

5-8-2019

Understanding Juuling Trends Among Differing Age and Gender Demographics Through Social Sensing

Thomas McFann
University of Mississippi

Follow this and additional works at: https://egrove.olemiss.edu/hon_thesis



Part of the [Computer Sciences Commons](#)

Recommended Citation

McFann, Thomas, "Understanding Juuling Trends Among Differing Age and Gender Demographics Through Social Sensing" (2019). *Honors Theses*. 1066.
https://egrove.olemiss.edu/hon_thesis/1066

This Undergraduate Thesis is brought to you for free and open access by the Honors College (Sally McDonnell Barksdale Honors College) at eGrove. It has been accepted for inclusion in Honors Theses by an authorized administrator of eGrove. For more information, please contact egrove@olemiss.edu.

Understanding Juuling Trends Among Differing Age and Gender Demographics Through
Social Sensing

by
Thomas McFann

A thesis submitted to the faculty of The University of Mississippi in partial fulfillment of
the requirements of the Sally McDonnell Barksdale Honors College.

Oxford
April 2019

Approved by

Naeemul Hassan

Advisor: Professor Naeemul Hassan

Kristin Davidson

Reader: Professor Kristin Davidson

Dawn E. Wilkins

Reader: Professor Dawn Wilkins

Copyright Thomas McFann 2019
ALL RIGHTS RESERVED

ABSTRACT

Middle schoolers, high schoolers, and college students have increasingly begun using the Juul, which is a type of e-cigarette. In fact, in 2017 a study showed that “2.1 million high schoolers and middle schoolers used e-cigarettes” (Richtel and Kaplan, 2018). The reason why my research specifically addresses the Juul is that it “is now the largest e-cigarette brand measured by retail sales companies” (Huang et al., 2018b). Studies have been done to show the short time it takes for a child to lose autonomy over tobacco use, ranging from 2 to 30 days of first inhaling a cigarette (DiFranza et al., 2007). What are the potential consequences of a nicotine addiction for developing youth? Nicotine is widely believed to inhibit normal mental development. (Goriounova and Mansvelder, 2012). Additionally, youths are much more susceptible to nicotine relative to adults (Goriounova and Mansvelder, 2012). Companies have claimed that they are unintentionally trying to market their product to youth and that a Juul’s primary purpose is to act as an alternative to currently smoking adults above the age of 35. Recently from The University of Mississippi Medical Center, a documentary came out about vaping in Mississippi (Remedy, Aired 08/30/18). Nicotine aside, is it healthy to breath in propylene glycol and vegetable glycerin after being heated? To summarize, more middle, high school, and college students are beginning to use e-cigarette products. More specifically, they are using the most popular e-cigarette product, the Juul. Because we know the dangers of nicotine addiction in developing brains coupled with the uncertainty of the safety of the chemicals within a vape device in general, strong concern for youth from associations such as the FDA and cancer society has succeeded in putting pressure on the Juul company to discourage young students from partaking in their products. Recently at the completion of this thesis, the FDA has been a force to be reckoned with. In order to understand the appeal of Juul, one can examine the tweets of those who mention the Juul.

Additionally, in order to understand the total appeal of the Juul from twitter, one should explore its perception in different age and gender demographics, a method which has not been widely explored to date. As far as I am aware, this is the first attempt to make a model which takes into account gender and age demographics. This is a complicated task considering that Twitter does not provide age and gender information about their users. Why is this problem important to my research, scientific, and academic community? Computer Scientists have the ability to dive into Large Data questions and have a moral responsibility to look into the safety of the younger generation. The Honors College at Ole Miss states that it seeks to prepare citizen scholars who are fired by the life of the mind, committed to the public good and driven to find solutions. This is a citizen question, because many of my contemporaries partake in Juuling and I am driven to help explore the nature of their subsequent nicotine addiction and the nicotine addiction of those younger than us. My computer science department states that it seeks to enable its undergraduate students to master the fundamental principles of computing and to develop the skill needed to solve practical problems using contemporary computer-based technologies and practices. Data analyses is a growing field in Computer Science, and I will have the opportunity to use these skills to help solve this problem. Lastly, the scientific community is eagerly trying to understand the effects of vaping and how to produce some oversight on this industry.

Understanding each demographic (age and gender) attraction to the product through Twitter content will allow one to compare which aspect of the Juul is most appealing to each age band. Further research into this may begin to uncover particular ways to help the FDA combat the current youth epidemic in terms of knowing where to cut off the appeal of the Juul at its source.

I have collected as many tweets which contain content regarding #Juul, normalize the data, determine whether or not a tweet comes from a person or an organization, identify the age and gender of each Twitter user, and implement LDA (Latent Dirichlet Allocation) Topic Modeling and Association Rule Mining on the processed data. From this information, I will

determine the most popular topics regarding Juul in each demographic as well as rules which reveal correlations between the Juul and other words from the tweets. This information will provide useful knowledge that can be utilized when considering how best to implement Juul policy while waiting to better understand its effect on health.

DEDICATION

This thesis is dedicated to my mom for the principles of learning that she instilled in me at such a young age and the constant encouragement and guidance that she provides me with on a daily basis.

ACKNOWLEDGEMENTS

First and foremost I would like to thank my advisor Dr. Naeemul Hassan for his constant work and energy that he has invested in me, his patience, and his guidance to equip me to complete this thesis.

I would like to thank the Fake Health News Lab Group, also led by Dr. Hassan, which gave me my first introduction into research and prepared me with the necessary background to successfully attempt my project and learn to work well with others.

I would like to thank Dr. Kristin E. Davidson and Dr. Professor Dawn E Wilkins for being my second and third reader and providing me with such excellent feedback.

I would like to thank the computer science department for creating such an inviting and intimate environment where I could learn the necessary skills to be successful in the work force and providing me with opportunities to research and uniquely apply my computer science skills in areas of personal interest.

I would like to thank the Honors College for funding so much of my education and providing an enriching atmosphere to engage with excellent professors, faculty, and students. Through the Honors College I have learned the necessary skills to become a citizen scholar and how to critically analyze and discuss essential topics that affect our society.

Lastly, I would like to thank all of my friends and family who are a constant source of encouragement and joy as I pursue my passion of learning.

TABLE OF CONTENTS

ABSTRACT	ii
DEDICATION	v
ACKNOWLEDGEMENTS	vi
LIST OF FIGURES	ix
HISTORY OF E-CIGARETTES	1
EFFECTS OF E-CIGARETTES ON YOUTH	4
HISTORY OF JUUL	8
LITERATURE AND HYPOTHESIS	13
METHODOLOGY	20
DATA COLLECTION	22
DATA PROCESSING	28
LDA TOPIC MODELING	33
ASSOCIATION RULE MINING	41
RESULTS	45
DISCUSSION OF RESULTS	52

METHODOLOGY LIMITATIONS	70
CONCLUSION	75

LIST OF FIGURES

1.1	Typical e-cigarette configuration. This shows a wick/heater as aerosol generator, gauze saturated with e-liquid, a microprocessor (optional) to control operations and an LED (optional) to imitate a burning coal. (Brown and Cheng, 2014)	2
3.1	Sales dollar of e-cigarettes in Nielsen-tracked retail channels: by brand 2011 through 2017. (Huang et al., 2018b)	9
3.2	Teens are more likely to use e-cigarettes than cigarettes. (Huang et al., 2018b)	10
3.3	What do teens say is in their e-cig? (Huang et al., 2018b)	10
3.4	High teen exposure to e-cig advertising. (Huang et al., 2018b)	11
4.1	Number of JUUL-related tweets on twitter 2015-2017 (Huang et al., 2018b) .	14
5.1	A Roadmap of My Code	21
6.1	JSON example	23
6.2	User object example	24
8.1	TF-IDF	36
8.2	LDADistribution (Algobeans, 05/30/2018)	40
10.1	HashtagFrequency	45
10.2	Data Collection	46
10.3	18 and Under TFIDF Model	46
10.4	18 and Under TFIDF Model Female	47
10.5	18 and Under TFIDF Model Male	47
10.6	19-24 TFIDF Model	47
10.7	19-24 TFIDF Model Female	47
10.8	19-24 TFIDF Model Male	48
10.9	25-35 TFIDF Model	48
10.10	25-35 TFIDF Model Female	48
10.11	25-35 TFIDF Model Male	48
10.12	35-55 TFIDF Model	49
10.13	35-55 TFIDF Model Female	49
10.14	35-55 TFIDF Model Male	49
10.15	Greater than 55 TFIDF Model	49
10.16	Greater than 55 TFIDF Model Female	50
11.1	Partying Imagery Example	56
11.2	Illegally selling Juul to High Schoolers	57
11.3	Discussing Those Who Do Juul	57
11.4	Male trying to quit smoking	58
11.5	Love and Juul Female example	59
11.6	Juul Versus Cigarettes	60
11.7	Introspection about the Juul Business	61

11.8 Juul and Smoking Cessation	62
11.9 Juul in the Bathroom	63
11.10 Teen Epidemic	65
11.11 Illegal selling of the Juul to minors	66
11.12 The Social Aspect	67
11.13 Juul in School	67
11.14 Need Juul	67
11.15 Juul pods for Christmas	67
11.16 Elders Juuling	68
11.17 Vaping Restrictions at Work	68
11.18 Quitting Cigarettes using Juul	69
11.19 Quitting Juul	69
12.1 Sharability of the tweet in a social context even in digital form.	71
12.2 Tweet shows that a Juul can often be found with a loko, the idea of ripping it in a bathroom and greek life.	71
12.3 Association that a juul addiction has with college, beer, and partying. 3 . . .	71
12.4 Tension of children trying to hide their Juul from their parents.	72
12.5 Year of different types of pods.	72
12.6 This represents a popular opinion of thinking of a Juul in the context of par- tying.	72
12.7 This shows a hidden aspect in which girls are likely to use flirtation to get what they want from a guy.	72
13.1 Pyramid Representation	77
13.2 Food Chain Representation (Amit, 5/24/2018)	78

CHAPTER 1

HISTORY OF E-CIGARETTES

Before exploring the Juul itself, we must first familiarize ourselves with e-cigarettes since a Juul is a type of cigarette. Once we set a foundation for the history of e-cigarettes, we can then begin to better understand the contents that are contained within a e-cigarette which will in turn further the conversation as to what exactly the epidemic concerning the Juul looks like. This is an important step, because topic modeling will be used to classify tweets according to specific topics in each age and gender group. Only through understanding the context of a Juul will one likely be able to most accurately interpret the data.

1.0.1 **Historic Origins of the Juul**

The idea for electronic cigarettes can be dated all the way back into the 1930s when the first patent for such a product was granted to Joseph Robinson (CASAA, 2016). However, the moment when e-cigarettes became most relevant to our culture began in 2007, when electronic cigarettes were introduced to the U.S (CASAA, 2016). According to Medical News Today, an “electronic cigarette is a battery-operated device that emits doses of vaporized nicotine, or non-nicotine solutions, for the user to inhale. It aims to provide a similar sensation to inhaling tobacco smoke, without the smoke” (Braier, 6/25/2018). It is important here to begin thinking about two potential forms of addiction. One coming from the nicotine itself and the other from the sensation of inhaling. At the time of its introduction, due to the newness of the product to public awareness, few FDA regulations were put in place. Southern Remedy suggests that “few regulations created a strong market for vaping which has resulted in a 2.5 billion dollar industry, a new youth culture, and concerned health advocates” (Remedy, Aired 08/30/18). In terms of the industry, each year it keeps growing.

When I mention youth culture, imagine a lounge with pool tables and dart boards. Imagine an area dedicated to openness and relaxation where fellow vapers can make connections and enjoy vaping together. Part of what makes a vape so appealing is the ability for one to make a hobby out of it. Let us consider the components of a vape to understand how people can use it as a hobby and also build community.

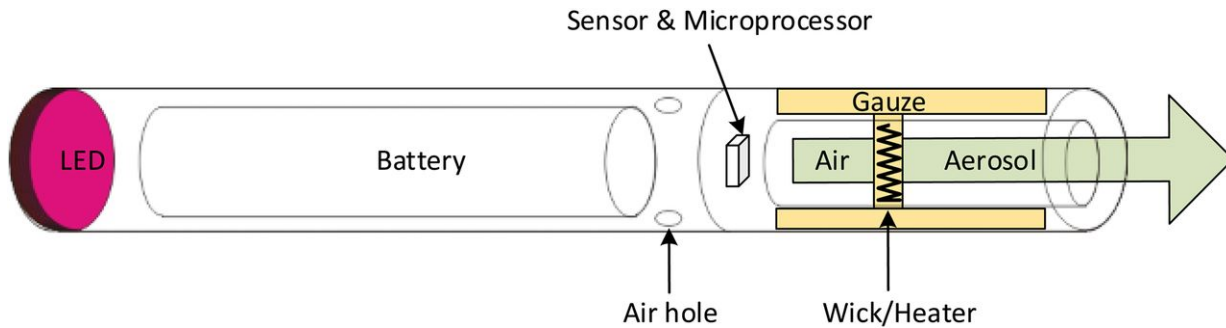


Figure 1.1: Typical e-cigarette configuration. This shows a wick/heater as aerosol generator, gauze saturated with e-liquid, a microprocessor (optional) to control operations and an LED (optional) to imitate a burning coal. (Brown and Cheng, 2014)

1.0.2 How Does a Vape Work?

As the source from which this diagram comes explains, the “user draws upon the e-cigarette, which activates an airflow sensor. The airflow sensor detects pressure changes and prompts the flow of power to an LED and a heating element. The e-liquid saturates a wick via capillary action and is then aerosolised by the heating element. The aerosolised droplets of e-liquid subsequently flow into the user’s mouth and lungs” (Brown and Cheng, 2014). According to the Cambridge Dictionary, aerosol is a container in which liquids are kept under pressure and forced out in a spray (mass of small drops). A vape, simply put, is “a coil wrapped around a cotton wick which heats up” according “to a certain voltage in order to vaporize the juice” (Remedy, Aired 08/30/18). This mechanism allows for various means of experimentation. For instance, one can experiment with different kinds of voltage to heat up the aerosol to ones liking. One can use different materials for both the wick and the coil. Varying amounts of the chemicals which I will touch on shortly can create different

sized clouds of vapor. Lastly, one can experiment with the cartridges themselves in order to determine the concentrations of nicotine as well as choose the flavor that one prefers. Building and modifying e-cigarettes has become a hobby for many friend groups, especially those who enjoy interacting with others at a vaping lounge. Thus far, I have identified three potential main attractions to the Juul: the nicotine, the sensation of smoking, and the creation of a social identity. The Juul is another way to pick up dates at the bar in that you can approach someone or they can approach you to take a puff of your Juul and it serves as a conversation starter. Having covered a brief history of the Juul and how it works, let us now consider its potential effects on youth.

CHAPTER 2

EFFECTS OF E-CIGARETTES ON YOUTH

Knowledge of the chemicals inside of a vape will give us a better understanding of its addictive substances and allow us to speculate as to whether such an addiction is worse than being addicted to another chemical such as caffeine. After all, a major reason that I took on this project is under the assumption that there are inherent dangers in youth being needlessly addicted to nicotine.

2.0.1 Dangers of Heating the Chemicals Inside of the Juul

Thomas Payne, PhD, a director at ACT Center for Tobacco Treatment, Education and Research at the University of Mississippi Medical Center cites five main components of the liquid. These are water, nicotine, propylene glycol, vegetable glycerin, and flavoring. “Vegetable glycerin creates the vapor while propylene glycol creates the feeling against the throat that you have inhaled something” (Remedy, Aired 08/30/18). A common statement that the documentary addresses is the idea that these chemicals on their own are not harmful in moderation to the human body. In fact it was mentioned that Propylene glycol is one of the vehicles used in puffer inhalers for medicine. The question is whether or not when heated at specific temperatures, if either the heated metal or the heated chemicals transform into something toxic.

Basically, this is an “incomplete combustion reaction says one on a chemistry answer exchange and the purpose is to get the propylene glycol, glycerin, and flavor molecules into vapor phase without chemically degrading them. Since the combustion is incomplete, you can get different types of compounds two of which are formaldehyde and acetaldehyde and other enhanced aldehydes” (Chemistry Stack Exchange, 6/08/2015). He makes sure to confirm

that the mechanism by which he describes this process is a theory. Another answer speaks further on the different chemicals that are potentially produced. Formaldehyde is stable. However, Acetaldehyde can dimerize to crotonaldehyde. Acrolein is the simplest unsaturated aldehyde (Chemistry Stack Exchange, 6/08/2015). The point is both formaldehyde and acrolein for example are known to be toxic. Payne, who was quoted earlier, is emphasizing the fact that although harmless to begin with, we need to be concerned with what these chemicals transform to when heated, many of which may be toxic (Remedy, Aired 08/30/18). Although researchers have not definitively proven the mechanism by which these chemicals transform, it is understandable where their concern lies.

2.0.2 The Affects of Nicotine on Brain Development

Since the results are not conclusive yet as to how toxic vaping is, for the sake of argument, I would like to briefly assume that vaping is a healthier alternative to smoking and look specifically at how nicotine alone would affect the minds of youth. What is nicotine? According to the documentary, nicotine is a highly addictive drug that leads to cravings and a physiological need for more. Nicotine naturally occurs in tobacco; however, the use of different e-cigarettes causes nicotine levels to vary. Nicotine stimulates the release of endorphins in the brain which one produces when you exercise, and also dopamine is released, a pleasure center simulator, when one inhales from a vape (Remedy, Aired 08/30/18). A study was done that examined the “Short-Term and Long-Term Consequences of Nicotine Exposure during Adolescence for Prefrontal Cortex Neuronal Network Function”. Already more inclined to making risky situations based upon impulse, the behavior “of adolescents to that of adults may point to an enhanced sensitivity of the adolescent brain to addictive properties of nicotine” (Goriounova and Mansvelder, 2012). Additionally the article sights how many studies ”indicate that smoking during adolescence is associated with disturbances in working memory and attention” (Goriounova and Mansvelder, 2012). Thus, aside from the addictive nature of nicotine, it effects brain development among adolescents.

2.0.3 Juuling As a Gateway To Tobacco Products

Additionally, there has been growing concern as to whether vaping can act as a gateway to tobacco products and even more potent addictive substances like drugs. One study was done that concluded that “increasing familiarity with nicotine could lead to the re-evaluation of both electronic and tobacco cigarettes and subsequently to a potential transition to tobacco smoking, hence the catalyst model” (Schneider and Diehl, 2015). There are two other models that support the idea of nicotine being a gateway drug. The first is the Gateway Hypothesis in which “young people become involved in drugs in stages and sequences” (Etter, 2018). For instance one can imagine a progression of addiction where specifically, “the use of tobacco or alcohol precedes the use of marijuana, which in turn precedes the use of cocaine and other illicit drugs” (Etter, 2018). The consequences of this is important. If it is established that vaping is a gateway to tobacco use and that in turn leads to the other drugs, vaping can be indirectly very dangerous. There is also the Common Liability Model in which “the use of multiple drugs reflects a common liability for drug use and that addiction, rather than the use of a particular drug, increases the risk of progressing to the use of another drug” (Etter, 2018). This idea is reminiscent of the common thought pattern that spells the end of a diet. As I sometimes think, “I already ate this food and lost control, I might as well lose control in this other area of my diet”.

