NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

HONOR AMONG THIEVES: ANALYZING LANGUAGE FEATURES OF DARKNET MARKET VENDORS

by

John E. Smith III and Nicholas E. Hughes

June 2020

Thesis Advisor: Neil C. Rowe
Co-Advisor: Vinnie Monaco

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE	*Form Approved OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503.

1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** June 2020	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE HONOR AMONG THIEVES: ANALYZING LANGUAGE FEATURES OF DARKNET MARKET VENDORS	**5. FUNDING NUMBERS**
6. AUTHOR(S) John E. Smith III and Nicholas E. Hughes	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.	**12b. DISTRIBUTION CODE** A

13. ABSTRACT (maximum 200 words)

The World Wide Web enables retailers to provide goods and services to consumers around the globe, including illegal products and services. The digital black market (Darknet) includes digital exploits, hacker-for-hire services, drugs, weapons, and other illicit goods and services. Appropriately classifying threats on the Dark Web is critical to military and law-enforcement cyber surveillance, reconnaissance, and defense. We used statistical and sentiment analysis to create a unique digital profile of Darknet market vendors. We also identified characteristics that indicate truthfulness, deception, credibility, and intent by analyzing product description language. The features of these profiles were used with a recommender system to track vendors across marketplaces. Our experiment achieved a rank-1 vendor identification accuracy of 74.7%. We also used semantic fingerprinting to identify vendors that deviated from the market average. Many of these anomalous vendors were discovered to have indicators of fraudulent and deceptive practices. We concluded that using a recommender system and sentiment analysis on vendor language in Darknet marketplaces helps law enforcement and national security professionals to track and disrupt the sale of illicit goods and services.

14. SUBJECT TERMS cyber, Darknet, markets, World Wide Web, machine learning, natural language processing, sentiment analysis, deception, truthfulness, Dark Web	**15. NUMBER OF PAGES** 99
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

THIS PAGE INTENTIONALLY LEFT BLANK

HONOR AMONG THIEVES: ANALYZING LANGUAGE FEATURES OF DARKNET MARKET VENDORS

John E. Smith III
Lieutenant Commander, United States Navy
BA, Prairie View A & M University, 2008

Nicholas E. Hughes
Lieutenant Commander, United States Navy
BA, University of Colorado at Boulder, 2009

Submitted in partial fulfillment of the
requirements for the degrees of

MASTER OF SCIENCE IN COMPUTER SCIENCE

and

MASTER OF SCIENCE IN CYBER SYSTEMS AND OPERATIONS

from the

NAVAL POSTGRADUATE SCHOOL
June 2020

Approved by: Neil C. Rowe
Advisor

Vinnie Monaco
Co-Advisor

Peter J. Denning
Chair, Department of Computer Science

Thomas J. Housel
Chair, Department of Information Sciences

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The World Wide Web enables retailers to provide goods and services to consumers around the globe, including illegal products and services. The digital black market (Darknet) includes digital exploits, hacker-for-hire services, drugs, weapons, and other illicit goods and services. Appropriately classifying threats on the Dark Web is critical to military and law-enforcement cyber surveillance, reconnaissance, and defense. We used statistical and sentiment analysis to create a unique digital profile of Darknet market vendors. We also identified characteristics that indicate truthfulness, deception, credibility, and intent by analyzing product description language. The features of these profiles were used with a recommender system to track vendors across marketplaces. Our experiment achieved a rank-1 vendor identification accuracy of 74.7%. We also used semantic fingerprinting to identify vendors that deviated from the market average. Many of these anomalous vendors were discovered to have indicators of fraudulent and deceptive practices. We concluded that using a recommender system and sentiment analysis on vendor language in Darknet marketplaces helps law enforcement and national security professionals to track and disrupt the sale of illicit goods and services.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

2-FA	2-factor authentication
BoW	bag-of-words
CSV	comma-separated values
DDoS	distributed denial of service
DNM	Darknet market
IP	Internet protocol
LIWC	linguistic inquiry word count
NLP	natural-language processing
OG	original gangster
PCA	principal component analysis
PGP	Pretty Good Privacy
SVM	support-vector machines
TF-IDF	term frequency-inverse document frequency

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

We would like to thank Dr. Neil Rowe and Dr. John “Vinnie” Monaco for inspiring us to embark on this work. In particular, we would like to thank Dr. Rowe for his guidance, encouragement and support throughout the past twelve months, and Professor Monaco for his ideas and inputs as we developed the thesis. We also would like to extend a special thank you to Nina Lopatina— her technical support and advice were crucial to our work —and the team at Naval Surface Warfare Center, Corona Division, for their help in guiding our research for this project.

Last but not least, this work is dedicated to our loving families who supported us every step of the way. To Jennifer, Troy, Bella, Hope and Rose, Lesley, Rowan, and Marshall, your love, prayers, and encouragement were a constant source of motivation and joy over the past two years. Without your support, none of this would have been possible.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

The Dark Web is the hidden part of the Internet that uses special software to provide anonymity. The threat landscape of the Dark Web raises serious issues of national security. In 2013, the National Security Agency revealed that for over a decade, coordination and communications between al-Qaeda leadership worldwide primarily took place within the Dark Web (Weimann, 2016). Also, these hidden layers of the Web allowed terrorists to fund-raise, spread propaganda for recruitment, and purchase illegal explosives and weapons. In fact, the weapons used in the 2015 Paris terrorist attack were bought on the Dark Web from a German Darknet vendor. Equally alarming is the proliferation of digital exploits through Dark Web channels. The WannaCry ransomware attack in May of 2017, which infected over 200,000 computers across 150 nations, and resulted in billions worth of damages, originated from the Dark Web (Leefeldt, 2017). By 2017, nearly 6,300 Dark Web marketplaces listed over 45,000 ransomware products for sale, increasing profits from $250,000 the year before to $6.2 million (Graham, 2017). Other services available for purchase on the Dark Web include botnets for distributed denial of service attacks, classified information, and hacker services. This part of the Internet attracts cyber-criminals, terrorists, hacktivists, and non-state actors who can use the anonymity of the Dark Web to avoid laws.

The public conducts most online activity on the “surface web” part of the Internet. It represents the part of the World Wide Web that is indexed by common search engines such as Google and Yahoo. These search engines use “Web crawler” software called spiders to categorize and index publicly available information which is then used to respond to keyword searches. However, indexed pages only represent a small percentage of the total Internet (Finklea, 2017). Most of the Internet lives on the Deep Web, with content only accessible by an authentication mechanism such as a username and password. Typical information found on the Deep Web are email databases, legal documents, and banking information. The Dark Web is a subsection of the Deep Web that employs anonymizing technology to hide user identities, locations, and usages. Thus, it is home to political activists in totalitarian nations, whistleblowers and journalists, as well as thousands of black-market sites. Figure 1 shows the different layers of the Internet and their typical usage.

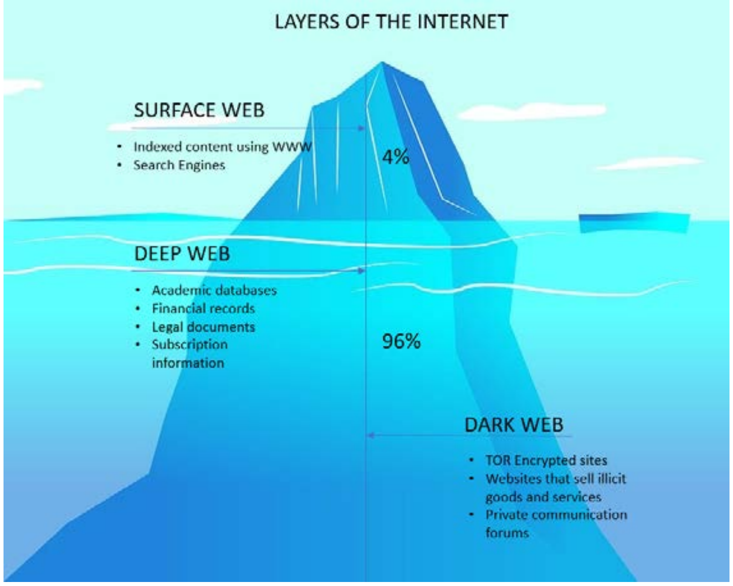


Figure 1. Iceberg Metaphor of the Internet. Adapted from Kang (2018).

The Dark Web is a refuge that offers malicious actors concealment to communicate and coordinate clandestine and illicit activity including the procurement of malware, weapons, child pornography, classified information, and narcotics. Transactions on the Dark Web were estimated to reach over $1 billion in 2019 (Kumar & Rosenbach, 2019). Tor (The Onion Router) is popular anonymizing software for accessing the Dark Web (The Tor project, 2020) and will be described in Chapter II.

Fueling the online underground economy are cryptocurrencies such as Bitcoin; decentralized digital monies that use peer-to-peer technology to accomplish instant payments (Finklea, 2017). Cryptocurrencies use public keys or electronic addresses to secure all transactions which are recorded into a blockchain, a publicly available immutable ledger (Finklea, 2017). Since these electronic addresses are not directly linked to personally identifiable information, users are afforded pseudonymity. However, when coupled with Tor’s anonymity technology, cryptocurrency addresses and node Internet Protocol (IP) addresses become unlinked thus creating untraceable transactions. Thus, Tor and cryptocurrency have created an environment from which Darknet marketplaces can thrive. Bitcoin transactions on the Dark Web grew from $250 million in 2012 to over $872 million in 2018 (Kumar & Rosenbach, 2019).

A market thrives when both buyer and seller benefit from an exchange (Soska & Christin, 2015). To that end, formality, trust, and appraisal can help achieve the conditions by which parties can arrange that money, goods, and services be transferred. These conditions also apply to the commerce of illicit goods and services offered on the Dark Web. In an environment inherently built on obfuscation and deception, how do participants in the digital black market establish the conditions to conduct e-commerce? They have several mechanisms to build trust (Soska & Christin, 2015). Silk Road, the original market for trading illicit items on the Dark Web, used a feedback system to rate buyers, vendors and escrow services to improve confidence in their transactions. This included a hedging service to account for the volatility of the bitcoin value. Indeed, brokering trust for the purchase of illegal goods and services is a primary responsibility of the Darknet marketplaces. The feedback system is even more important in maintaining individual reputations than with the open Web since it decreases the perceived risk from scamming and law-enforcement intervention. So, the language and etiquette used on Darknet markets to enable illegal commerce can give valuable insights into this anonymous ecosystem.

Data analytics is helpful to understand the inner workings of Darknet marketplaces and to quantify characteristics and relationships unique to vendors and customers engaged in the trade of illicit goods and services. Natural-language processing, the ability of computers to read, translate language, and derive meaning behind human communications, is another useful method for analyzing the Dark Web. Natural-language processing can identify patterns in textual communication that indicate intent and meaning (SAS Institute, Inc., 2020). Applying sentiment and syntactic analysis to Darknet product listings will help us understand the semantics of discourse in this underground economy, as well as how a vendor’s reputation is maintained and the circumstances in which deceptive language is used. Appropriately classifying threats on the Dark Web is critical to cyber reconnaissance, surveillance, and defense. We hypothesize that vendors on the Darknet use distinct language to advertise goods and services and that applying a sentiment analysis tool and a recommender system will enable fingerprinting and tracking sellers across marketplaces.

Chapter II provides background on Darknet markets and law-enforcement efforts to disrupt the exchange of illicit goods and services on the Dark Web. We examine the mechanisms used by Darknet marketplaces and vendors to foster trust and establish reputations. Finally, we explore the linguistic features that suggest truthfulness and deception.

In Chapter III, we further define the challenges that law enforcement and intelligence professionals face in combating the sale of illicit goods on the Dark Web, as well as those challenges faced internally among Darknet vendors and consumers. This chapter also discusses similar studies that investigated Darknet linguistic features.

Chapter IV discusses our methodology for conducting our research, whereby we first selected datasets that would be valuable for studying Darknet market language. We then conducted quantitative and qualitative analysis to pinpoint interesting subsets of the data that warranted further exploration. We used a recommender system and sentiment-analysis tool on vendor product descriptions to measure the language composition of Darknet markets.

In Chapters V and VI, we report on the results of our experiment with the recommender system and sentiment-analysis software and highlight the value these tools can provide to law enforcement and intelligence professionals.

II. RELATED WORK

Reviewing the history and characteristics of the digital underground provides meaningful context and is the foundation of our research. This includes discussing the methods buyers and sellers use to establish a community of trust while also attempting to circumvent law enforcement.

A. DARKNET MARKET CHARACTERIZATION

Darknet marketplace users trust the anonymizing technology provided by services such as Tor (The Tor project, 2020). Tor is open-source (freely-available) software that prevents attribution, and by extension, tracking and collecting digital information on user activity (Soska & Christin, 2015). The technology was originally developed by the Naval Research Laboratory to anonymize U.S. intelligence communications and is a general tool to send and get information anonymously. It is used by people concerned about privacy, but it has also enabled cyberspace criminals to create the online anonymous marketplaces for illicit goods and services.

Instead of routing network traffic directly to a server, Tor uses trusted nodes located across the globe to create secure tunnels that route packets to their ultimate destination. The trusted nodes, also known as relay points, are linked using session keys and transport layer security handshakes. Traffic is re-encrypted at each node until it reaches the exit relay. Due to the multiple layers of encryption and the decentralization of nodes, network traffic using Tor proxies is nearly impossible to trace to individual users. Figure 2 shows a Tor network:

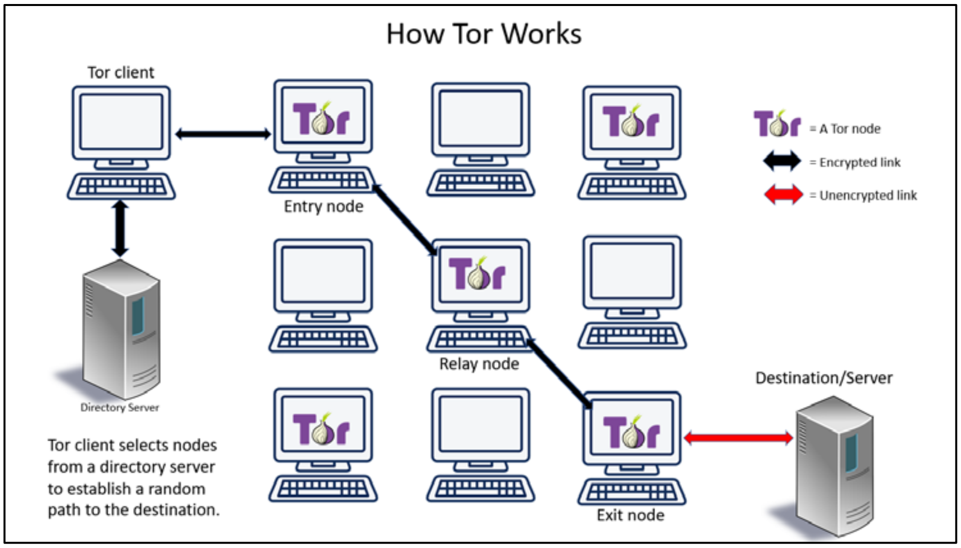


Figure 2. Tor Peer-to-Peer Network. Adapted from The Tor Project (2020).

Almost all Darknet marketplaces use cryptocurrency such as Bitcoin for monetary transactions. Cryptocurrency provides a digital form of payment and can give an additional level of security and trust for transactions on the Dark Web (Kirkpatrick, 2017). Cryptocurrencies use blockchain technology with encryption to control the creation of currency units and to verify transactions by private keys, thus obviating the need for a centralized authority (Bitcoin Project, 2020). These transactions are tracked in a digital public ledger, but the accounts or "wallets" that hold the currency at both ends of the transaction can be established anonymously. Also, Bitcoin addresses are disposable: a user can easily create a new account, making it difficult to identify and link entities with specific financial activities. By distancing a digital wallet from its physical owner, law enforcement and government agencies have great difficulty tracking the sale of illicit goods in the Darknet.

Obscuring digital currency flows can also increase anonymization used by Darknet market patrons. For instance, CoinJoin, combines multiple Bitcoin payments into one (Elendner et al., 2016). With CoinJoin, all inputs and outputs are listed under the Bitcoin address of the aggregated transaction, and it is very difficult to correlate individual profiles with specific Bitcoin sums. To provide further trust, anonymity, and reliability, Darknet marketplaces also use an escrow system that temporarily holds funds until all parties are satisfied with their transaction. This step allows patrons to overcome the natural uncertainty anonymity creates between individual users.

The first successful Darknet market using the anonymizing features of the Dark Web was Silk Road. Within a few years of its launch in 2011, Silk Road became the most prominent and sophisticated criminal marketplace in the world. For nearly three years, Silk Road generated millions of dollars in revenue through the trade of illicit goods and services (Bearman & Hanuka, 2015). Its success spawned other underground black markets. Typical items found for sale on Darknet markets include:

- Prescription and recreational drugs
- Weapons
- Official documents
- Credit cards and card data
- Identities and credentials
- Child pornography
- Cyber exploits and cyber-attack tools and services

Law enforcement agencies, cyber security teams, and academia have observed and characterized Darknet marketplaces (Bearman & Hanuka, 2015). This includes surveying descriptions of products available for purchase or investigating sales among the product and service categories. They also observed market prices with sales to determine what products and services are producing the greatest revenue.

For just the cyber-attack market, researchers saw a wide range of products and services for sale (Meland & Sindre, 2019); the most abundant of which are hackers for hire, account and password crackers, and stealers and grabbers (clipboard data exploiters). In 2019, password crackers and phishing kits were predominant, although only a few marketplaces sold most of these products (Meland & Sindre, 2019). Sales figures can give another accurate view of marketplaces. Phone-hacking tools, hacking packages, and stealers and grabbers are the three most prominently sold cyber-attack items. However, hacker-for-hire services generate the most revenue (Meland & Sindre, 2019).

B. LAW ENFORCEMENT INTERVENTION

In the past decade, law enforcement around the globe has ended several major Darknet marketplaces. Operation Marco Polo, which began in 2011, brought together seven U.S. federal agencies and culminated in October 2013 with the arrest of the Silk Road 1.0 administrator, Ross Ulbricht. Law enforcement agencies had infiltrated the inner circles of Silk Road by purchasing many goods on the market and using flipped administrative accounts to gain Ulbricht's trust. FBI surveillance teams also saw him logging off at the same time Dread Pirate Roberts, the Silk Road administrator profile, went offline (Bearman & Hanuka, 2015). Although this operation was conducted by U.S. agencies, the Darknet community was affected on the international level. In 2014, Operation Onymous operated on an even larger scale when agencies from 17 countries worked to arrest 17 vendors and to shut down more than 410 hidden services from both the surface Internet and the Dark Web (Reitano et al., 2015), including Cloud9, Hydra and Silk Road 2.0 (van Wegberg & Verburgh, 2018). In 2017, Operation Bayonet was yet another law enforcement action in which the FBI and the Dutch Nation High Tech Crime Unit coordinated to end two of the most notorious marketplaces, AlphaBay and Hansa Market. While some interdiction operations affected the markets in predictable ways, other effects were less readily apparent.

Despite law-enforcement intervention, the overall tone across the Darknet marketplace after these raids remained generally positive (Lacson & Jones, 2016). While trust in the viability of the Darknet remained high, sentiment analysis of forums on marketplaces such as Agora and Evolution indicated a negative opinion of Silk Road. Correlating user sentiment with profile experience showed that experienced and inexperienced users held different opinions following the Silk Road disruption. Trust in Darknet markets expressed by "newbies" was unchanged since they were still trying to understand the social norms. Trust expressed by "junior members" was higher since they had gained some experience with the site and were actively networking with peers. Trust of “full members” decreased while trust of “senior members” remained high. Although the closure of Silk Road did not create widespread distrust in the Darknet, most users began to seek alternate marketplaces to conduct business, and the more senior their membership, the more likely they were to take their business elsewhere. Only a month following the FBI closure of the original Silk Road, Silk Road 2.0 came online. However, nearly 50% fewer vendors registered for Silk Road 2.0 compared to the original Silk Road (Lorenzo-Dus & Cristofaro, 2018).

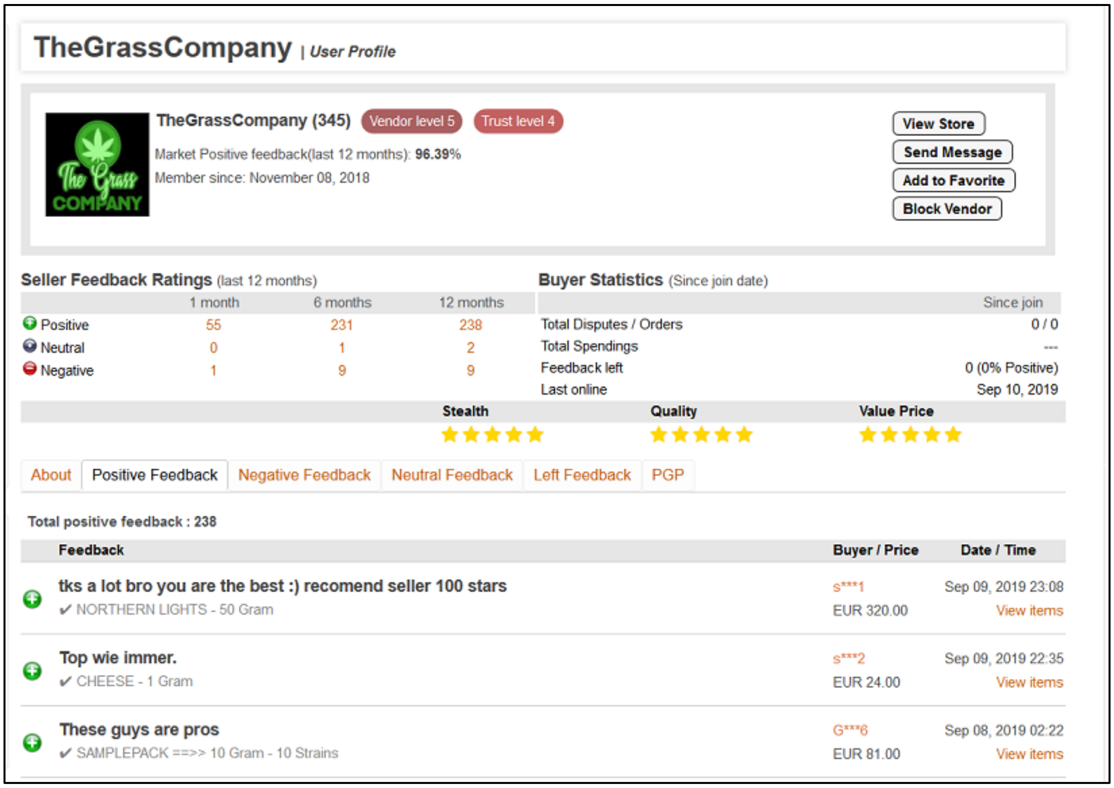
Despite the raids, overall Darknet market activity has continued to grow. The total number of drugs listed for sale increased from 18,000 to 46,000 in the two years following the closure of the original Silk Road site, and the number of marketplaces that listed drugs for sale also increased (Lacson & Jones, 2016). The fact that senior members maintained a generally positive sentiment while saying that they were looking to leave Silk Road seems contradictory. Although senior members of Silk Road may have hoped to salvage the health of their community through positive sentiment and reaffirmation to junior members, they clearly anticipated an unstable future in which vendors and buyers alike would seek new marketplaces.

After Operation Bayonet, researchers investigated its effects on Darknet marketplaces (van Wegberg & Verburgh 2018). While the Federal Bureau of Investigation ended AlphaBay, the Dutch National High-Tech Crime Unit waited to seize control of Hansa Market servers. The goal was to force an exodus of vendors and customers from AlphaBay to Hansa Market, where user account information and intelligence could be gathered before an eventual shutdown of the Hansa Market.

The operation had interesting impacts on the Darknet marketplace community. The AlphaBay intervention resulted in typical migration patterns seen in most marketplace raids: Vendors mainly relocated, copying their usernames and PGP public keys to maintain their established reputations. However, Hansa Market was hijacked by law enforcement and continued to run for almost a month before being shut down. Even though new user registration on Dream Market increased greatly after Operation Bayonet, only 2% of new users appeared to come from Hansa Market, while 40% came from AlphaBay. The researchers concluded that new vendors decided that the risk was too high to continue conducting business under their established names (van Wegberg & Verburgh, 2018).

C. TRUST

In a virtual world built on secrecy, how do buyers and sellers trust one another? Several factors contribute, the most important is online discourse (Lorenzo-Dus & Cristofaro, 2018). In the same way Amazon allows buyers to comment on customer service and the quality of products purchased, Darknet marketplaces provide a mechanism to rate and comment on transactions (Figure 3). A social hierarchy within the digital underground is established through assignment of "karma points" or experience levels (Lacson & Jones, 2016). Experience level reflects how long a vendor has been active as well as their overall involvement in the marketplace, while their trust level is determined by consumer-feedback ratings, comments, and use of escrow services. These scores indicate the vendor's degree of knowledge, status and trustworthiness within the community. A sense of community through the exclusivity of the Dark Web also enhances high levels of cohesiveness and comradeship among its users.



