plot a separate figure for each of the binary class for clarity. In this case SVM seems to perform the best with scores 23%, 28% and 45% for the intervals of 2014-15 , 2015-16 and 2016-17 respectively. We get an average F1 score of 32.4% using the SVM model.

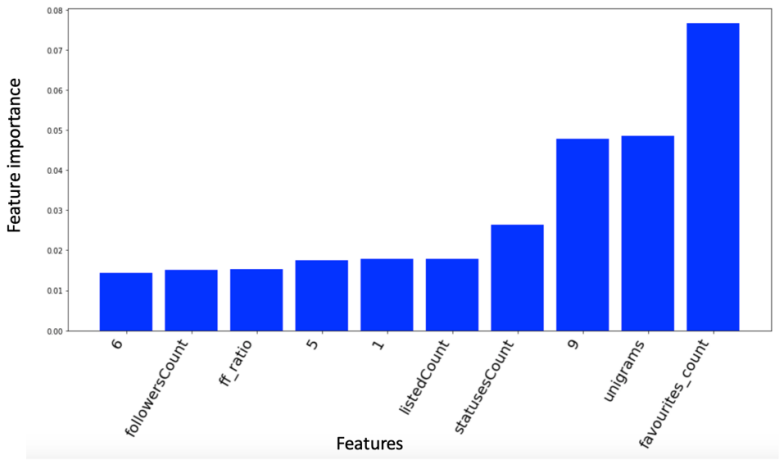


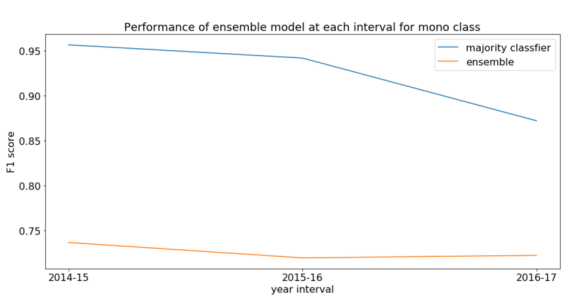
Figure 18: We plot the relevance of 10 most predictive feature using the weights XGBoost classifier trained using simple-combination of user and text features trained on 2016-17 data. The text features are which were the words from the tweets obtained by TF-IDF matrix are reduced 100 dimensions using SVD. Text features are numbered from 0 to 99. Features are arranged from left to right, with rightmost being the most relevant features. We can see that user features like favourites count, unigrams have higher relevance as compared to text features represented numerically.).

on user features), with 3 text-based classifiers (including SVM, rf, XGBoost trained on text features) in a parallel fashion. Ensemble tends to yield better results when there is significant diversity among the model [101]. We create an ensemble of 6 classifiers for the first stage of the ensemble. For the second stage, we use SVM classifier to make the final prediction. We experimented with all above 3 mentioned classifiers as meta learner but we got better results with SVM classifier. We utilize the predictions of the base learner from the first stage to create a new dataset with predictions for the second stage classifier. We utilize k-fold cross-validation to work in the case of the 2-stage ensemble model as follows (mentioned by Zhou et al [96] :

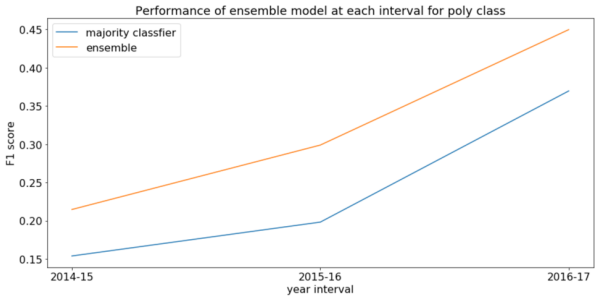
Algorithm for cross-validation for ensemble models

1. Split the train set into k-folds
2. Fit first stage stacked models on k-1 folds and predict on the kth fold
3. Repeat 2 to predict for each fold
4. We obtain (out-of-folds) prediction of the k folds to create a new dataset for the second stage
5. We split these out-of-folds predictions into p folds
6. Fit the second stage model on the p-1 folds and predict on the pth fold
7. Repeat 6 to predict on each fold
8. Obtain the mean of all fold to get an average score.