2.0.4 Juuling As It Relates To Smoking Cessation

Alternatively, it is also important to cite that people who use e-cigarettes as a nicotine replacement therapy (NRT) to quit smoking often succeed. One such study found that “among smokers who have attempted to stop without professional support, those who use e-cigarettes are more likely to report continued abstinence than those who used a licensed NRT product bought over-the-counter or no aid to cessation” (Brown et al., 2014). That being said, medical associations are not saying that there is absolutely no benefit to vaping such as for smoking cessation. The question is, just how well is this cessation of smoking

mission being carried out when many youth are becoming addicted to nicotine from vaping which, in turn, makes them potential future customers of tobacco products.

2.0.5 The Difference Between Being Addicted to Nicotine Versus Caffeine

Lastly, is it possible to associate nicotine as an addiction similar to caffeine? Stanton A. Glantz, PhD and part of the Center for Tobacco Control Research and Education addresses this question by stating 5 conclusions related to the effects of nicotine which would differ from the affects of caffeine (Stanton A. Glantz, 10/3/2014).

1. “The evidence is sufficient to infer that at high-enough doses nicotine has acute toxicity.
2. The evidence is sufficient to infer that nicotine activates multiple biological pathways through which smoking increases risk for disease.
3. The evidence is sufficient to infer that nicotine exposure during fetal development, a critical window for brain development, has lasting adverse consequences for brain development.
4. The evidence is sufficient to infer that nicotine adversely affects maternal and fetal health during pregnancy, contributing to multiple adverse outcomes such as preterm delivery and stillbirth.
5. The evidence is suggestive that nicotine exposure during adolescence, a critical window for brain development, may have lasting adverse consequences for brain development.
6. The evidence is inadequate to infer the presence or absence of a causal relationship between exposure to nicotine and risk for cancer.”

Now that we have gone through the history of e-cigarettes and the way in which its chemicals are thought to affect teens, we can now look specifically into the current state of affairs of Juul which will finally give the context needed to fully explore my large data model.

CHAPTER 3

HISTORY OF JUUL

3.0.1 Who Invented the Juul?

Now we come to the subject matter of my research, the Juul. According to their website, “James Monsees and Adam Bowen co-founded JUUL Labs when they applied their background in product design to the challenge of finding a true alternative to smoking. They had been smokers for many years, but when they could find no acceptable alternative to cigarettes, James and Adam recognized a groundbreaking opportunity to apply industrial design to the smoking industry, which had not materially evolved in over one hundred years. As smokers, they knew a true alternative to cigarettes would have to offer a nicotine level found in no other alternative on the market. It would also have to invite its own ritual. The result was JUUL” (Labs, 2019). Thus we see that the stated intent of this product was to aid fellow smokers in quitting from their addiction. A Juul is a type of e-cigarette that has steadily become the most dominant e-cigarette in the market. Figure 3.1 shows just how fast Juul has been growing by sales. When looking at the red vertical bar, it is clear that its influence is growing exponentially.

3.0.2 The FDA’s Concern and Teen Statistics

Victor Luckerson wrote an article called “Is Juul the Startup World’s Greatest Long Con?” Although the article is not extremely credible by scholarly standards, in terms of information, it provides an accurate synopsis of the history of Juul, which I compared to other news articles. While the article’s claim that the Juul is a con may be strong, it does provide some food for thought on potential motives which must be taken into account.

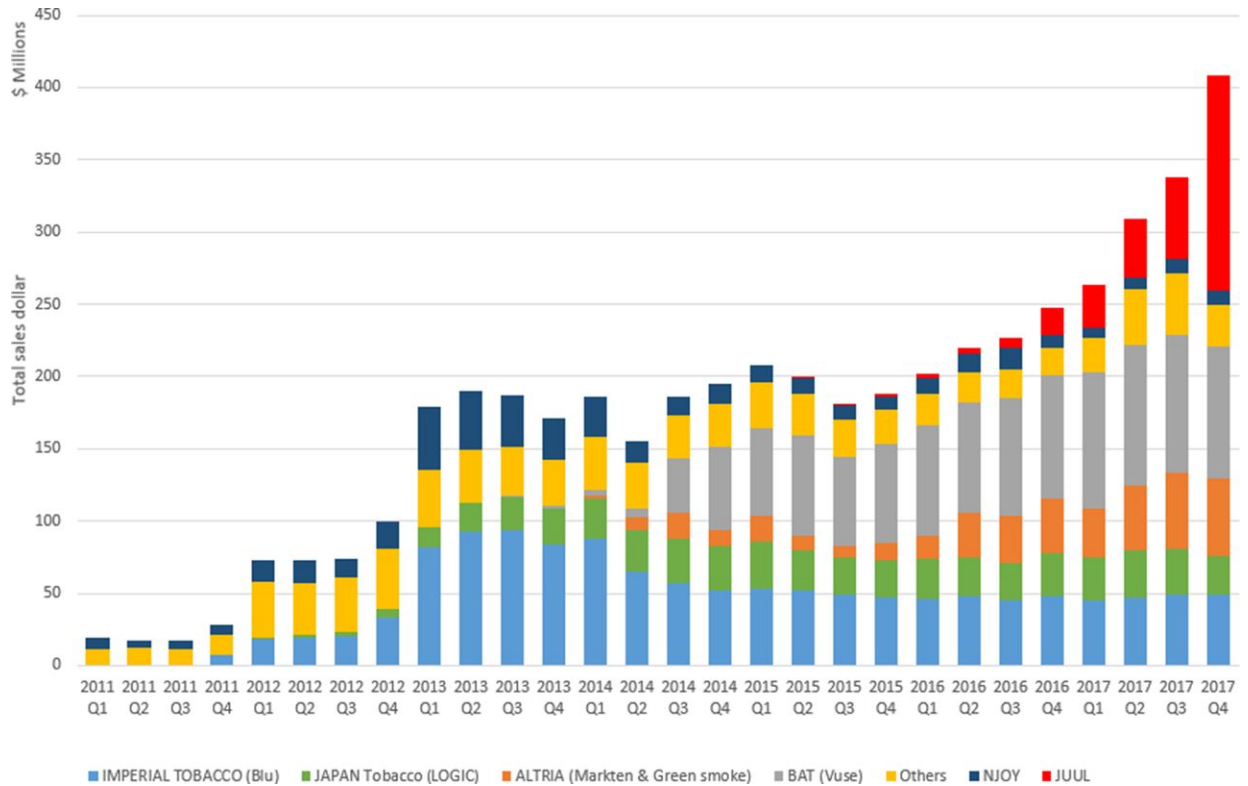


Figure 3.1: Sales dollar of e-cigarettes in Nielsen-tracked retail channels: by brand 2011 through 2017. (Huang et al., 2018b)

Juul founders Adam Bowen and James Monsees developed their product to help themselves to quit smoking. However, after teens started getting addicted to vaping, the FDA became concerned and eventually labeled Juuling as an epidemic. Juul, however, claimed that it was not intentionally targeting the youth, but youth usage kept rising. Around the same time, the FDA began an investigation and shortly after, Juul shut down all social media accounts and began limiting the flavors of pods.

Certain figures were released from the National Institute on Drug Abuse related to teens and e-cigarettes, which by extension, would relate to the Juul. They relate to topics such as how likely it is that teens will use a e-cigarette over a cigarette, the knowledge that teens have of what is in their e-cigarette, and the exposure teens have had to Juul advertising through social media such as twitter. From the three different figures below, some conclusions can be made.

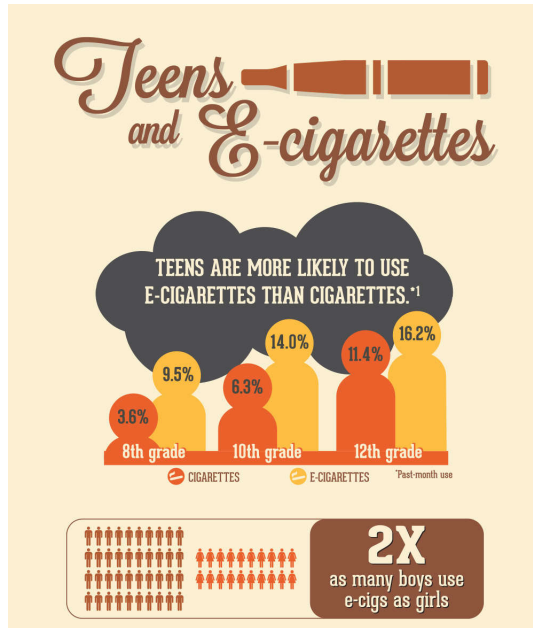


Figure 3.2: Teens are more likely to use e-cigarettes than cigarettes. (Huang et al., 2018b)

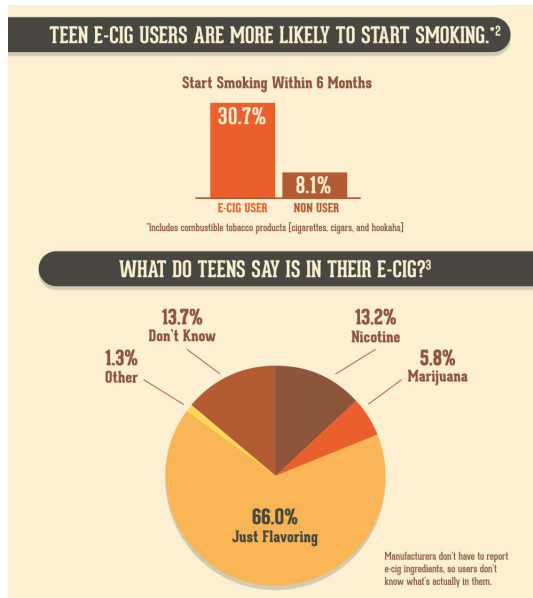


Figure 3.3: What do teens say is in their e-cig? (Huang et al., 2018b)

Figure 3.2 shows that “30.7 percent of e-cig users started smoking within 6 months while 8.1 percent of non users started smoking,” figure 3.3 shows that “66 percent of teens think there is just flavoring” in their e-cig/juul, and figure 3.4 shows that “7 in 10 teens are exposed to e-cig adds” (NIDA, 2/01/2016).

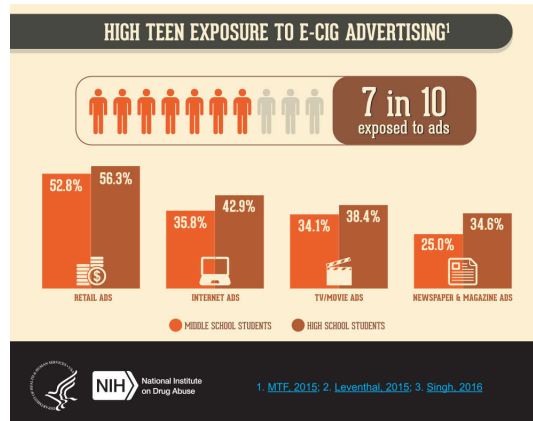


Figure 3.4: High teen exposure to e-cig advertising. (Huang et al., 2018b)

Eventually, the FDA took initiative to make it very difficult for youth to get their hands on flavored vapes after the FDA and CDC revealed that “3.6 million high school and middle school students are using e-cigarettes” (NIDA, 2018). “That’s 1.5 million more students using these products than the previous” year said the FDA Commissioner Scott Gottlieb” (NIDA, 2018). Interestingly enough, this statement came out two years after the previous figures shown. Because Juuls are the most popular e-cigarette, especially among youth, it is clear that Juuls are exploding in popularity.

3.0.3 Controversial Question Surrounding the Juul

Based upon the article, “Is Juul the Startup World’s Greatest Long Con?”, an interesting development can be found. Although the implications of this development may or may not be true, the facts are interesting to consider. Bowen and Monsees, the founders of Juul, sold stake in an earlier vaping product that they made to Ploom, which is a Japanese version of Philip Morris. In essence, Altria bought a 35 percent stake in Juul (Rivas, 12/20/2018). Let’s pick apart what this might mean by learning a little about this company. According to the Philip Morris site, in 1847 Mr. Philip Morris opened a shop on London’s Bond Street, selling tobacco and ready-made cigarettes. After much company growth, they eventually came out with Marlboro in 1924, which was to become the company’s most famous cigarette brand. In 2000, Philip Morris International (PMI) called for regulation of the tobacco in-

dustry at the World Health Organization’s public hearings on the Framework Convention for Tobacco Control in Geneva, Switzerland. In 2013, this led to in 2013, PMI establishing a strategic framework with Altria Group, Incorporated (Morris, 2/13/2019). Altria emerged from Philip Morris during the onset of re-branding where Altria is “moving toward a future of lower-risk products” says (PMI) Chief Financial Officer Martin King, and, in doing so, has confirmed that it is “taking a 35 percent stake in Juul” (Rivas, 2018). Looking at its website, PMI, highly brands itself as working toward “designing a smoke-free future.” Although they are selling tobacco products, research has gone into creating better alternatives. From my perspective, it makes sense that PMI would acquire so much stock in Juul since it is a potential helpful alternative for adults. However, this confuses the message of Juul since PM still continues to sell tobacco products. There is a tension between this product being supposedly designed for quitting yet so much of it is owned by a company that still sells tobacco products. I am curious if these companies will be mentioned in any of the tweets. It is important to keep track of how the tweets change over time as this subject begins to evolve.

CHAPTER 4

LITERATURE AND HYPOTHESIS

4.1 Literature

In light of the history of the Juul, especially more recently and its effect on the younger population, I have found it of utmost importance to understand this problem on a deeper level. In order to do so, I am delving into the world of twitter. Why twitter? I like to think of twitter as the pulse taker of the world. Already, through the use of twitter, improvements in understanding world problems such as when the “US geological survey was using tweet data to track earthquakes”, “predicting flu outbreaks”, “following Ebola”, tracking and responding to “flood damage in Jakarta”, monitoring “civil unrest in Egypt,” and “predict crime in the US” have been accomplished (Hutchinson, 3/18/2018). In a similar way, Juul’s have been a topic of much conversation on twitter. Already, some research has been done on Juul through the use of social media.

4.1.1 Literature Research Study 1

In a research paper call “Vaping versus Juuling: how the extraordinary growth and marketing of Juul transformed the US retail e-cigarette market”, a variety of data sources were used to examine Juul retail sales in the USA and its marketing and promotion (Huang et al., 2018b). One of the social media platforms used was Twitter. Their method for collecting twitter data involved collecting Juul-related tweets from the Twitter Historic Power Track, “which provides access to 100 percent of all archived tweets, as well as meta data associated with each tweet” (Huang et al., 2018b). These were collected from January 2015 to December 2017. A figure of their results for twitter sales data is posted below. I am about

to discuss several studies which will be highly relevant in guiding the discussion of my own research.

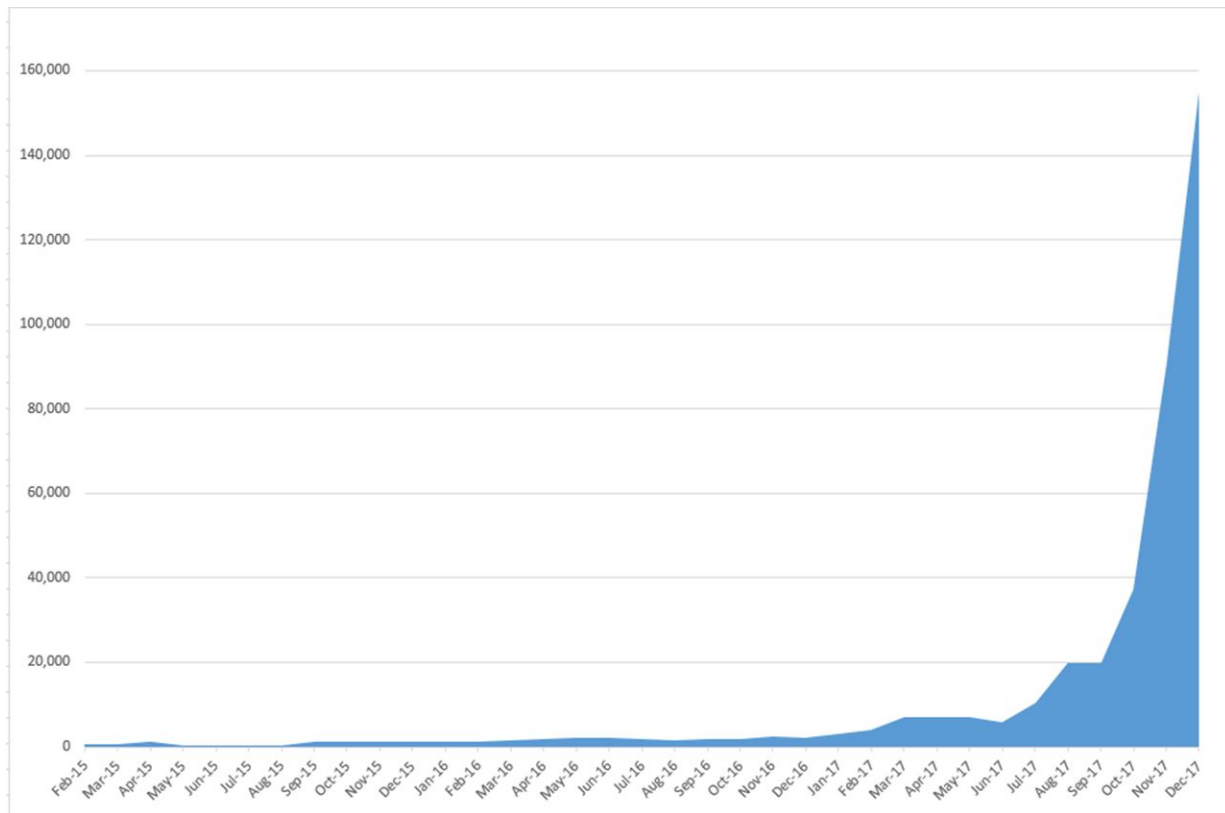


Figure 4.1: Number of JUUL-related tweets on twitter 2015-2017 (Huang et al., 2018b)

These results basically confirmed for the researchers that “the growth trend in Juul tweets noticeably tracks well with the growth in Juul retail sales; the two data series were highly correlated, with a correlation coefficient of 0.968” (Huang et al., 2018a).

4.1.2 Literature Research Study 2

Another study which is similar in scope is the study called “Characterizing JUUL-related posts on Twitter.” I will discuss the high points of this study in terms of their strategy and conclusion, so that it can be compared to the strategy and conclusion of my own work. After cleaning the twitter posts, which included removing the word “Juul” in contexts which are unrelated to the “Juul” that is being studied and attempting to identify tweets produced by a social bot they had a final analytical sample of 81,689 tweets from

52,098 unique users. They also utilized basic normalization in that they removed lower case text, stop word removal, normalization of Twitter user mentions (“@johnsmith” converted to “@person”, lemmatization, and non-printable character removal). I will revisit these topics when discussing my own data normalization process. The relevance of this study stems from its attempt to identify the major topics associated with Juul tweets. Some popular topics deduced from this study includes, “Flavors” like mango or mint, “pods” such as the juul refill cartridges, “person tagging” such as using @username to tag someone, “school” such as seeing one in a school setting, “college” indicative of being a college student, “buying”, “Lost and Found”, “wanting a Juul”, and “charger” (Allem et al., 2018).