TheGrassCompany | *User Profile*

TheGrassCompany (345) Vendor level 5 Trust level 4

Market Positive feedback(last 12 months): **96.39%**

Member since: November 08, 2018

Seller Feedback Ratings (last 12 months)

	1 month	6 months	12 months
Positive	55	231	238
Neutral	0	1	2
Negative	1	9	9

Buyer Statistics (Since join date)

	Since join
Total Disputes / Orders	0 / 0
Total Spendings	---
Feedback left	0 (0% Positive)
Last online	Sep 10, 2019

Stealth	Quality	Value Price
★★★★★	★★★★★	★★★★★

Total positive feedback : 238

Feedback	Buyer / Price	Date / Time
tks a lot bro you are the best :) recomend seller 100 stars ✔ NORTHERN LIGHTS - 50 Gram	s***1 EUR 320.00	Sep 09, 2019 23:08 View items
Top wie immer. ✔ CHEESE - 1 Gram	s***2 EUR 24.00	Sep 09, 2019 22:35 View items
These guys are pros ✔ SAMPLEPACK ==>> 10 Gram - 10 Strains	G***6 EUR 81.00	Sep 08, 2019 02:22 View items

Figure 3. Example of a Darknet Marketplace Feedback System. Source: DarknetStats (2019).

D. DECEPTION

Societies thrive when everyone abides by the established rules of conduct, norms, and social behavior. However, people purposefully mislead and deceive one another for individual gain. On the Internet, detecting deception through nonverbal cues such as body language or voice inflection is rarely possible. For that reason, identifying indicators that suggest misleading language in text is especially important.

Online dating profiles are useful for studying deceptive text. Unsurprisingly, according to (Toma & Hancock, 2010), deception is often used to improve online personas to attract a mate. More specifically, their research showed that words and sentence structure could indicate the degree of truthfulness. They used a popular model for text deception including the frequency of first-person pronouns, exception words, negative-emotion words, and action words (Newman et al., 2003). A lack of first-person pronouns may indicate a user is trying to psychologically distance themselves from the act of lying because dishonesty creates negative emotions such as shame and guilt (Toma & Hancock, 2010). Similarly, exclusive words (i.e., “except,” “but,” and “without”) are used to categorize information; a false narrative will typically contain fewer exclusive words, and instead use simple action words that are easier to construct (Newman, Pennebaker et al., 2003). False narratives also tend to use fewer words overall. Negative-emotion words like “worthless,” “enemy,” and “hate” are clues to deception because they can suggest feelings of guilt and shame by an author engaged in deception (Hancock, Curry & Goorha, 2008). However, (Toma & Hancock, 2010) found that deceivers tend to use fewer negative-emotion words than those in truthful profiles. Simple motion words can be indicators of deception, since they remove some effort of having to fabricate information.

The same deception model was also applied to a large corpus of Enron corporate email by using a singular-value-decomposition (SVD) technique (Gupta, 2007). In Figure 4, each dot represents an email from the Enron corpus. Email (dots) closer to the origin of the fan (point D), have more indicators of deceptive language. Cue-words identified by the Pennebaker model were given weighted values determined by context, frequency, and sentence structure to better characterize the author’s sentiment and motivation. Of the four deception-cue attributes, first-person pronouns and exclusive words were the most useful for this corpus (Gupta, 2007). Messages in the Figure are ranked according to the Pennebaker Deception Model. Point A indicates higher use of first-person pronouns, point B indicates higher use of the exclusive word “or,” point C indicates higher use of the exclusive word “but.” A decrease in the presence of these cue words places an email closer to point D, indicating a higher likelihood of deception.

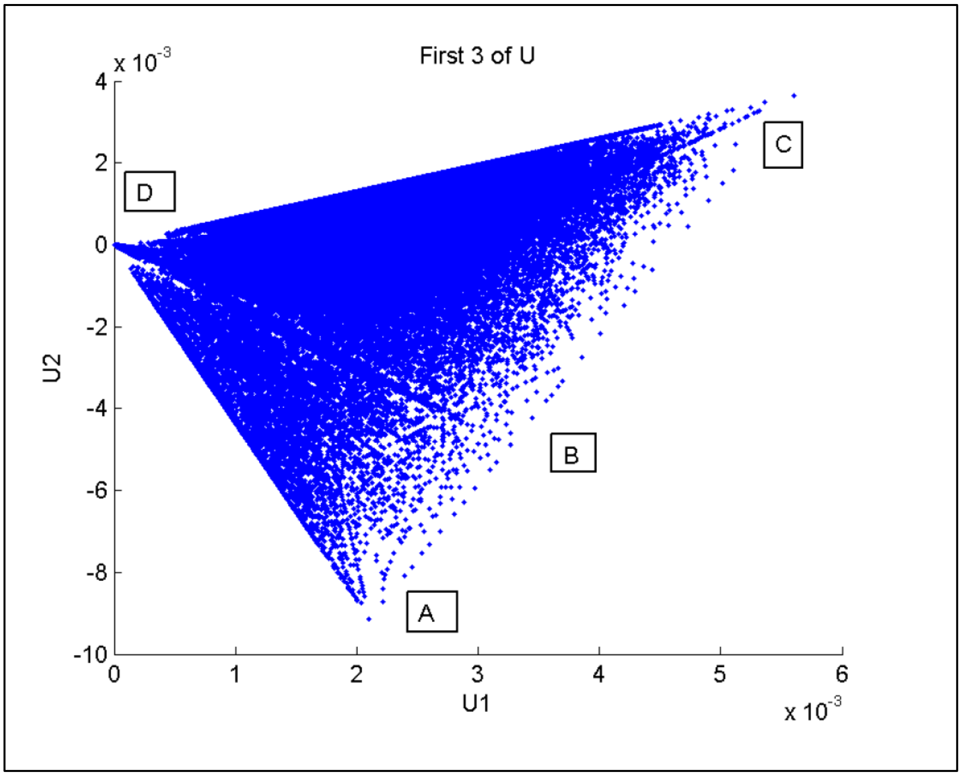


Figure 4. SVD Plot in Two Dimensions with One Point for Each Email. Source: Gupta (2007).

THIS PAGE INTENTIONALLY LEFT BLANK

III. PROBLEM DEFINITION AND ASSUMPTIONS

Given the clandestine nature of the Dark Web, identifying crime and threats and collecting evidence about them is challenging for law-enforcement and intelligence agencies. To accurately assess cyber threats, a taxonomy of markets in the digital underground and analysis of their capabilities is necessary.

Our research analyzed Dark Web samples with semantic and sentiment analysis focusing on vendors. Our goal was to find relationships in specific markets by examining vendor language to capture indications of truthfulness and deception. Deception exists on the Dark Web in many forms. Vendors who cheat buyers by not delivering goods or services are known as “rippers.” Rippers often try to deceive potential buyers by mimicking reputable established vendors, and forge accounts and feedback data to attract business to them and discourage business with competitors (Holt et al., 2016). Hacking attacks and exit scams also can occur when marketplaces go offline, taking with them all bitcoins held in escrow (Van Buskirk et al., 2016). On the regular Web, entities such as the Federal Trade Commission and the Better Business Bureau work to ensure trust between consumers and sellers by enforcing regulations, prosecuting online fraud, and providing a mediation service. These consumer protections are absent on the Dark Web, and thus both buyers and sellers are at an elevated risk of falling prey to scams. For this reason, Dark Web markets rely on measures of trust and vendor experiences, determined by buyer feedback and forum presence, to foster trust (Holt et al., 2016).

We assumed that each market category has probabilistic semantic fingerprints and that the product descriptions among top-ranked vendors in specific markets have similar semantic fingerprints; and that the vocabulary is different between vendors who participate in different product categories.

A related study used machine learning to automatically categorize products for sale on the Dark Web (Graczyk & Kinningham, 2015). They used term frequency and inverse document frequency to rate items, and principal component analysis (PCA), and support-vector machines (SVMs) to find patterns. Their findings gave a useful taxonomy of vendors and patrons of drugs on the Dark Web.

Another study collected and classified data from 12,542 Tor domains on the Dark Web into six product categories by applying a naïve Bayes classification method (Takaaki & Atsuo, 2019). Uniform resource locators (URLs) were extracted from each domain and a directed graph was built to capture relationships between product categories. The study found features and attributes occurred unique to each market.

Another study did natural-language processing for sentiment analysis of extremist and terrorist groups that use the Dark Web for communication (Chen, 2008). They used support-vector regression and support-vector machine learning. They identified and measured positive, neutral, and negative sentiments, thereby providing insight into the opinions and emotions of Jihadists and their radicalization processes (Chen, 2008).

IV. METHODOLOGY

This thesis combined descriptive and qualitative methods to assess vendors soliciting illicit goods and services on the Dark Web. Figure 5 shows our methodology. First, we selected and prepared datasets for analysis. Next, we scoped our data based on associations and relationships discovered during our analysis. This not only allowed us to gain a better understanding of the markets, but also helped us to identify interesting data subsets for use in our natural language processing experiments. We took an exploratory approach to identify similar characteristics among vendors and products based on trust level, vendor level, market revenue, feedback data, and cross-dataset vendor presence. Our goal was to assess a vendor’s trustworthiness to distinguish legitimate and illegitimate product listings. We used linguistic tools to analyze sentiment and truthfulness.

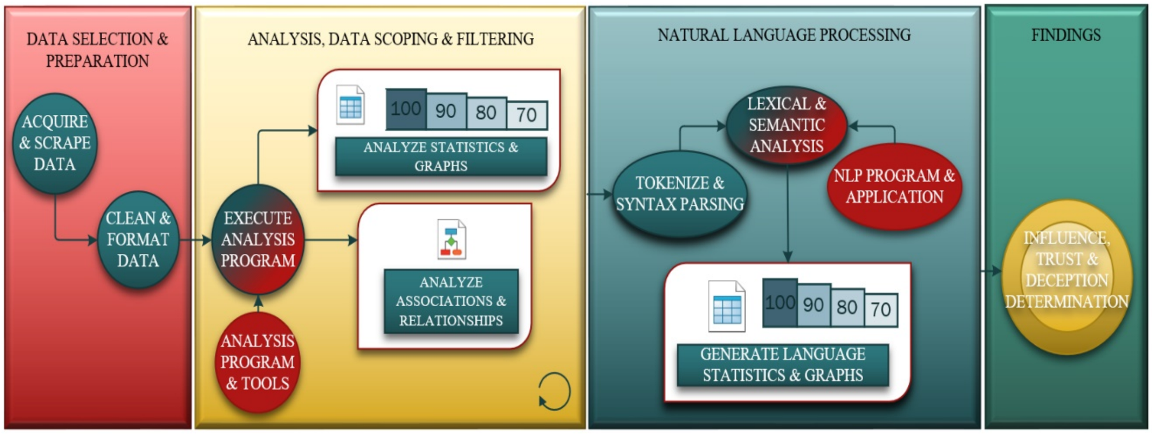

DATA SELECTION & PREPARATION
ACQUIRE & SCRAPE DATA
CLEAN & FORMAT DATA
ANALYSIS, DATA SCOPING & FILTERING
EXECUTE ANALYSIS PROGRAM
ANALYSIS PROGRAM & TOOLS
100 90 80 70
ANALYZE STATISTICS & GRAPHS
ANALYZE ASSOCIATIONS & RELATIONSHIPS
NATURAL LANGUAGE PROCESSING
TOKENIZE & SYNTAX PARSING
LEXICAL & SEMANTIC ANALYSIS
NLP PROGRAM & APPLICATION
100 90 80 70
GENERATE LANGUAGE STATISTICS & GRAPHS
FINDINGS
INFLUENCE, TRUST & DECEPTION DETERMINATION

Figure 5. Methodology Flowchart with Four Components

A. DATA SELECTION AND PREPARATION

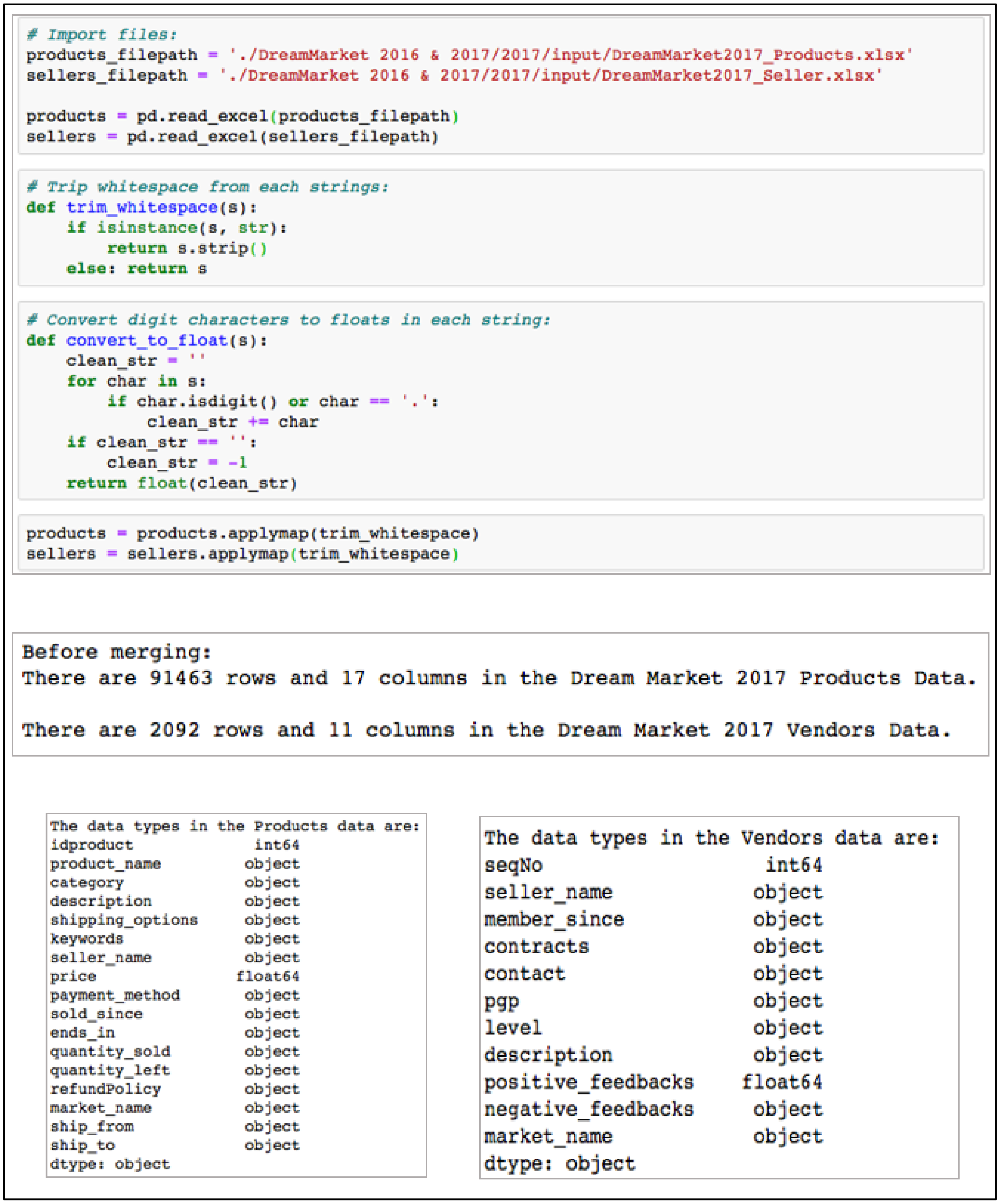
We used 2016 and 2017 Dream Market datasets provided by the AZSecure project (Alsayra, 2012) and the AlphaBay dataset hosted through the Gwern Darknet Market Archives (Branwen et al., 2019). These datasets serve as a reference for academic and professional researchers interested in exploring the Dark Web. They also provide analysis tools, recommendations for future exploration, and external links to organizations and individuals studying this data.

The AZSecure project also offers open-source datasets and tools to study Darknet markets, such as Web scrapes of Dream Market, a popular Darknet market (Alsayra, 2012). This data was collected using a special-purpose crawler developed in Python. The HTML scrapes were parsed using the "Beautiful Soup" Python package into files containing product-listings and vendor profiles. Each entry in the Dream Market product dataset contains product names, distinct categories, product descriptions, shipping information, price, and payment methods. Vendor profiles included seller usernames, member starting date, PGP public key, seller's descriptions, and feedback ratings. The datasets were available as a mySQL dump file which we exported into both CSV and Excel file formats.

The Gwern DNM Archives hosts an AlphaBay market dataset collected in late January 2017 using The Pirate Bay bit torrent software and ten separate AlphaBay accounts (Branwen et al., 2019). Once collected, McKenna and Goode used the "Beautiful Soup" Python package to compile the AlphaBay dataset into distinct CSV files.

The datasets we selected contained most elements required to make quantifications, categorizations, and comparisons, but a few things were missing. For example, the AlphaBay dataset lacked the vendor usernames for each product. However, we used Octoparse, an open-source Web-scraping tool, to get this information from the original HTML files.

The AlphaBay dataset had two files: product listings and customer feedback with sizes of 124 MB and 93.3 MB, respectively. We used the Pandas library (McKinney, 2010) to join the dataframes on the item numbers under which products and feedback were listed. Similarly, we joined the Dream Market data on vendor name. It was also divided into products and vendors with sizes of 101 MB and 22.5MB respectively. Figures 6 and 7 show how we used Pandas to join the Dream Market dataframes on vendor name. Most listings in both datasets were in the English language with a few exceptions. Our study used only entries written in English.



```python
# Import files:
products_filepath = './DreamMarket 2016 & 2017/2017/input/DreamMarket2017_Products.xlsx'
sellers_filepath = './DreamMarket 2016 & 2017/2017/input/DreamMarket2017_Seller.xlsx'

products = pd.read_excel(products_filepath)
sellers = pd.read_excel(sellers_filepath)
```

```python
# Trip whitespace from each strings:
def trim_whitespace(s):
    if isinstance(s, str):
        return s.strip()
    else: return s
```

```python
# Convert digit characters to floats in each string:
def convert_to_float(s):
    clean_str = ''
    for char in s:
        if char.isdigit() or char == '.':
            clean_str += char
    if clean_str == '':
        clean_str = -1
    return float(clean_str)
```

```python
products = products.applymap(trim_whitespace)
sellers = sellers.applymap(trim_whitespace)
```

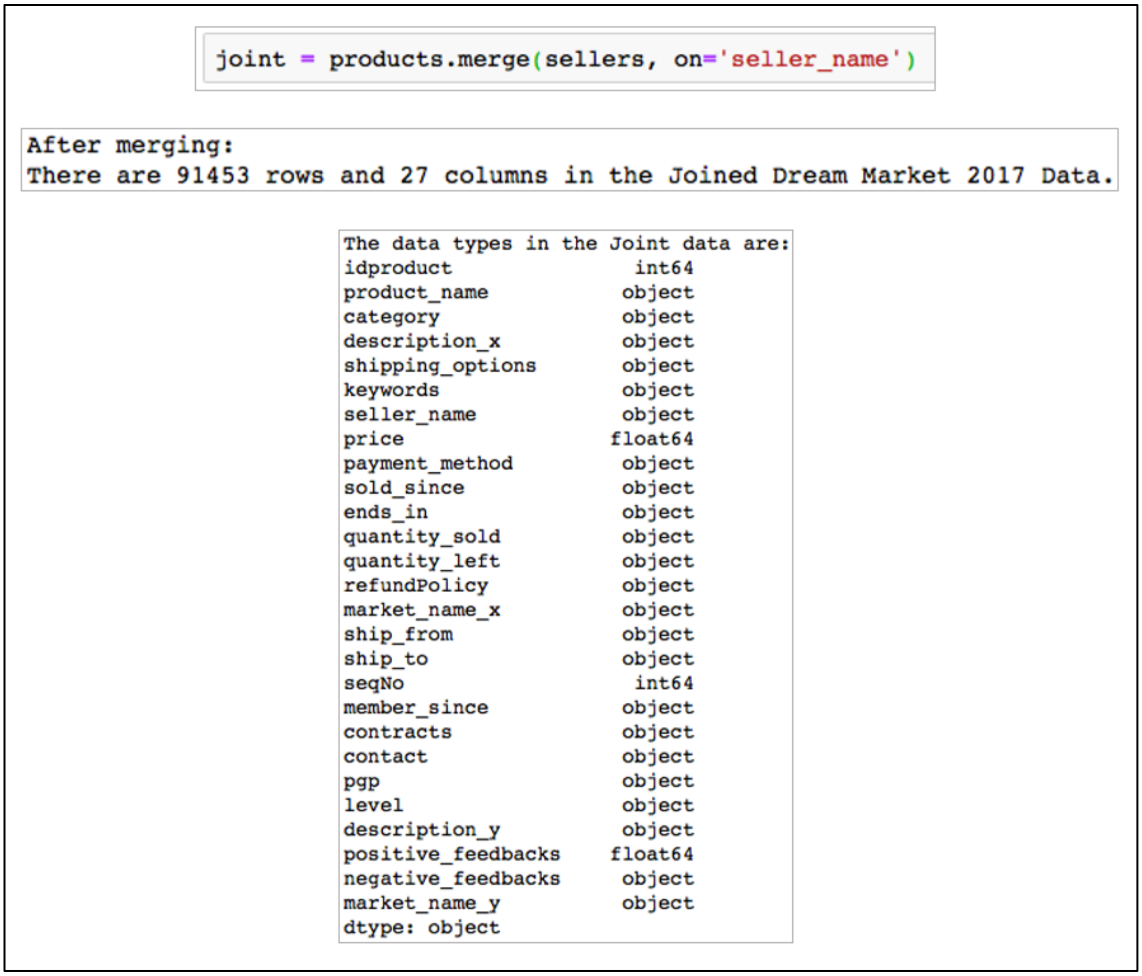
```
Before merging:
There are 91463 rows and 17 columns in the Dream Market 2017 Products Data.

There are 2092 rows and 11 columns in the Dream Market 2017 Vendors Data.
```

```
The data types in the Products data are:
idproduct                  int64
product_name              object
category                  object
description               object
shipping_options          object
keywords                  object
seller_name               object
price                    float64
payment_method            object
sold_since                object
ends_in                   object
quantity_sold             object
quantity_left             object
refundPolicy              object
market_name               object
ship_from                 object
ship_to                   object
dtype: object
```

```
The data types in the Vendors data are:
seqNo                   int64
seller_name            object
member_since           object
contracts              object
contact                object
pgp                    object
level                  object
description            object
positive_feedbacks    float64
negative_feedbacks     object
market_name            object
dtype: object
```

Figure 6.   Example Pandas Dataframe Cleaning and Merging: Before



```
joint = products.merge(sellers, on='seller_name')
```

```
After merging:
There are 91453 rows and 27 columns in the Joined Dream Market 2017 Data.
```

```
The data types in the Joint data are:
idproduct                int64
product_name            object
category                object
description_x           object
shipping_options        object
keywords                object
seller_name             object
price                  float64
payment_method          object
sold_since              object
ends_in                 object
quantity_sold           object
quantity_left           object
refundPolicy            object
market_name_x           object
ship_from               object
ship_to                 object
seqNo                    int64
member_since            object
contracts               object
contact                 object
pgp                     object
level                   object
description_y           object
positive_feedbacks     float64
negative_feedbacks      object
market_name_y           object
dtype: object
```

Figure 7. Example Pandas Dataframe Cleaning and Merging: After

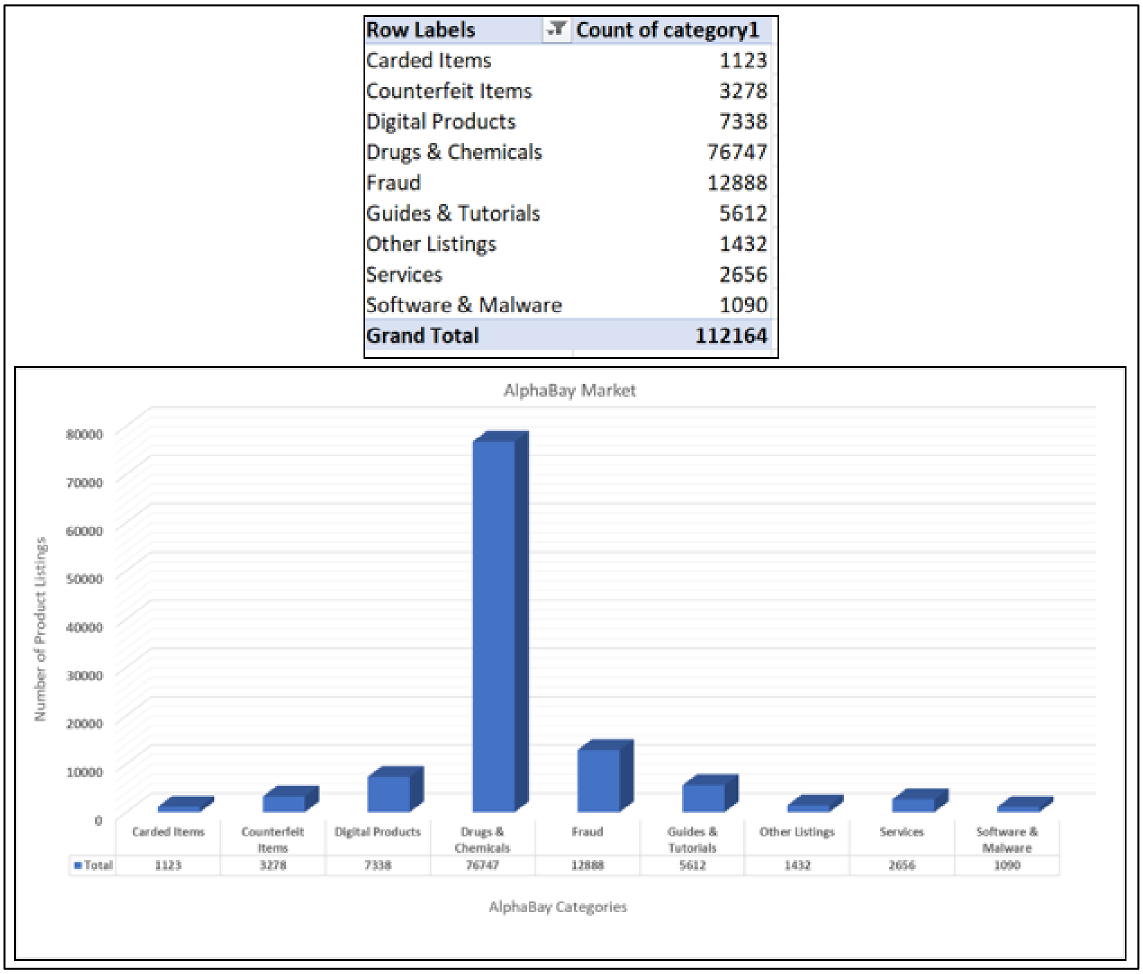
B. ANALYSIS PROGRAMMING, DATA SCOPING AND FILTERING

Once our data was cleaned and consolidated, we scoped and filtered it. We used several programs and tools to analyze our AlphaBay and Dream Market data, and to identify noteworthy data points, categories, vendors, and relationships.