In the ensemble stacking model, the last stage is similar to the k-fold cross-validation setup as explained above. We use the predictions of the base learners to make accurate predictions by utilizing



(a) prediction performance on mono users



(b) prediction performance on user who turn to poly in next year

Figure 19: We plot the weighted F1 performance for each of the classifiers using ensemble approach. We combine 2 sets of classifiers including the 3 text based classifier (svm, xgBoost and random forest trained on text features) and 3 user based classifier including (svm, xgboost and random forest trained on user features) in a parallel based manner. We use SVM for the second level ensemble to make the final prediction. We plot the average 5 fold cross validation scores. We achieve scores of 21.4%,29.8% and 44.9%, with an average of 32.03% scores

the meta or second stage learner (i.e. SVM in our case). We plot the average F1-scores of each interval as shown in Figure 19. We plot the average 5 fold cross-validation scores. We achieve scores of 21.4%,29.8%, 44.9%, with an average of 32.03% scores. We see that the ensemble approach performs little better than individual text and user features and almost similar to the simple-combination model.

We plot the relevance of meta (2nd stage) classifier of the ensemble model. We look at the most predictive features, which in the case would represent the classifiers trained on the user and text features. We can see in Figure 20 that the classifiers ”XGB user” and ”SVM user” trained on user features seem to be more predictive in comparison to the classifiers trained on text features, leading us to believe user features to be more predictive in comparison to text features.

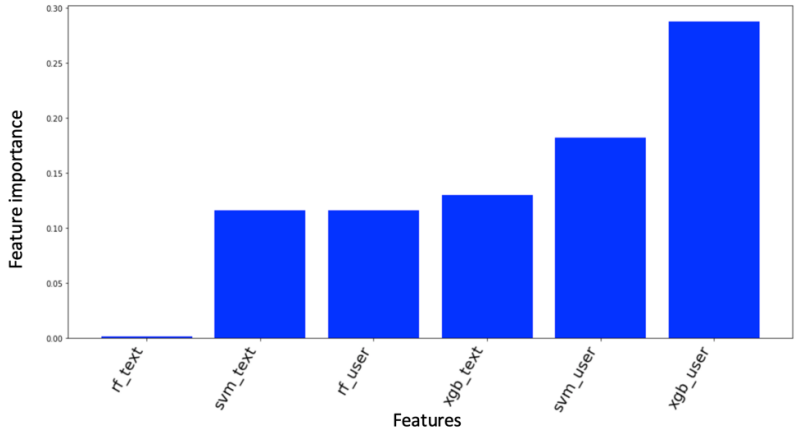


Figure 20: We obtain the weights from the trained SVM from the 2nd stage of the ensemble model using the input data of 2016-17. We order the top 10 features from left to right with rightmost being the most predictive feature. We see that classifiers trained on the user feature (i.e. "XGB" and "SVM") to be more predictive in comparison to the text features.

6.2.5 Comparison of text and user features

We utilized the text features and user features in isolation to understand the escalation of users. We also looked at the features in combination by using ensemble and simple-combination approach. We saw that while looking at the relevance of simple combination of user and text features Figure 18 that the user features seem to be more predictive in comparison of text features and the similar idea was viewed in the ensemble relevance Figure 20. We plot each of the best classifiers for the relevance of each user, text, simple combination and ensemble approach (there is only one classifier for the second stage) as shown in Figure 21. A close look at the plot reveals that ensemble and simple-combination approach seem to perform fairly similar with 32.4% and 32.1% respectively. We can further see a similar idea emerge in Figure 21 that the individual user features are more predictive in comparison to text features.

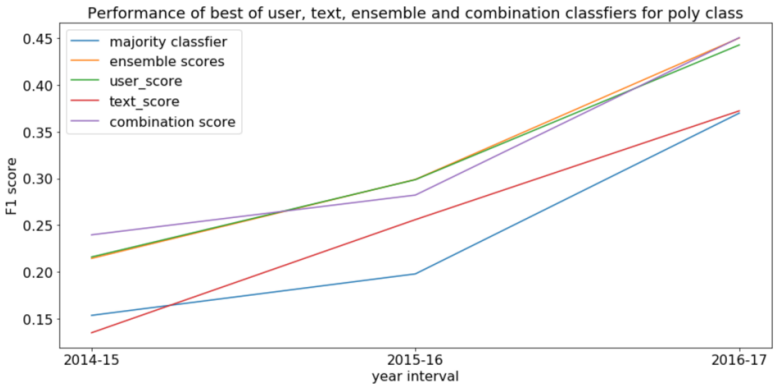


Figure 21: We look at the best classifiers for each of the individual text and user, simple combination and ensemble manner. We can see that ensemble and simple combination approach seem to perform fairly similar with an average of 32.4% and 32.1%. We also see that individual user features(labeled as user scores) are more predictive in comparison of text features.

The majority of classifier performs with a mean of 24.05%. Although not huge our best classifier in case of simple-combination of user and text classifier help us achieve 33% improvement over the majority classifier.