4.1.3 Literature Research Study 3

Another study was done called “A content analysis of JUUL discussions on social media: Using Reddit, a social media platform, to understand patterns and perceptions of JUUL use” in which the “content analysis utilized social media discussions posted between January 2015 through May 2017. Public posts on Reddit were gathered and coded. Posters of discussions relevant to both Juul and youth were included for analysis” (Brett et al., 2019). Some findings included the strong role social norms play in youth use as well as how age restrictions and public health interventions can curb youth use. Additionally, some reasons for use were identified as ”popularity“, ”social/friends“, “buzz”, “taste”, “easily available”, “price”, “stealthy/discreet”, “addiction”, “use to quit smoking”, “throat hit”, other (Brett et al., 2019). Barriers to use include “price”, “concern about health”, “concern about addictiveness”, “age restrictions”, “lack of availability” (Brett et al., 2019).

4.1.4 Description of Demographics Being Studied

A critical part of my research involves validating the methodology used to inference age and gender from the Twitter users. Not only do I want to understand the main topics related to the Juul as these articles discuss, but I also want to find a way to study how these main topics interact with each demographic. For example, I would like to create five age

brackets: under the age of 18, between 18 and 24, between 24 and 35, between 35 and 55, and 55 and older. I also would like to look at this from the standpoint as a whole as well as differentiating between male and female. Such an analysis will hopefully provide a way to see which demographics interact with which topics. Additionally, I am also looking to see which words or topics tend to be correlated together. For example, does the word “Juul” and “stress” seem to have a significant probability of being seen together within a tweet. In this study, I expect to see some consistency in the types of topics found related to the Juul as the previous studies showed.

4.2 Hypothesis

The following includes my hypothesis about which I am expecting to find based upon my research into the context of the emergence of the Juul, the findings of others, and personal anecdotes.

4.2.1 Hypothesis of Expected Topics to Find Within Different Age Groups

1. Hypothesis One: A great deal of concern as to the health of the Juul for youth which was discussed earlier.
2. Hypothesis Two: Polarization of topics in that a tweet will either be related to an adult or entity concerned about the health effects of the Juul for the youth or else see the juul as a normal every day word in the context of juvenile conversation.
3. Hypothesis Three: Social life or to personal well being to help manage the stress of school.

Due to the fact that I am getting the data from twitter, I find it more likely that I will see more of a correlation towards the Juul and social protocol. Yet, in a sense this is one of the questions at the heart of my study. Below are other questions I would like answered.

4.2.2 Question To Be Answered 1

As a youth, does the insatiable desire to feel included within a social group outweigh the biological effects of addiction? As the previous research would indicate, however, if Juuling can be seen as a gateway drug, would not the desire for popularity be the gateway craving to a gateway drug?

4.2.3 Question To Be Answered 2

Which are the most common 3 topics for each age and gender demographic as revealed through WordCloud?

1. **Under 25 Category Expectation:** For those under the age of 25, I expect to see more slang terminology that is casual and incorporates Juuling as a way of life. I would also expect to see the word “Juul” more in the context of social pleasure and partying where alcohol is involved. Lastly, I would expect to see the Juul with relation to school.
2. **Middle Age Category Expectation:** In the middle aged category, I expect to see something different. For one, I expect to see more conversation related to the cessation of smoking. Due to the recent branding approach of Pax Labs, a successful promotion campaign would involve the popularity of adults discussing the importance of the Juul in helping them to quit. I would also expect a lack of discussion about the Juul as a part of their every day life. For instance, although I would not expect anybody to be ashamed of Juuling, especially if it is being used to help them quit, I would not imagine them wanting to draw too much attention to the fact that they vape for professional reasons as well as to encourage their kids to not start. However, I could see one talking about it if they were an activist in favor of using Juuling to cease smoking.
3. **Above 55 Age Category Expectation:** Among those 55 and older, I do not imagine much will be said about the Juul among them. If anything, I would expect greater

retweets from them about the health concerns of the Juul as it potentially relates to their older children and eventual grand children. I am curious to find in which age demographic to most people discuss the health concerns of the Juul for youth.

4.2.4 Question To Be Answered 3

Which will be the most common Association Mining Rules found?

1. **Under 25 Category Expectation:** I expect to find larger associations between Juuling, flavors, and friends which would indicate the social aspect to Juuling. In a similar manner, I expect to see the Juul associated with stress, especially in a college and high school settings.
2. **Middle Age Category Expectation:** Juuling and smoking should be very popular when there is so much talk of the health comparisons. Likely, one will see quitting as a major theme as well as the word “teenager” indicating concern with the health epidemic.
3. **Above 55 Age Category Expectation:** Most likely those over 55 would be fascinated by the Juul itself having been familiar with cigarettes their entire lives. Therefore, there is most likely a curiosity concerning the Juul

4.2.5 Question To Be Answered 4

Is my coding methodology a successful one that is capable of being duplicated and refined for similar tasks in different topic areas as well as improved to find more subtle trends in Juul twitter data?

4.2.6 Question To Be Answered 5

Even underlying this data analytic question is the nature of addiction. As I said, I cannot but help ponder how social acceptance may be the gateway to the Juul gateway. The scholarly article “Social Identities as Pathways into and out of Addiction” speaks about

how “peer influence (both indirectly through modeling substance use, and directly through provision of substances and encouragement to use) is widely considered to be the most consistent and influential factor (Newcomb and Bentler, 1989)” related to addiction for adolescents (Marino et al., 2016). So, regardless of whether the data shows among youth a greater propensity to talk about the Juul in the context of their social spectrum, to the addictive nature of the Juul itself, hesitancy, or some combination of a three, it will be important to contemplate the nature of addiction as it relates to Twitter, the greatest social connector. The next section will deal primarily with the methodology by which I approached this study.

4.2.7 Question To Be Answered 6

Are there any factors that would be helpful to consider as the FDA goes about making policy decisions related discouraging teens to Juul? My thought is that perhaps by examining those topics and word associations which are most popular in the under 25 age groups, clues might be uncovered as to the root cause of the Juul epidemic and ways to address the problem.

CHAPTER 5

METHODOLOGY

From a bird's eye view, the code begins with collecting Twitter data from the Twitter API based upon tweets that contain #Juul. The state of data is represented by circles whereas each of the squares represent either an API or an algorithmic black box which transforms the data into a different form. So once the raw twitter data has been collected, my code sends it to an API called Humanizr which determines whether or not a tweet is from an individual or from an organization. The data is then sent through a function that I wrote which normalizes the text making it fit to send through the different algorithms. Then the user's of the tweet url image for their profile picture is then sent to the Face Plus Plus API which returns facial characteristics of the image. The facial information allows one to inference the age and gender of the Twitter which then is used to categorize the tweets according to age and gender demographics. This facial information for each tweet is then combined with the Twitter API information. The processed tweets are then sent to a function which performs LDA (Latent Dirichlet Allocation) topic modeling on the data and to a function which performs Association Rule mining. Consequently, the generated topic model and association rules are then consolidated into one area as the final result. Chapter 6 is dedicated to the details of data collection, chapter 7 deals with the data normalization/processing portion, chapter 8 addresses LDA topic modeling, and chapter 9 addresses the Association Rule Mining.

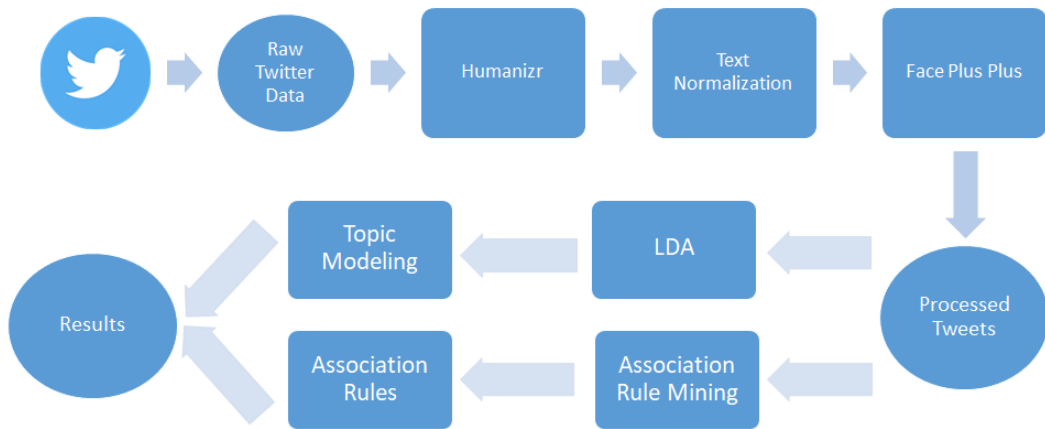


Figure 5.1: A Roadmap of My Code

CHAPTER 6

DATA COLLECTION

6.1 Data Collection sections

My programming language of choice for this project was python since it is a multi-paradigm programming language which supports object-oriented programming, structured programming, and functional programming patterns. This means that Python can be used for almost any task. Already, Python has many free data analyses libraries such as Pandas, SciPy, Numpy, Scikit-Learn, and Gensim among others.

6.1.1 Twitter API #Juul Detection

Twitter is “a social network where people post short, 140-character, status messages called tweets” (Dataquest, 2016). API stands for Application Programming Interface which is “a software intermediary that allows two applications to talk to each other” (MuleSoft, 2019). This means that the Twitter API allows me to interact with Twitter data which would otherwise be hidden from my application. In fact, it allows me to stream tweets real time from the Twitter API according to a topic. In my case, I streamed all tweets related to hashtag Juul. This is a rudimentary case. Given more time I could have collected tweets related to Juul and tweets related to other hashtags associated with the Juul. As part of my results, I hope to run a script to visualize the most popular tweets found. However, in the mean time, I found a website called “best-hashtags.com” which tells you the most popular hash tags associated with a specific topic. For instance, the top 10 case hashtags popular on Instagram, Twitter, Facebook, and Tumblr are “juul, memes, fortnite, vape, funny, dankmemes, meme, dank, edgymemes, memesdaily” (BestHashtagsWebsite, 1/31/2019). So I could have also

done various test with memes such as these but, unfortunately, did not have the time to test them all. So in real time, for every tweet that contains hash tag Juul, I am returned a JSON (JavaScript Object Notation) object of information regarding that tweet. Copterlabs in an article “JSON: What It Is, How It Works, and How to Use It” describes JSON as “a way to store information in an organized, easy-to-access manner” (copterlabs, 2019). The article goes on to say by enclosing the variable’s value in curly braces, we’re indicating that the value is an object. Inside the object, we can declare any number of properties using a “name”: ”value” pairing, separated by commas. Using this same terminology, I am returning an object which contains many “key”:“value” pairs which tell me a lot about not only the tweet but also the user. This is includes the hashtags, the user name, user name id, followers, follower count, profile image url, the tweet text, and much more. Below is a picture of a sample tweet collected from my data. This includes the id, id str, text, source, user, re-tweeted status, entities, extended entities, filter, level, lang, timestamp, and much more.

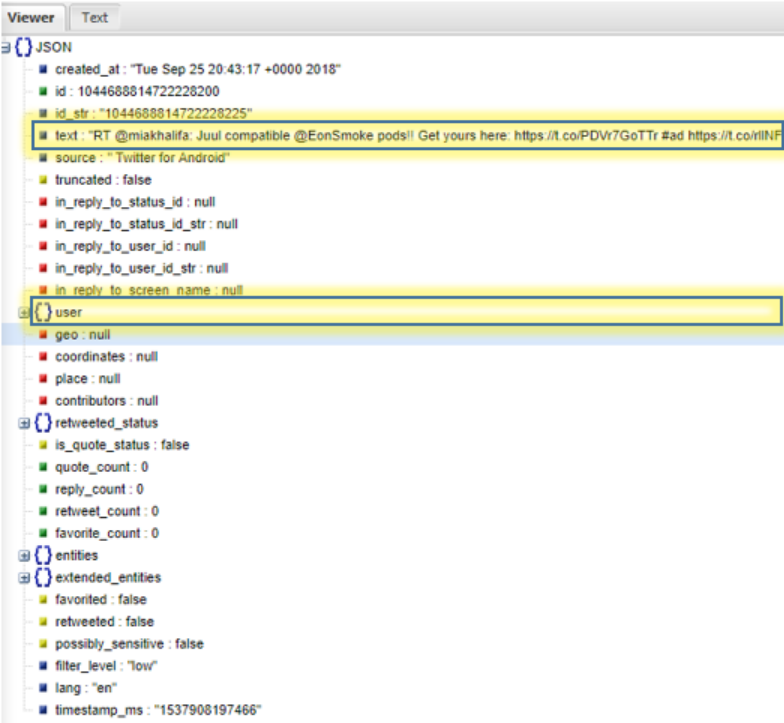


Figure 6.1: JSON example

Pay attention to the highlighted area of the object key “text” which contains the text of the tweet and the user, which holds even more key value pairs, important to the methodology. The following image below is a screen shot of some of the data contained within the user object.

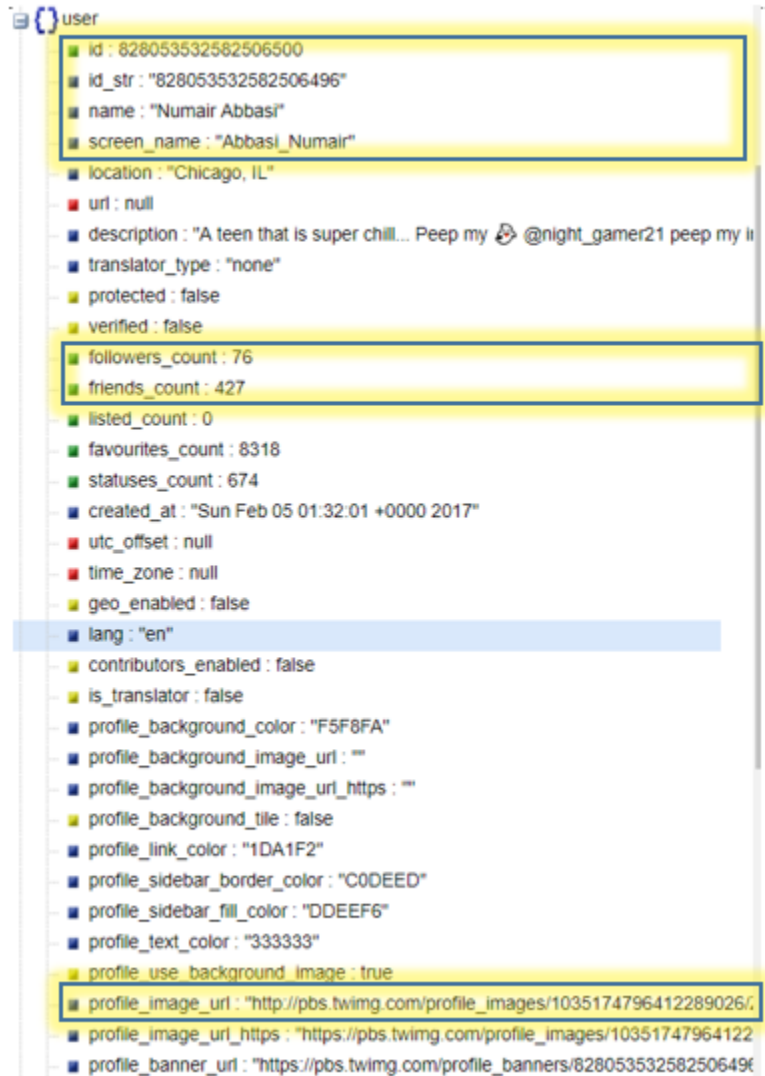


Figure 6.2: User object example

These highlighted lines are just some of the “key”:“value” pairs of which I am interested. Most of the information will help me to categorize the data later on. However, at the moment, the most important line is the “profile image url” which gives me a link to the tweeter’s profile picture on the internet. In such a manner, I collected 32,399 twitter objects

Data Collection	
Date	Number of Tweets Collected
9/25/2018	167
9/25/2018	53
9/27/2018	2323
10/04/2018	593
10/08/2018	89
11/05/2018	1063
12/12/2018	1680
12/13/2018	3665
12/14/2018	5476
12/16/2018	1094
1/09/2019	5711
1/11/2019	5720
1/15/2019	4765
Total	32399
Tweets with valid image	17000
Faces Recognized	11984
Ratio of Faces Recognized	.370

Table 6.1: Table of Data Collection

over the time frame September 25, 2018 to January 16, 2019 from the Twitter API.

6.1.2 Human Account Detection versus Organization Account Detection

As the Twitter logo indicates on the code road map in chapter 5, the 32,399 returned twitter objects were sent to the Humanizr API. The Humanizr software on GitHub “is a software library and off-the-shelf tool for classifying Twitter accounts as belonging to either an individual or an organization (e.g., a corporation or social group)” (McCorriston, 2015). Now each tweet is labeled as either coming from an individual or an organization. It turns out that all but two were labeled as coming from individuals. An assumption must be made here that Humanizr was successful in its delegation. However, it is possible that it erred in labeling so many as individuals. Another obvious possibility is that these tweets are in fact composed solely of individuals. This can perhaps be explained by the fact that the FDA highly discouraged Pax Labs from promoting their product on social media. This would explain a lack of corporations tweeting about it. However, this would not explain why more

organizations such as medical societies or news groups were not represented as much. Since data collection was over a three month span, it is possible that they did not tweet regarding this topic. So, it will be assumed that this data set is composed only of individuals moving forward in the procedure. Since Text Normalization has an entire dedicated chapter, I will briefly skip this step to discuss the step involving the Face Plus Plus API. Face Plus Plus analyzes “face related attributes, including age, gender, emotion, head pose, eye status, ethnicity, face image quality and blurriness” (Plus, 2019).

6.1.3 Profile Picture Changing Discussion

Going back to Figure 6.2 where the profile image url his highlighted, Python captures each of these images, re sizes them to a visual format, and stores them in a folder which is then sent to the Face Plus Plus API, collectively. If there is an error in the collection of the image, an error is logged and the program continues executing. The folder was filled with approximately 17,000 pictures. I spent a while contemplating why this might be that only 17,000 out of 32,399 images were collected. When I would follow the error, I was given a message of an invalid picture. Therefore, I took a few sample user names from the tweet images that logged an error and went to their Twitter page. Indeed, they did have a current picture of themselves on their profile pick. I then compared the url of their current picture to the picture returned to me in the data and found these were different in each case. I then researched whether or not Twitter has a way of storing previous profile pictures and it was found that they do not in fact keep track of previous profile pictures. Therefore, by default, if a user changes his or her profile picture before I am able to collect their images, I will be thrown an error. There was about a 1-2 month gap between when I collected the data and when I stored the images. Therefore, it is reasonable to assume that in a two month period, about 40 percent of users have changed their profile pictures. 40 percent is about the amount of data that was lost due to this issue. If I have time to rerun this test, I will instantaneously collect the images after having collected the next set of twitter data to first

of all prove my hypothesis as to what went wrong and to hopefully get a larger proportion of pictures.

6.1.4 Face Plus Plus API facial features detection

I then sent the collection of images to Face Plus Plus which returns a JSON object of facial attributes. The attributes that I collected which were of most interest to me is the age and gender of each individual. Out of all of the images sent, a total of 11,984 returned with an age and a gender. This was more of an expected drop off because not everybody who has a twitter profile image has their face as the user profile. I scrolled through the images and saw many logos, cat images, pictures of more than one person, and pictures where there is a person but their face is not clearly seen. All told, 12,000 out of 17,000 seems like a reasonable drop off. So at this moment, my data collection consists of 12,000 tweets and I also have the age and gender associated with each tweet. Before the data can be sent to different algorithms, it must first be pre-processed.