Microsoft Excel has “pivot tables” which allow users to organize and summarize data from a more complex table. Also, “pivot charts” allow users to visualize statistics, patterns, and trends of pivot tables (Microsoft, 2020). Each row in our original AlphaBay and Dream Market datasets represented individual product listings with its own features and attributes. With the Excel pivot tables, we generated statistics across all products in the datasets. The pivot chart feature provided graphics of these statistics. We used these tools to familiarize ourselves with both datasets. Figure 8 is an example of a pivot table and its associated pivot chart, showing the number of listings by category in the AlphaBay dataset. Note that drugs and chemicals dominated the listings, and this was an important consideration during our natural-language processing phase. We made inquiries regarding:

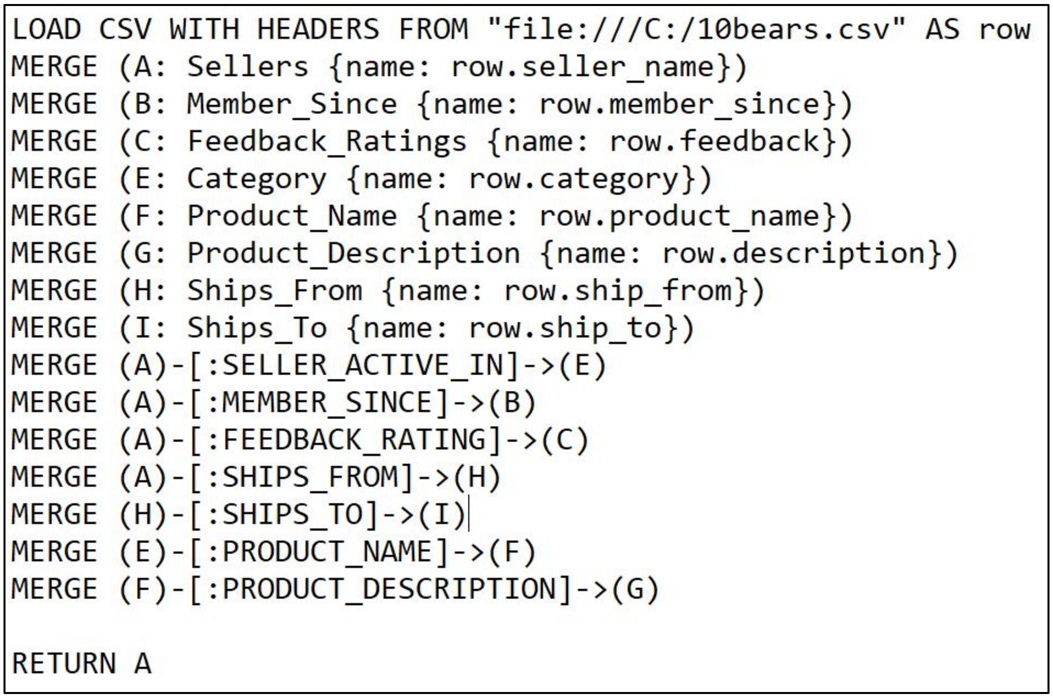
- The number of vendors in each category
- The revenue produced in specific categories and by specific vendors
- The vendor trust levels across categories
- The vendors likely using the same username in both marketplaces

NEO4J is a database-management and graphing system designed to organize data into nodes and relationships (Neo4j, 2020). Users create a schema template specific to the data features, so information is arranged for optimal exploration. We used NEO4J to identify features suggesting further investigation. We developed a data schema for the Alpha Bay and the Dream Market data scrapes. Subsequently, we matched vendor data by identifying common attributes and properties such as vendor trust level and product descriptions. Figure 9 shows the syntax for creating and querying the NEO4J database for a single vendor, and Figure 10 shows the database schema used for the datasets.



Row Labels	Count of category1
Carded Items	1123
Counterfeit Items	3278
Digital Products	7338
Drugs & Chemicals	76747
Fraud	12888
Guides & Tutorials	5612
Other Listings	1432
Services	2656
Software & Malware	1090
Grand Total	**112164**

Figure 8. AlphaBay Market Product Listings per Category



```
LOAD CSV WITH HEADERS FROM "file:///C:/10bears.csv" AS row
MERGE (A: Sellers {name: row.seller_name})
MERGE (B: Member_Since {name: row.member_since})
MERGE (C: Feedback_Ratings {name: row.feedback})
MERGE (E: Category {name: row.category})
MERGE (F: Product_Name {name: row.product_name})
MERGE (G: Product_Description {name: row.description})
MERGE (H: Ships_From {name: row.ship_from})
MERGE (I: Ships_To {name: row.ship_to})
MERGE (A)-[:SELLER_ACTIVE_IN]->(E)
MERGE (A)-[:MEMBER_SINCE]->(B)
MERGE (A)-[:FEEDBACK_RATING]->(C)
MERGE (A)-[:SHIPS_FROM]->(H)
MERGE (H)-[:SHIPS_TO]->(I)
MERGE (E)-[:PRODUCT_NAME]->(F)
MERGE (F)-[:PRODUCT_DESCRIPTION]->(G)

RETURN A
```

Figure 9. NEO4J Cypher Syntax Example for Querying the Database

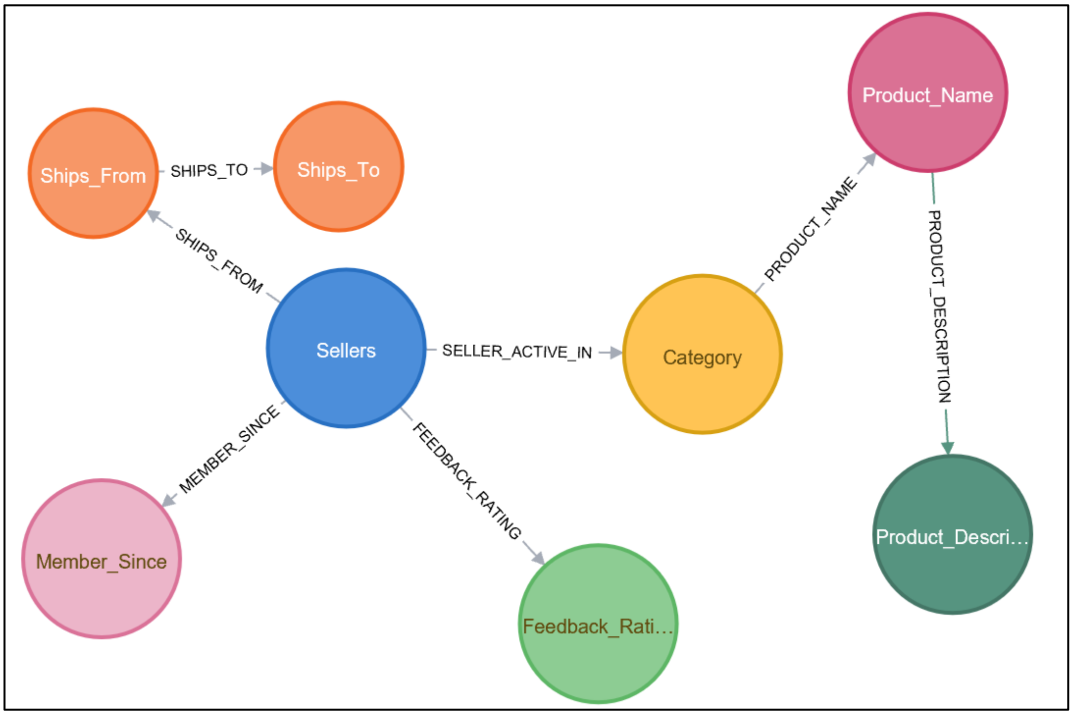


Figure 10. NEO4J Data Model Used for AlphaBay and Dream Market Datasets

We also used the graphing tools in the Python Matplotlib package (Hunter, 2007) to show marketplace characteristics such as:

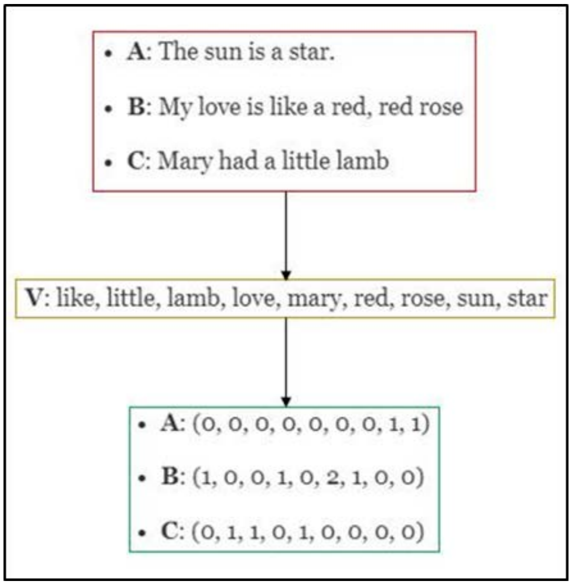
- Vendor feedback scores across data subsets (including categories, trust levels, and vendor levels). While the Dream Market gave an overall vendor feedback score, the AlphaBay feedback scores were for individual products based on customer reviews, similar to Amazon's star rating system. However, we could still determine a vendor's overall feedback score in AlphaBay by calculating the mean and median score across all products offered by that vendor. This was important to establish comparable rating standards to identify the poor, average, and high-quality vendors in our datasets.
- Products per vendor. We investigated to what degree vendors in subsets defined by categories, trust levels, and vendor levels sought to diversify their listings.
- Products versus the vendor's feedback score. We correlated the business model (limited versus diversified) and the vendor's feedback score, trust level, and vendor level.
- Feedback comments versus sales. We investigated why some vendors received more feedback than others, considering whether they used different marketing language or specifically requested feedback.

C. NATURAL-LANGUAGE PROCESSING

The first step in natural-language processing of the data was to create histograms of key terminology within a marketplace. We had a bag-of-words model, which uses the frequency of a word in documents to estimate that word's value to the documents (Brownlee, 2017). It required extracting words from the product and vendor pages. Additionally, we removed punctuation, and filtered out stop words which are common English words (i.e., "the," "is," and "or") not helpful in market language.

We tried to identify unique vendors active in both AlphaBay and Dream Market by matching vendors across markets based on their product descriptions. Content-based recommender systems use a bag-of-words approach and sentiment analysis to quantitatively compare bodies of text and measure their similarity. This works best when each word is assigned a numerical value, or “weight,” that represents its importance to a corpus. CountVectorizer and TF-IDFVectorizer (term frequency – inverse document frequency) are two such methods for calculating weights on words. We tested both methods to decide which was best for vendor profiling.

CountVectorizer (Pedregosa et al., 2011) counts all distinct non-stop words used across all documents, in our case, words used in product descriptions, and placed them in a dictionary. This dictionary was used to generate word counts for each product description found in the corpus (Banik, 2018). Figure 11 shows CountVectorizer’s application to three short “documents,” A, B, and C. The red square contains three documents, A, B, and C. The gold square contains the vector dictionary created from A, B, and C, after the removal of stop words. The green square is the count-vectorized representation of each document.



- **A**: The sun is a star.
- **B**: My love is like a red, red rose
- **C**: Mary had a little lamb

V: like, little, lamb, love, mary, red, rose, sun, star

- **A**: (0, 0, 0, 0, 0, 0, 0, 1, 1)
- **B**: (1, 0, 0, 1, 0, 2, 1, 0, 0)
- **C**: (0, 1, 1, 0, 1, 0, 0, 0, 0)

Figure 11. CountVectorizer Applied to Three Documents.
Source: Banik (2018).

TF-IDFVectorizer (Pedregosa et al., 2011) was another tool we used to analyze vendor descriptions. The term frequency of a term in a document is the count of that term. Inverse document frequency is the number of documents in which a term occurs. TF-IDFVectorizer used the following weighting formula (Banik, 2018):

$$w_{i,j} = tf_{i,j} \times log(\frac{N}{df_i})$$

where:

$w_{i,j}$ is the weight of word i in document j

$tf_{i,j}$ is the frequency of term i in document j

df_i is the number of documents that contain the term i

N is the total number of documents

We created histograms of the top 25 words according to average TF-IDF score across product descriptions from the software, malware, drug, and chemical markets. The top words that are unique to specific categories and vendors reveals the language of those subsets. We used pairwise cosine similarity to calculate the angle between vectors of different products, which allowed us to measure how similar the products were. Figure 12 and the formula that follows explain cosine similarity.

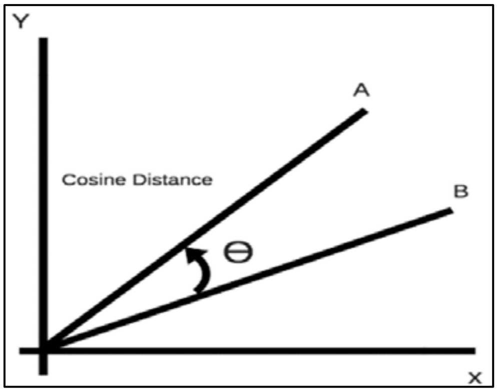


Figure 12. Cosine Distance of Two Vectors. Source: Banik (2018).

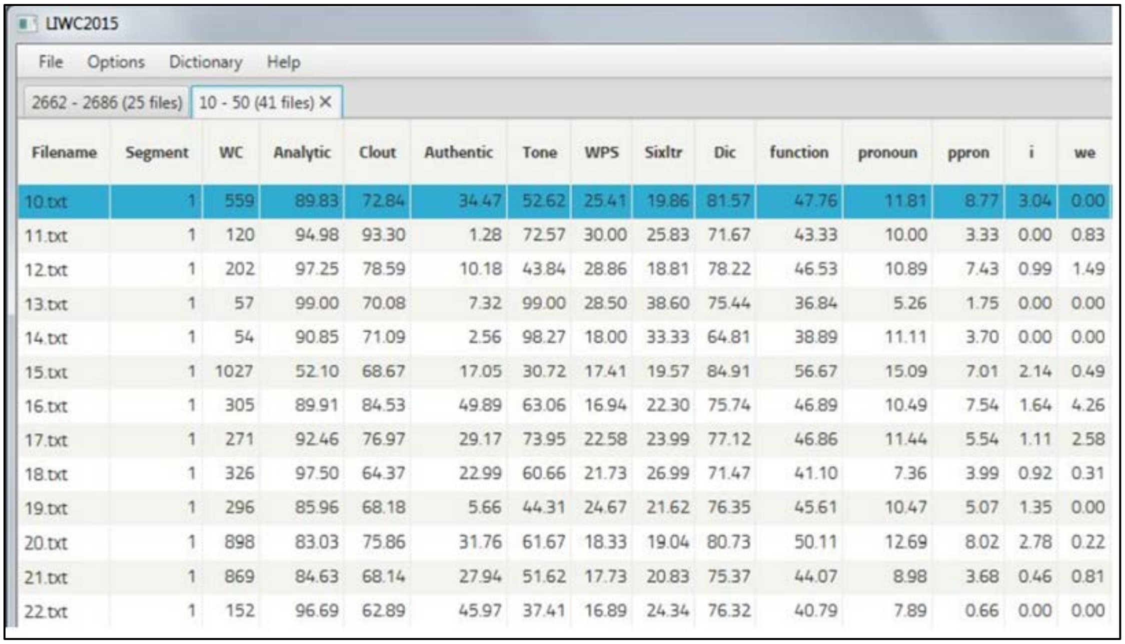
$$cosine\ \theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\|\ \|\vec{b}\|}$$ (Perone, 2013)

Cosine similarity of document vectors ranges 0 and 1. Identical documents have a cosine similarity of 1, and documents with completely different words have a cosine similarity of 0 (Banik, 2018).

We wrote a program in Python that computed the top ten Dream Market product-vendor pairs with the highest cosine similarity. We then manually examined the results to determine how well this method identified products sold by the same vendor on different marketplaces.

D. TEXT CLASSIFICATION

Additional features were identified in the product descriptions using the Linguistic Inquiry and Word Count (LIWC) tool, a text-analysis tool that estimates the emotional, cognitive, and structural features of an author's written text (Pennebaker et al., 2015). LIWC compared words in the product descriptions to tailored word lists for psychologically meaningful categories. These category word lists included negative and positive emotions, use of personal pronouns, conjunctions, verbs, and prepositions. The LIWC2015 word lists were developed from sources including blog posts, forums, Facebook, novels, and student writing with numbers, punctuation, and emoticons (Pennebaker et al., 2015). Figure 13 shows an example output for multiple text files analyzed by LIWC. Results in each category are a percentage of total words used in each document. The 10.txt document has a total word count ("WC") of 559. 11.81% of all words in the 10.txt document were identified as "pronouns."



Filename	Segment	WC	Analytic	Clout	Authentic	Tone	WPS	Sixltr	Dic	function	pronoun	ppron	i	we
10.txt	1	559	89.83	72.84	34.47	52.62	25.41	19.86	81.57	47.76	11.81	8.77	3.04	0.00
11.txt	1	120	94.98	93.30	1.28	72.57	30.00	25.83	71.67	43.33	10.00	3.33	0.00	0.83
12.txt	1	202	97.25	78.59	10.18	43.84	28.86	18.81	78.22	46.53	10.89	7.43	0.99	1.49
13.txt	1	57	99.00	70.08	7.32	99.00	28.50	38.60	75.44	36.84	5.26	1.75	0.00	0.00
14.txt	1	54	90.85	71.09	2.56	98.27	18.00	33.33	64.81	38.89	11.11	3.70	0.00	0.00
15.txt	1	1027	52.10	68.67	17.05	30.72	17.41	19.57	84.91	56.67	15.09	7.01	2.14	0.49
16.txt	1	305	89.91	84.53	49.89	63.06	16.94	22.30	75.74	46.89	10.49	7.54	1.64	4.26
17.txt	1	271	92.46	76.97	29.17	73.95	22.58	23.99	77.12	46.86	11.44	5.54	1.11	2.58
18.txt	1	326	97.50	64.37	22.99	60.66	21.73	26.99	71.47	41.10	7.36	3.99	0.92	0.31
19.txt	1	296	85.96	68.18	5.66	44.31	24.67	21.62	76.35	45.61	10.47	5.07	1.35	0.00
20.txt	1	898	83.03	75.86	31.76	61.67	18.33	19.04	80.73	50.11	12.69	8.02	2.78	0.22
21.txt	1	869	84.63	68.14	27.94	51.62	17.73	20.83	75.37	44.07	8.98	3.68	0.46	0.81
22.txt	1	152	96.69	62.89	45.97	37.41	16.89	24.34	76.32	40.79	7.89	0.66	0.00	0.00

Figure 13. LIWC2015 Output Variable Information. Source: Pennebaker et al. (2015).

THIS PAGE INTENTIONALLY LEFT BLANK

V. RESULTS

A. DATA ANALYSIS AND SCOPING

1. AlphaBay

The AlphaBay dataset contains information on over 6,000 vendors of over 270,000 illicit goods and services on the Dark Web. It includes vendor feedback ratings, product summaries, product categories and sub-categories, vendor experience level, quantity sold, price, and vendor trust level. However, it does not include the vendor usernames. To get these, we used Octoparse, a publicly available Web-scraping tool, on the HTML text of each product page. The tables and charts in this chapter were generated with Microsoft Excel pivot table and charts; the relational graphs were generated with NEO4J.

The AlphaBay dataset includes the number of products sold, allowing us to show the distribution of trafficked goods and services (Figure 14): 40% of the transactions involved drugs and chemicals and 25% involved fraud. As seen in Figure 15, of the 6,083 vendors active in the AlphaBay market, 4,455 (73%) advertised drugs and drug-related paraphernalia. While 75% of vendors specialized their goods and services within one category, 15% were active in two categories, and 10% were active in three or more categories.

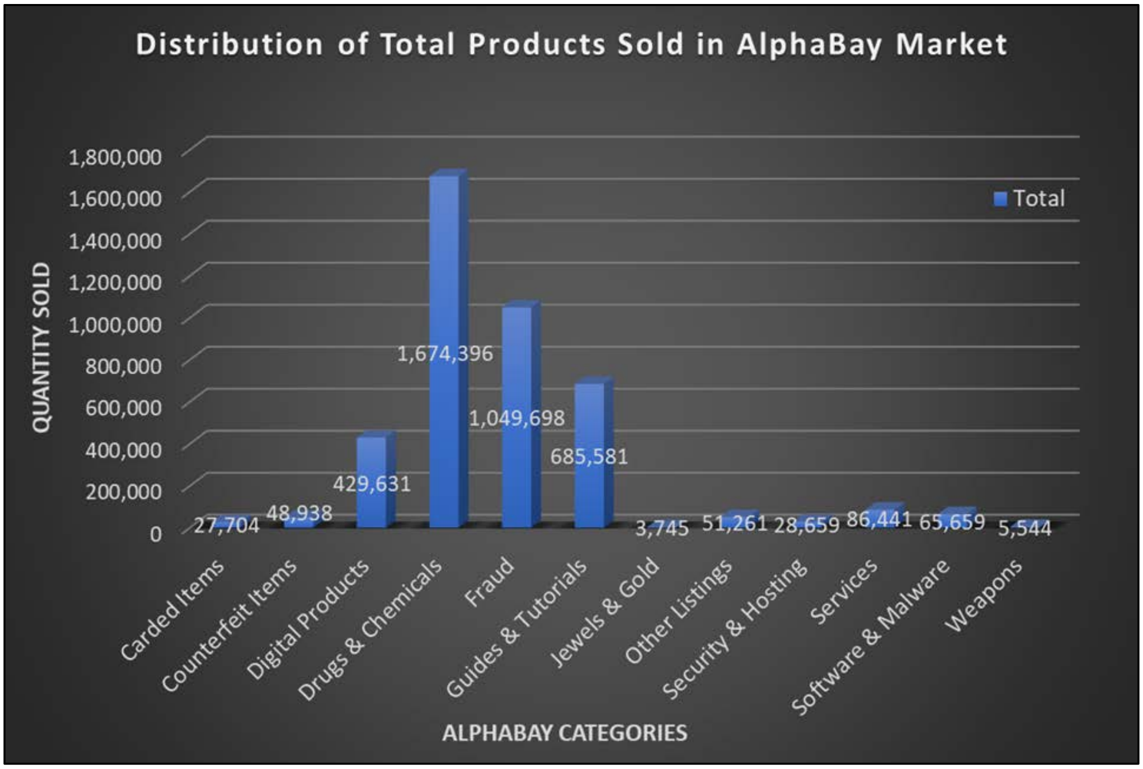


Figure 14. Distribution of Total Products Sold in AlphaBay Market

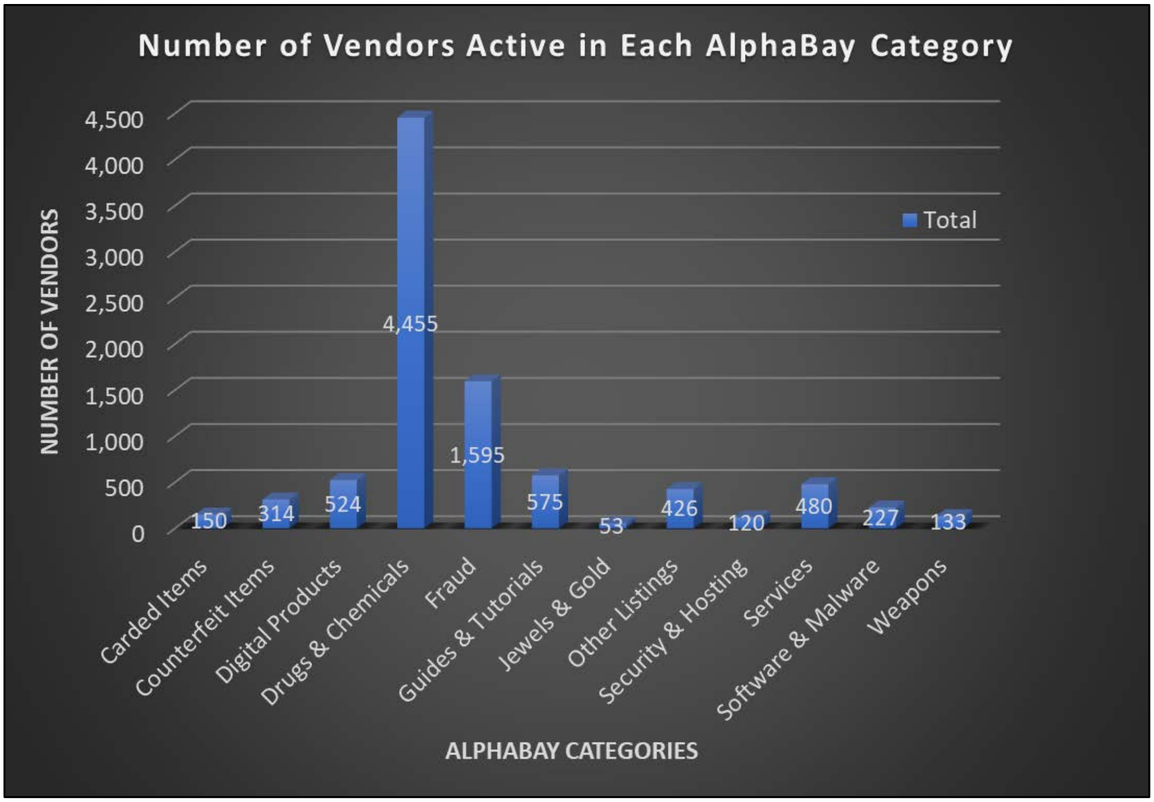


Figure 15. Number of Vendors Active in Each AlphaBay Category

Each market is typically dominated by a few of the vendors. For instance, Figure 16 shows that 227 vendors active in the software and malware category accounted for over 65,000 transactions, but ten vendors had 64.7% of the transactions (42,482). Similarly, the right side of the Figure shows that only ten vendors in the fraud market accounted for over 48% of total transactions. Curiously, 1,139 vendors have items listed for sale in the fraud category but have not recorded any transactions. Despite this inactivity, the average vendor level and trust level for these vendors were 2 and 5, respectively, demonstrating that these numbers were earned through activity and sales in market categories other than fraud.