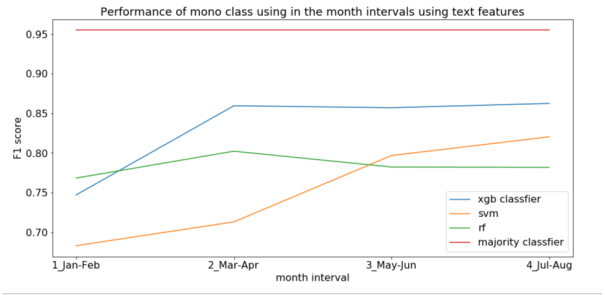
6.3 Escalation by month interval

In the previous sections, we looked at data in the interval by year. In this section, we aim to look at the escalation of mono users to poly in monthly intervals. As we saw in the previous chapter that the growth of dataset containing Juul hashtags has been maximum in the year of 2018. Therefore, we select the data for the 2018 year interval. We take the data for the users who start in Jan 2018 and will escalate to poly in September 2018. The users who don't follow this rule are filtered out for this analysis. We label the user based on the activity after September 2018. We end up with 10,760 users who fall under

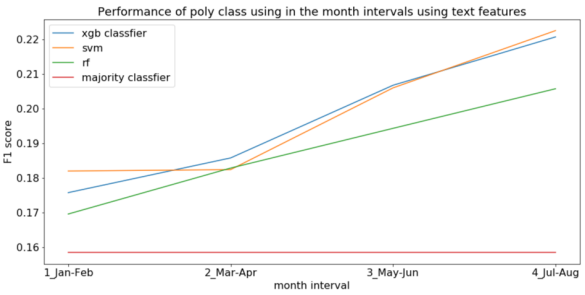
this query and out of those 929 of the users will escalate after September. Our input period is within a span of two-months:

- Jan- Feb
- Mar-Apr
- May -June
- July August

In this scenario there is no growth of users, they remain the same for each of the intervals and only the amount of tweet text change with every interval. We try to understand how accurately can we predict the escalation of users from mono to poly by looking at the same users and varying the amount of tweet data. We use all of the ML models mentioned above except bi-LSTM as they were not giving better results similar to above even after tuning various hyper-parameters. For preprocessing and parameters for ML models we follow the same experimental setup as explained above in predicting in the year interval. In this case, we only look at the text features and not at user features, as the users remain the same for each of the intervals. We follow the methodology as before using the same set of classifiers. We plot the average 5 fold cross-validated F1-score for each of the models as shown in Figure 22. Although the performance difference in the performance of best classifier (19.9% F1 score with SVM) and majority classifier(15.8%) is not huge we see that each of classifier performance increases with successive intervals. Increase in the performance of each of the classifier against the majority classifier in case of the poly class leads us to believe that more text is predictive of escalation while keeping the number of users fixed.



(a) prediction performance on mono users



(b) prediction performance on user who turn to poly in next year

Figure 22: We plot the average F1 performance for each of the classifiers in each month interval for the year 2018. The plots are represented using the same setup as before. It illustrates which of the mono user will turn to poly after September 2018. The users remain the same in each interval and only the text changes with each interval. We achieve 19.9% mean F1-score with SVM in comparison to 15.8% with majority classifier. There is an increase in the performance of the classifier with each successive interval, leading us to believe that more text is predictive for escalation of mono to the poly user.

CHAPTER 7

CONCLUSION

In this thesis, we presented a predictive approach that predicted the escalation of e-cigarette user to cannabis by utilizing each of the user and text-based features associated with the user. We looked at the performance of predicting with a combination of features using a simple and an ensemble approach. We found that although not a huge improvement but by utilizing both the text features and user features we could achieve 32.4% average F1 score (33.33% improvement over the majority classifier) on predicting escalation. As the data is quite imbalanced with an average of 14.03% of the poly user, this task is harder to predict utilizing only the text and user features. We also found that by keeping the users same as in the case of the month interval while utilizing the text features with a span of 2 months that there was an increase in the performance of each classifiers leading us to believe that more textual data is predictive of escalation.

Some of the previous works have suggested that social influence is a contributing factor in the substance abuse problem [102; 103; 104]. In future work, we aim to look at the network properties of the user. One direction could be to utilize network-based properties of poly-substance and e-cigarette users, e.g. using Node2vec to learn the network embeddings of the following network of mono and poly user to make accurate escalation predictions.