CHAPTER 7

DATA PROCESSING

7.1 Data Processing

Here, I will talk about data processing and normalization in a series of steps. Before doing so, let us first quickly explore the possibility of bot nets and whether or not it is likely to greatly affect the data.

7.1.1 Discussion of Bot Detection

According to Galvin, “researchers estimate there are tens of millions of bots automating accounts sometimes posing as real people on Twitter, with their presence also felt on Facebook and other social media platforms” (Galvin, 2018). It talks about the danger of bots if “harnessed to promote certain products- such as e-cigarettes, diet pills and supplement” (Galvin, 2018). An inflation in numbers can potentially give people a false understanding of reality which therefore can influence their decision making. A study was done using feature rankers on a full dataset which determined an “individual’s likelihood to interact with a social bot” (Wald et al., 2013). It was found that the “klout score and number of friends are the strongest predictors of whether or not an individual will interact with a social bot” (Wald et al., 2013). Klout score is a metric that determines one’s overall social influence on respective social media accounts. Currently, social bots are difficult to spot. However, research is continuously being done on ways to predict this. For example, there is an article that shows one example of an algorithm to begin classifying different fake accounts. The algorithm is an iterative search algorithm with dozens of fake Twitter accounts previously identified. The program iterates over friends and followers of each of these fake users, looking

for other “accounts displaying similar traits (e.g., similar description, including an URL to a sex-website called “Dirty Tinder”)” (van der Laken, 2018). Due to time constraints, there was not enough time to implement a bot check on my data. If I can duplicate the process, I would include a type of bot checking implementation. However, according to a CNBC article which cites a University of Southern California and Indiana University, up to “15 percent of Twitter accounts are in fact bots rather than people” (Michael, 2017). From what I understand on previous research, 10-15 percent is about the consensus number. So overall, worst case scenario, no more than 15 percent of the data has is likely to be misleading.

7.1.2 What is Natural Language Processing and How Do I Plan to Use It?

Natural Language Processing plays heavily to my ability to process such a large data set. Natural language processing (NLP) is “an area in machine learning and artificial intelligence that deals with understanding the meaning of human-style language and its interactions with machines” (nlp, 2019). This is helpful in that it allows me to draw accurate conclusions without having to individually parse through each document by hand. Think of it in terms of a certain metaphor. There has been ongoing debate as to which technique is more useful in treating depression whether it be cognitive therapy or medication. Both methods, however, are very prevalent in society and both have had their share of successes. Many have experienced success by using both. I feel the same with NLP and manually perusing the tweets. For instance, consider a person trying to find and reach a destination. Sign posts point you in the right direction and after that, you carefully consider the geography around you to reach your destination. Similarly, NLP points me in the right direction as to what are the major trends in the data, and then human inference is used in contextualizing the evidence. NLP also not only gives ways to understanding the data but also to normalize it.

7.1.3 Normalization Step 1: Tokenization, Stemming, and Lemmatizing

The first step in normalizing the data involved tokenizing the text which means parsing the data into tiny parts. Upon these little pieces I converted them all to lowercase,

removed all stop words, stemmed and lemmatized each word, and lastly applied my own personal stop words based upon the context of the twitter data. Stop words are commonly used words like “the”, “a”, “an” which do not provide much meaning for a sentence. The goal of both stemming and lemmatization is to “reduce inflectional forms and sometimes derivationally related forms of a word to a common base form” (Standford.edu, 2009). This allows you take words such as “am”, “are”, and “is” and convert them to “be”, or “car”, “cars”, “car’s”, and “cars” to car. This will allow such a sentence as “the boy’s cars are different colors” to translated as “the boy car be differ color” (Standford.edu, 2009). This is important for the maximum effectiveness of other algorithms used in NLP.

7.1.4 Normalization Step 2: Remove “HTTPS” and “@”

Because there are a lot of links in the tweets, I removed all lines within a sentence with “https” in it as well as words that start with @.

7.1.5 Normalization Step 3: Pros and Cons of Removing Retweeted Tweets

I removed re-tweeted tweets, because popular user names tend to choke out other relevant words. This is where some experimentation is involved in cleaning up the tweets. One of the complicated aspects of normalizing the tweets are the sheer number of re-tweets which give certain tweets much greater weight than a whole host of others. To experiment with what I would find, I have tried removing certain very popular tweets that did not provide a whole lot of meaning or provided a lot of meaning. The benefit of leaving the re-tweets is that one will see the popular opinion of the many who deemed a tweet agreeable enough to re-tweet. On the other hand, these tend to choke out more relevant and diverse tweets. For example, a guy got 4.1 K re tweets on this post: “If u shotgun a four loko then rip a juul in a bathroom a frat boy will appear in the mirror.” The reason I found this tweet, is that I will systematically search for words that seem oddly popular in my results. For instance, I searched this super popular but odd word and this user and thus was able to take out this tweet which made words like shot gun and frat so dominantly popular. Thus

I did two types of normalization in addition to the previous steps. A summary of some of the left out re-tweets and how they would have affected the data if left in the Methodology Limitation section.

7.1.6 Normalization Step 4: Remove Re-tweeted Tweets and Create Another Personal Stop Word List

I removed all re-tweeted tweets. Lastly, I made a list of personal stop words to take out other irrelevant terms in the data. I determined whether or not a word was relevant by doing a find-all search in the CSV file that contained all of my consolidated data. I then scanned through many of the tweets which contains a specific word and comparing it to the context of the problem. For instance, I would remove a word such as “think” since it is a word universally used in conversations. I also would remove words that were lemmatized and stemmed to the point of being hard to understand. For instance “bacco” was a word that I found as a popular word. In this case, after further research I realized that this stood for tobacco but there were other words more difficult to grasp. Some words were associated with names that I had to remove as well.

7.1.7 Normalization Step 5: Data Consolidation

After pre-processing all of the data, I then combined the important information from my two Age, Gender, and Tweet JSON objects into one main JSON object. This JSON object was then transformed into a CSV file (like an excel file) which is easier to read and sort.

7.1.8 Normalization Step 6: Data Categorization

Lastly, I categorized the data according to five overall age groups and then analyzed each age group as a whole and split it up into male and female. This resulted in a total of 15 subcategories derived from the five original categories. These categories are under the age of nineteen (smaller than 19), nineteen to twenty four (19-24), twenty five to thirty five (25-35), thirty five to fifty five (35-55) and fifty five and above (greater than 55). For each

category there is a distinction of gender. For instance there are two other subcategories of (19-24), being “19-24 Female” and “19-24 Male”. I chose these age groups, because they represent key groups in the population without becoming too specific. The results of this grouping are shown in Figure 10.2 in the results section of chapter 10.

7.1.9 Normalization Step 7: Data Sent to LDA Topic Modeling and Association Mining Rules

From here, a list of each category is sequentially sent through the Latent Dirichlet Allocation (LDA) Topic Modeling algorithm and Association Rule Mining via Apriori Algorithm to return to me the most popular topics as well as specific rules for each age and gender category. The next two chapters are dedicated to explaining how LDA topic modeling and the Apriori Algorithm work.

CHAPTER 8

LDA TOPIC MODELING

8.1 LDA Topic Modeling

8.1.1 Example of Processed Tweets

Before describing the LDA Topic Modeling Algorithm, it is useful to understand the models on which LDA works.

1. Document 1 contains:

- ‘need’, ‘savori’, ‘juul’, ‘podschicken’, ‘beer’

2. Document 2 contains:

- ‘need’, ‘loan’, ‘juul’, ‘haha’

When all of the tweets are pre-processed, a list of documents are returned with each document itself being a list of the important words from the document. A document in this case is a tweet and what we have is a list of lists. Above is an example of what two documents out of many may contain.

8.1.2 Bag of Words Data Set is Created

Next a Bag of Words Data set is used to create a dictionary which “contains the number of times a word appears in the training set” (Li, 05/30/2018). This means that every unique word found in the pre-processed text is added together to make a dictionary and it keeps track of how many times each word is used out of the entire collection of documents. In the below table is an example of the frequency for some of the words across the documents.

Dictionary	
Frequency	Word
296	sale
2604	match
2070	american
1266	daili
1421	neck
3698	never
1596	ticket
121	podcast
2397	pleaseau
3766	chickfila
658	creepi

Table 8.1: Example Dictionary

8.1.3 Filter Out Extreme Extreme Frequency of Tokens

In order to filter out different extremes of tokens in my dictionary, I filtered out tokens that appeared in less than 15 tweet documents or more than 50 percent of the tweet documents. After these two initial steps, I kept the first 100,000 most frequent tokens.

8.1.4 The Bow Corpus

Next, for each tweet document I created a “dictionary reporting how many words and how many times those words appear” (Li, 2018). I will call this the “Bow Corpus”. For instance, document 43 of my corpus has the following dictionary.

$[(0,1), (47,1), (48,1), (49,1), (50,1),(51,1)]$

This translates to:

Word 0 (“juul”) appears 1 time.

Word 47 (“bathroom”) appears 1 time.

Word 48 (“class”) appears 1 time.

Word 49 (“go”) appears 1 time.

Word 50 (“kill”) appears 1 time.

Word 51 (“rip”) appears 1 time.

The reason these words are in order probably has to do with the fact that document 43 is early on in the corpus and this must be the first time most of these words are found so they are added sequentially to the overall dictionary.

8.1.5 TF-IDF Model

Another method that I used was creating a TF-IDF model using a python package called “models.TfidfModel” on the “Bow Corpus“ and applying this model to the entire corpus thus transforming my model into a TF-IDF model. The reason that I used TF-IDF is that there is a weakness with scoring word frequency. That is that “highly frequent words start to dominate in the document (e.g. larger score), but may not contain as much ”informational content” to the model as rarer but perhaps domain specific words” (Brownlee, 2017). So using TF-IDF will give a higher weight to words which are inherently more meaningful to a sentence. For instance if I said that ”The girl over there is ...” The final word of the sentence is critical to the type of sentence. For instance using the word “ugly” would be drastically different from “beautiful”. Therefore, there should be a lot of weight to certain words as apposed to others. The formula for the TF-IDF Model is:

$$TFIDF = TF * IDF$$

where TF = Term Frequency and IDF = Inverse Document Frequency. The formula for Term Frequency is

$$(NumOccurrenceWordDictionary/NumWordsDocument)$$

And Inverse Document Frequency is equal to

$$\log((NumDocuments)/(NumDocumentsContainedWord))$$

In plain English, term frequency “is a scoring of the frequency of the word in the current document” and inverse document frequency is a “scoring of how rare the word is across documents” (Brownlee, 10/09/2017). When you multiply TF and IDF together for each word you get a score for that word within each document. Below is a showing of the TF-IDF for the first couple of documents.

```
[(0, 0.02350542239809432), (7, 0.6909149494194787), (25, 0.722553823453012)]  
[(15, 0.4124545647071117), (27, 0.6874248944055067), (28, 0.5977693924950028)]  
[(0, 0.02340315320126786), (29, 0.6583025767191054), (30, 0.7523895333570398)]  
[(0, 0.03762572089125441), (31, 0.9992919018622203)]  
[(0, 0.023831421628203318), (32, 0.7661579635819065), (33, 0.6422102756755024)]
```

Figure 8.1: TF-IDF

For each word in each document, there is a tuple that contains which word in the dictionary that the word belongs to and its TF-IDF score. Thus, I use two models, Bag of Words and TF-IDF, on which to run the LDA model. Next, I ran LDA using the Bag of Words model.

8.1.6 How Does LDA Topic Modeling Return?

As stated in the data science article “Topic Modeling and Latent Dirichlet Allocation (LDA) in Python,” topic modeling is a type of statistical algorithm used for “discovering the abstract ‘topics’ that occur in a collection of documents” (Li, 05/30/2018). LDA specifically is a type of topic model that is “used to classify text in a document to a particular topic” and it builds a “topic per document model and words per topic model” (Li, 05/30/2018). To give a very simple example of LDA in action, imagine four sentences about Juul.

1. I **lost** my Juul today and I am **heartbroken**, because it helps me not **need** cigarettes.
2. The newest pod **flavor** is **mango** and it is my favorite.
3. There is an **epidemic** among youth as children become **addicted** to **nicotine** in increasingly high numbers as a result of Juul use.
4. When I was taking a hit of my Juul, I **misplaced** it and am very sad because it was **mint** mint flavored.

LDA might derive three topics from these four sentences. Topic L might be labeled “losing” by LDA from the red words, topic F might be labeled “flavor” by LDA from the blue words, and topic E might be labeled “epidemic” from LDA by the green words. LDA “defines each topic as a bag of words”, so it is the user’s responsibility to label topics as they deem fit (Algobeans, 05/30/2018). I will explain things such as the Bag of Words Model and TF-IDF momentarily after explaining how LDA works. LDA allows one to do two things, namely in can “infer the content spread of each sentence by word count” (Algobeans, 05/30/2018). This means that the four sentences could be labeled as the following.

Sentence 1: 70 percent topic L and 30 percent Topic E

Sentence 2: 100 percent topic F

Sentence 3: 100 percent Topic E

Sentence 4: 50 percent Topic L and 50 percent Topic F

This shows that a sentence can be comprised of more than one topic. Secondly, one can “derive the proportions that each word constitutes in given topics” (Algobeans, 05/30/2018). For example topic F may be comprised of words in the following proportions: 40 percent flavor, 20 percent mango, 20 percent mint, and so on. The LDA achieves its results in three steps. Step 1: Telling the algorithm how many topics are expected, Step 2: involves assigning “every word to a temporary topic”, and topic 3: where the algorithm checks and updates the topic assignments (Algobeans, 05/30/2018).

What is LDA Topic Modeling? LDA or latent Dirichlet allocation is a “generative probabilistic model” of “a collection of composites made up of parts” (Lettier, 2/23/2018). In this context, a tweet document is a composite and the parts are made of words. The model consists of two matrices or tables worth of information. The first table “describes the probability or chance of selecting a particular part when sampling a particular topic (category)” (Lettier, 2/23/2018). This means, what is the probability of finding a specific word when exploring a specific category. For instance, what is the probability that I will find the word ”mango” when I sample the topic “flavor”? The second table “describes the chance

of selecting a particular topic when sampling a particular document or composite” (Lettier, 2/23/2018). This means, what is the probability that a document or tweet will be part of a specific category. For instance, what is the likelihood that a specific tweet will be in the flavor category? How does LDA come to this conclusion? First I would like to use a metaphor that I was inspired by based off of an article called ”Introduction to Latent Dirichlet Allocation.”

8.1.7 How Does LDA Topic Modeling Work?

Pretend that I move to a new city and being a Nintendo Mario Kart fan as well as a philosophy major, I want to know where these type of people hang out. Because I am shy I am not going to outright ask. So I go to a bunch of different establishments across town and I make note of the people hanging out at each of them. In this case, think of the establishments as a different document (tweet), each person as a word, and a typical interest group as a topic. For instance, I go to a coffee shop and I meet Mario, Wario, Luigi, and Peach. I go to Toy’s R Us and I meet Bowser and Daisy, but I also see Mario there from the coffee shop. I then go to the park and I meet the Yoshi Turtles but I also see Peach there from the coffee shop. I do not know the typical interest groups of any of the establishments. So I pick some number K categories to learn the K most important kinds of categories people fall into. In this case, I decide four. For example, I initially guess that Mario was at the coffee shop because people who are at a coffee shop like to eat. I guess that people at the park all like to walk outdoors and I guess that people who go to Toy’s R Us have a fascination with video games. I do this on and on. Because these are off the cuff and random guess, I am likely to be incorrect. So I want to improve my labeling accuracy. I do this by picking a place and a person. So I go to the coffee shop and meet Mario. As I am talking to him, I notice that he is holding a book in his hand and so are several other people. So I guess that the thing Mario has in common with the rest of the people in the coffee shop is that they are all holding books. In other words, “the more people with interests in X there are at the coffee shop and the stronger” Mario “is associated with interest X (at all the

other places” he “goes to), the more likely it is that” Mario is at “the coffee shop “because of interest X” (Chan, 8/22/2011). As I am in the park, I notice Mario walking by and he is holding a book again with another friend. I then make a new guess that Mario goes where he goes according to whether or not it is a peaceful place to read. I use this information to further classify Mario and use what I know of him to make better guesses about the people around him and so on. All of my guesses are based upon my probability of being correct. In a similar way LDA does the following. It randomly assigns each “word in the document to one of the K topics” (Chan, 8/22/2011). This brings me to step 1.

1. In step 1, I tell LDA how many topics I believe are in the set. In my case, since I did not find an example in literature which did the same experiment, I had to use trial and error. I tried several topic numbers until I reached a desired level of interpret ability. In my case, I would try four topics, then 6 topics, then 10 topics, then 15 topics. In each case, I would examine the topics and see what patterns are revealed. If I felt that I was beginning to see the same topics re-appear over and over, I would know to begin removing the topics. When a topic that I found interesting suddenly disappeared from the results, I new to increase the topic number until I found a good balance between redundancy and disappearance of topics. I ultimately found that expecting three topics provided this balance.
2. Step 2 involves assigning “every word to a temporary topic” (Algobeans, 05/30/2018). This means that topics are assigned to each word according to a Dirichlet distribution. Above is an example of a Dirichlet Distribution that is taken from an article that I referenced. The distribution itself is 3D and every dot in the distribution represents some mixture of the three topics. In my case, a dot would represent a tweet or document and its location in the distribution is based on a value such as (.4, .3, .3) where each value correlates to a specific topic and all numbers add up to one. The distribution takes a number called alpha. A low alpha which is a value less than one one means that

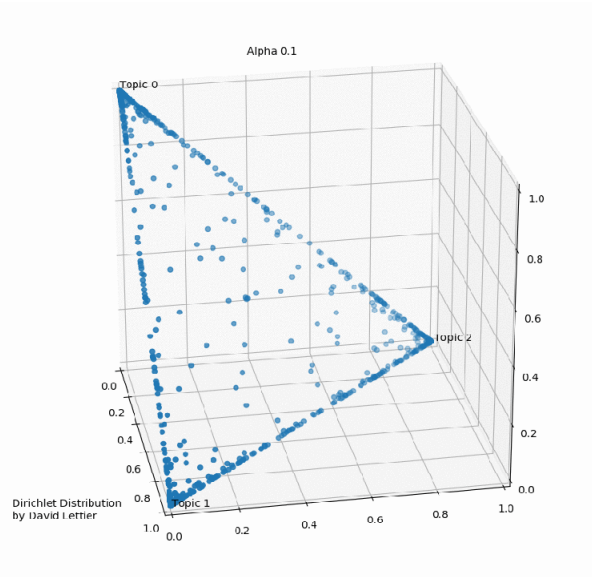


Figure 8.2: LDADistribution (Algobeans, 05/30/2018)

the dots will be more distributed towards the corners. This implies that you expect for your document to be categorized very distinctly. For instance, a low alpha would imply that I expect to see topics that are distinct rather than tweets which might contain more than one topic. An alpha of one means that the dots are uniformly distributed. I chose this alpha as a default since it seems a bit presumptuous to do otherwise for the data. And of course an alpha greater than one would mean that I expect the topics to be mixed. So in essence, the “alpha controls the mixture of topics for any given document” (Chan, 8/22/2011).