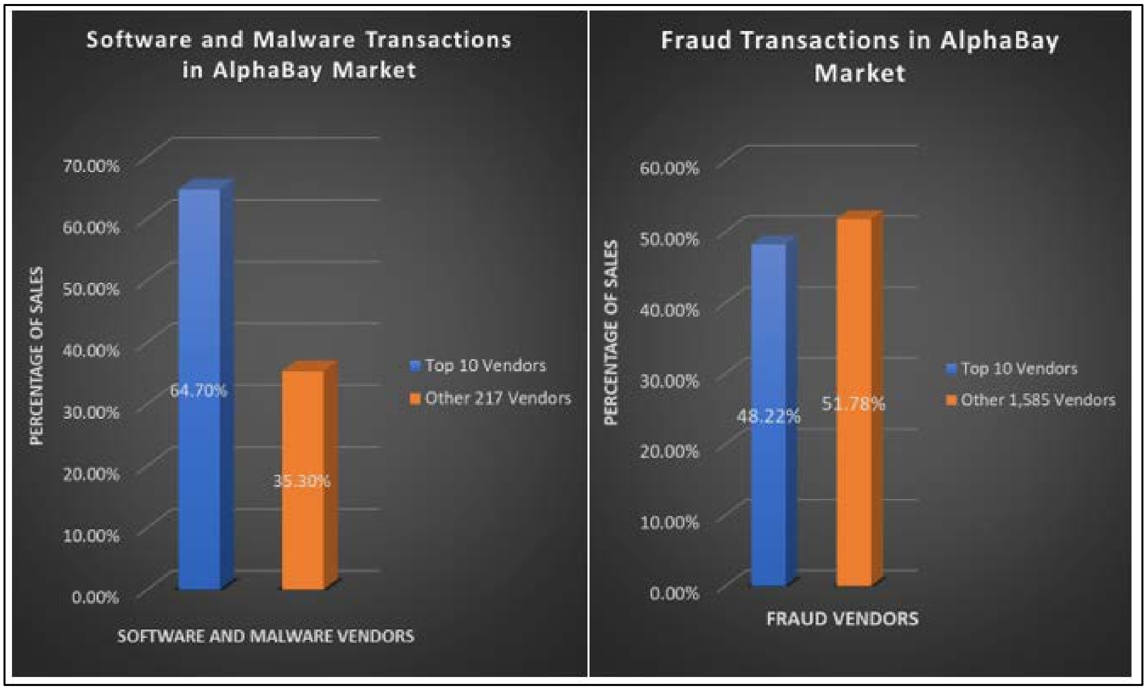


Figure 16. Software, Malware, and Fraud Transactions in AlphaBay Market

Over U.S. $338 million was spent for illicit goods and services on the AlphaBay. Narcotics alone produced nearly $240 million in revenue (Figure 17); fraud produced close to $82 million in revenue. The top five vendors in the AlphaBay market overall (Figure 18) specialized in either drugs or fraud, profiting close to $78 million. One vendor “somecvvvendor” grossed over $33 million selling services for fraud.

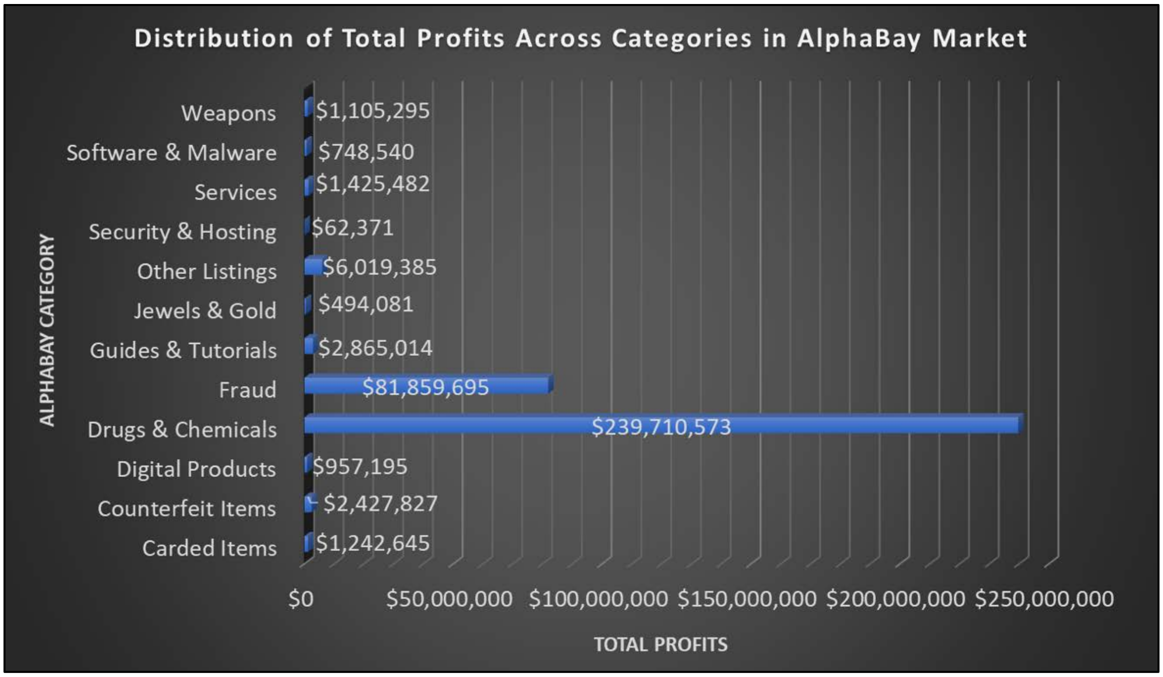


Figure 17. Distribution of Total Profits across Categories in AlphaBay Market

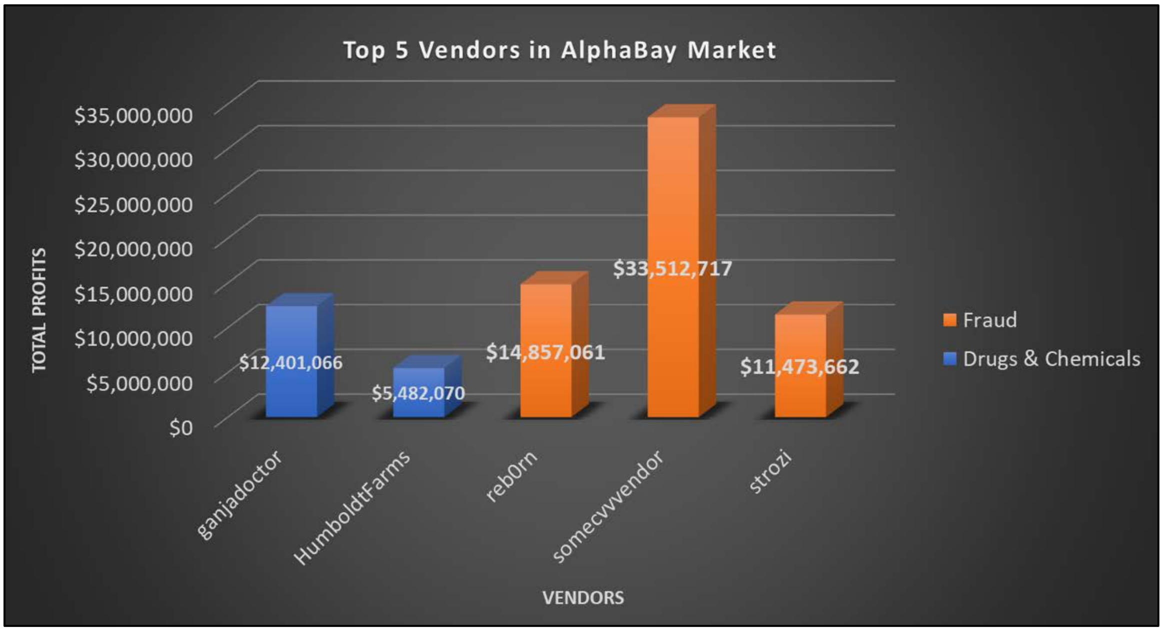
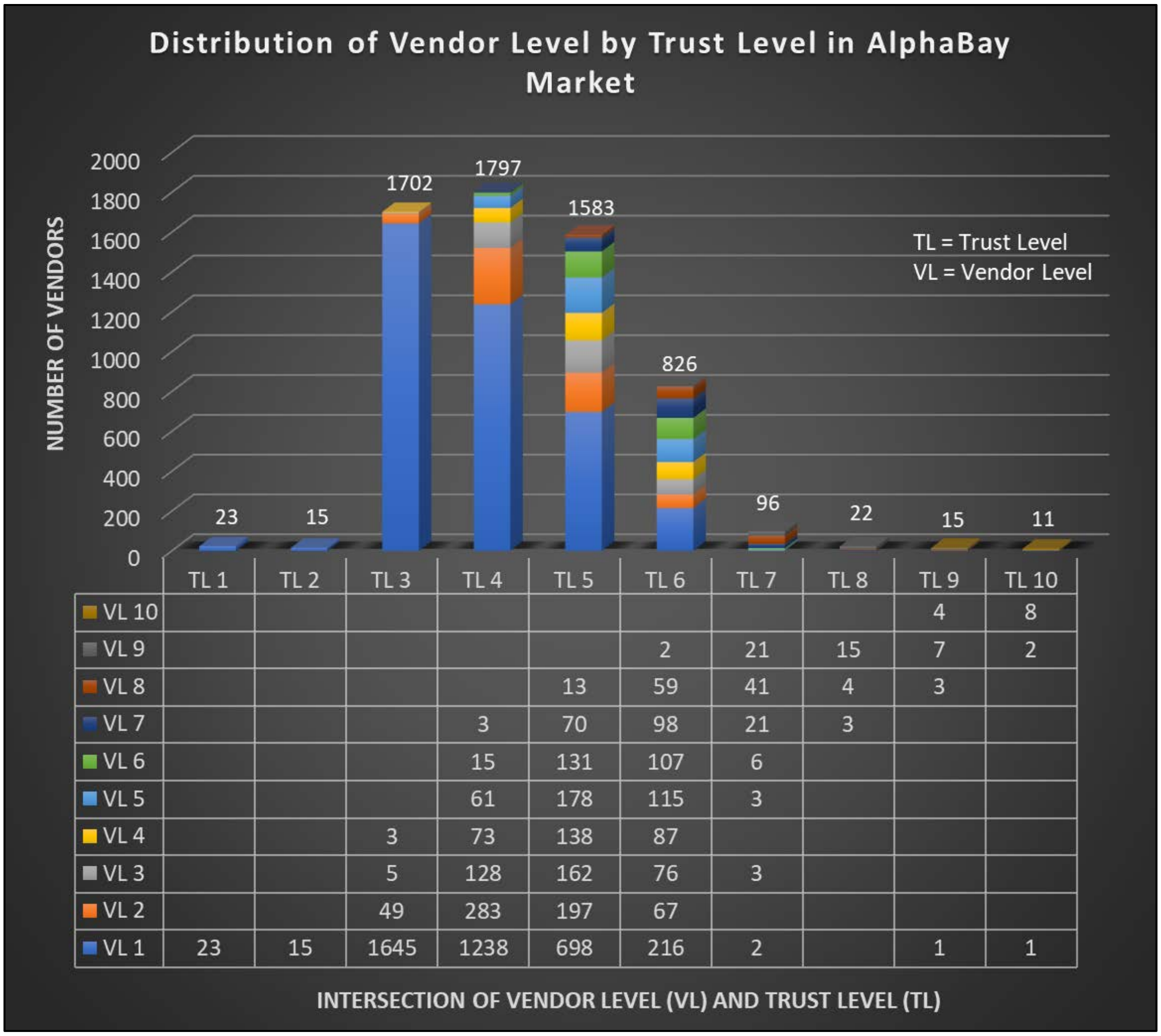


Figure 18. Top Vendors in AlphaBay Market

Each vendor in the AlphaBay market is rated with both a trust level and experience level. High values for these scores suggest legitimacy, reliability, and quality of advertised products and services.

Figure 19 shows the distribution of trust levels and experience levels for vendors in the AlphaBay market. Each column represents a trust level, and colors represent the experience levels. The figure shows that 97% of all vendors (5,908) had a trust level between three and six; of those, 62% (3,797) had a vendor level of only one. This suggests that most vendors are either new or unsuccessful at attracting buyers. Moreover, this graph shows that trust in the 3–6 range is not a good indicator of vendor credibility. A vendor with only a few sales can get a decent trust level without a long history of credible transactions because trust level is based on average feedback ratings, and a high feedback rating on a few sales can artificially inflate trust levels. Taking a closer look at trust levels 8–10, we find 48 total vendors. Of those vendors, only 8 have both a trust level and experience level of 10. Vendors with both a high trust and experience levels were almost exclusively active in the fraud and drug categories.



	TL 1	TL 2	TL 3	TL 4	TL 5	TL 6	TL 7	TL 8	TL 9	TL 10
VL 10									4	8
VL 9						2	21	15	7	2
VL 8					13	59	41	4	3	
VL 7				3	70	98	21	3		
VL 6				15	131	107	6			
VL 5				61	178	115	3			
VL 4			3	73	138	87				
VL 3			5	128	162	76	3			
VL 2			49	283	197	67				
VL 1	23	15	1645	1238	698	216	2		1	1

Figure 19. Distribution of Vendor Level by Trust Level in AlphaBay Market

Using the “create” and “merge” functions of the NEO4J cypher syntax, we connected the main product category nodes to their subcategory nodes based on the primary, secondary, and tertiary category labels assigned to each product. The AlphaBay dataset had 12 main categories and 49 subcategories as seen in Figure 20. Through graphical depiction, we discovered relationships across different subcategories. For instance, in Figure 20, both the “services,” and the “guides and tutorial” subcategories advertised products and services related to “social engineering.” Likewise, clothing products were listed in both the “counterfeit” and “carded items” subcategories. In this case, “carded items” refers to clothing purchased with a stolen credit card or hacked pay account. Similar affiliations were discovered when graphically depicting vendor specific data.

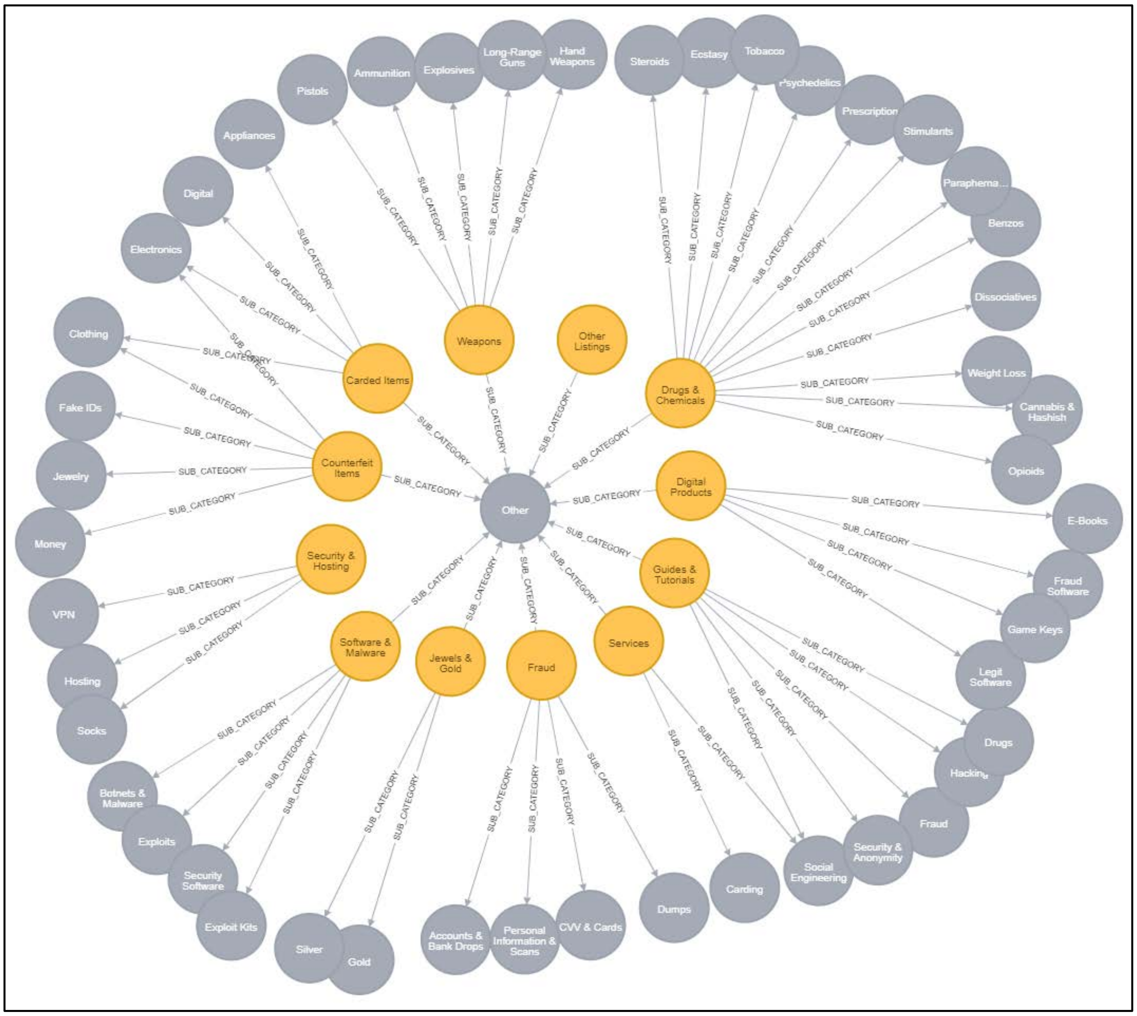


Figure 20. NEO4J Illustration of How Products and Services Were Categorized in AlphaBay

2. Dream Market

The Dream Market dataset covers 2016 and 2017 including nearly 131,000 products from over 2,000 vendors. This period includes Operation Bayonet which seized and shut down the AlphaBay market. The Dream Market 2017 Seller and Product data contain information on vendor names, PGP keys, dates joined, market categories, market descriptions, product descriptions, shipping information, prices, and feedback ratings, though quantity sold is missing.

The distribution of vendors and products in the Dream Market resembled that of AlphaBay: 89% (1,853/2,093) of all Dream Market vendors advertised drug-related products. Figure 21 shows the top 10 Dream Market sectors by number of vendors.

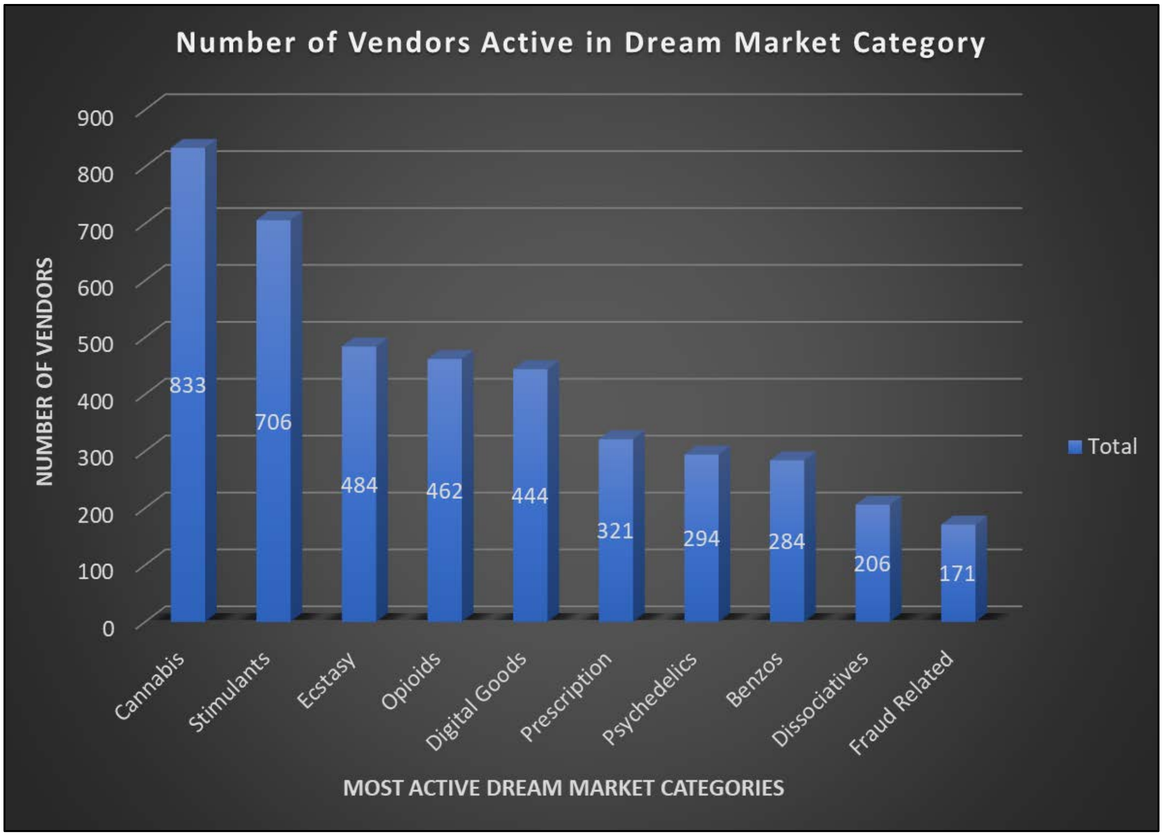
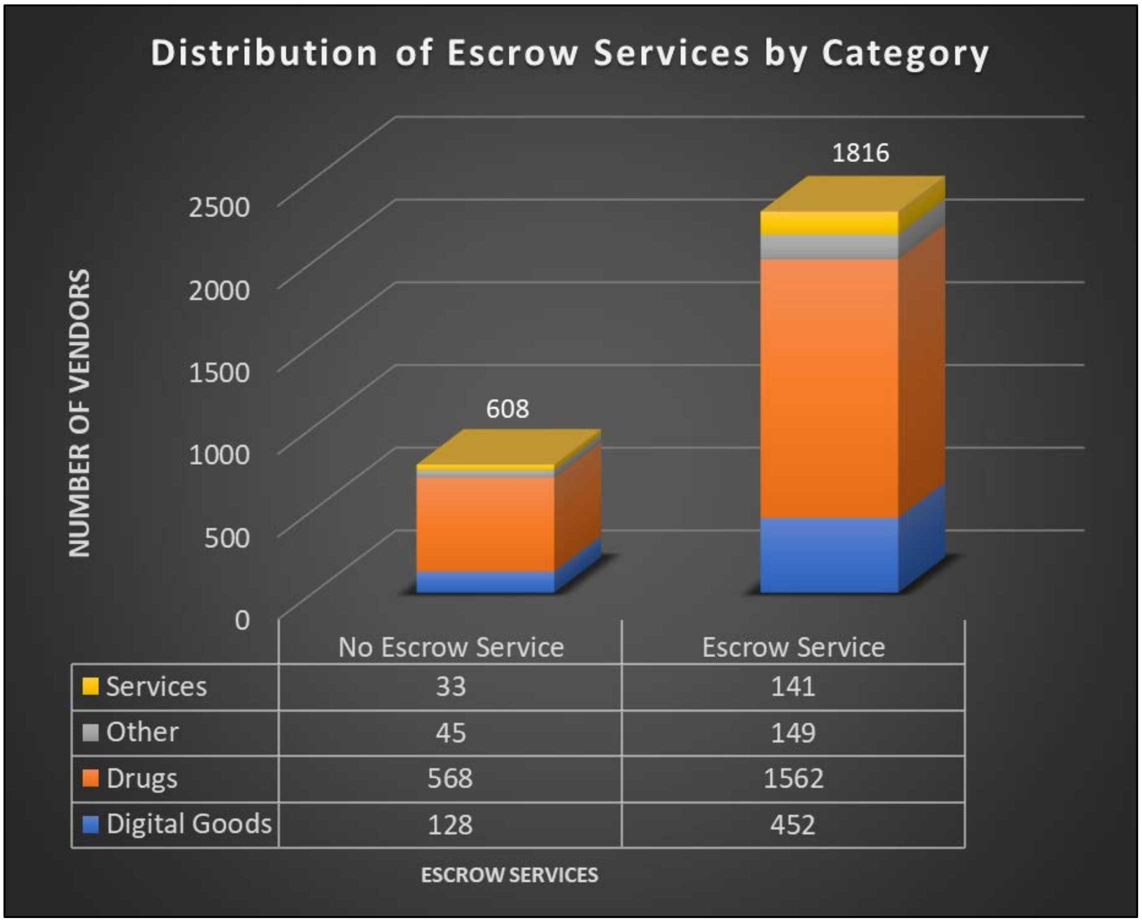


Figure 21. Number of Vendors Active in each Dream Market Category

The Dream Market equivalent of trust level is the positive feedback rating, an average of the scores given by consumers to a vendor. The high averages could be due to Dream Market incentivizing good customer service as well as Dream Market's vetting process for new vendors (Crooks, 2020). Another factor could be the better escrow services offered for the goods and services; of the 91,461 items listed on the market, 75,780 included an escrow service to ensure customer satisfaction. Figure 22 shows the number of vendors offering escrow services per category.