CITED LITERATURE

1. Anderson, A., Huttenlocher, D., Kleinberg, J., Leskovec, J., and Tiwari, M.: Global diffusion via cascading invitations: Structure, growth, and homophily. In Proceedings of the 24th International Conference on World Wide Web, pages 66–76. International World Wide Web Conferences Steering Committee, 2015.

2. Demick, B.: A high-tech approach to getting a nicotine fix. Los Angeles Times, 25, 2009.

3. Vansickel, A. R. and Eissenberg, T.: Electronic cigarettes: effective nicotine delivery after acute administration. Nicotine & Tobacco Research, 15(1):267–270, 2012.

4. Bullen, C., McRobbie, H., Thornley, S., Glover, M., Lin, R., and Laugesen, M.: Effect of an electronic nicotine delivery device (e cigarette) on desire to smoke and withdrawal, user preferences and nicotine delivery: randomised cross-over trial. Tobacco control, 19(2):98–103, 2010.

5. Barrington-Trimis, J. L., Berhane, K., Unger, J. B., Cruz, T. B., Urman, R., Chou, C. P., Howland, S., Wang, K., Pentz, M. A., Gilreath, T. D., et al.: The e-cigarette social environment, e-cigarette use, and susceptibility to cigarette smoking. Journal of Adolescent Health, 59(1):75–80, 2016.

6. the fashion law.

7. Chaker, A. M.: Schools and parents fight a juul e-cigarette epidemic.

8. Robert K. Jackler, Cindy Chau, B. D. G. M. M. W. J. L.-H. A. M. B. S. H. K.-O. Z. A. H. L. M. J. D. R.: Juul advertising over its first three years on the market. Technical report, Stanford Research into the Impact of Tobacco Advertising, Stanford University School of Medicine, Stanford Research into the Impact of Tobacco Advertising 801 Welch Road, Stanford, CA, 94305, January 2019.

9. Johnston, L., OMalley, P., Miech, R., Bachman, J., and Schulenberg, J.: Monitoring the future national survey results on drug use: 1975-2013: Overview, key findings on adolescent drug use. institute for social research, the university of michigan. Ann Arbor, page 84, 2014.

10. Moss, H. B., Chen, C. M., and Yi, H.-y.: Early adolescent patterns of alcohol, cigarettes, and marijuana polysubstance use and young adult substance use outcomes in a nationally representative sample. Drug and alcohol dependence, 136:51–62, 2014.

11. Andlin-Sobocki, P.: Economic evidence in addiction: a review. The European Journal of Health Economics, 5:s5–s12, 2004.

12. Dishion, T. J. and Loeber, R.: Adolescent marijuana and alcohol use: The role of parents and peers revisited. The American journal of drug and alcohol abuse, 11(1-2):11–25, 1985.

13. Donovan, J. E. and Jessor, R.: Adolescent problem drinking. psychosocial correlates in a national sample study. Journal of Studies on Alcohol, 39(9):1506–1524, 1978.

14. Robins, L. N.: Deviant children grown up, a sociological and psychiatric study of sociopathic personality. 1966.

15. Buchmann, A. F., Blomeyer, D., Jennen-Steinmetz, C., Schmidt, M. H., Esser, G., Banaschewski, T., and Laucht, M.: Early smoking onset may promise initial pleasurable sensations and later addiction. Addiction biology, 18(6):947–954, 2013.

16. Johnson, E. O. and Novak, S. P.: Onset and persistence of daily smoking: the interplay of socioeconomic status, gender, and psychiatric disorders. Drug and alcohol dependence, 104:S50–S57, 2009.

17. Kendler, K. S., Myers, J., Damaj, M. I., and Chen, X.: Early smoking onset and risk for subsequent nicotine dependence: a monozygotic co-twin control study. American Journal of Psychiatry, 170(4):408–413, 2013.

18. Morrell, H. E., Song, A. V., and Halpern-Felsher, B. L.: Earlier age of smoking initiation may not predict heavier cigarette consumption in later adolescence. Prevention Science, 12(3):247–254, 2011.

19. Westling, E., Andrews, J. A., and Peterson, M.: Gender differences in pubertal timing, social competence, and cigarette use: a test of the early maturation hypothesis. Journal of Adolescent Health, 51(2):150–155, 2012.

20. Ehlers, C. L., Slutske, W. S., Gilder, D. A., and Lau, P.: Age of first marijuana use and the occurrence of marijuana use disorders in southwest california indians. Pharmacology Biochemistry and Behavior, 86(2):290–296, 2007.

21. Gillespie, N. A., Neale, M. C., and Kendler, K. S.: Pathways to cannabis abuse: a multi-stage model from cannabis availability, cannabis initiation and progression to abuse. Addiction, 104(3):430–438, 2009.