3. Lastly, step 3 involves the algorithm checking and updating topic assignments. For each word in every document, a word is updated based upon the following criteria. “How prevalent is that word across topics and how prevalent are topics in the document” (Algobeans, 05/30/2018)? The next algorithm to be discussed is Associative Rule Mining.

CHAPTER 9

ASSOCIATION RULE MINING

9.1 Association Rule Mining

As talked about in “Association Rule Mining via Apriori Algorithm in Python,” association rule mining is a “technique to identify underlying relations between different items” (Malik, 8/09/2018). Let us think of specific use cases. For instance, college students consume a variety of different products. For one they may buy beer, burgers, and baseball tickets. Kids might buy toys and fast food. In marketing, firms are constantly looking for trends in customer behavior so that they can better predict how to present their other items to them. In a similar way, I would like to find the underlying relations between Juul and different items. This would mean thinking about the different topics of conversation. For instance, when people mention Juul, do they mention flavors? Do they talk about addiction? Do mothers associate Juul with a teen epidemic? All of these may be potential rules or underlying associations. The profit of finding such associations may add to the conversation regarding the FDA. For instance, does the FDA truly know how to go about trying to get kids to stop Juuling? Or if they knew some of the true incentives underlying the teen epidemic, they might better be able to come up with a protocol that better discourages or takes away the incentives of Juuling.

Before one can understand the Association Rule Mining, one must be able to understand the Apriori Algorithm and how it relates to my research. The Apriori Algorithm consists of three major components: support, confidence, and lift.

As a use case to explore these three topics, let us suppose that there are 1000 tweets collected from the API and that they have already been pre-processed as described before-

hand. Now let us explore if we wanted to find the Support, Confidence, and Lift for the two words being “Juul” and “Flavor”. Let’s suppose that out of the 1000 tweets, 100 tweets contained the word flavor while 150 tweets contained the word Juul. Out of the 150 tweets where the word Juul was found, 50 of those same tweets contained the word flavor.

9.1.1 Support

Support refers to the “default popularity of an item and can be calculated by finding number” of tweets containing a particular word “divided by total number of transactions”, where each tweet is a transaction (Malik, 8/09/2018). The formula looks as such.

$$Support(B) = (TweetsContaining(B))/(TotalTweets)$$

Thus, if we are looking at the 1000 tweets, and 100 tweets contain the word “flavor”, then the support of the word “flavor” is calculated as:

$$Support(Flavor) = (TweetsContainingFlavor)/TotalTweets)$$

$$Support(Flavor) = \frac{100}{1000} = 10\%$$

9.1.2 Confidence

Confidence refers to the likelihood that a word B is also in a tweet if A is in that tweet. It is calculated by finding the number of tweets where words A and B are in it together divided by the total number of tweets where A is present. This can be represented as follows:

$$Confidence(A \rightarrow B) = (TweetsContainingboth(AandB))/(TweetsContainingA)$$

So if there are 50 tweets where “Juul” and “Flavor” are both present then one can find the likelihood that “Flavor” will be found in a tweet that already contains the word

“Juul” as the confidence of association rule “Juul” \rightarrow “Flavor”. In other words, given A what is your confidence that B will also be present. Mathematically this would look like:

$$\text{Confidence}(\text{“Juul”} \rightarrow \text{“Flavor”}) = (\text{ContainBoth}(\text{“Juul” and “Flavor”})) / (\text{Contains}(\text{“Juul”}))$$

$$\text{Confidence}(\text{“Juul”} \rightarrow \text{“Flavor”}) = 50/150 = 33.3\%$$

9.1.3 Lift

Lift (A \rightarrow B) refers to the “increase in the ratio” of the presence of B when A is present. The Lift(A \rightarrow B) can be calculated as dividing the Confidence(A \rightarrow B) by Support(B). Mathematically, this looks like:

$$\text{Lift}(\text{“Juul”} \rightarrow \text{“Flavor”}) = (\text{Confidence}(\text{“Juul”} \rightarrow \text{“Flavor”})) / (\text{Support}(\text{“Flavor”}))$$

$$\text{Lift}(\text{“Juul”} \rightarrow \text{“Flavor”}) = 33.3/10 = 3.33\%$$

Thus, the likelihood that “Juul” and “Flavor” will be in a tweet together is 3.33%

Each of the 11,000 tweets is composed of roughly 4-15 words. The Apriori algorithm “tries to extract rules for each possible combination of words” (Malik, 8/09/2018). Since there are so many possible combinations, it is important to set a specific threshold for the support and confidence. In other words, I am only interested in rules that have a minimum of a specific support and confidence. For instance, the threshold that I set for my association rules varied from .003 to .005. I did the calculation and this is a reasonable setting for the minimum support of my association rules. By “reasonable”, I mean a number not so high that no word rules appear and not so low that so many appear that it becomes nonsensical. Out of 11,000 tweets or documents, if 44 tweets are found that contain a specific word, this

would mean that the word is relatively popular but not so popular that it is most likely a word devoid of meaning. My minimum confidence was set to .2 meaning that there needs to be at least a 20 percent confidence in the presence of word B, given the presence of word A. Likewise, the likelihood that the ratio at which word A goes up if word B is present is at least by a factor of 3. Lastly, I set the minimum length to two which means that there must be at least two words in my rules. In pre-processing the data, there is an extra step needed for the Association Rule Mining. The twitter data needed to be formatted in table in such a way that each row represents an individual tweet and each column is represented by a word. The Apriori algorithm is then applied to the data set.

CHAPTER 10

RESULTS

10.0.1 Most Common Hash-tags From My Data

The following section contains every figure and graph which collectively comprise the results. A discussion of these results will take place in the next chapter.