	No Escrow Service	Escrow Service
🟨 Services	33	141
⬜ Other	45	149
🟧 Drugs	568	1562
🟦 Digital Goods	128	452

Figure 22. Distribution of Escrow Services by Category

NEO4J enabled us to store, query and visualize vendor-specific information. For instance, we queried the database for the vendor “10Bears.” Figure 23 shows which markets 10Bears was active in and the products and services they solicited. The database also gives information such as shipping policy, feedback rating, and the date of membership. Product descriptions written by 10Bears were displayed and linked by NEO4J to the proper products and services. NEO4J can be used to query vendors linked by data such as ratings.

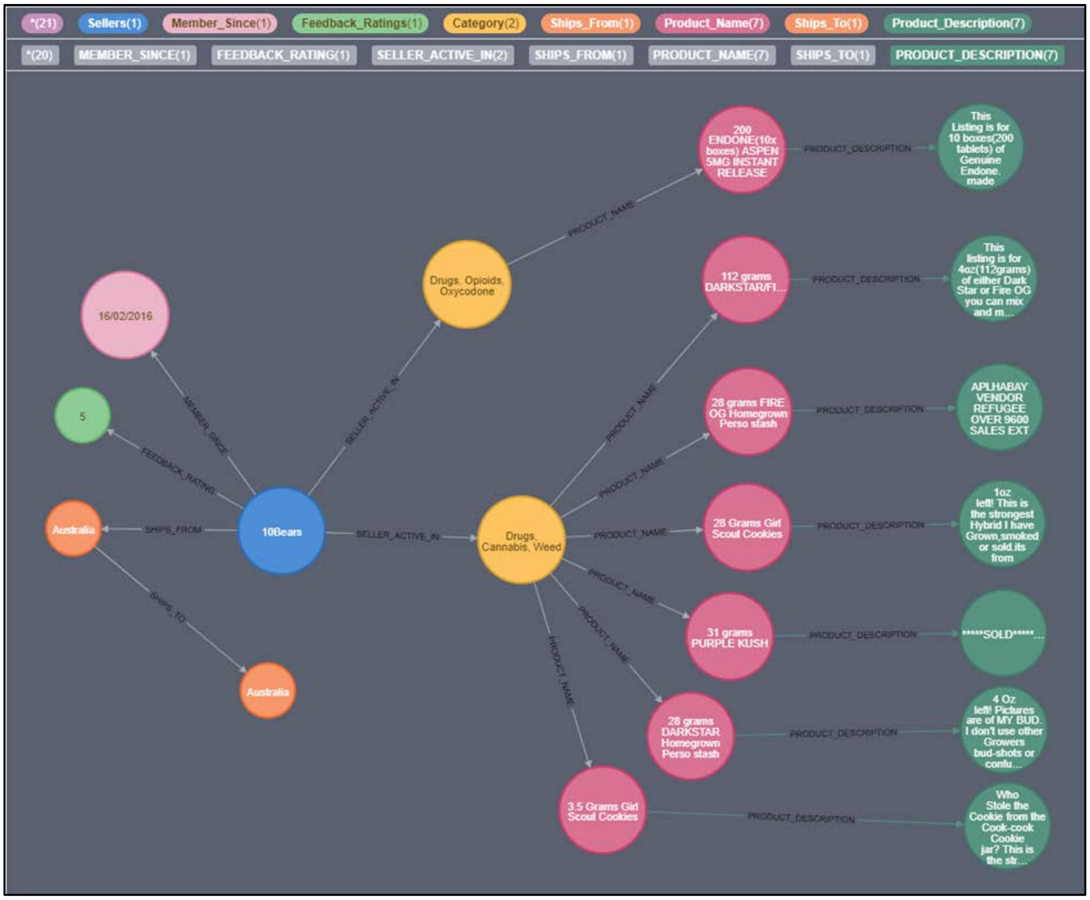

*(21)
Sellers(1)
Member_Since(1)
Feedback_Ratings(1)
Category(2)
Ships_From(1)
Product_Name(7)
Ships_To(1)
Product_Description(7)
*(20)
MEMBER_SINCE(1)
FEEDBACK_RATING(1)
SELLER_ACTIVE_IN(2)
SHIPS_FROM(1)
PRODUCT_NAME(7)
SHIPS_TO(1)
PRODUCT_DESCRIPTION(7)
16/02/2016
5
Australia
10Bears
Australia
MEMBER_SINCE
FEEDBACK_RATING
SHIPS_FROM
SHIPS_TO
SELLER_ACTIVE_IN
SELLER_ACTIVE_IN
Drugs, Opioids, Oxycodone
Drugs, Cannabis, Weed
PRODUCT_NAME
200 ENDONE(10x boxes) ASPEN 5MG INSTANT RELEASE
112 grams DARKSTAR/FI…
28 grams FIRE OG Homegrown Perso stash
28 Grams Girl Scout Cookies
31 grams PURPLE KUSH
28 grams DARKSTAR Homegrown Perso stash
3.5 Grams Girl Scout Cookies
PRODUCT_DESCRIPTION
This Listing is for 10 boxes(200 tablets) of Genuine Endone. made
This listing is for 4oz(112grams) of either Dark Star or Fire OG you can mix and m…
APLHABAY VENDOR REFUGEE OVER 9600 SALES EXT
1oz left! This is the strongest Hybrid I have Grown,smoked or sold.its from
*****SOLD*****…
4 Oz left! Pictures are of MY BUD. I don't use other Growers bud-shots or confu…
Who Stole the Cookie from the Cook-cook Cookie jar? This is the str…

Figure 23. NEO4J Illustration of a Vendor Node and Relationships with Category Nodes

3. The 198 Migrant Vendors

Figure 24 plots the number of new vendors to Dream Market over time. This Figure indicates that membership increased at a slow and steady rate before a sudden jump by 65% (815) in the 3rd quarter of 2017 with the FBI shutdown of AlphaBay. Of the 815 new Dream Market registrants, 198 vendor names exactly matched those of AlphaBay vendors, whom we dubbed the "198 migrants," although it is likely that other vendors migrated from AlphaBay while changing names. However, vendor-name matches gave the most reliable mechanism for identifying migrants based on our dataset. This was supported by nearly 25% of the 198 migrant vendors referring to their previous AlphaBay business in their Dream Market product descriptions. The Quarter 4 decrease reflects the truncation of data collection, not a drop in the new-member registration rate. Of those migrants from AlphaBay, 149 were in the drugs and chemicals market (Figure 25).

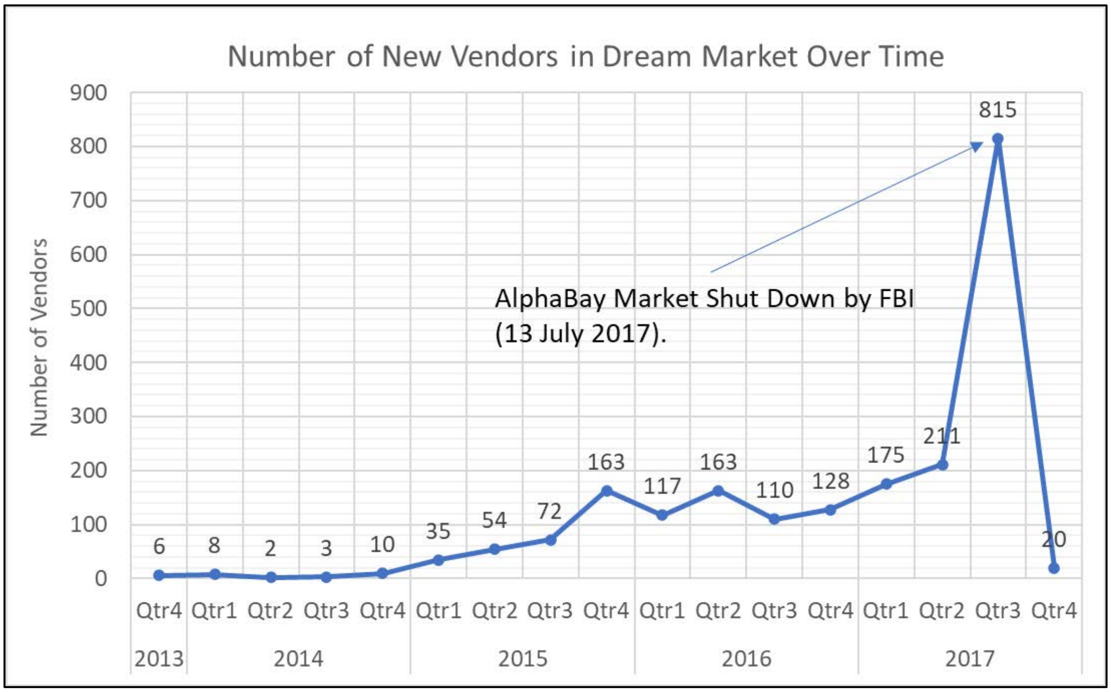
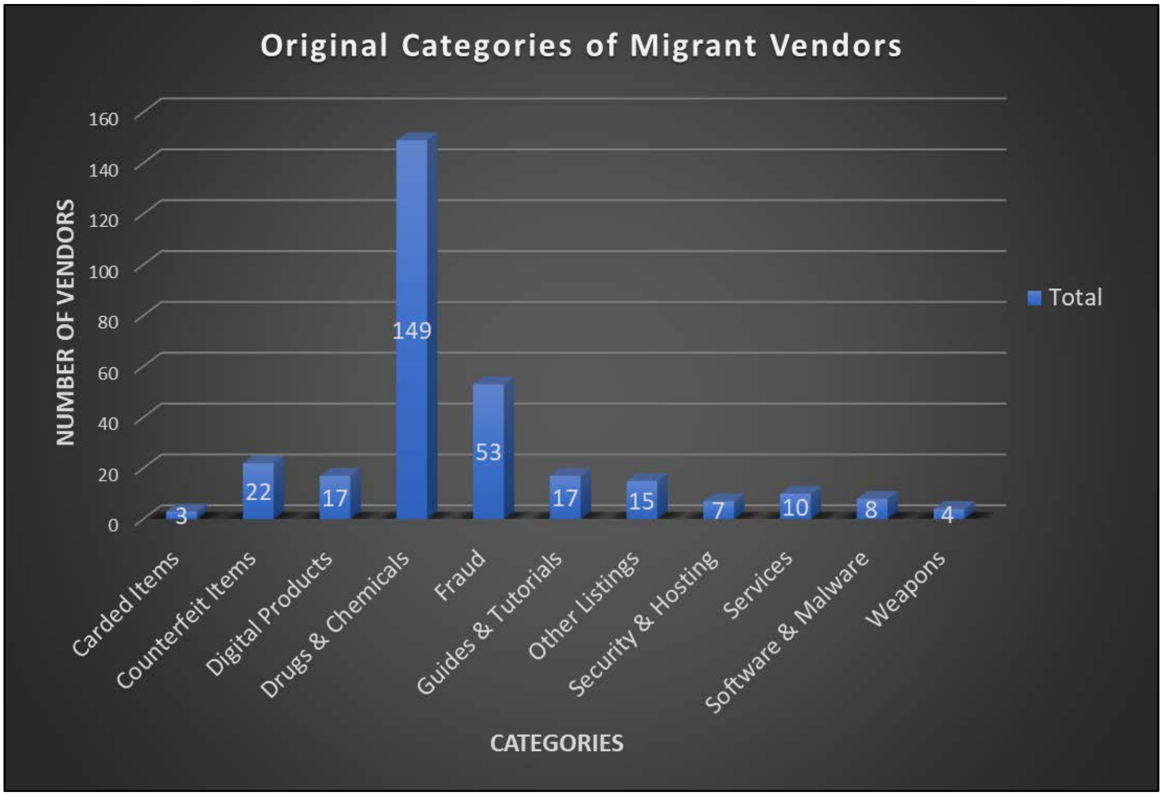


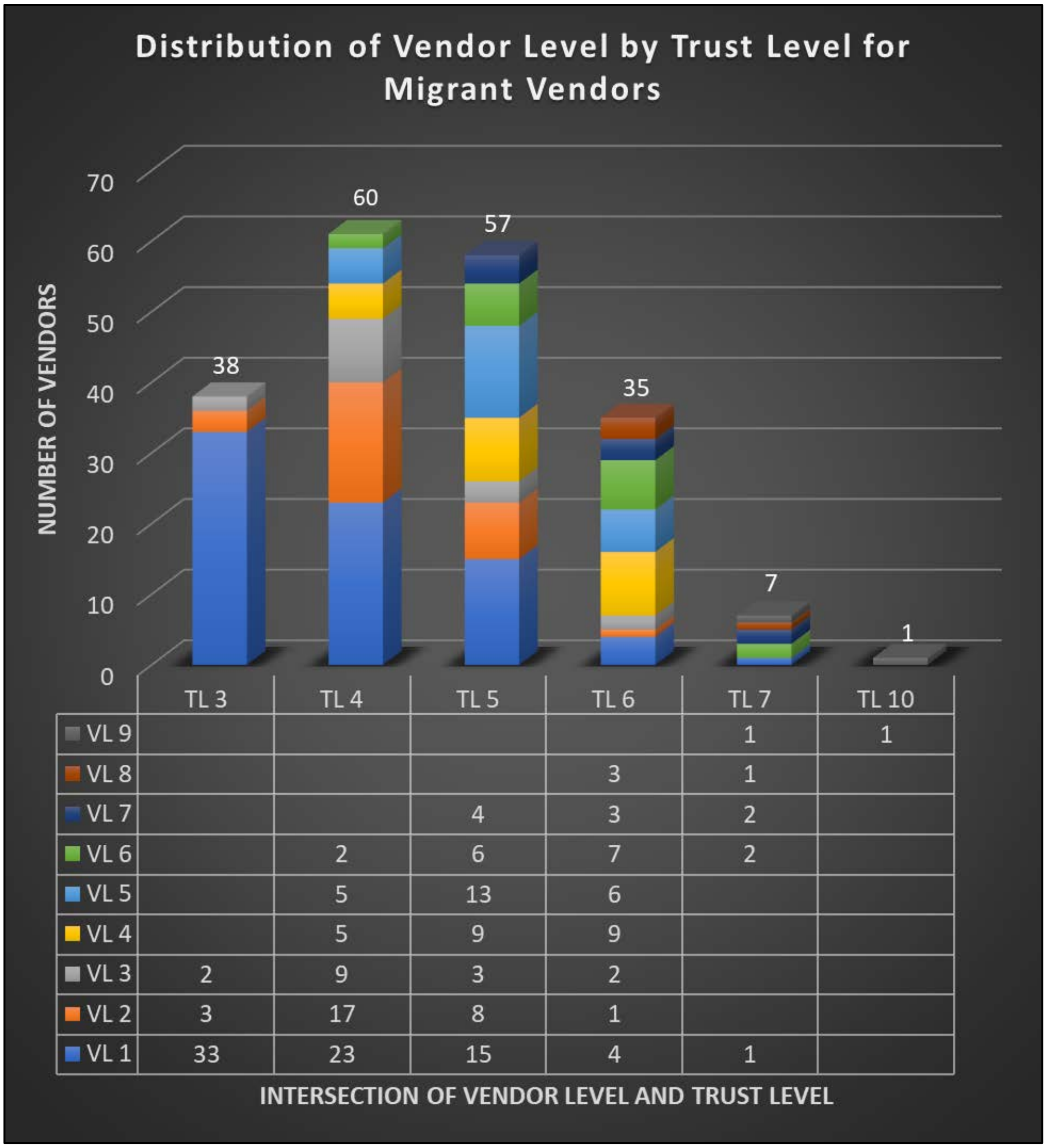
Figure 24. Number of New Vendors in Dream Market over Time



This figure shows category activity of migrant vendors before moving to Dream Market

Figure 25. Original Categories of Migrant Vendors

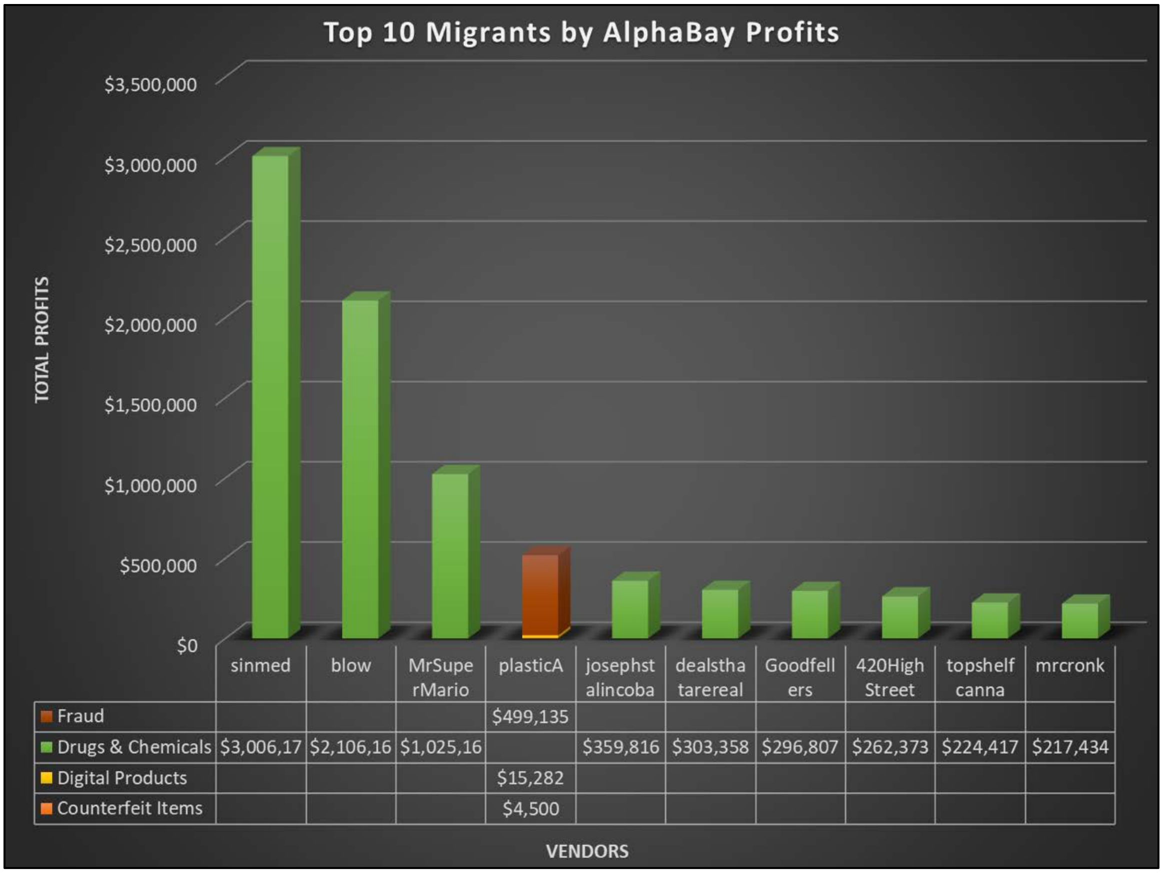
We examined the relationship between trust level and vendor level for the 198 migrants with the same username on both markets. Figure 26 shows that 78% (155/198) of these vendors maintained a trust level of 5 or below from the AlphaBay Market. Of these, 128 also had a low vendor experience status, level 3 and below. Only a few identifiable migrant vendors with high trust level and high vendor level migrated to the Dream Market with the same username.



	TL 3	TL 4	TL 5	TL 6	TL 7	TL 10
VL 9					1	1
VL 8				3	1	
VL 7			4	3	2	
VL 6		2	6	7	2	
VL 5		5	13	6		
VL 4		5	9	9		
VL 3	2	9	3	2		
VL 2	3	17	8	1		
VL 1	33	23	15	4	1	

Figure 26. Distribution of Vendor Level by Trust Level for Migrant Vendors

The migrant vendors grossed over $13.5 million during their time at AlphaBay. 85% of migrant profits were primarily from drug and chemical transactions. Comparatively, non-migrant vendors made $325 million. 70% of these profits were from drug and chemical sales. Of these, “sinmed,” was the 9th highest grossing vendor, earning $3 million before moving to the Dream Market (Figure 27). The vendor “placticA” was also a high earner but specialized in fraud services.



	sinmed	blow	MrSupe rMario	plasticA	josephst alincoba	dealstha tarereal	Goodfell ers	420High Street	topshelf canna	mrcronk
🟫 Fraud				$499,135						
🟩 Drugs & Chemicals	$3,006,17	$2,106,16	$1,025,16		$359,816	$303,358	$296,807	$262,373	$224,417	$217,434
🟨 Digital Products				$15,282						
🟧 Counterfeit Items				$4,500						

Figure 27. Top Ten Migrants by AlphaBay Profits

B. NATURAL LANGUAGE PROCESSING

1. Language Histograms

We also studied what words were important in product descriptions. We were particularly interested in the words of the 198 migrants; we created separate histograms of the top 25 words according to TF-IDF score for the software-and-malware and drug-and-chemical categories. Figures 28 and 29 show the top 25 words for in each respective category.