22. Jessor, R.: Problem-behavior theory, psychosocial development, and adolescent problem drinking. British journal of addiction, 82(4):331–342, 1987.

23. Kandel, D.: Stages in adolescent involvement in drug use. Science, 190(4217):912–914, 1975.

24. Hingson, R. W., Heeren, T., and Winter, M. R.: Age of alcohol-dependence onset: associations with severity of dependence and seeking treatment. Pediatrics, 118(3):e755–e763, 2006.

25. Degenhardt, L., Chiu, W., Conway, K., Dierker, L., Glantz, M., Kalaydjian, A., Merikangas, K., Sampson, N., Swendsen, J., and Kessler, R.: Does the gatewaymatter? associations between the order of drug use initiation and the development of drug dependence in the national comorbidity study replication. Psychological medicine, 39(1):157–167, 2009.

26. Patton, G. C., Coffey, C., Carlin, J. B., Sawyer, S. M., and Lynskey, M.: Reverse gateways? frequent cannabis use as a predictor of tobacco initiation and nicotine dependence. Addiction, 100(10):1518–1525, 2005.

27. Sartor, C., Agrawal, A., Lynskey, M., Duncan, A., Grant, J., Nelson, E., Madden, P., Heath, A., and Bucholz, K.: Cannabis or alcohol first? differences by ethnicity and in risk for rapid progression to cannabis-related problems in women. Psychological medicine, 43(4):813–823, 2013.

28. Stenbacka, M., Allebeck, P., and Romelsjö, A.: Initiation into drug abuse: the pathway from being offered drugs to trying cannabis and progression to intravenous drug abuse. Scandinavian journal of social medicine, 21(1):31–39, 1993.

29. Tarter, R. E., Kirisci, L., Mezzich, A., Ridenour, T., Fishbein, D., Horner, M., Reynolds, M., Kirillova, G., and Vanyukov, M.: Does the gateway sequence increase prediction of cannabis use disorder development beyond deviant socialization? implications for prevention practice and policy. Drug and Alcohol Dependence, 123:S72–S78, 2012.

30. Timberlake, D. S., Haberstick, B. C., Hopfer, C. J., Bricker, J., Sakai, J. T., Lessem, J. M., and Hewitt, J. K.: Progression from marijuana use to daily smoking and nicotine dependence in a national sample of us adolescents. Drug and alcohol dependence, 88(2-3):272–281, 2007.

31. Vanyukov, M. M., Tarter, R. E., Kirillova, G. P., Kirisci, L., Reynolds, M. D., Kreek, M. J., Conway, K. P., Maher, B. S., Iacono, W. G., Bierut, L., et al.: Common liability to addiction and gateway hypothesis: theoretical, empirical and evolutionary perspective. Drug and alcohol dependence, 123:S3–S17, 2012.

32. Hicks, B. M., Iacono, W. G., and McGue, M.: Index of the transmissible common liability to addiction: heritability and prospective associations with substance abuse and related outcomes. Drug and alcohol dependence, 123:S18–S23, 2012.

33. Iacono, W. G., Malone, S. M., and McGue, M.: Behavioral disinhibition and the development of early-onset addiction: common and specific influences. Annu. Rev. Clin. Psychol., 4:325–348, 2008.

34. Tarter, R. E., Kirisci, L., Mezzich, A., Cornelius, J. R., Pajer, K., Vanyukov, M., Gardner, W., Blackson, T., and Clark, D.: Neurobehavioral disinhibition in childhood predicts early age at onset of substance use disorder. American Journal of Psychiatry, 160(6):1078–1085, 2003.

35. Stahr, A.: Demographics of key social networking platforms.

36. Li, H., Mukherjee, A., Liu, B., Kornfield, R., and Emery, S.: Detecting campaign promoters on twitter using markov random fields. In 2014 IEEE International Conference on Data Mining, pages 290–299. IEEE, 2014.

37. Bollen, J., Mao, H., and Pepe, A.: Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In Fifth International AAAI Conference on Weblogs and Social Media, 2011.

38. Valente, T. W., Gallaher, P., and Mouttapa, M.: Using social networks to understand and prevent substance use: A transdisciplinary perspective. Substance use & misuse, 39(10-12):1685–1712, 2004.

39. Fetta, A., Harper, P., Knight, V., and Williams, J.: Predicting adolescent social networks to stop smoking in secondary schools. European Journal of Operational Research, 265(1):263–276, 2018.