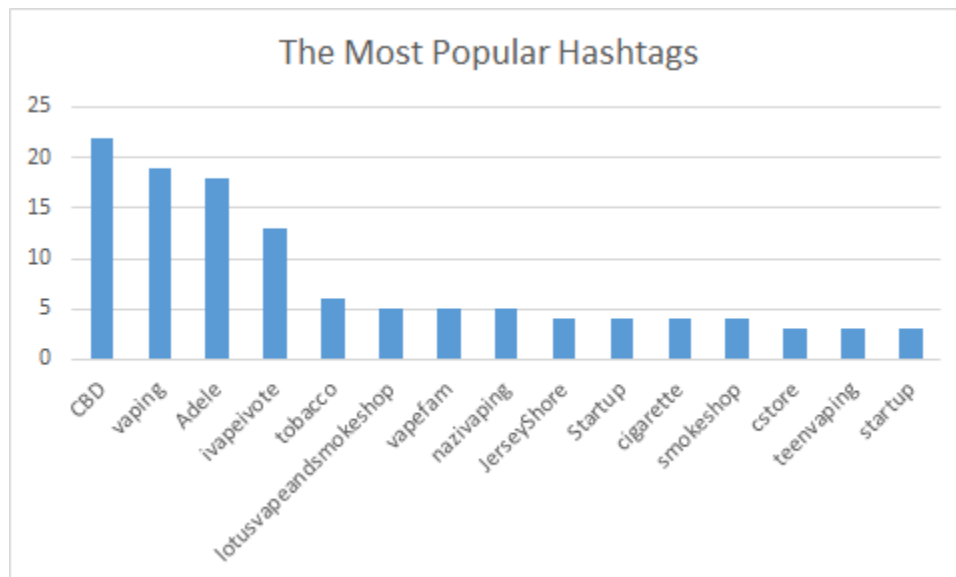


Figure 10.1: HashtagFrequency

10.0.2 Age and Gender Group Categories Results

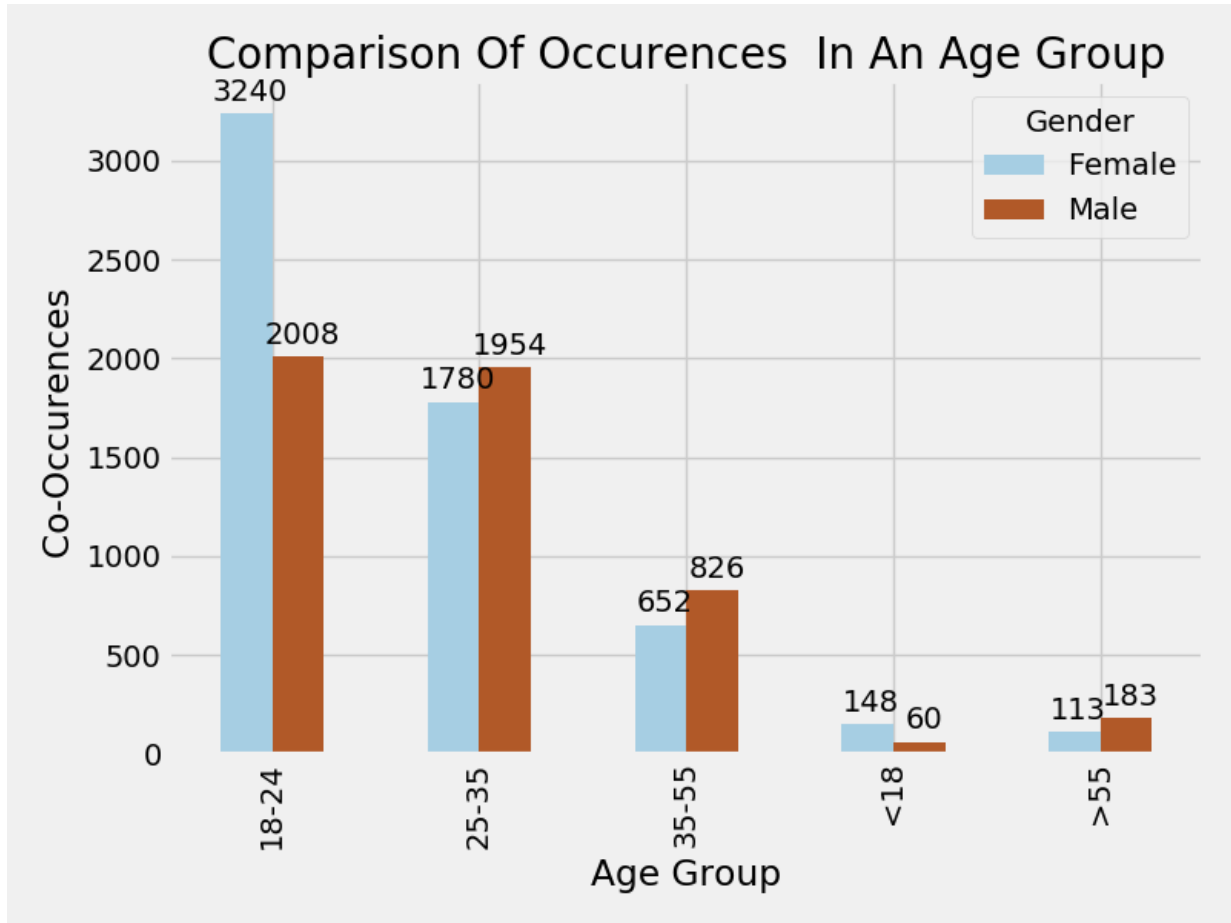


Figure 10.2: Data Collection

10.0.3 LDA Topic Modeling

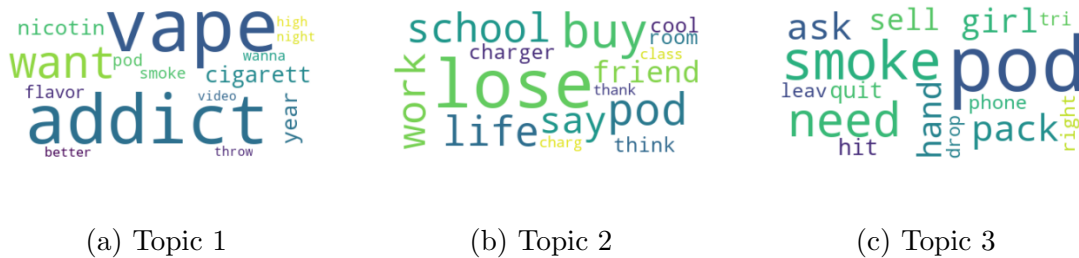


Figure 10.3: 18 and Under TFIDF Model

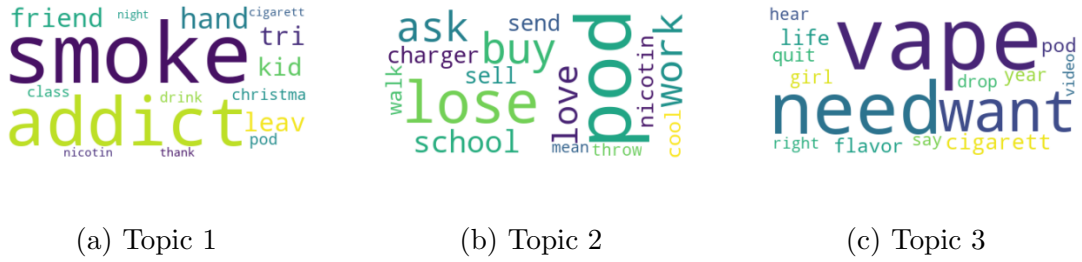


Figure 10.4: 18 and Under TFIDF Model Female

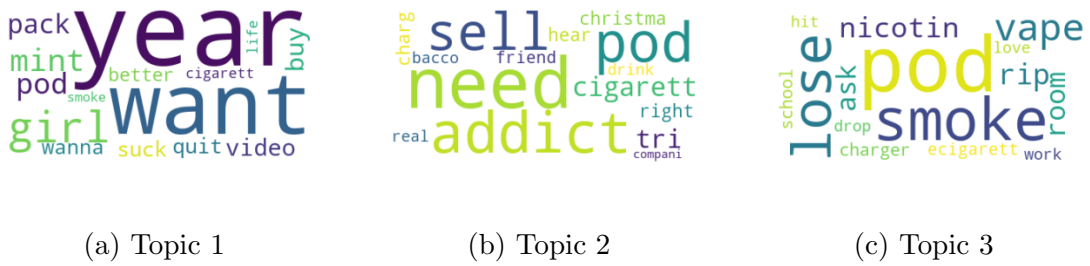


Figure 10.5: 18 and Under TFIDF Model Male

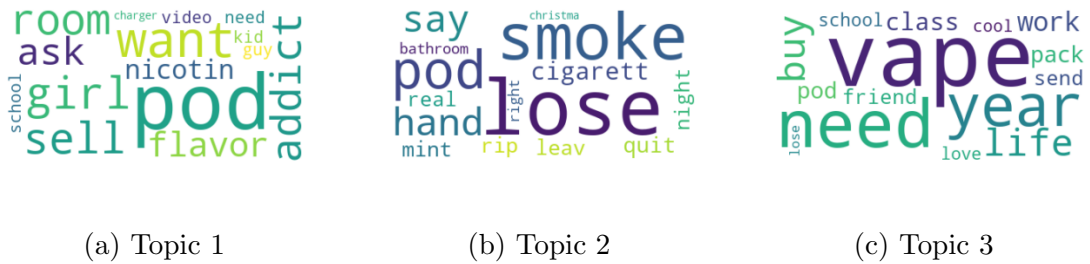


Figure 10.6: 19-24 TFIDF Model



Figure 10.7: 19-24 TFIDF Model Female

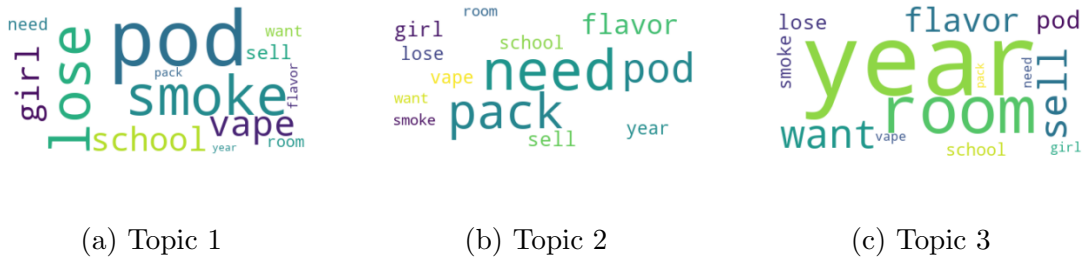


Figure 10.8: 19-24 TFIDF Model Male

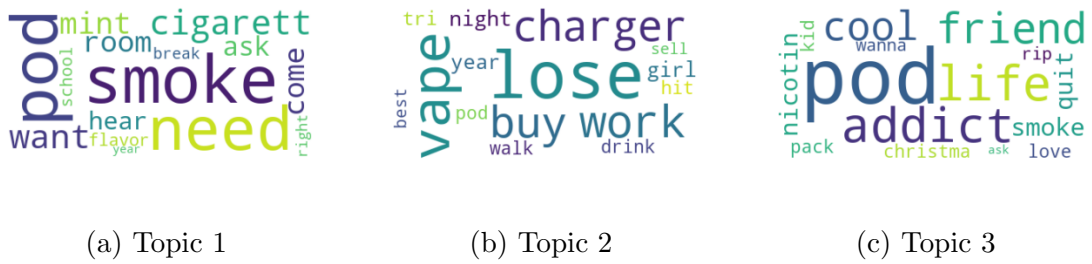


Figure 10.9: 25-35 TFIDF Model

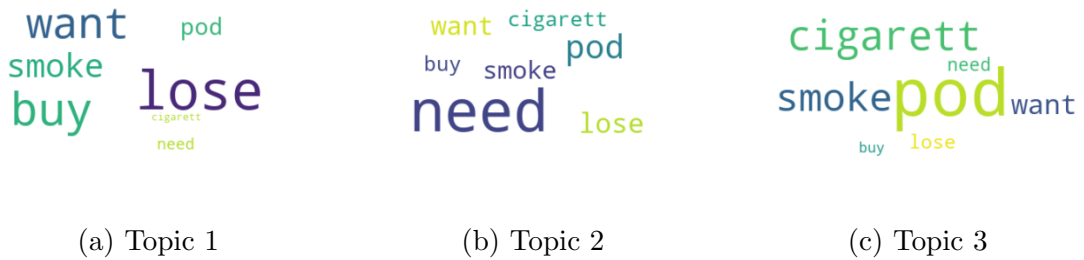


Figure 10.10: 25-35 TFIDF Model Female

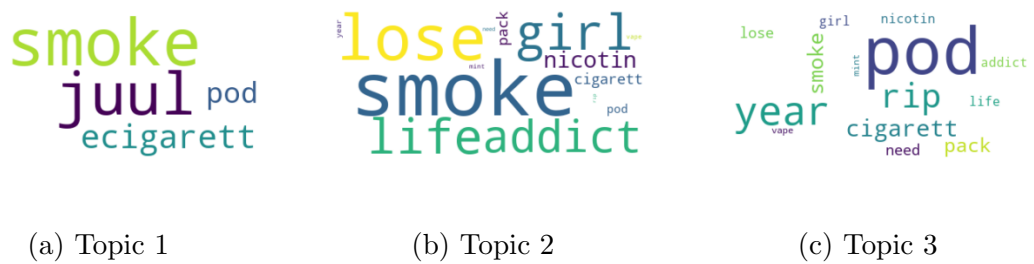


Figure 10.11: 25-35 TFIDF Model Male

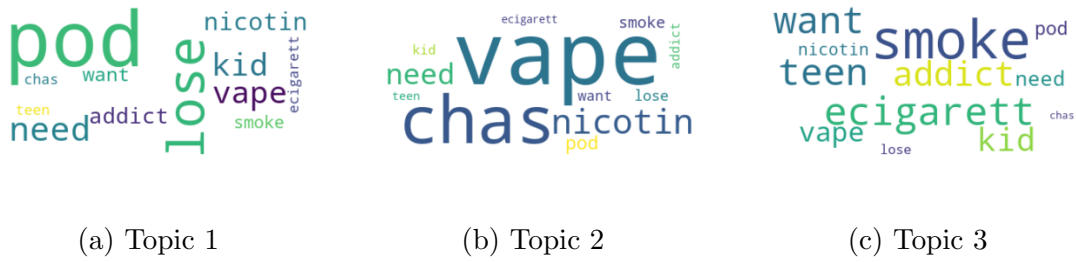


Figure 10.12: 35-55 TFIDF Model

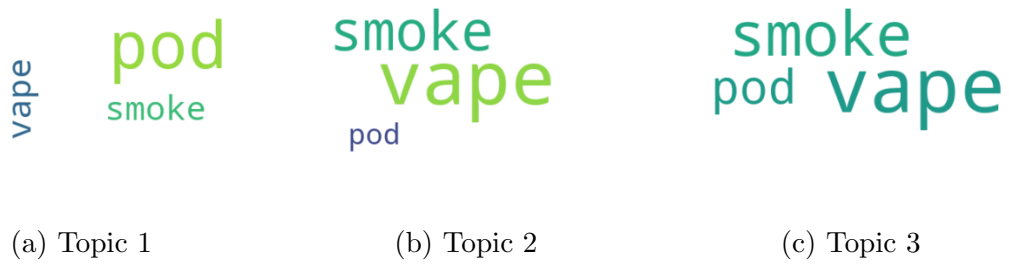


Figure 10.13: 35-55 TFIDF Model Female



Figure 10.14: 35-55 TFIDF Model Male

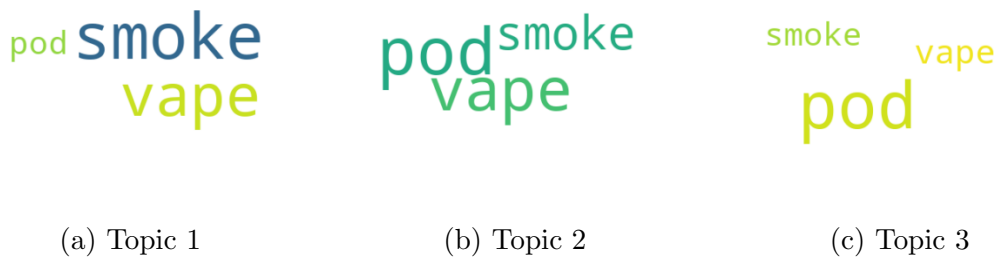


Figure 10.15: Greater than 55 TFIDF Model

pod pod pod

(a) Topic 1

(b) Topic 2

(c) Topic 3

Figure 10.16: Greater than 55 TFIDF Model Female

10.0.4 Association Rule Mining

Association Rules 18 and Under			
Rule	Support	Confidence	Lift
juul - > cigarette	.006	0.381	12.590
juul - > pack	0.004	0.579	6.050
Association Rules 18 and Under Notable Female			
sell - > pod	0.005	0.500	4.869
juul - > nicotine	0.004	0.385	33.709
juul - > lose	0.005	0.500	9.438
christma - > juul	0.004	0.500	4.869
juul - > sell	0.004	0.545	5.311
Association Rules 19-24			
christma - > pod	0.005	0.500	3.984
mint - > pod	0.004	0.412	3.282
pack - > pod	0.011	0.818	6.520
juul - > nicotine	0.005	0.348	21.955
christma - > juul	0.005	0.533	4.251
juul - > cigarette	0.006	0.323	8.483
juul - > flavor	0.006	0.391	3.118

mint - > juul	0.004	0.412	3.281
juul - > pack	0.011	0.818	6.520
juul - > sell	0.008	0.538	429
Association Rules 19-24 Notable Female			
juul - > school	0.005	0.444	3.43
need - > pod	0.012	0.417	3.13
Association Rules 19-24 Notable Male			
juul - > pod	0.004	0.6	29.480
Association Rules 25-35			
react - > elder	0.004	1.0	231.75
juul - > nicotine	0.004	.333	30.900
juul - > cigarett	0.008	.412	14.137
vape - > juul	0.004	1.0	57.947
juul - > room	0.004	0.666	56.182
Association Rules 25-35 Notable Female			
night - > lose	0.005	0.333	7.235
juul - > anymor	0.005	0.5	13.178
Association Rules 35-55			
vape - > desk	0.005	1.0	29.4
juul - > flavor	0.005	1.0	3.243
mint - > juul	0.005	1.0	3.243
juul - > quit	0.007	1.0	3.243

CHAPTER 11

DISCUSSION OF RESULTS

Discussion of results is divided similarly into how the results are displayed. Starting with the figure for Hash-tag Frequency.

11.0.1 Hash-tag Frequency

Referencing Figure 10.1, one thing I was surprised by is that there were not very many hash-tags used compared to the number of tweets. Out of the 30,000 tweets, only 280 returned an actual hash-tag. Either there was an error in my method of collecting the hash-tags or people simply do not use as many hash-tags as I thought. However, the hash-tags themselves I find somewhat insightful. For instance, “CBD”, the second most popular hash-tag with 22 mentions as well as “cbdshop” stands for cannabidiol, “a compound found in weed or marijuana plants that relieves pain, reduces anxiety, helps sleep, etc.” (Shiwnarain, 5/01/2018). I also did some research and found that there are CBD Juul pods which brings a whole new aspect of Juul use to light. That is, using a Juul can easily be associated with trying to reduce some form of anxiety. Additionally, the Juul may potentially be linked to marijuana use. Then of course there are some expected hash-tags in that there are variations of the topic. Words such as “e-cigarettes”, “nicotine”, “tobacco”, and “vape” are seen. Since a Juul is a type of e-cigarette and one can vape using the Juul, it would make sense to see these hash-tags associated with Juul. Associating Juul with nicotine and tobacco is interesting in the fact that nicotine addiction among youth is a popular topic found in the media. Referencing earlier chapters, there is also disagreement in the medical community as to their willingness to say that a Juul is a much healthier alternative to traditional tobacco products such as cigarettes. One also sees hash-tags related to shops meaning that there

most likely is some remnants of promotion whether directly by the shop itself or a consumer of their product. This also speaks to the fact mentioned before that vaping communities do indeed spend a lot of time together in vape shops, especially when it has the extra features such as lounges, pool tables, etc.

11.0.2 Questions Generated from Hash-tag Frequency

Politics and country are briefly mentioned with hash-tags such as “America” and “ivapeivote”. Although not a huge presence, it does cause one to ponder the effects that constituent demand for Juul may have on the way politics function. Soon, Juul policy might need to be part of a candidates platform. Or perhaps there may be a correlation between political party and Juuling. Juuling may even be useful in galvanizing certain voters who otherwise would not be interested in policy. Then of course, there is the health sector as it relates to teen addiction and addiction in general. You see things such as “UNTobaccoControl”, “teenvaping”, “lunghealth”, “marketing”, “inspection”, and “teenhealth”. Again, there is large discourse in the idea of Juul intentionally marketing to youth. Different control centers, especially the FDA are very interested in bringing about Juuling cessation among youth. The topic of lung health can also be applied to the discourse regarding whether or not it is better to smoke or vape. For instance, tags like “smokefree”, “quitting”, “RealityCheck”, “BigLie”, “smokers”, “marketing”, and “quitsmoking” all relate to several key issues based on a similar topic. It is known that the Juul company has begun to re-brand their image to exclusively target people 25 and older, especially to help them to stop smoking. Many people, as can be seen from the tweets which will be talked about later, are frustrated with the emphasis put on the danger of the Juul when they consider it to be a healthy alternative to cigarettes. Many people are using the Juul to help them quit smoking. A few interesting things to look at in the future are things such as “NintendoSwitch”, “DevPro”, “Fortnite”, “SoundCloud”, and “JerseyShore” where they may be a link that clues us in on a high correlation with people who like to play video games, watch a lot of TV, listen to a lot

of music and Juul. There as of now is no evidence to suggest this claim besides the hash-tags themselves. This reminds me though of the research done that says Juuling is a great hobby for many people in that you can mix and match the mechanisms. So for some demographics, there is a draw to the technological aspect of the Juul. The last other hash-tag group refers to components of the Juul themselves such as “juulcartridge”.

11.0.3 Age Group Category

Referencing Figure 10.2, as expected, when grouping each age and gender demographic into different categories, the older the demographic the fewer the tweets with the exception of those less than 18. This would make sense, because one’s social media activity tends to be greatest when one gets to college. The 18-24 category had the highest number of tweets with 5,248 tweets altogether. I was surprised to find that in this category, there is significantly more female tweets than male tweets. However, among further research, it seems that females do tweet more than males. There is an interesting read called “Gender Divide: How Men and Women Use Twitter Differently” by COMPLEX. A few rules states here which may be helpful in interpreting some of the results involve the following.

1. “Women use Twitter more than men” (COMPLEX, 11/29/2013).
2. “Men are more likely to swear, and keep things short” (COMPLEX, 11/29/2013).
3. “Girls are more likely to express emotions” (COMPLEX, 11/29/2013).
4. “Women talk about each other” (COMPLEX, 11/29/2013).
5. “Men talk about themselves” (COMPLEX, 11/29/2013).
6. “Women are able to gain a larger following than men” (COMPLEX, 11/29/2013).
7. “But men aren’t as likely to follow other people in the first place” (COMPLEX, 11/29/2013).

This makes it interesting that outside of the most popular group, men seem to tweet slightly more than females. Perhaps men may be in certain age groups more likely to tweet about the Juul due to a possible stigma associated with Juuling that they may be more

willing to accept than women.

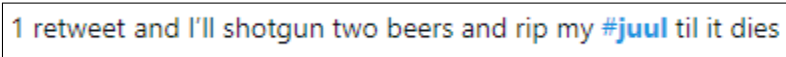
11.1 LDA Topic Modeling

Some interesting trends emerged from the LDA model. After doing a few test runs, I ended up deciding to display the results based upon the TFIDF model exclusively due to its advantages over the Bag of Words model. After a few test runs with LDA, it seemed that limiting the model to three topics produced results in which distinct topics were found with very little redundancy. However, as the age groups increase and consequently their twitter use declines three topics tends to be too much. However, in order to stay consistent with the majority of the model, I kept to three topics for each age demographic and the male and female categories within each age group. As a reminder, when I say "three topics" I mean that I tell the LDA algorithm to generate three topics in the result which I will then classify. Any time a quotation is used, it refers to a word within the WordCloud. Each tweet picture is from either the data set or a tweet that has very similar content to a tweet in the data set.

11.1.1 18 and Under TFIDF Model (Figure 10.3)

The theme of 18 and under year old tweeting patterns involves a duality between how the Juul relates to partying and pleasure versus the Juul's use in casual every day life as well as the Juul as it relates to the product itself. Topic 1 seems to correlate with partying and pleasure. I came to this conclusion based upon the connotation of the words. In this WordCloud, the words that immediately stand out as having a connection involve "addiction", "cigarette", "vape", "nicotine", "smoke", "flavor", "want", and "wanna". There are several ways to understand these particular cluster of words. For one, after perusing the different under 18 tweets personally, very little of the tweets were concerned with the Juul's use in smoking cessation. Based upon my research, this association could be highlighting the hypothesis that the Juul is a gateway drug to other tobacco products. Nicotine is the underlying attraction to both Juul's and cigarettes and plays a large part in attracting customers. However, there is also the partying aspect in which it is the pleasure of the

moment where the Juul is a useful tool in elevating the atmosphere with gadgets such as “party mode” in which the Juul will start flashing rainbow lights. The reason I think of the word “party” are the other words in the WordCloud. For instance, “video”, “high”, and “night” often reference things that happen in the dark. At a party, a video would be likely be seen filming an embarrassing moment for a friend or a party trick that they do to show the amount of fun that they are having. The word “high” may be referring to the high having to do with the smoking of other substances. I also infer these conclusions based upon my study of the tweets themselves. In studying the tweets, I found that the under 18 tweets often speak of those elements which are, at their base, associated with physical pleasure. In summary, topic 1 produces the imagery in which one sees the combination of words associated with night activities and the addictive nature of Juul in heightening an experience as other tobacco products might leading to an overall label of partying.



1 retweet and I'll shotgun two beers and rip my #juul til it dies

Figure 11.1: Partying Imagery Example

Topic 2 is associated with Juul’s use in every day life. For instance, words such as “school”, “friend”, “life”, “cool”, “work”, “thank”, “room”, and “charger” paints a unique image. Activities such as “school” and “work” are daily events in which people will often take Juuling breaks in specific “room” like a bathroom. Juuling breaks can be thought similar to smoking breaks at work except for the fact that Juuls are more subtle, and one can get away with vaping underneath their sleeve. The word “cool” is a casual one in which cool things are generally considered enjoyable and an acceptable part of society. For instance, it is cool to go and get a coffee with a friend. “Life” implies the arch encompassing these vaping breaks as well as the gadgets which are associated with the events. When one may “say” things about the “pod”, one is often discussing the flavors that a pod might contain. “Chargers” would be more likely used throughout the day, especially since a Juul can be charged on a laptop. People seem to for some reason lose their Juul all the time. Losing the

Juul is one of the most prevalent and popular topics in Juuling. This begs the question if something about the activity itself causes one to be forgetful or ,if the Juul being so small, is just easy to lose. However, we often do not hear much about people losing their USB drive which looks just like a Juul. In summary, Topic 2 represents casual life.

@sheetz ordered up on flavored #juul pods before they stopped making them available to retail. This way #sheetz can keep up pushing to high schoolers and make a buck. #badcustomerservice

Figure 11.2: Illegally selling Juul to High Schoolers

Topic 3 is associated with the Juul product itself and common discussion points. For instance, people like to talk about how one "pod" is equal to a "pack" of cigarettes. Juuling is often referenced when people talk about quitting smoking ("quit", "smoke"). As seen in topic 2, people will often talk about losing their juul ("lose"). As can be seen in different gender demographics, Juul's are often seen as a flirtation device where a "girl" might "ask" a cute guy for a "hit" of Juul. A guy might be interested in profit and discuss "sell"ing his Juul.

I'm not going to flat out judge anyone for using a #juul but if you consistently lose it and then make everyone you're with stop what they're doing and help you look for it I will then PROCEED TO JUDGE!

Figure 11.3: Discussing Those Who Do Juul

11.1.2 18 and Under TFIDF Model Female and Male (Figures 10.4 and 10.5)

Instead of going through each topic individually for male in female, I find it more useful to explore the main differences in the subtleties of the topics between male and females since the topics will be relatively the same and include the same themes of addiction and partying, everyday life, and the topic of Juul itself. Females seem to be interested in the social aspect of Juuling more so than males. For instance, words that stand out include "vaping", "pods", and "flavor" which are words that are seen more at a vape-shop where fellow Juuling

connoisseurs can discuss those elements related to their hobby. They are more likely to talk about “kid’s and “quit”ing which imply that women are more concerned with a child’s well being. I think of social interactions for women with words such as “hand” being more popular. You will see girls handing their Juul to someone else to try. Even more importantly, flirtation can be used to get what you want. So in these cases, Juuling among a peer group is a way of building deeper connections. Girls also mention the word “love” more often which is a possible association with the activity of dating and how it can be used as a bonding experience. Alternatively, “love” can also be used as a more useful description from a girl’s perspective to express her appreciation for the Juul itself.

Guys are more interested in the business side of the Juul with words like “company”, “year”, and “sell”. To talk about a company’s profitability require one looking back at the past sells and trends of the year and introspectively drawing conclusions from the data. A guy uses the word “sell” more often than the girl who uses ”buy” more often. This points the possibility that men are more likely to sell their own Juul or pods as a business whereas girls are more likely to be a customer. A guy’s “need” or “want” is shown more often. This need can be seen in their more popular words of “video”, “rip”, “suck”, and “hit”. Now these words are all very graphic representations of the act itself of inhaling from a Juul. So it would seem guys talk more about the physical act itself which produces the sensation whereas a girl would more likely talk about the components surrounding a Juul which make it a socially connective piece. Additionally, the word video is a word found with other words associated with partying. This would also imply, that we would see a guy more likely to use a Juul in a party like atmosphere. Lastly guys are more likely to talk about girls in the context of the Juul rather than a girl talk about a guy.

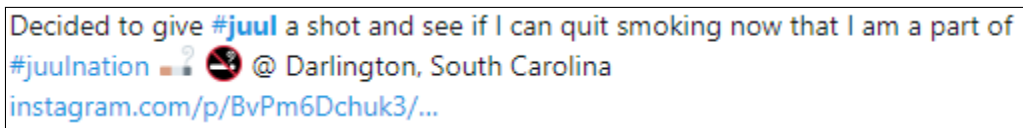


Figure 11.4: Male trying to quit smoking

Still waiting for a guy to come into my life to hit the juul while he's in the bed with me. I want him to scream "I'm hitting the juul while hitting the jewel" #lovemaking #juul

Figure 11.5: Love and Juul Female example

11.1.3 19-24 TFIDF Model (Figure 10.6)

Topic 1 I see as referencing the eventual hobby adults might develop and their discussion habits. For instance, in topic 1, “pod”, “flavor” are some of the most weighty words. Both pods and flavors can be altered to fit personal preference. Here we see the reference to “room” where people might Juul. These behaviours are related to nicotine addictions. Additionally, words like “sell” show how this product is highly desired and widely used. The word “ask” highlights the social aspect again where one could pass the Juul around to friends who “want” or “ask” for it. Also, the word “school” exists showing a slowly increasing popularity of Juuling in school whether in High School or college either to relieve stress or to fit in with one’s peers. The words “kid” and “guy” show a new awareness of people in terms of age groups. Perhaps in these age groups, people set different standards as to whether it is okay to Juul or smoke. However, words associated with addiction are less defined than in the 18 and under category, implying a greater maturity in young professionals not wishing to draw too much attention to their Juuling habits.

Topic 2 seems to involve the conversation of smoking versus the Juul as well as losing it. For instance, words such as “smoke”, “cigarette”, and “quit” are popular words. This shows the first minor example of how Juul use may be linked to smoking cessation and a main example of how interlinked the Juul is with its tobacco counterpart. The words that one would associate with partying are also less defined. Part of the comparison between Juul and smoking involves the discussion of how the “mint” “flavor”, for instance, of the Juul may be desired to the tobacco taste. The word “rip” and “smoke” are juxtaposed to each other with “smoke” being the more frequent word. Juuling and smoking has always been associated with work or school breaks. This is where words like “bathroom” produces an

example of the diverse areas where one can take a Juuling break instead of having to stand outside. “Quit” is also present and the central focus of whether or not one would choose to Juul in the first place to perhaps quit smoking.