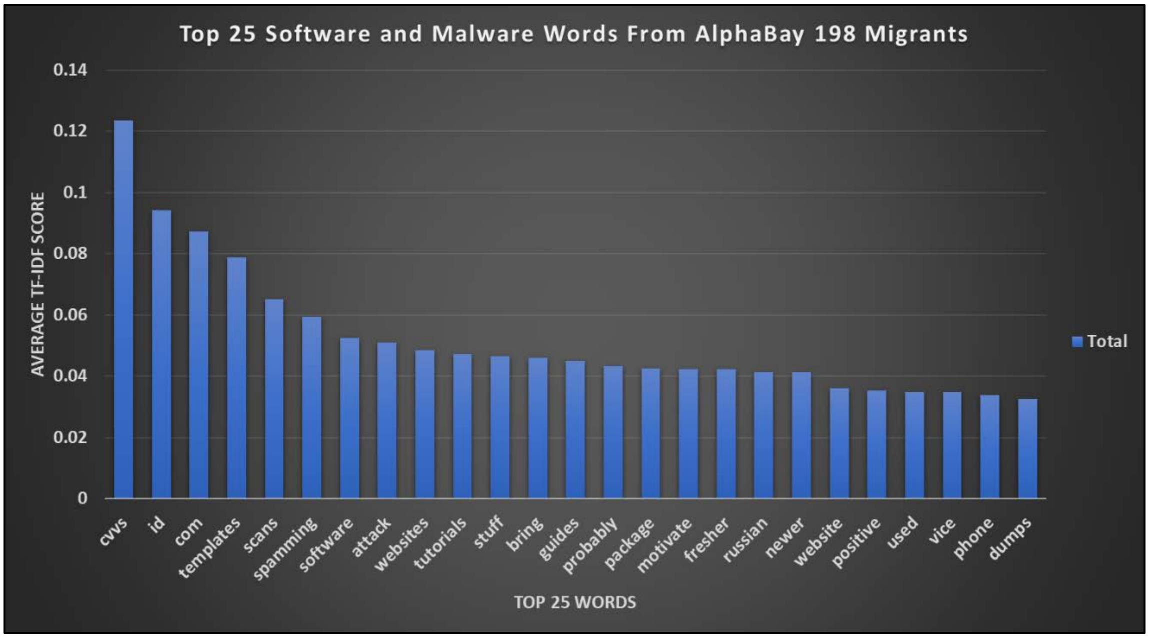


Figure 28. Top 25 Software and Malware Words by Average TF-IDF Score from AlphaBay 198 Migrants

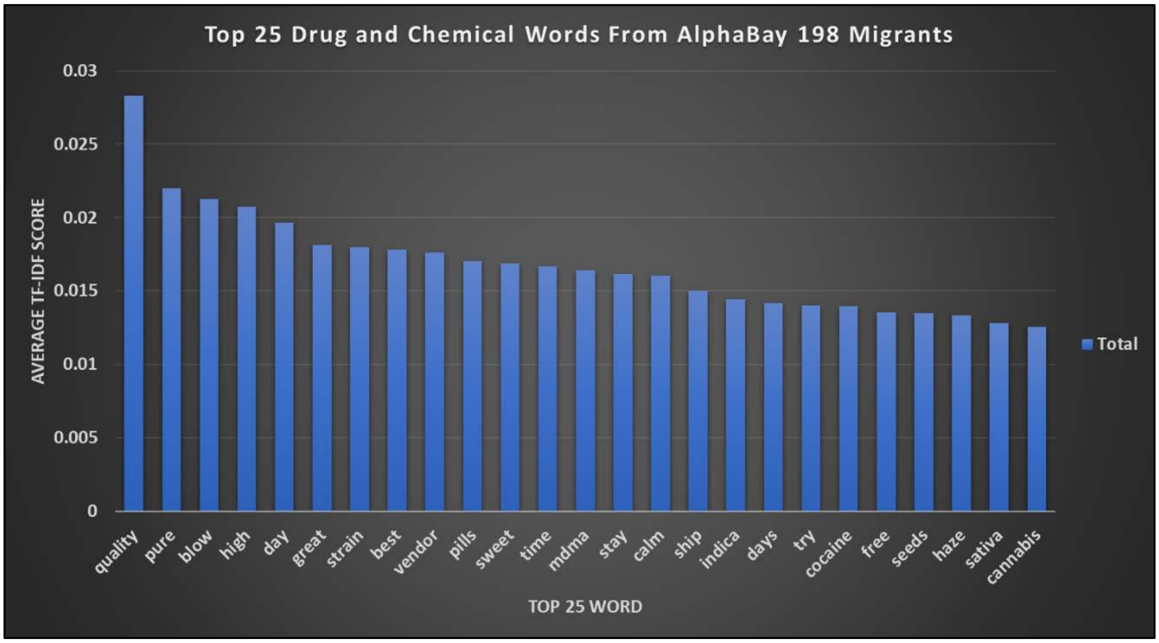


Figure 29. Top 25 Drug and Chemical Words by Average TF-IDF Score from AlphaBay 198 Migrants

We created histograms from a sample of eBay product descriptions to compare AlphaBay and Dream Market language to a conventional market environment (Figures 30 and 31). Illegal drugs and malware are not advertised on eBay, so we identified categories we felt were most suited for language comparison: Software, and vitamins-and-supplements. We correctly assumed that top words germane to each market and category would be different. While AlphaBay’s drug market contained top words like, “blow, strain, pills, and cannabis,” ebay’s vitamins and supplement market contained top words like, “vitamin, oil, capsules, and organic.” We learned that although these markets contained distinct vocabulary, the general sentence structure in similar categories across Darknet and surface markets were similar.

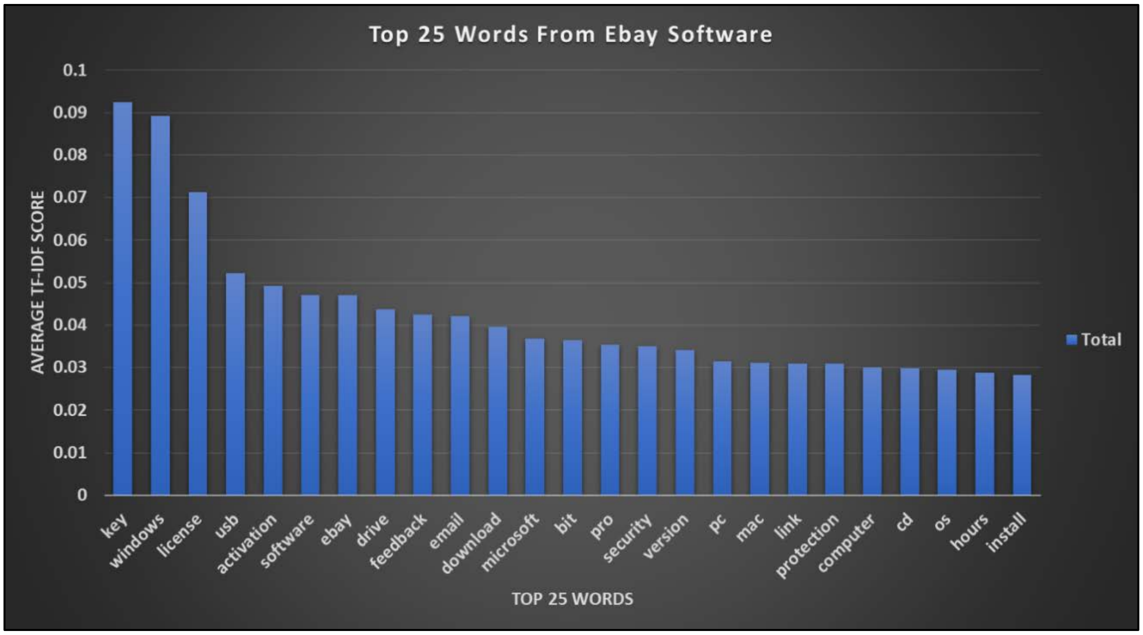


Figure 30. Top 25 Words by Average TF-IDF Score from Ebay Software

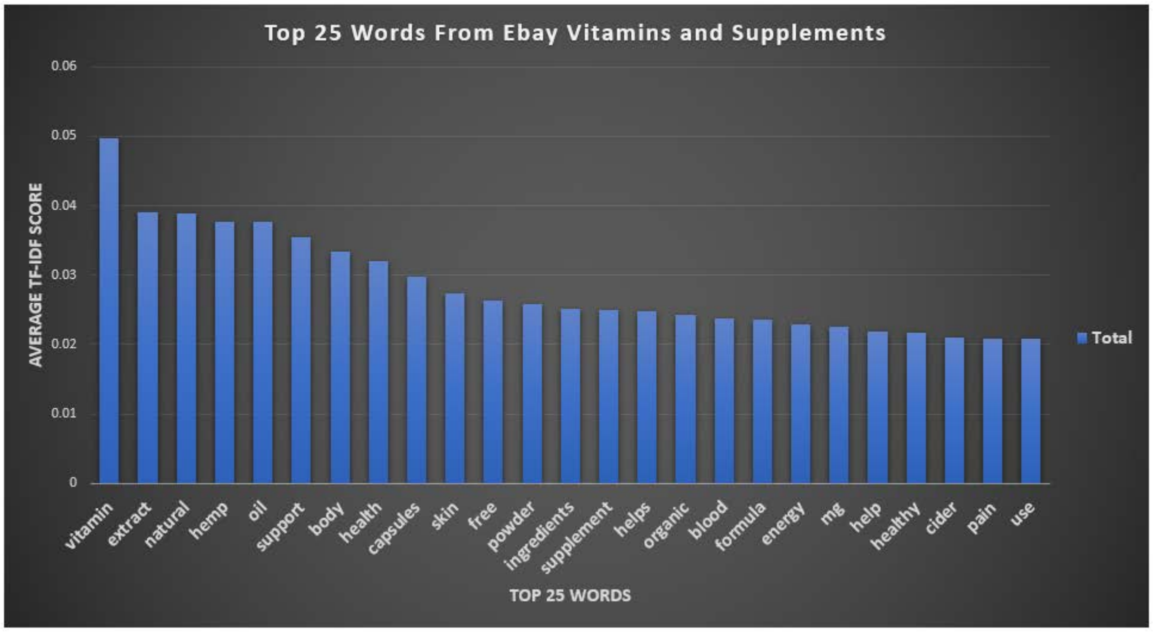


Figure 31. Top 25 Words by Average TF-IDF Score from eBay Vitamins and Supplements

2. Comparison of File Word Distributions

The 198 migrants had 4,436 product listings in the AlphaBay corpus and 3,266 product listings in the Dream Market corpus. After computing TF-IDF cosine similarities, we used a simple recommender system based on cosine similarity for each product from the AlphaBay corpus, querying it to return the ten closest product matches from the Dream Market corpus. In all, 44,360 matches were returned by the system (4,436 AlphaBay products times the 10 recommendations each). Of these, the system matched the vendor name as its first recommendation 1,901 times, between the second and fifth recommendation 7,410 times, and between the sixth and tenth recommendation 6,794 times.

However, the matching did not consider how many products a vendor listed. For example, a vendor with only two product listings on both markets could receive a name match on those two products but would receive an additional eight recommendations from different vendors. The resulting 20% match rating would appear as a poor result although

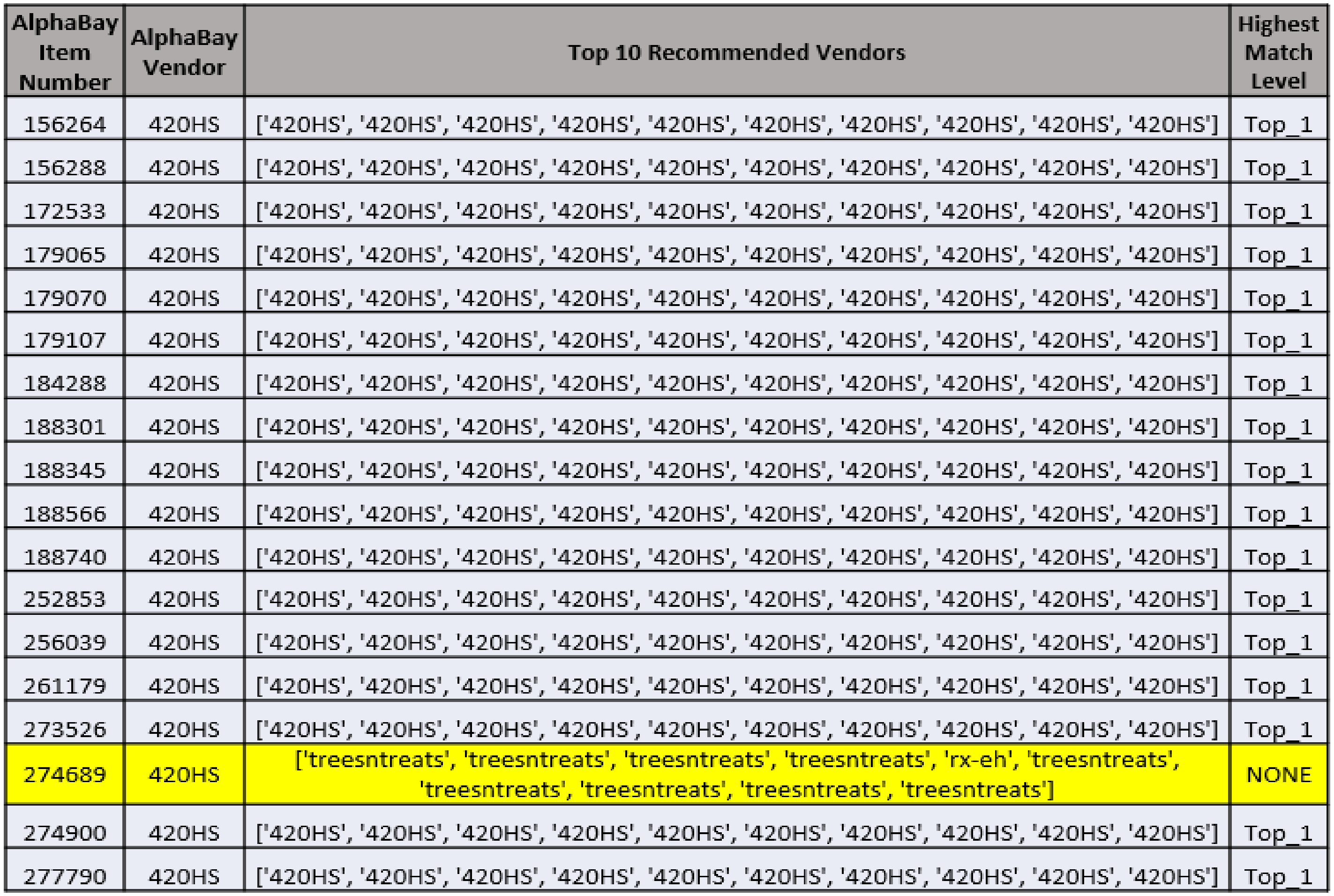
100% of the vendor’s listings matched. Conversely, vendors that list many products could account for many vendor name matches.

Therefore, it was better to use the ranking of the best match of the products. We matched 148 vendors (74.7% of the 198) with the first recommendation at least once (i.e., a rank-1 vendor identification accuracy of 74.7%); 17 vendors (8.6%) matched no worse than the second to fifth; 1 vendor (0.5%) matched in the sixth to tenth range; and 32 vendors (16.2%) did not match in any of their top ten recommendations.

In a separate experiment, we used CountVectorizer to calculate the vectors before measuring their cosine similarities. However, this approach was not as successful in matching product descriptions. Therefore, we decided to proceed to our final experiment with only the TF-IDF vectors.

Our final experiment compared the product descriptions of the 198 migrants from AlphaBay to Dream Market, including all 815 new Dream Market vendors that joined during the 2017 third quarter. With this expanded dataset, 119 vendor names (60.1%) were returned as the first recommendation at least once; 13 vendors (6.6%) matched no worse than the second to fifth; 7 vendors (3.5%) matched no worse than the sixth to tenth; and 59 vendors (29.8%) did not match in any of their top ten best matches.

Our matching method also revealed vendor-specific behaviors. For instance, the drug vendor “420HighStreet” migrated from AlphaBay to Dream Market in the third quarter of 2017. Before migrating, this vendor actively advertised 18 products on AlphaBay. The product descriptions matched across all top ten recommended Dream Market products except for one (Table 1). The yellow highlights the recommendation anomaly which led us to investigate the 420HighStreet products that matched across markets, and those sold by the primary anomalous vendor, “treesntreats”.



AlphaBay Item Number	AlphaBay Vendor	Top 10 Recommended Vendors	Highest Match Level
156264	420HS	['420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS']	Top_1
156288	420HS	['420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS']	Top_1
172533	420HS	['420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS']	Top_1
179065	420HS	['420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS']	Top_1
179070	420HS	['420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS']	Top_1
179107	420HS	['420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS']	Top_1
184288	420HS	['420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS']	Top_1
188301	420HS	['420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS']	Top_1
188345	420HS	['420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS']	Top_1
188566	420HS	['420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS']	Top_1
188740	420HS	['420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS']	Top_1
252853	420HS	['420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS']	Top_1
256039	420HS	['420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS']	Top_1
261179	420HS	['420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS']	Top_1
273526	420HS	['420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS']	Top_1
274689	420HS	['treesntreats', 'treesntreats', 'treesntreats', 'treesntreats', 'rx-eh', 'treesntreats', 'treesntreats', 'treesntreats', 'treesntreats', 'treesntreats']	NONE
274900	420HS	['420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS']	Top_1
277790	420HS	['420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS', '420HS']	Top_1

Table 1. Simple Recommender System Top Ten Recommendations for "420HighStreet"

A manual examination of 420HighStreet’s products that matched the vendor’s own name across marketplaces revealed that this vendor used very similar language and writing style in both descriptions. In fact, some portions of the language were a direct copy from AlphaBay to Dream Market. The following paragraph was present word-for-word in 420HighStreet’s AlphaBay product #156264 and its top recommended Dream Market product, #362297:

> “Due to security issues and high order volumes, I do not provide tracking numbers unless there is a problem with the shipment. Please do not request a tracking number unless it has taken more than 5 days (excluding Sundays) from the day your order was marked shipped. If you do have to request a tracking number, be sure to include the sale number in the message with your tracking request.”

Additional word-for-word and style similarities existed between these products, which are shown in their full product descriptions provided in Appendix B.

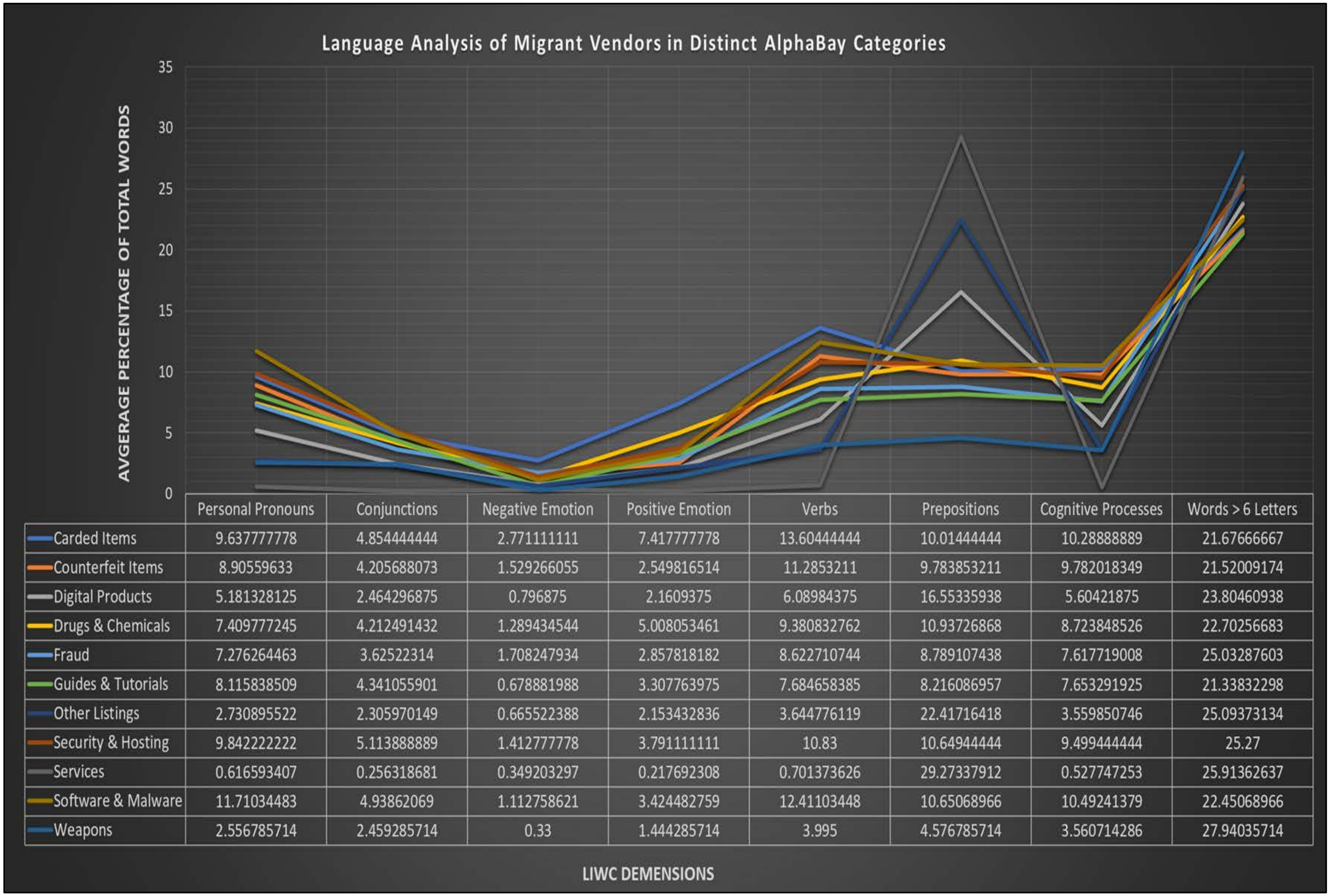
One drug product (#274689 in the AlphaBay dataset) had no 420HighStreet name matches in the top ten recommended Dream Market products. We found that this anomalous product was the only one that 420HighStreet did not move to Dream Market. However, a new Dream Market vendor established business with that particular drug and whose descriptions matched on the words “pack” and “gummy,” so this was the likely connection to the AlphaBay vendor.

Overall, our matching approach was effective in matching vendors across the AlphaBay and the Dream Market corpuses, producing 148 first-recommendation matches out of 198 migrants (74.7% rank-1 identification accuracy) in the one-to-one vendor name experiment and 119 first-recommendation matches against the larger Dream Market 2017 quarter three corpus. This makes it a useful tool for tracking vendors from market to market. It also may be used to discover comparable vendors that deserve the attention of law enforcement and intelligence professionals. However, the recommender system had trouble in matching short or vague product descriptions. Sometimes the closest match was a vendor from a different market category because these product descriptions gave little more than information on accepted payment forms and delivery methods.

3. Analysis of Special Clue Words

The LIWC tool can be used to analyze the psychological meaning behind an author’s words. For our study, we used eight dimensions provided by LIWC to measure truthfulness and deception. The frequency of conjunctions, verbs, prepositions, cognitive processing, and words over six letters long measured cognitive complexity while the frequency of personal pronouns, negative words, and positive words indicated an author’s emotional state. We again focused this tool on the 198 migrant vendors because we could compare their language across Darknet markets, AlphaBay and Dream Market. Figure 32 shows LIWC’s count ratios to total words used for the eight language dimensions for AlphaBay product descriptions.

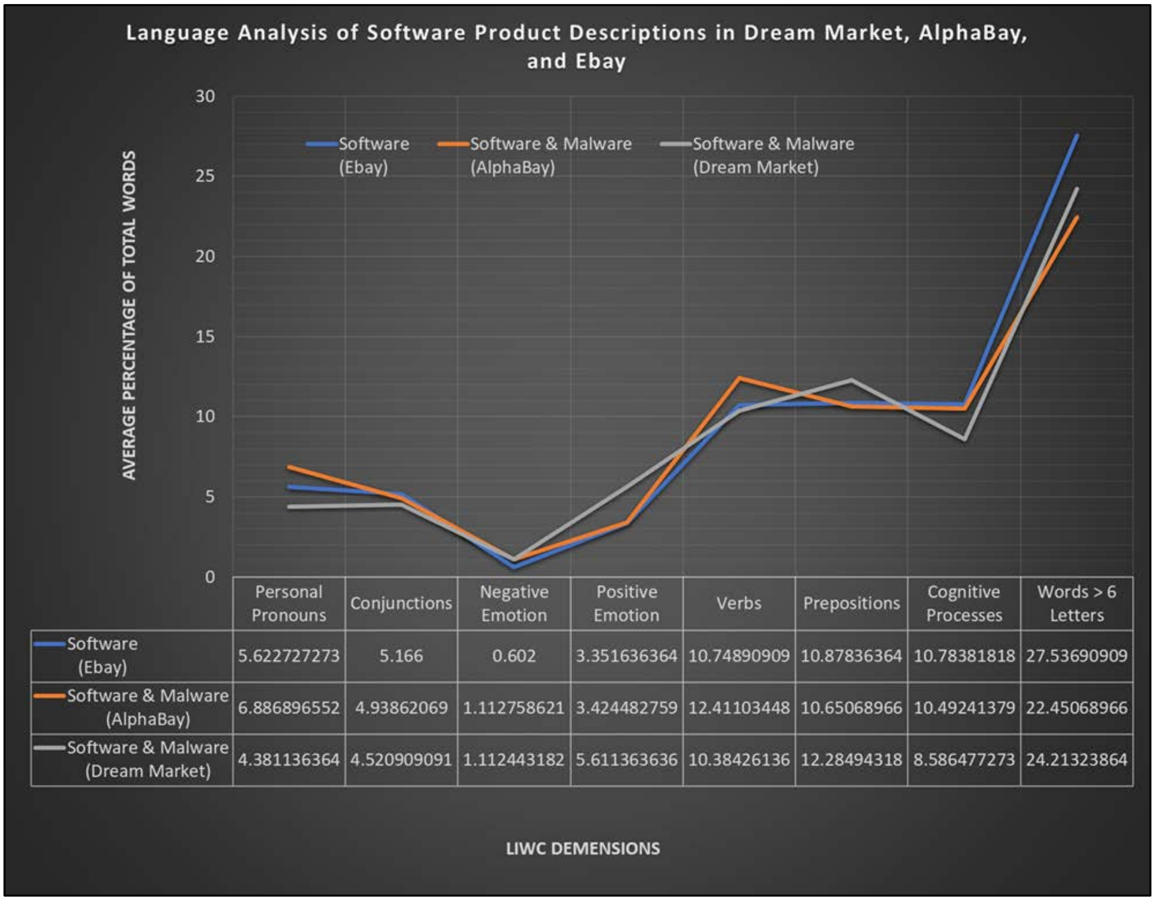
Each market category showed a different semantic fingerprint. For example, in Figure 32, prepositions were more common in listings for services than in any other category. Since advertising services can require a more complex explanation than a physical item, they likely use more prepositions.