40. Cortese, D. K., Szczypka, G., Emery, S., Wang, S., Hair, E., and Vallone, D.: Smoking selfies: Using instagram to explore young womens smoking behaviors. Social Media+ Society, 4(3):2056305118790762, 2018.

41. Huang, J., Kornfield, R., Szczypka, G., and Emery, S. L.: A cross-sectional examination of marketing of electronic cigarettes on twitter. Tobacco control, 23(suppl 3):iii26–iii30, 2014.

42. Jo, C. L., Kornfield, R., Kim, Y., Emery, S., and Ribisl, K. M.: Price-related promotions for tobacco products on twitter. Tobacco control, 25(4):476–479, 2016.

43. Dai, H. and Hao, J.: Mining social media data for opinion polarities about electronic cigarettes. Tobacco control, 26(2):175–180, 2017.

44. Cheng, J., Kleinberg, J., Leskovec, J., Liben-Nowell, D., Subbian, K., Adamic, L., et al.: Do diffusion protocols govern cascade growth? In Twelfth International AAAI Conference on Web and Social Media, 2018.

45. Rosenbaum, P. R. and Rubin, D. B.: The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55, 1983.

46. Zhang, Y., Ramesh, A., Golbeck, J., Sridhar, D., and Getoor, L.: A structured approach to understanding recovery and relapse in aa. In Proceedings of the 2018 World Wide Web Conference on World Wide Web, pages 1205–1214. International World Wide Web Conferences Steering Committee, 2018.

47. Hu, T., Lin, Z., and Keeler, T. E.: Teenage smoking, attempts to quit, and school performance. American Journal of Public Health, 88(6):940–943, 1998.

48. Jiménez, R., Anupol, J., Cajal, B., and Gervilla, E.: Data mining techniques for drug use research. Addictive behaviors reports, 8:128–135, 2018.

49. Palmer, A., Jiménez, R., and Gervilla, E.: Data mining: Machine learning and statistical techniques. In Knowledge-oriented applications in data mining. IntechOpen, 2011.

50. Martinez, L. S., Hughes, S., Walsh-Buhi, E. R., and Tsou, M.-H.: okay, we get it. you vape: An analysis of geocoded content, context, and sentiment regarding e-cigarettes on twitter. Journal of health communication, 23(6):550–562, 2018.

51. Cortes, C. and Vapnik, V.: Support-vector networks. Machine learning, 20(3):273–297, 1995.

52. Cherkassky, V., Shao, X., Mulier, F. M., and Vapnik, V. N.: Model complexity control for regression using vc generalization bounds. IEEE transactions on Neural Networks, 10(5):1075–1089, 1999.

53. Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J., and Vapnik, V.: Support vector regression machines. In Advances in neural information processing systems, pages 155–161, 1997.

54. Vinodhini, G. and Chandrasekaran, R.: Sentiment analysis and opinion mining: a survey. International Journal, 2(6):282–292, 2012.

55. Zhu, Y., Wang, X., Zhong, E., Liu, N. N., Li, H., and Yang, Q.: Discovering spammers in social networks. In Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012.

56. Breiman, L.: Random forests. Mach. Learn., 45(1):5–32, October 2001.

57. Kohavi, R. and Kunz, C.: Option decision trees with majority votes. In ICML, volume 97, pages 161–169, 1997.

58. Bauer, E. and Kohavi, R.: An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine learning, 36(1-2):105–139, 1999.

59. Maclin, R. and Opitz, D.: An empirical evaluation of bagging and boosting. AAAI/IAAI, 1997:546–551, 1997.

60. Ho, T. K.: Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition, volume 1, pages 278–282. IEEE, 1995.

61. Tang, Y., Krasser, S., He, Y., Yang, W., and Alperovitch, D.: Support vector machines and random forests modeling for spam senders behavior analysis. In IEEE GLOBECOM 2008-2008 IEEE Global Telecommunications Conference, pages 1–5. IEEE, 2008.

62. Qi, Y.: Random forest for bioinformatics. In Ensemble machine learning, pages 307–323. Springer, 2012.

63. Xu, P. and Jelinek, F.: Random forests in language modelin. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pages 325–332, 2004.

64. Friedman, J. H.: Stochastic gradient boosting. Computational statistics & data analysis, 38(4):367–378, 2002.

65. Chen, T. and Guestrin, C.: Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pages 785–794. ACM, 2016.

66. Li, P.: An empirical evaluation of four algorithms for multi-class classification: Mart, abc-mart, robust logitboost, and abc-logitboost. arXiv preprint arXiv:1001.1020, 2010.

67. Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J., and Khudanpur, S.: Recurrent neural network based language model. In Eleventh annual conference of the international speech communication association, 2010.