With its unique satisfaction profile, simple interface and flavor variety, JUUL was designed with smokers in mind. By accommodating cigarette-like nicotine levels, JUUL provides satisfaction to meet the standards of smokers looking to switch from smoking cigarettes. #juul #e-cig pic.twitter.com/kQpdCOfgpm

Figure 11.6: Juul Versus Cigarettes

Topic 3 seems to be about a heightened sense of introspection regarding the Juul within the context of the “year”. You see people talking about “life”, “love”, “buy”, “pod”, and “friend”. All of these words come in even amounts. Not only do they come in even amounts, but they also are broad words with many ways in which it can be used, providing a central framework by which to discuss all possible topics. “School”, “class”, and “work” now emerge as specific places and contexts to Juul. Now we see the Juul intersecting even more into every day life. One may “need” to take his Juul to class and then to work. In class, it is now considered “cool” to Juul’, this is a phrase I have seen in the data. Being so popular in the daily sector, the word “buy” would naturally appear indicating a rise in sales to supply the demand. Even in normal discourse. Then the regular discussion topics still exist such as “lose”ing the Juul or how a “pack” of cigarettes is equal to a “pod”.

In essence, topic 3 is further developed in its discussion of the topic of Juul as compared to the 18 and under group.

11.1.4 19-24 TFIDF Model Male and Female (Figures 10.7 and 10.8)

Females seem to be more focused on conversation topics regarding health which involve “life”, “addiction”, “cigarette”, and “nicotine”. For instance, a life of “nicotine addiction” would be cause of concern of women seeking a healthy life as they grow older. Their use of the word “say” and “ask” shows respect for one another’s opinion most likely involving testimonials concerning the Juul. Women in this group seem to mention losing their Juul

Every so often, a **business** comes along and reminds everyone that if you create a product that makes smoking seem cool, you can make a lot of money. So, yeah, **Juul** is telling investors it expects sales of \$3.4 billion this **year**



Juul Expects Skyrocketing Sales of \$3.4 Billion, Despite Flavored Va...

The e-cigarette maker forecasts revenue that would nearly triple from last year.

[bloomberg.com](https://www.bloomberg.com)

Figure 11.7: Introspection about the Juul Business

more often.

Men in the group seem more inclined to talk about Juuling in the context of school. Words are seen such as “room” and “need” which refers to moments where one might need a room for a Juul break. The trend of males using the word “sell” and “girl” continue. One interesting switch up is the male’s talking about flavor becoming more common.

As age increases, it becomes rather difficult to disassociate men from women as their maturity levels meet. Based upon my study of the data, words such a “addiction”, “smoke”, and “cigarette” are used almost exclusively regarding the comparison of Juuling to smoking instead of as it relates to pleasure and partying.

11.1.5 25-35 TFIDF Model (Figure 10.9)

At the age of 25-35, I begin to see a significant shift in subject matter. Topic 1 is associated with the debate regarding cigarettes and the Juul, topic 2 deals with Juul's use in the work force and one's tendency to lose the Juul, and lastly topic 3 deals with the addictive appeal of Juuling.

In topic 1, we again see words such as “need”, “want”, “smoke”, and “pod”. These words are not quite such a large proportion of the WordCloud as they were in the earlier age group. I can think of two reasons. One, this age group is not the main consumers of Juul and therefore are less likely to be addicted to it, hence the need for Juul increasing its target sales on the older demographic. They would probably be more likely to smoke cigarettes. Option two involves the lack of incentive for an older person to make public their use of the Juul. It may indirectly affect their public image for their career or they simply just not care to post about it. Words such as “teen”, “hear”, “ask”, and “school” are frequent. Here we see the first mention of a teen which will play heavily into the following topics as discussions regarding the teen epidemic appears. Really topic 1 and topic 2 are very similar in content. One more heavily favors the nature of addiction, but it is addiction itself that makes the question so intriguing as to whether a Juul or cigarette is more useful in satisfying that crave while trying to remain healthy.

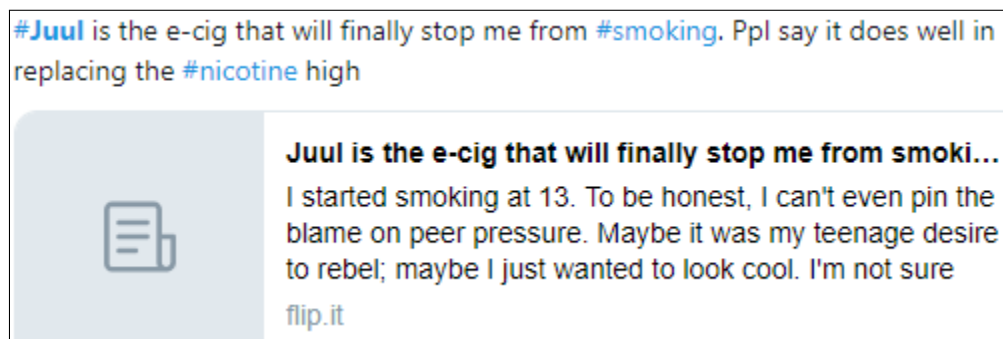


Figure 11.8: Juul and Smoking Cessation

This brings me to topic 2 where words such as “work”, “vape”, and “lose” are the main words. This points to the popularity of the Juul not only in schools but in the work

force as well. The 25-35 age group is the threshold where tweets turn more heavily into more mature topics except for the fact that even this group talks about losing the Juul which is surprising to me. At this point words such as “best”, “year”, “sell”, “walk”, “pod”, “try” appear. These words are casual terms which are a good indication that in this age group, Juuling is less of a taboo subject and people accept that it is part of a person’s daily routine. These words although capable of being used in a social context, have begun to morph into more individualistic words. By individualistic, I mean that people are more interested in describing the Juul as it relates to wide range issues regarding Juul trends during the year, or how the Juul improves their daily life, especially as it relates to smoking cessation.

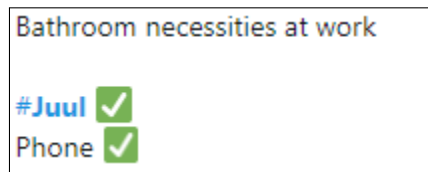


Figure 11.9: Juul in the Bathroom

Topic 3 has words such as “smok”, “cool”, “friend”, “pod”, “addiction”, and “quit”. When put together, these words are often associated together in the midst of a debate. For instance one will use testimonies about a friend to encourage one to Juul in order to quit smoking. However, a smoker might bring up the topic of pods and claim that pods pack a greater punch of nicotine than do cigarettes. Again, these debates all center around the need to satisfy a craving. The reappearance of “friend” also reemphasizes the social aspect of Juuling even in this age group. Words such as “rip” and ”love” which appeared in earlier groups still reminds us that even this age group likes to retain characteristics of the younger tweeters. Also, “kid” appears again showing that as one reaches the age of parenthood, concern for the environment in which a kid will grow increases.

11.1.6 25-35 TFIDF Model Male and Female (Figures 10.10 and 10.11)

At this point, there is not much distinction between male and females other than the fact that a guy is more likely to mention the word “girl” than a female is likely to.

In this age bracket, interestingly enough, males seem to have a wider distribution of words that incorporates many of the words seen in earlier male groups. Words such as mentioning “girl” in a tweet, losing the Juul, “smoke”, “life”, “addiction” and the others incorporates words which in earlier groups had male characteristics but now also contain words which previously were in female characteristics. For instance, now there is much more mentioning of addiction and smoking which females earlier on seemed more likely to do. This proves my earlier observation that now it is hard to distinguish between male and female. Almost clearly now, instead of three topics, two types of topics are steadily beginning to emerge in both groups, namely, nicotine addiction within the context of discussing smoking versus vaping and the societal implications of Juul.

11.1.7 35-55 TFIDF Model (Figure 10.12)

At this stage, expecting two topics may be more accurate than three since Topic 1 seems to contain not very many words with a high frequency but hints at older topics seen in early groups such as losing the juul, pods, and the need for nicotine. Buried in there is the word “kid” which is a foreshadowing of the main contents of topic 3. Topic 2 associates well with Topic 1 where it seems to be a continuation of the discussion of how addicting a Juul might be which is a type of “vape”. However, a new topic that appears is a huge emphasis on the youth epidemic. Here we see the words “teen”, “kid”, “addiction”, “smoke”, and “e-cigarette” mentioned. Upon parsing through many of the tweets, I found that a large sample of the tweets did involve discussing ways to combat the epidemic. This would make sense, since I would expect there to be many parents in this age group who are concerned about the well-being of their teenage child. Male and females are indistinguishable at this point.

11.1.8 35-55 TFIDF Model Male and Female (Figures 10.13 and 10.14)

At this point, there is really no distinguishing between male and females and one begins to see the topics slowly streamlining exclusively into ones concerning smoking, vaping,

Flash drive? No, it's a vaping device that teens smoke in school bathrooms & classrooms. One pod delivers as much #nicotine as a pack of #cigarettes. Join @CalHealthline Thursday for a #FacebookLive about e-cigs/e-juices that appeal to young people: facebook.com/events/2002655... #juul

Figure 11.10: Teen Epidemic

and pods.

11.1.9 55 and above TFIDF Model (Figure 10.15)

It seems the only real topic that people over the age of 55 talk about is the Juul as it relates to smoking. This would make sense that this would be the most interesting aspect to this age demographic since this group would have been smoking for the longest time period and probably least likely to quit now. Since there are not as many tweeters of this age, there is not as great a sample size to pass through the LDA topic modeling.

11.2 Association Rule Mining

In order to discuss the results from the association rules, I will note the main rules associated with the word “Juul” for each age demographic and note new rules that appear in specifically male or female groups.

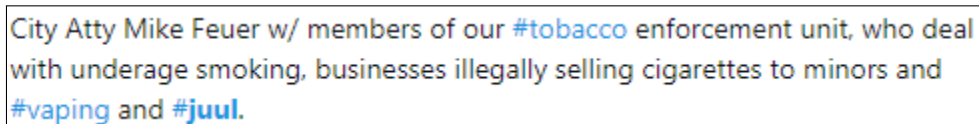
11.2.1 Association Rules 18 and under

For a reminder on how to interpret the graph starting with the rule “juul - > cigarette”, the support value for the first rule is .006. This number is calculated by dividing the number of tweets containing cigarette divided by the total number of tweets. The confidence level for the rule is .381 which shows that out of all the tweets containing “Juul”, 38.1% of the tweets also contain “cigarettes”. Finally, the 12.590 tells us that cigarettes are 12.59 times more likely to be seen in tweets where “Juul” is present than the default likelihood of the appearance of cigarettes. For example, one tweet which states, “I added a video to a @YouTube playlist: Convincing My Mom to Quit Cigarettes With The Juul”

where he places the link in the tweet (<https://t.co/kfisgCxMqV>). In the video he discusses how he hopes that his mom will begin Juuling instead of smoking cigarettes.

There is a 57.9% chance that pack is mentioned in a post where Juul is present and a 60.5% chance in this age group that if Juul is mentioned, that pack will be mentioned. These two rules does much to show how closely linked the conversation is between smoking and juuling.

In the female category, we see rules that associate the juul with selling pods, nicotine, losing it, Christmas, and selling the Juul itself. These rules highlight again the strong association that Juul has with nicotine. Additionally, the sharing capability of the Juul is highlighted in which it makes a good gift, or one may ask a friend for it for Christmas. Selling pods highlights the variability of the different flavor which provide nuanced experiences with the Juul. Losing the Juul is a very common occurrence and others to make some money can sell Juuls to their peers.



City Atty Mike Feuer w/ members of our #tobacco enforcement unit, who deal with underage smoking, businesses illegally selling cigarettes to minors and #vaping and #juul.

Figure 11.11: Illegal selling of the Juul to minors

11.2.2 Association Rules 19-25

The 19-24 age group emphasizes the main themes regarding Juul. Pods, namely mango pods are spoken of in regard to Juul and even Christmas. Granted, a lot of data was gathered during the Christmas season. Also, one can see the social aspect.

The female section notably we see the first occurrence of Juul being in school. We see reliance upon it with interesting tweets. These rules emphasize the need for the Juul which would lead to many breaks to go and Juul, many of which happen in the bathroom.



Figure 11.12: The Social Aspect

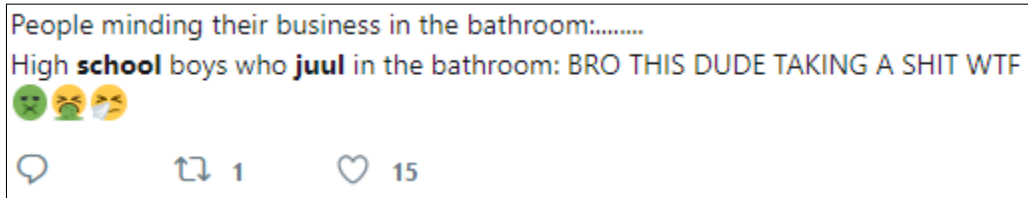


Figure 11.13: Juul in School

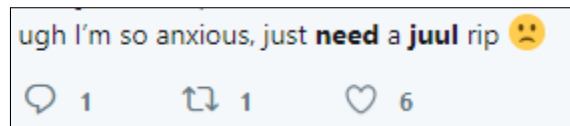


Figure 11.14: Need Juul

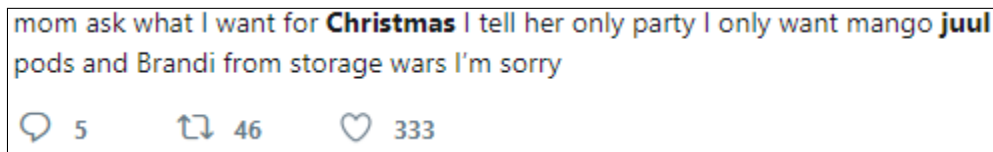


Figure 11.15: Juul pods for Christmas

11.2.3 Association Rules 25-35

A notable change found in the 25-35 category involves “react” and “elder”. The tweet shown in Figure 11.17 contains a funny video showing many elders trying a Juul for the first time. This again, shows the constant tension between cigarettes and Juuls.

Interestingly enough, even though the WordClouds show this age categories concern with the youth epidemic, it doesn’t show through in the association rules. Rather, again the Juul and its association with nicotine and cigarettes are highlighted. The word “room” associated with Juul highlights the environment in which one can be found Juuling.

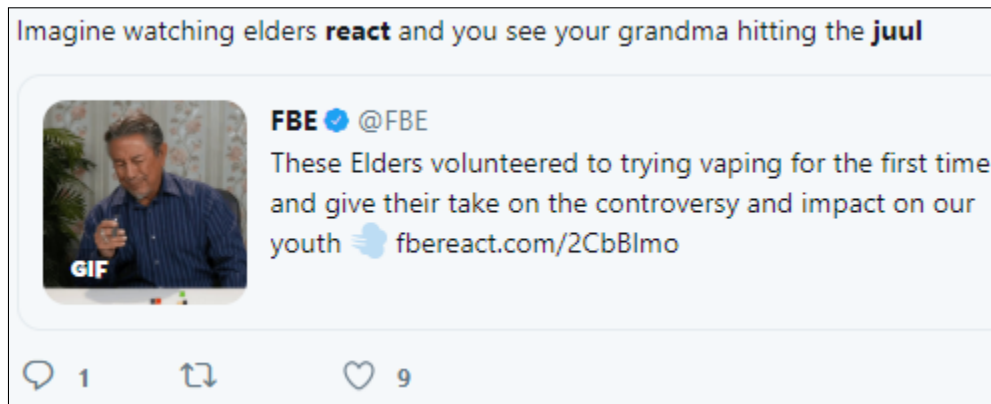


Figure 11.16: Elders Juuling

11.2.4 Association Rules 35-55

The 35-55 age category has two very interesting rules, “vape – > desk” and “juul – > quit”. Vaping and desks seem to relate to tweets regarding the prohibition of Juuls in the work force. There is perceived frustration regarding the rules that prohibit one from Juuling at work.



Figure 11.17: Vaping Restrictions at Work

Most likely quitting is referencing to smoking cessation rather than quitting the Juul itself. For instance, there are many who post studies on the matter, often including doctors.

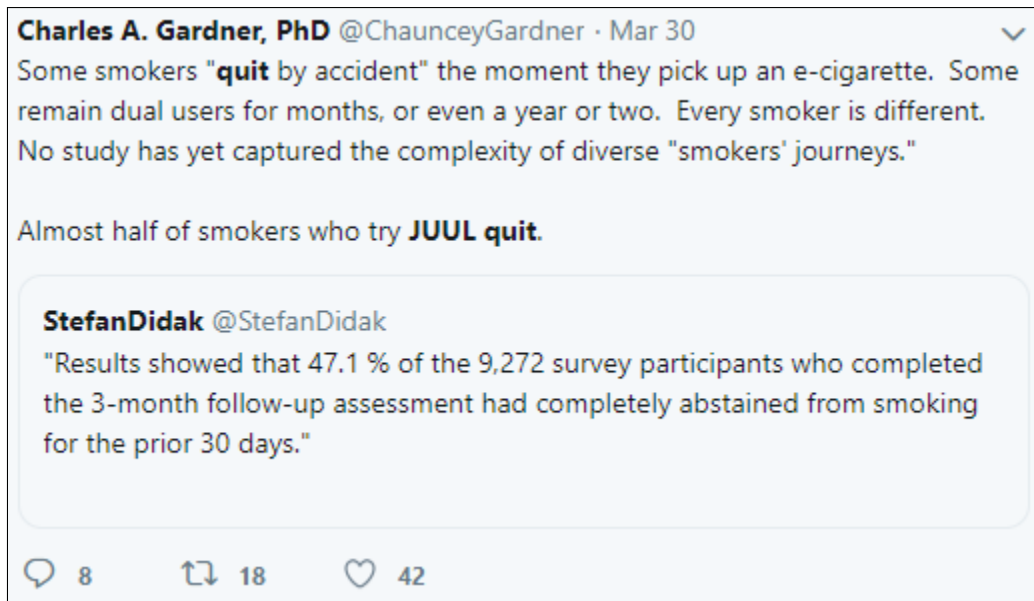


Figure 11.18: Quitting Cigarettes using Juul

However, people also do often talk about trying to quit the Juul itself due to its addictive nature. After this, the tweet sample size was not great enough to generate meaningful results.



Figure 11.19: Quitting Juul

CHAPTER 12

METHODOLOGY LIMITATIONS

Methodology Limitations.

12.0.1 **Weight of a re-tweet vs individual tweet**

One of the greatest difficulties is accounting for the re-tweets. For instance, the LDA and Association Rule models would emit skewed results due to the sheer numbers of certain tweets. However, I could not help but notice that the very fact that a tweet is re-tweeted is that it resonates with others. Therefore, a re-tweet to a large degree would be impact. Therefore, I will briefly mention a few notable re-tweets that were discarded through the normalization process but gives some insight. If there is not a picture of the tweet shown, either the account no longer exists or the user keeps the tweets protected.