	Personal Pronouns	Conjunctions	Negative Emotion	Positive Emotion	Verbs	Prepositions	Cognitive Processes	Words > 6 Letters
Carded Items	9.637777778	4.854444444	2.771111111	7.417777778	13.60444444	10.01444444	10.28888889	21.67666667
Counterfeit Items	8.90559633	4.205688073	1.529266055	2.549816514	11.2853211	9.783853211	9.782018349	21.52009174
Digital Products	5.181328125	2.464296875	0.796875	2.1609375	6.08984375	16.55335938	5.60421875	23.80460938
Drugs & Chemicals	7.409777245	4.212491432	1.289434544	5.008053461	9.380832762	10.93726868	8.723848526	22.70256683
Fraud	7.276264463	3.62522314	1.708247934	2.857818182	8.622710744	8.789107438	7.617719008	25.03287603
Guides & Tutorials	8.115838509	4.341055901	0.678881988	3.307763975	7.684658385	8.216086957	7.653291925	21.33832298
Other Listings	2.730895522	2.305970149	0.665522388	2.153432836	3.644776119	22.41716418	3.559850746	25.09373134
Security & Hosting	9.842222222	5.113888889	1.412777778	3.791111111	10.83	10.64944444	9.499444444	25.27
Services	0.616593407	0.256318681	0.349203297	0.217692308	0.701373626	29.27337912	0.527747253	25.91362637
Software & Malware	11.71034483	4.93862069	1.112758621	3.424482759	12.41103448	10.65068966	10.49241379	22.45068966
Weapons	2.556785714	2.459285714	0.33	1.444285714	3.995	4.576785714	3.560714286	27.94035714

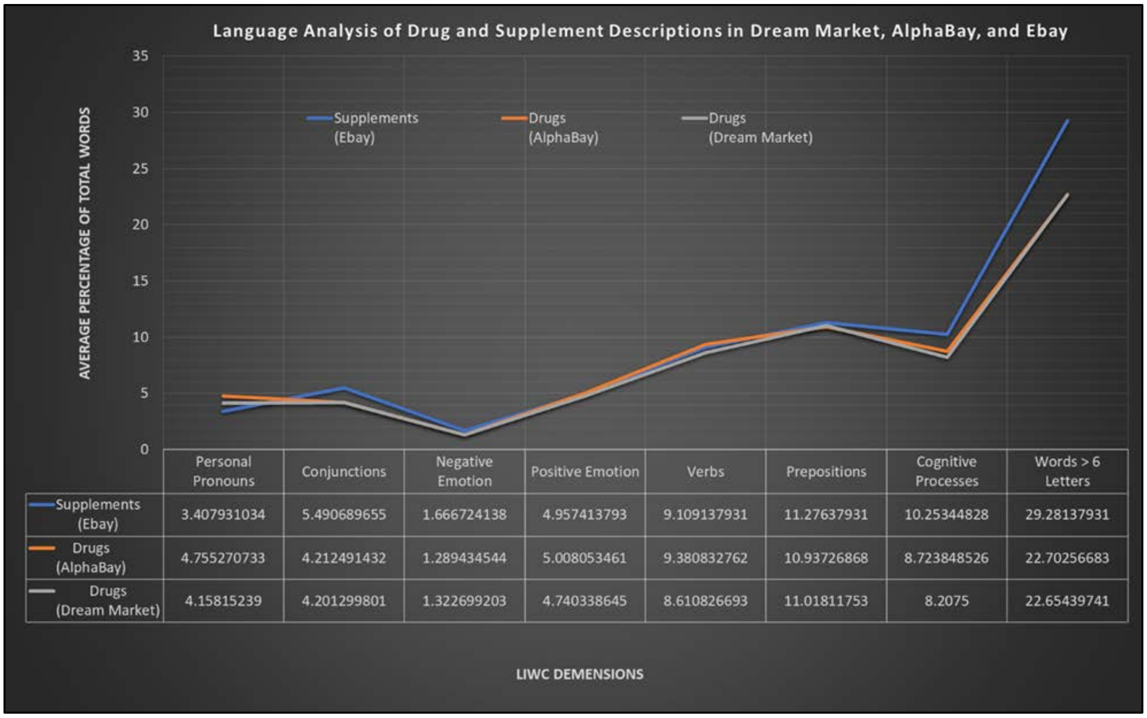
Figure 32. Language Analysis of Migrant Vendors in Distinct AlphaBay Categories

Figure 33 and Figure 34 show how the averages of the eight LIWC dimensions were consistent in similar categories across AlphaBay, Dream Market, and eBay. This suggests that vendors on the Darknet are not any more deceptive than vendors on eBay.



	Personal Pronouns	Conjunctions	Negative Emotion	Positive Emotion	Verbs	Prepositions	Cognitive Processes	Words > 6 Letters
Software (Ebay)	5.622727273	5.166	0.602	3.351636364	10.74890909	10.87836364	10.78381818	27.53690909
Software & Malware (AlphaBay)	6.886896552	4.93862069	1.112758621	3.424482759	12.41103448	10.65068966	10.49241379	22.45068966
Software & Malware (Dream Market)	4.381136364	4.520909091	1.112443182	5.611363636	10.38426136	12.28494318	8.586477273	24.21323864

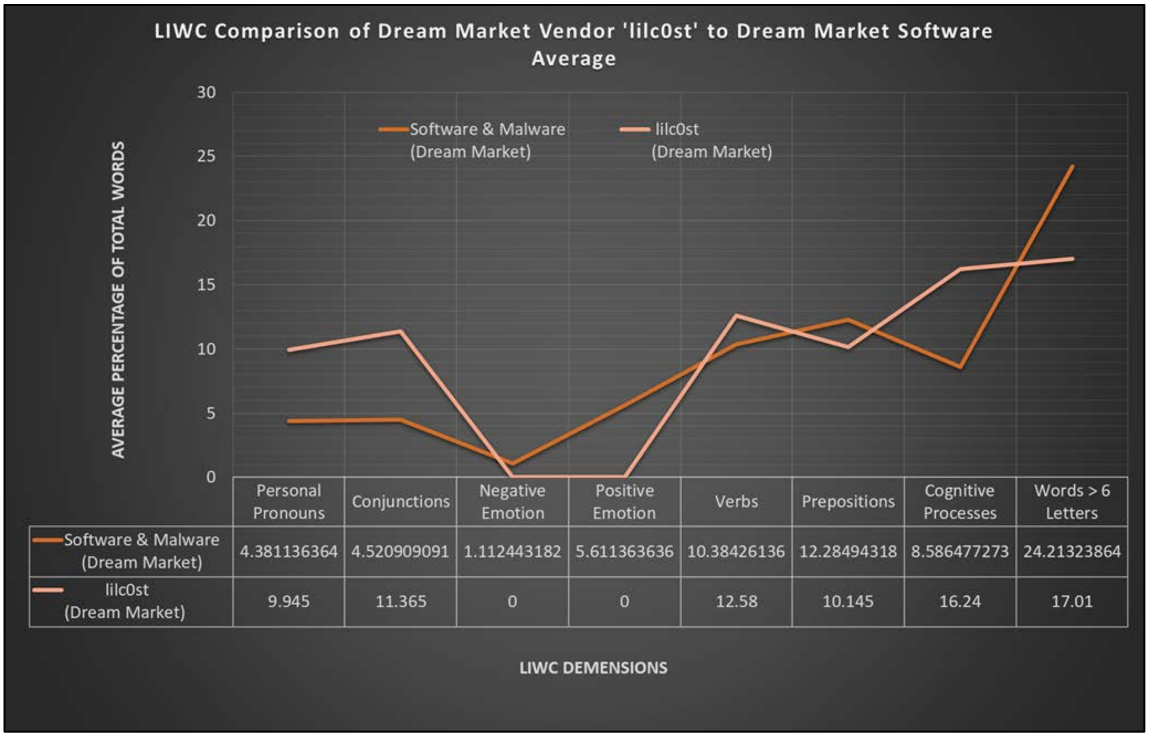
Figure 33. Language Analysis of Software Product Descriptions in Dream Market, AlphaBay, and eBay



	Personal Pronouns	Conjunctions	Negative Emotion	Positive Emotion	Verbs	Prepositions	Cognitive Processes	Words > 6 Letters
Supplements (Ebay)	3.407931034	5.490689655	1.666724138	4.957413793	9.109137931	11.27637931	10.25344828	29.28137931
Drugs (AlphaBay)	4.755270733	4.212491432	1.289434544	5.008053461	9.380832762	10.93726868	8.723848526	22.70256683
Drugs (Dream Market)	4.15815239	4.201299801	1.322699203	4.740338645	8.610826693	11.01811753	8.2075	22.65439741

Figure 34. Language Analysis of Drug and Supplement Descriptions in Dream Market, AlphaBay, and eBay

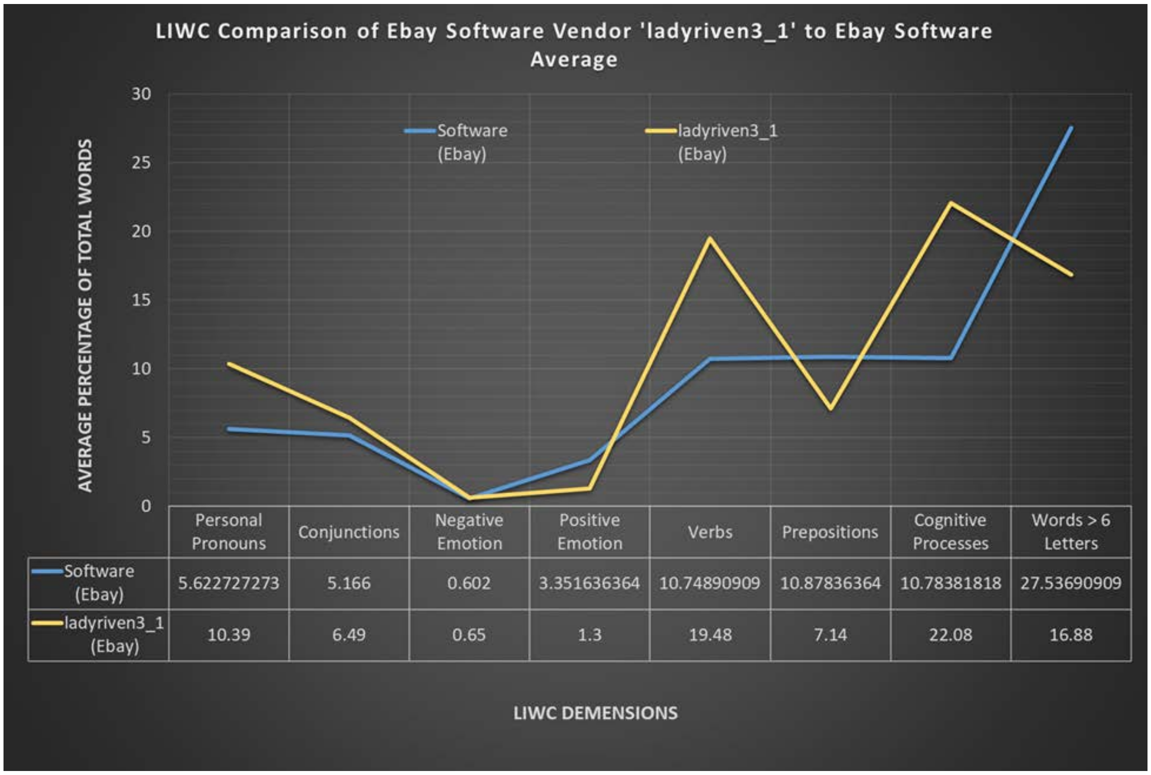
With our calculated LIWC baselines, we identified vendors who deviated significantly from the norm of their categories. For instance, Figure 35 compares the average LIWC output for the Dream Market malware and software category and compares it to that of software vendor “lilc0st.” We found several indicators that suggested “lilc0st” was engaged in fraud. Our statistical analysis showed that vendors with significant profits and both high trust and high vendor levels tended to specialize in just a few markets. However, before migrating to the Dream Market, “lilc0st” had 44 active listings in seven categories: counterfeit items, drugs and chemicals, fraud, guides and tutorials, software and malware, weapons, and other listings in AlphaBay. When active on AlphaBay, “lilc0st” had only $26 in profits, and gained a vendor level of 1 and a trust level of 3, both relatively low scores.



	Personal Pronouns	Conjunctions	Negative Emotion	Positive Emotion	Verbs	Prepositions	Cognitive Processes	Words > 6 Letters
Software & Malware (Dream Market)	4.381136364	4.520909091	1.112443182	5.611363636	10.38426136	12.28494318	8.586477273	24.21323864
lilc0st (Dream Market)	9.945	11.365	0	0	12.58	10.145	16.24	17.01

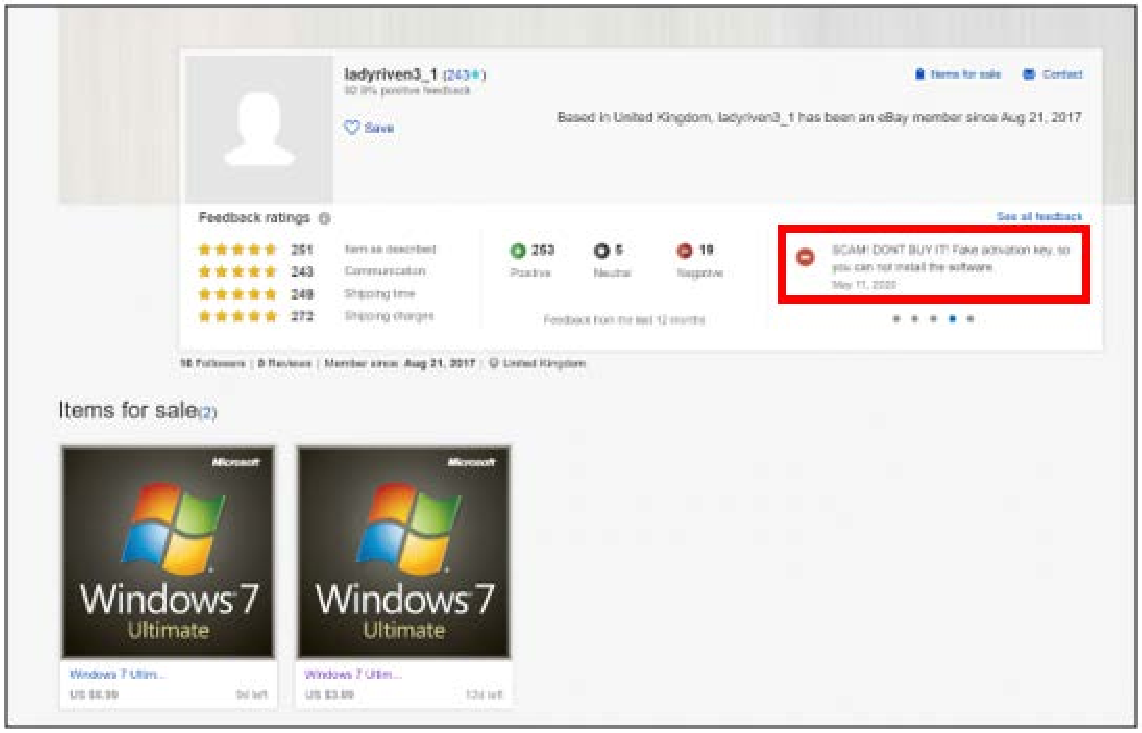
Figure 35. LIWC Comparison of Dream Market Vendor "lilc0st" to Dream Market Software Average

We applied the same approach to eBay listings. Figures 36 and 38 show the LIWC output of vendors who deviated significantly from the average LIWC output. Using the advanced search option on eBay, we discovered that "ladyriven3_1" and "debbme_0" both specialized in the sale of deeply discounted Microsoft Windows products (Figures 37 and 39). Specifically, vendor "debbme_0" advertised Windows Server 2019 Standard and Datacenter software for only $3.99, for products that normally sell for hundreds of dollars. Interestingly, both "ladyriven3_1" and "debbme_0" had just a few followers who were interested in their listings but had incomplete profiles and appeared to be ghost accounts. Although eBay's masking of product-reviewer usernames prevents us from confirming our theory, we suspect that many positive product reviews these vendors received were from Sybil accounts, which are accounts created under pseudonymous identities to influence the reputation of a fraudulent account.



	Personal Pronouns	Conjunctions	Negative Emotion	Positive Emotion	Verbs	Prepositions	Cognitive Processes	Words > 6 Letters
Software (Ebay)	5.622727273	5.166	0.602	3.351636364	10.74890909	10.87836364	10.78381818	27.53690909
ladyriven3_1 (Ebay)	10.39	6.49	0.65	1.3	19.48	7.14	22.08	16.88

Figure 36. LIWC Comparison of eBay Software Vendor "Ladyriven3_1" to eBay Software Average



ladyriven3_1 (243★)
92.9% positive feedback

♡ Save

Items for sale | Contact

Based in United Kingdom, ladyriven3_1 has been an eBay member since Aug 21, 2017

Feedback ratings

See all feedback

★★★★★	251	Item as described
★★★★★	243	Communication
★★★★★	248	Shipping time
★★★★★	272	Shipping charges

263	5	19
Positive	Neutral	Negative

Feedback from the last 12 months

SCAM! DONT BUY IT! Fake activation key, so you can not install the software
May 11, 2020

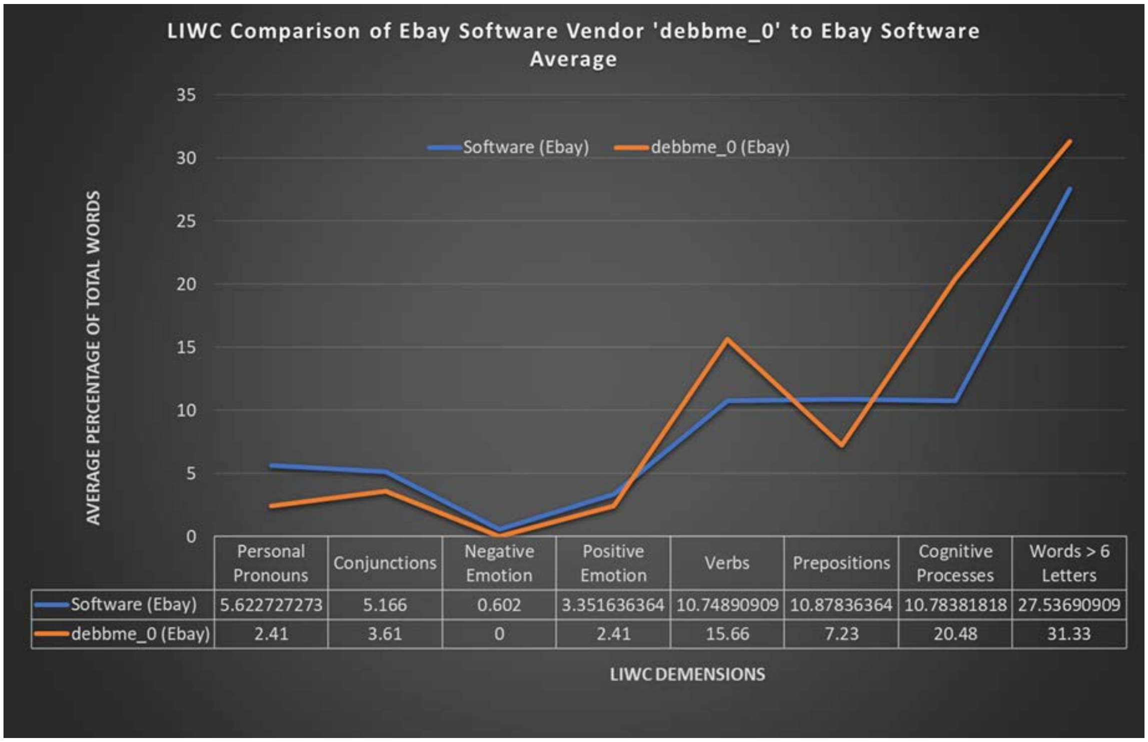
16 Followers | 0 Reviews | Member since: Aug 21, 2017 | United Kingdom

Items for sale(2)

Windows 7 Ultim...
US $8.99 9d left

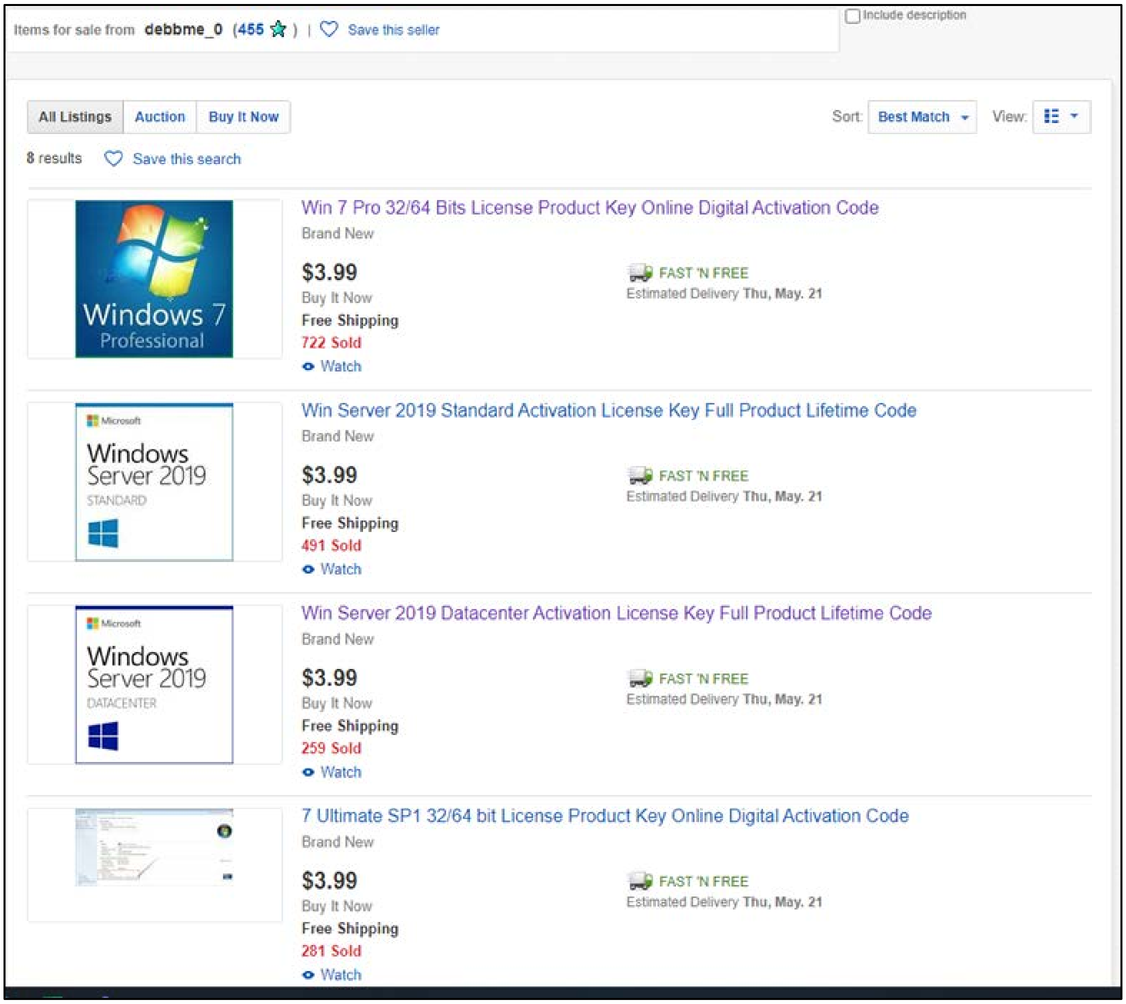
Windows 7 Ultim...
US $3.89 13d left

Figure 37. “Ladyriven3_1” Fraudulent Product Listings. Source: eBay (2020).