68. Hochreiter, S. and Schmidhuber, J.: Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

69. Zhang, Y., Jin, R., and Zhou, Z.-H.: Understanding bag-of-words model: a statistical framework. International Journal of Machine Learning and Cybernetics, 1(1-4):43–52, 2010.

70. Huang, Z., Xu, W., and Yu, K.: Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991, 2015.

71. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E.: Hierarchical attention networks for document classification. In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, pages 1480–1489, 2016.

72. Wang, Y., Huang, M., Zhao, L., et al.: Attention-based lstm for aspect-level sentiment classification. In Proceedings of the 2016 conference on empirical methods in natural language processing, pages 606–615, 2016.

73. Wöllmer, M., Metallinou, A., Eyben, F., Schuller, B., and Narayanan, S.: Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In Proc. INTERSPEECH 2010, Makuhari, Japan, pages 2362–2365, 2010.

74. Volkova, S., Shaffer, K., Jang, J. Y., and Hodas, N.: Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 647–653, 2017.

75. Chen, W., Wang, C., and Wang, Y.: Scalable influence maximization for prevalent viral marketing in large-scale social networks. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1029–1038. ACM, 2010.

76. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., and Glance, N.: Cost-effective outbreak detection in networks. In Proceedings of the 13th ACM

SIGKDD international conference on Knowledge discovery and data mining, pages 420–429. ACM, 2007.

77. He, X., Song, G., Chen, W., and Jiang, Q.: Influence blocking maximization in social networks under the competitive linear threshold model. In Proceedings of the 2012 siam international conference on data mining, pages 463–474. SIAM, 2012.

78. Ye, M., Liu, X., and Lee, W.-C.: Exploring social influence for recommendation: a generative model approach. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, pages 671–680. ACM, 2012.

79. Twitter financial.

80. business insider.

81. twitter stats.

82. Stahr, A.: New product: Pax labs introduces e-cigarette juul.

83. Sundermeyer, M., Schlüter, R., and Ney, H.: Lstm neural networks for language modeling. In Thirteenth annual conference of the international speech communication association, 2012.

84. Loper, E. and Bird, S.: Nltk: the natural language toolkit. arXiv preprint cs/0205028, 2002.

85. Cohn, M. A., Mehl, M. R., and Pennebaker, J. W.: Linguistic markers of psychological change surrounding september 11, 2001. Psychological science, 15(10):687–693, 2004.

86. Hutto, C. J. and Gilbert, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Eighth international AAAI conference on weblogs and social media, 2014.

87. Preoţiuc-Pietro, D., Sap, M., Schwartz, H. A., and Ungar, L.: Mental illness detection at the world well-being project for the clpsych 2015 shared task. In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 40–45, 2015.

88. Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

89. Lai, S., Xu, L., Liu, K., and Zhao, J.: Recurrent convolutional neural networks for text classification. In Twenty-ninth AAAI conference on artificial intelligence, 2015.

90. Manning, C., Raghavan, P., and Schütze, H.: Introduction to information retrieval. Natural Language Engineering, 16(1):100–103, 2010.

91. Deerwester, S. C., Dumais, S. T., Furnas, G. W., Harshman, R. A., Landauer, T. K., Lochbaum, K. E., and Streeter, L. A.: Computer information retrieval using latent semantic structure, June 13 1989. US Patent 4,839,853.

92. Pennington, J., Socher, R., and Manning, C.: Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543, 2014.

93. Gers, F. A., Schmidhuber, J., and Cummins, F.: Learning to forget: Continual prediction with lstm. 1999.

94. Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., et al.: Api design for machine learning software: experiences from the scikit-learn project. arXiv preprint arXiv:1309.0238, 2013.

95. Quinlan, J. R.: Induction of decision trees. Machine learning, 1(1):81–106, 1986.

96. Zhou, Z.-H.: Ensemble learning. Encyclopedia of biometrics, pages 411–416, 2015.

97. Wolpert, D. H.: Stacked generalization. Neural networks, 5(2):241–259, 1992.

98. Sun, Y., Wong, A. K., and Kamel, M. S.: Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence, 23(04):687–719, 2009.

99. Bird, S., Klein, E., and Loper, E.: Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.", 2009.

100. Pennington, J., Socher, R., and Manning, C. D.: Glove: Global vectors for word representation. 2014. URL: https://nlp. stanford. edu/projects/glove/[accessed 2018-01-11][WebCite Cache ID 6wOYSJxnU], 2018.

101. Kuncheva, L. I. and Whitaker, C. J.: Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine learning, 51(2):181–207, 2003.