12.0.2 Sample Tweets

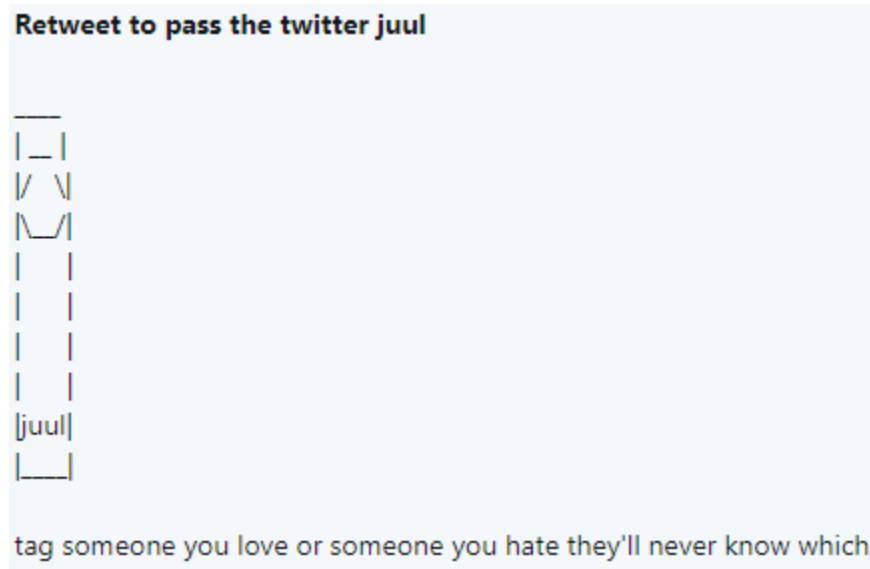


Figure 12.1: Sharability of the tweet in a social context even in digital form.

if u shotgun a four loko then rip a juul in a bathroom a frat boy
will appear in the mirror

Figure 12.2: Tweet shows that a Juul can often be found with a loko, the idea of ripping it in a bathroom and greek life.



Figure 12.3: Association that a juul addiction has with college, beer, and partying. 3

Perhaps the greatest thing that is lost without the re-tweets is the repetitive nature of the re-tweets and its association with partying almost as if it is one the big appeal to Juul.



Figure 12.4: Tension of children trying to hide their Juul from their parents.

airpod juul pod

5 2 39

Figure 12.5: Year of different types of pods.

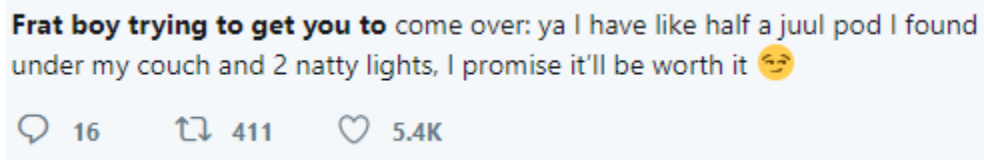


Figure 12.6: This represents a popular opinion of thinking of a Juul in the context of partying.

i will never forget the time bridget thought this guy was trying to **flirt** w her so she asked to use his **juul** and when he said no she flat out said "do you know how many boys wish my lips would touch their **juul**?"

Figure 12.7: This shows a hidden aspect in which girls are likely to use flirtation to get what they want from a guy.

However, after removing these tweets and other re-tweets similar to these, all of the other contents come out of the different models.

12.0.3 **Not Sure if Humanizr Worked**

Because Humanizr only returned a total of two organizations out of the numerous tweets, either Humanizr was not as helpful as hoped in determining whether a tweet was from a person or an organization or that organizations simply avoid tweeting about the Juul.

12.0.4 **Assumptions Need to be Made About the Age and Gender**

Due to the use of APIs like Face ++ and Humanizr, different assumptions have to be made such as that the vast majority of tweets are labeled correctly as being the right age and gender. Although the results seem to validate the procedure, it is possible that some errors are embedded in the labeling of data in certain instances.

12.0.5 **I lost some data because people changed their posts**

Some of the data collected was unable to be processed since between the time that I collected the data and ran them through the models, roughly 40% of the users changed their profile pictures which made it impossible to determine their age and gender. So some of the samples had to be excluded.

12.0.6 **Inability to account for potential bots**

Although in general it is hard to detect potential bots in the tweets, some sort of bot detection mechanism would be helpful in perhaps finding the estimated 5-10% tweets made by bots.

12.0.7 **There are further robust implementations of the different algorithms**

If given more time, there are different variations of the algorithms used which could expose further nuances of the different trends already found in the data.

12.0.8 Use different hashtags that also relate to Juul

In collecting the data, I could have used various hashtags other than #Juul to collect different types of tweets. For instance, I could have queried for tweets that contained #VapeLife or I could look at party hashtags and see how often a Juul is mentioned.

12.0.9 Larger sample size

As always a larger data set over an extended period of time would probably reveal new trends especially as the conversations regarding the Juul ebb in flow as scientists make new findings and youth use of the Juul increases.

12.0.10 Lack of time

Due to time restrictions, certain aspects of collection and modeling could not be repeated as much as I would have liked in order to ensure as normalized results as possible.

12.0.11 Hashtag collection mistake

The hashtag table is based on all of the tweets rather than only the tweets that are labeled an age and a gender. Unfortunately, when collecting the initial data, I did not collect the most popular tweets within each category. By the time I realized this, if I re-ran the process, even more users would have changed their profile pictures further restricting the sample size of tweets.

CHAPTER 13

CONCLUSION

Conclusion:

13.0.1 Methodology Experimentation

A large part of this thesis has involved the validity of using such a method to try and categorize tweets by age and gender to unveil more complex and nuanced patterns within the data. Based upon the data generated, a lot of the trends I was curious to see were validated and others that I did not think to expect were found as well. Due to the cohesive trends seen throughout this process, I suspect that although certain assumptions have to be made with this methodology, they are reasonable assumptions to make regarding the use of Humanizr and Face ++. Further experiments under this methodology will need to be undergone with varied data sets over longer periods of time to see how effective this methodology is in understanding other twitter trends.

13.0.2 Conclusions based upon data

I would like to discuss the conclusions I came to based upon the cumulative evidence of reading through inferences made by reading many of the tweets myself, utilizing LDA topic modeling, and Association Rule Mining. I like to think of the results as resembling a pyramid-like structure similar to the food chain. To understand this, let me briefly explain what I mean by the food chain graph. The sun being the primary source of energy to the food chain is transferred into the ecosystem from the bottom up. The primary producers at the bottom take "only around 1% of the total available energy from the sun" is transferred to the next level (Amit, 5/24/2018). After photosynthesis, only 10% of their energy is deposited

in their tissues to be available for grazing herbivores. So what we see at each respective level, is a much lower level of the original energy being transferred up than is present in the lower levels. Similarly, I think of my normalized data being passed into the LDA topic modeling and Association Rules as being my sum total available amount of energy. At the bottom of the Pyramid represented mostly by those 18 and under where we get a base level of topics that are frankly rather crude but descriptive in terms of experience where we see Juul being mentioned in the context of nicotine addiction, losing it, partying, flirting to fuel the addiction, beer, the types of flavors, its association with life itself, and contexts in which to Juul such as in school bathrooms. In this category, most of the tweets are rather raw and deal a lot with juvenile interactions and jokes. Partial aspects of this baseline are present in some of the levels above it. However, as the pyramid levels increase, the frequency of these base associations decrease slowly which allows new topics to arise to the surface which define that level although these newer topics are less common overall than the topics found in below levels. At the next level, there is a slightly greater air of maturity in which there is a little more self-awareness. Rather than the Juul being used as much in such experiential terms, we begin to see the Juul being mentioned more in the context of school and college. Students are going to bathrooms to Juul out of necessity in school. Because many are in the 19-25 age group, a lot of these tweets most likely also reference college when mentioning the word "school". You also begin to see more discussion related to the context of the Juul. For instance, I might see a tweet talking about their opinion about smoking versus the Juul, some will mention Altria's acquisition of a large portion of Juul, others will discuss the amount of money that the Juul company makes, and still others will give personal testimonies. At the next level, we see a large emphasis of talking about the Juul in the context of work. At one point in the discussions, we see tweets talking about how it is not allowed to Juul in the office of Pax Labs itself. Others like to often mention in the context of smoking cessation. At this point in the chain, people will talk about the Juul from a more individualistic perspective rather than more collectively as it relates to their peers. It is more of a tool to either find an

alternative to smoking or the means to quit. At the next level, we see concern to overflow regarding the teen epidemic. Many are frightened that the Juul is hooking many teenagers to nicotine. This discussion happens in older audiences presumably because they have kids which would make them more concerned about the health affects of Juul. Lastly at the highest level, in the age category above 55 with what little tweet information or "energy" is left is constituted as tweets mentioning smoking or cigarettes. This would make sense since the oldest demographic would have been living around cigarettes for most of their lives and if they smoke and are still alive, they may be slower to trust potential benefits of transitioning to the Juul.

So in conclusion, we see a transition over time from a broad range of topics usually regarding base topics or more general topics and slowly it transitions to less frequent but revealed more complex topics as you ascend the different age demographics.

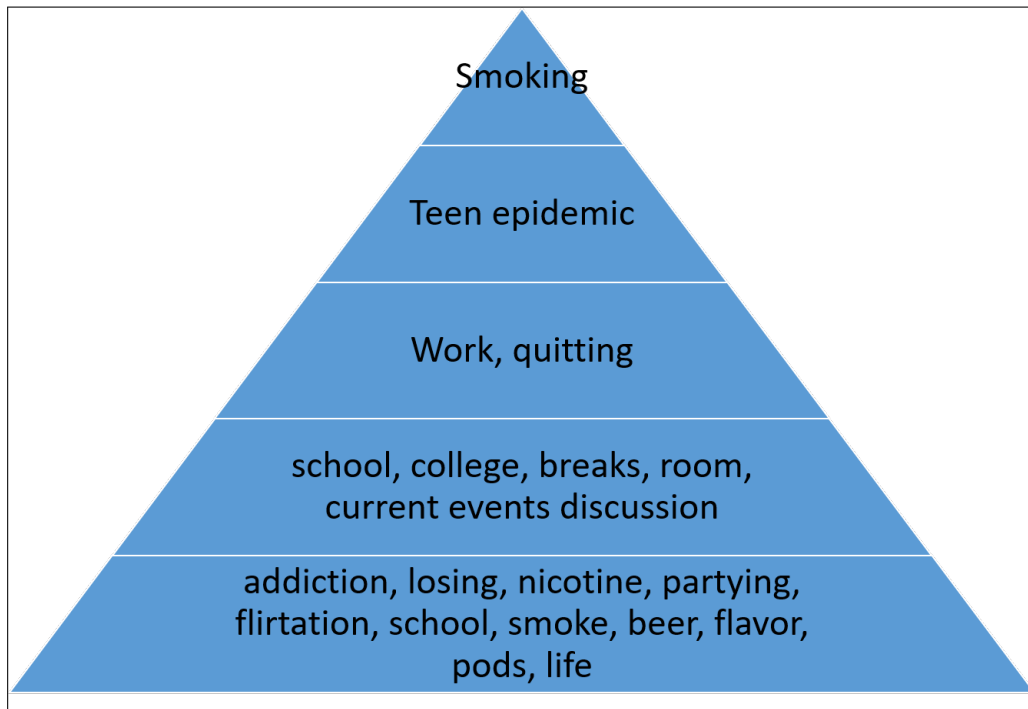


Figure 13.1: Pyramid Representation

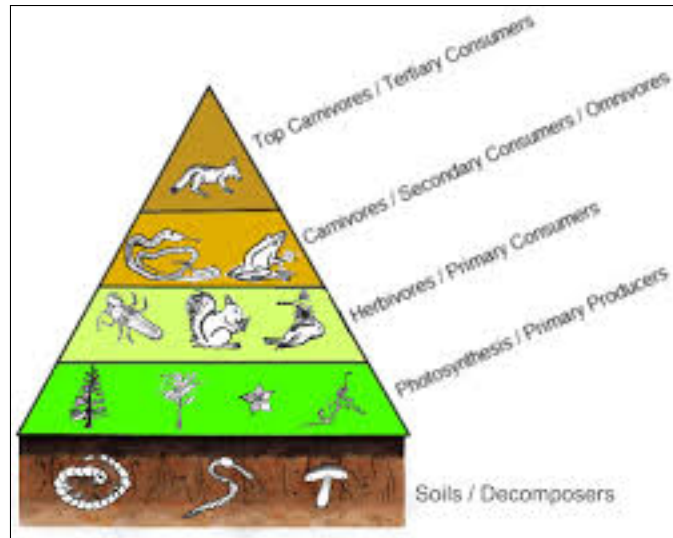


Figure 13.2: Food Chain Representation (Amit, 5/24/2018)

13.0.3 Future wishes

In the future for this study, I would like to begin looking into the trends that are revealed over the course of a year rather than just a couple months. I also can further perfect my data normalization and implementation of LDA topic modeling and Association Rules. Lastly, varying the different hash tags used to collect data will also provide more variability and the quantity of the data set.

13.0.4 Final Thoughts

The Juul is such a complex subject and great work has been done by the FDA and Medical associations in attempting to discourage youth from unnecessarily becoming addicted to nicotine. As far as policy is concerned, the gateway moment of a teen trying the Juul goes deeper than simply his or desire for a specific flavor. Juuling is firmly established in the social construct of a youth's experience. In this case, unless the fad dies down on its own, policy itself does not seem like it will be a long term solution to a teen's use of the Juul. Rather, much work will need to be done in capturing the hearts and the minds of the younger generation. Children will need to be educated on the dangers and long term effects of nicotine addiction, the cost benefit of spending so much money on pods when

one could be using for other more profitable things, and the creation of self awareness to understand how peer pressure is influencing so much of their malleable minds. Education coupled with self confidence to know that one does not need to try a Juul to appear cool is a mandatory prerequisite to avoiding such a nicotine addiction. If nothing else, a well laid out presentation of the financial cost that a Juul represents may help a child understand that Juuling is not worth its imagined benefits. Unfortunately, it will be years before we know the long term effects of Juuling. However, the subject of the Juul does open up a platform for the discussion of the nature of addiction itself. Therefore, we must take this opportunity to apply what we learn about nicotine addiction and use it to combat all forms of addiction.

Bibliography

- Algobeans (05/30/2018), Topic modeling with lda introduction, <https://algobeans.com/2015/06/21/laymans-explanation-of-topic-modeling-with-lda-2/>.
- Allem, J.-P., L. Dharmapuri, J. B. Unger, and T. B. Cruz (2018), Characterizing juul-related posts on twitter, *Drug and alcohol dependence*, 190, 1–5.
- Amit, S. (5/24/2018), Energy pyramid, <https://www.toppr.com/bytes/energy-pyramid/>.
- BestHashtagsWebsite (1/31/2019), Best juul hashtags, <http://best-hashtags.com/hashtag/juul/>.
- Braier, Y. (6/25/2018), Are e-cigarettes a safe alternative to smoking?, <http://www.casaa.org/historical-timeline-of-electronic-cigarettes/>.
- Brett, E. I., E. M. Stevens, T. L. Wagener, E. L. Leavens, T. L. Morgan, W. D. Cotton, and E. T. Hébert (2019), A content analysis of juul discussions on social media: Using reddit to understand patterns and perceptions of juul use, *Drug and alcohol dependence*, 194, 358–362.
- Brown, C. J., and J. M. Cheng (2014), Electronic cigarettes: product characterization and design considerations, *Tobacco control*, 23(suppl 2), ii4–ii10.
- Brown, J., E. Beard, D. Kotz, S. Michie, and R. West (2014), Real-world effectiveness of e-cigarettes when used to aid smoking cessation: a cross-sectional population study, *Addiction*, 109(9), 1531–1540.
- Brownlee, J. (10/09/2017), A gentle introduction to the bag-of-words model, <https://machinelearningmastery.com/gentle-introduction-bag-words-model/>.
- CASAA (2016), A historical timeline of electronic cigarettes, <http://www.casaa.org/historical-timeline-of-electronic-cigarettes/>.
- Chan, E. (8/22/2011), Introduction to latent dirichlet allocation, <http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/>.
- COMPLEX (11/29/2013), Gender divide: How men and women use twitter differently, <https://www.complex.com/pop-culture/2013/11/men-women-twitter-differences-gender/>.

- DiFranza, J. R., et al. (2007), Symptoms of tobacco dependence after brief intermittent use: the development and assessment of nicotine dependence in youth-2 study, *Archives of pediatrics & adolescent medicine*, 161(7), 704–710.
- Etter, J.-F. (2018), Gateway effects and electronic cigarettes, *Addiction*, 113(10), 1776–1783.
- Goriounova, N. A., and H. D. Mansvelder (2012), Short-and long-term consequences of nicotine exposure during adolescence for prefrontal cortex neuronal network function, *Cold Spring Harbor perspectives in medicine*, p. a012120.
- Huang, J., Z. Duan, J. Kwok, S. Binns, L. E. Vera, Y. Kim, G. Szczypka, and S. Emery (2018a), Vaping versus juuling: how the extraordinary growth and marketing of juul transformed the us retail e-cigarette market, *Tobacco Control*, pp. tobaccocontrol-2018, doi:10.1136/tobaccocontrol-2018-054382.
- Huang, J., Z. Duan, J. Kwok, S. Binns, L. E. Vera, Y. Kim, G. Szczypka, and S. L. Emery (2018b), Vaping versus juuling: how the extraordinary growth and marketing of juul transformed the us retail e-cigarette market, *Tobacco control*, pp. tobaccocontrol-2018.
- Hutchinson, A. (3/18/2018), Here’s why twitter is so important, to everyone, <https://www.socialmediatoday.com/social-networks/heres-why-twitter-so-important-everyone>.
- Lettier (2/23/2018), Your easy guide to latent dirichlet allocation, <https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24>.
- Li, S. (05/30/2018), Topic modeling and latent dirichlet allocation (lda) in python, <https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24>.
- Malik, U. (8/09/2018), Association rule mining via apriori algorithm in python, <https://stackabuse.com/association-rule-mining-via-apriori-algorithm-in-python/>.
- Marino, C., A. Vieno, M. Pastore, I. P. Albery, D. Frings, and M. M. Spada (2016), Modeling the contribution of personality, social identity and social norms to problematic facebook use in adolescents, *Addictive behaviors*, 63, 51–56.
- Morris, P. (2/13/2019), Philip morris website, <https://www.pmi.com/>.
- NIDA (2/01/2016), National institute on drug abuse: Teens and e-cigarettes, <https://www.drugabuse.gov/related-topics/trends-statistics/infographics/teens-e-cigarettes>.
- Remedy, S. (Aired 08/30/18), Juul vaping documentary, <https://www.pbs.org/video/vaping-clouded-by-controversy-lrdmfx/>.
- Richtel, M., and S. Kaplan (2018), Did juul lure teenagers and get ‘customers for life’?, *The New York Times*.

- Rivas, T. (12/20/2018), What altria's juul investment says about tobacco's future, <https://www.barrons.com/articles/why-altria-is-investing-in-juul-51545329470>.
- Schneider, S., and K. Diehl (2015), Vaping as a catalyst for smoking? an initial model on the initiation of electronic cigarette use and the transition to tobacco smoking among adolescents, *Nicotine & Tobacco Research*, 18(5), 647–653.
- Shiwnarain, M. (5/01/2018), What does cbd stand for?, <https://sciencetrends.com/what-does-cbd-stand-for/>.
- Stanton A. Glantz, P. (10/3/2014), Nicotine is not caffeine, <https://tobacco.ucsf.edu/nicotine-not-caffeine>.
- Wald, R., T. M. Khoshgoftaar, A. Napolitano, and C. Sumner (2013), Predicting susceptibility to social bots on twitter, in *2013 IEEE 14th International Conference on Information Reuse & Integration (IRI)*, pp. 6–13, IEEE.