102. Moreno, M. A. and Whitehill, J. M.: Influence of social media on alcohol use in adolescents and young adults. Alcohol research: current reviews, 36(1):91, 2014.

103. Simons-Morton, B.: Social influences on adolescent substance use. American Journal of Health Behavior, 31(6):672–684, 2007.

104. Wood, M. D., Read, J. P., Mitchell, R. E., and Brand, N. H.: Do parents still matter? parent and peer influences on alcohol involvement among recent high school graduates. Psychology of Addictive Behaviors, 18(1):19, 2004.

105. Cortes, C. and Vapnik, V.: Support-vector networks. Mach. Learn., 20(3):273–297, September 1995.

106. Olteanu, A., Varol, O., and Kiciman, E.: Distilling the outcomes of personal experiences: A propensity-scored analysis of social media. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, pages 370–386. ACM, 2017.

107. Gervilla, E., Cajal, B., and Palmer, A.: Quantification of the influence of friends and antisocial behaviour in adolescent consumption of cannabis using the zinb model and data mining. Addictive behaviors, 36(4):368–374, 2011.

108. Lam, M. W., Leung, N. W., Bat, B. K., Leung, G. K., and Tsoi, K. K.: real-time data capture with electronic cigarettes for smoking cessation programme: a cloud platform for behavioural research. 2017.

109. Charlton, N., Kingston, J., Petridis, M., and Fletcher, B. C.: Using data mining to refine digital behaviour change interventions. In Proceedings of the 2017 International Conference on Digital Health, pages 90–98. ACM, 2017.

110. Hon, L.: Electronic atomization cigarette, March 12 2013. US Patent 8,393,331.

111. Ngwenya, M. and Bankole, F.: Mining and representing unstructured nicotine use data in a structured format for secondary use. In Proceedings of the 52nd Hawaii International Conference on System Sciences, 2019.

112. Pechmann, C., Pan, L., Delucchi, K., Lakon, C. M., and Prochaska, J. J.: Development of a twitter-based intervention for smoking cessation that encourages high-quality social media interactions via automessages. Journal of medical Internet research, 17(2):e50, 2015.

113. McNeill, A., Brose, L., Calder, R., Hitchman, S., Hajek, P., and McRobbie, H.: E-cigarettes: an evidence update. Public Health England, 3, 2015.

114. Leventhal, A. M., Strong, D. R., Kirkpatrick, M. G., Unger, J. B., Sussman, S., Riggs, N. R., Stone, M. D., Khoddam, R., Samet, J. M., and Audrain-McGovern, J.: Association of electronic cigarette use with initiation of combustible tobacco product smoking in early adolescence. Jama, 314(7):700–707, 2015.

115. Lazard, A. J., Wilcox, G. B., Tuttle, H. M., Glowacki, E. M., and Pikowski, J.: Public reactions to e-cigarette regulations on twitter: a text mining analysis. Tobacco control, 26(e2):e112–e116, 2017.

116. Tan, P.-N., Steinbach, M., and Kumar, V.: Data mining cluster analysis: basic concepts and algorithms. Introduction to data mining, pages 487–533, 2013.

117. Müller, K.-R., Smola, A., Rätsch, G., Schölkopf, B., Kohlmorgen, J., and Vapnik, V.: Using support vector machines for time series prediction. Advances in kernel methodssupport vector learning, pages 243–254, 1999.

118. Ma, Y. and Guo, G.: Support vector machines applications. Springer, 2014.

119. Pang, B., Lee, L., and Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pages 79–86. Association for Computational Linguistics, 2002.

120. Zhang, B., Qian, Z., and Lu, S.: Structure pattern analysis and cascade prediction in social networks. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 524–539. Springer, 2016.

121. Liu, L., Qu, B., Chen, B., Hanjalic, A., and Wang, H.: Modelling of information diffusion on social networks with applications to wechat. Physica A: Statistical Mechanics and its Applications, 496:318–329, 2018.

122. Colas, F. and Brazdil, P.: Comparison of svm and some older classification algorithms in text classification tasks. In IFIP International Conference on Artificial Intelligence in Theory and Practice, pages 169–178. Springer, 2006.

123. Breiman, L.: Bagging predictors. Machine learning, 24(2):123–140, 1996.

124. Youth and tobacco.

125. Smith, A. and Anderson, M.: Social media use in 2018.

126. Crimson hexagon.

127. Twitter.

128. tweepy reference.

VITA

NAME: Akshay Uppal

EDUCATION: MS, Computer Science, University of Illinois at Chicago, Chicago 2019.

B.tech., Computer Science, GGSIPU, Delhi, 2014.