	Personal Pronouns	Conjunctions	Negative Emotion	Positive Emotion	Verbs	Prepositions	Cognitive Processes	Words > 6 Letters
Software (Ebay)	5.622727273	5.166	0.602	3.351636364	10.74890909	10.87836364	10.78381818	27.53690909
debbme_0 (Ebay)	2.41	3.61	0	2.41	15.66	7.23	20.48	31.33

Figure 38. LIWC Comparison of eBay Software Vendor “debbme_0” to eBay Software Average



Items for sale from debbme_0 (455 ☆) | ♡ Save this seller

□ Include description

All Listings | Auction | Buy It Now

Sort: Best Match ▾ View: ☷ ▾

8 results ♡ Save this search

Win 7 Pro 32/64 Bits License Product Key Online Digital Activation Code

Brand New

$3.99

Buy It Now

Free Shipping

722 Sold

Watch

FAST 'N FREE

Estimated Delivery Thu, May. 21

Win Server 2019 Standard Activation License Key Full Product Lifetime Code

Brand New

$3.99

Buy It Now

Free Shipping

491 Sold

Watch

FAST 'N FREE

Estimated Delivery Thu, May. 21

Win Server 2019 Datacenter Activation License Key Full Product Lifetime Code

Brand New

$3.99

Buy It Now

Free Shipping

259 Sold

Watch

FAST 'N FREE

Estimated Delivery Thu, May. 21

7 Ultimate SP1 32/64 bit License Product Key Online Digital Activation Code

Brand New

$3.99

Buy It Now

Free Shipping

281 Sold

Watch

FAST 'N FREE

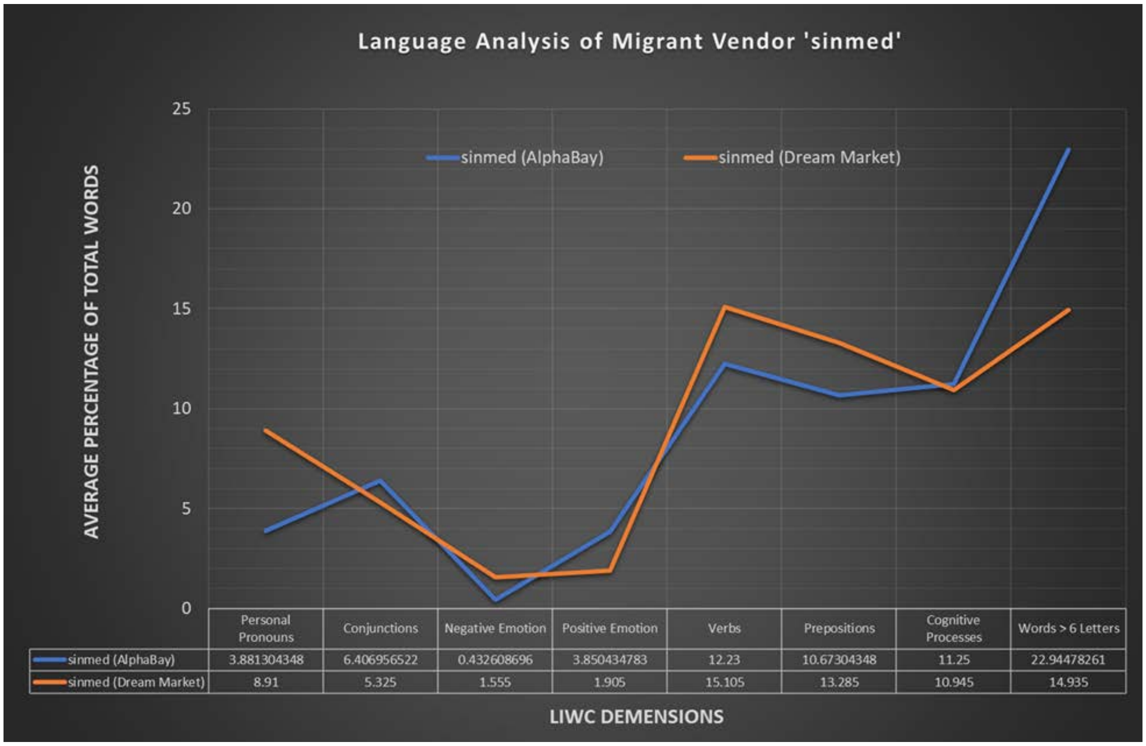
Estimated Delivery Thu, May. 21

Win Server 2019, when sold by legitimate sources, can cost upwards of $1000. Here it is listed for $3.99.

Figure 39. Suspiciously Priced “debbme_0” Product Listings. Source: eBay (2020).

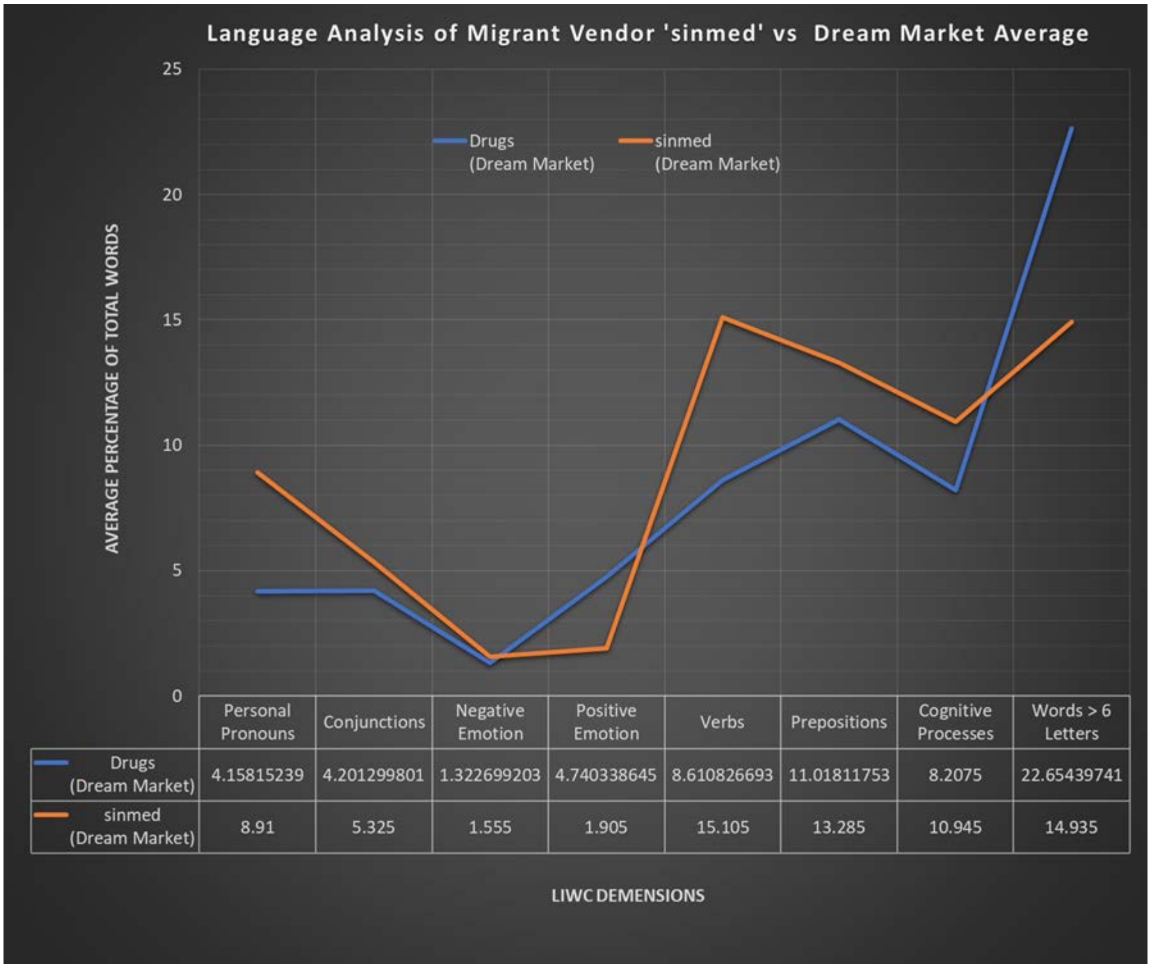
LIWC also could compare the same vendor across different sites. As indicated in Figure 40, vendor “sinmed” who specialized in the sale of drugs and chemicals increased the frequency of personal pronouns, verbs, and prepositions in product descriptions in going from AlphaBay to the Dream Market. In fact, compared to fellow drug sellers that migrated to the Dream Market, he used personal pronouns and verbs almost twice as often (Figure 41). According to the Pennebaker model, this shift is an indication of truthfulness. On AlphaBay “sinmed” had attained a vendor level of 8 and a trust level of 7, but these scores were not transferrable to Dream Market. We suspected that “sinmed” changed his language to communicate trustworthiness to appeal to a new customer base. An excerpt from a “sinmed” product description advertising Adderall supports this:

> ““PUT YOUR MONEY WHERE YOUR MOUTH IS” limited time and qty promotion
>
> Ok... so I am 100% OG when it comes to my presence on DMs however this is my first foray into the counterfeit Adderall market. I see I am going to have to do something to set myself apart. Hmmmmm. I have a hunch that alot [*sic*] of the well know competitors can say they have “the real” ingredients in there [*sic*] product but how would you ever know for sure? Well better yet if you do not know how can your friend, and customers then know what you have to offer is the real deal with actual Dextroamphetamine and Amphetamine all uncut an unadulterated. With the next 10 orders of my 100 x cor 135 20mg adderall listing I am going send you a sample of two SEPERATE [*sic*] bags one containing pure dextroamphetamine and the other pure amphetamine. Feel free to send it out , do a presumptive test,or sniff it all . I don’t care. All I care about is that as a requirement you certify that I sent it to you as promised in feedback. I know will still sleep good at night no matter who gets my samples. How about you other vendors put up or shut up?! SiN;)
>
> This my friends is not pressed amphetamine paste slapped in a pill. It was 2 years + in the works. I am awaiting 1H-NMR results to post. This tablet is certified Dextroamphetamine/Amphetamine mixed salts. I always try my best to deliver something superior than what is currently on the markets. It has been released in freebie mode twice and passed the lithmus [*sic*] test. As always what you see is what you get and you do get what you pay for. Amphetamine is readily available on the markets but Dextroamphetamine is certainly not. THe [*sic*] dextroamphetamine synergistic effect is what allows you to achieve that intense focus while not getting wired. This is the actual photo line-up of product. I think the work speaks for itself. SiN;)”



	Personal Pronouns	Conjunctions	Negative Emotion	Positive Emotion	Verbs	Prepositions	Cognitive Processes	Words > 6 Letters
sinmed (AlphaBay)	3.881304348	6.406956522	0.432608696	3.850434783	12.23	10.67304348	11.25	22.94478261
sinmed (Dream Market)	8.91	5.325	1.555	1.905	15.105	13.285	10.945	14.935

Figure 40. Language Analysis of Migrant Vendor “sinmed”: AlphaBay versus Dream Market



	Personal Pronouns	Conjunctions	Negative Emotion	Positive Emotion	Verbs	Prepositions	Cognitive Processes	Words > 6 Letters
Drugs (Dream Market)	4.15815239	4.201299801	1.322699203	4.740338645	8.610826693	11.01811753	8.2075	22.65439741
sinmed (Dream Market)	8.91	5.325	1.555	1.905	15.105	13.285	10.945	14.935

Figure 41. Language Analysis of Migrant Vendor “sinmed” versus Dream Market Average

THIS PAGE INTENTIONALLY LEFT BLANK

VI. CONCLUSION

This thesis focused on classifying activity in Darknet marketplaces through graph and linguistic analyses. Using our methodology, we discovered 198 vendors who moved from one market site to another due to raids by law enforcement. This provided us with an opportunity to analyze language attributes and tendencies across Darknet marketplaces. We saw that vendors will often use similar or identical language when forced to migrate their products to a new market. This makes language histograms useful in tracking vendors across marketplaces and in identifying comparable vendors.

Although we could not find clear indications of deceptive language, we saw shifts in individual vendor language tendencies following the FBI shutdown of the AlphaBay market and subsequent migration to the Dream market. We showed that LIWC can create a semantic fingerprint which can be used by law enforcement to track or mimic a vendor. The terminology used in Dark Web and Surface Web product descriptions differed with the categories and markets in which they were sold. However, averages for sentiment and general sentence structure in similar markets were similar in Dark and Surface Web environments.

Darknet market sales did not stop when AlphaBay and Dream Market were shut down; the Dark Web community continues to evolve and adapt to survive. The two most popular Darknet Markets today are Empire Market and White House Market. Empire Market was the successor to Dream Market and has the user-friendly features found on AlphaBay. It uses multiple cryptocurrencies to enable the exchange of illicit goods and services, uses mirror sites to counteract denial-of-service attacks, and uses two-factor authentication (2-FA) to encrypt and decrypt messages (Empire Market, 2020). White House Market is considered one of the most secure Darknet Markets today. It uses the unique cryptocurrency Monero to implement transactions, enforces mandatory encryption, and makes direct payments between sellers and buyers to minimize the risk of exit scams (Sedgwick, 2020). Nonetheless, these new sites are still vulnerable to the methods and tools that we have described in this thesis. They should help law enforcement and national security professionals survey the threat landscape and track the evolutionary changes in the Dark Web.

Towards future work, our product-description matching could be improved by introducing a category check and by including the product title when computing of TF-IDF vectors. Also, the effectiveness of semantic fingerprinting could be tested against a corpus of marketplace forum posts and vendor profiles. These techniques could additionally be adapted for other languages besides English, with the detection of security threats in mind.

APPENDIX A. DISTRIBUTION OF PRODUCT SUBCATEGORIES SOLD IN ALPHABAY DRUG, SOFTWARE, AND FRAUD MARKETS

The figures in this appendix provide additional details about the types of products sold within each major category of the AlphaBay market.

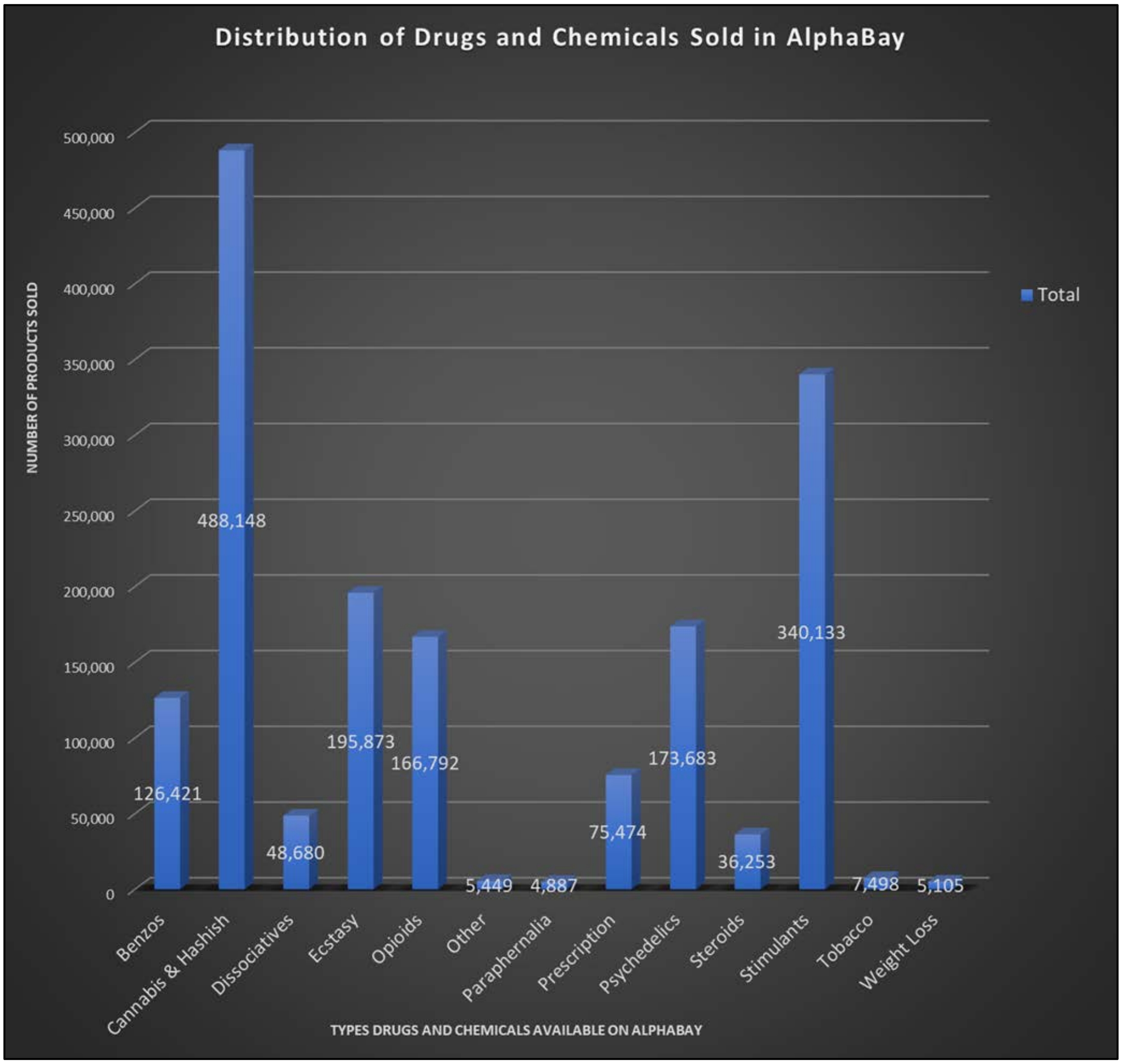


Figure 42. Distribution of Drugs and Chemicals Sold on AlphaBay

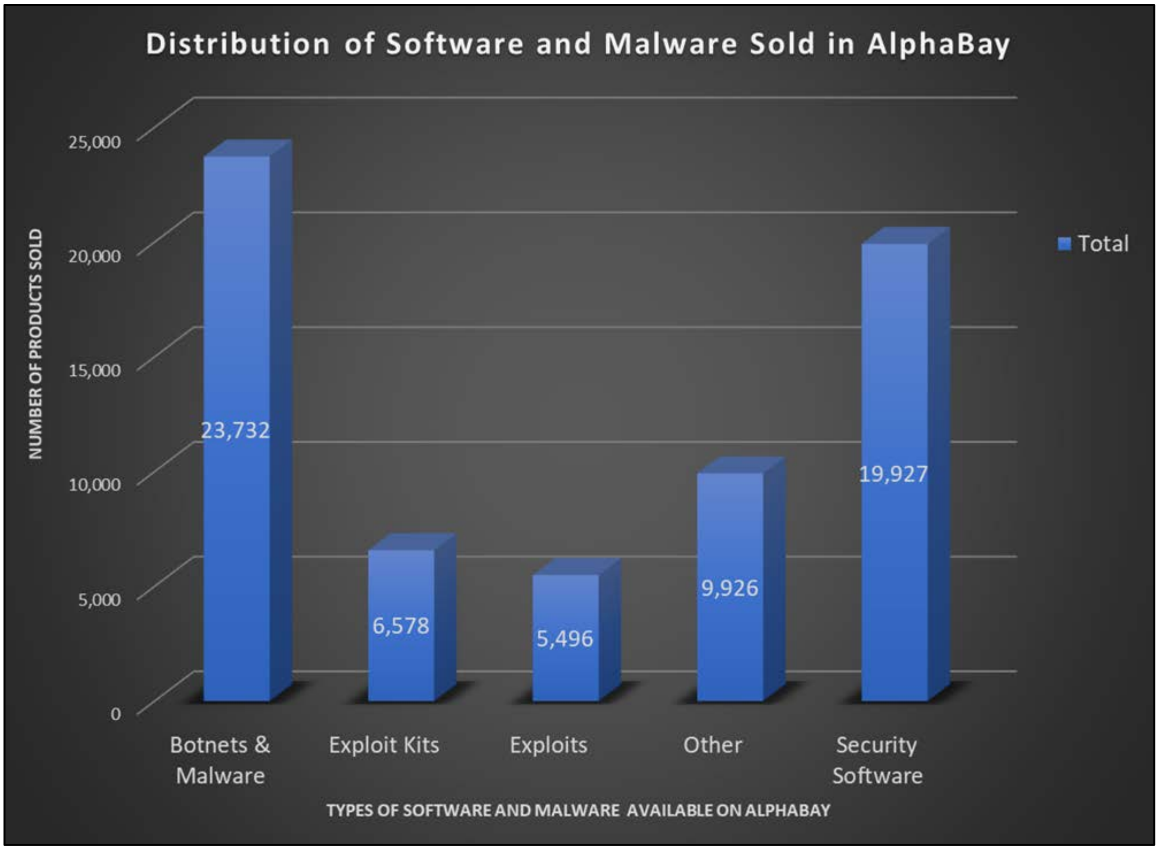


Figure 43. Distribution of Software and Malware Sold on AlphaBay

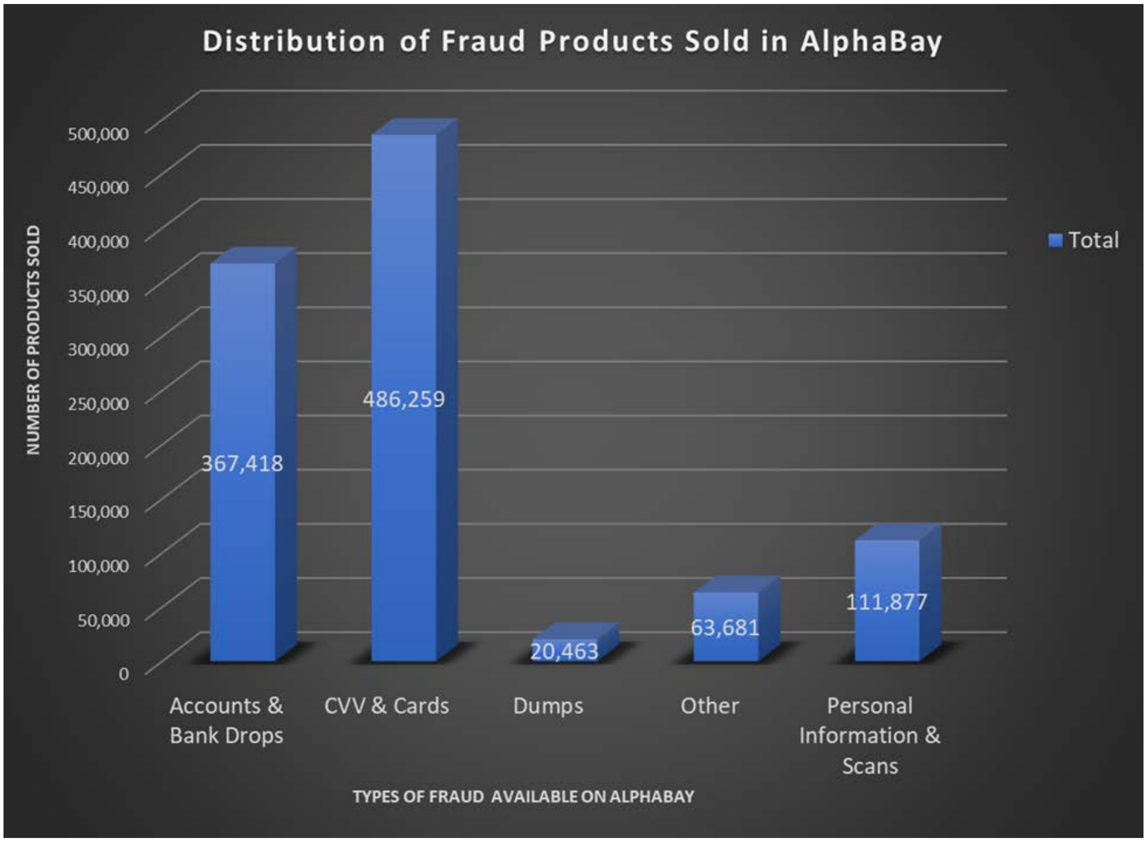


Figure 44. Distribution of Fraud Products Sold on AlphaBay

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX B. “420HIGHSTREET” ALPHABAY PRODUCT DESCRIPTION COMPARISON

This appendix shows the full descriptions from the “420High Street” products discussed in the recommender system section of Chapter V, English errors and all.

A. ALPHABAY PRODUCT # 156264

I am offering a low intruductory [*sic*] rate on 10-gram orders of top-shelf weed. Prices will be rising soon, so don’t miss out on this deal! All product in this listing is top shelf medical grade outdoor weed. Please also see my indoor listings. Available strains are listed below. Just specify which strain you want in the buyer notes for your order. I will do my best to keep my inventory up to date; but I reserve the right to substitute strains if I am out of the selected strain. You may want to mention a second-choice strain in case the one you want runs out before I’m able to update the listing.

****

LIST OF STRAINS

Woody Kush

Power Nap

Moonshine Haze

Mendo Montage

Southfork Kush

Girlscout Cookies

Animal Cookies

-------SHIPPING-------

*** Free USPS shipping***

Please formate [*sic*] the address according to the following guidelines:

--All capitals (entire address)

--Do not include country (USA, United States, etc)

--Abbreviate words like Street, Road, Lane, etc., in the usual way: St., Rd., Ln., etc.

--Put any unit numbers, apartment numbers, etc., on a separate line after the street address

--Abbreviate state names, do no [*sic*] spell out the entire state, e.g: AL, MI, NY, etc.

SAMPLE ADDRESS:

JOHN DOE

585 MAIN ST

APT 284

CHICAGO, IL 48392

Due to security issues and high order volumes, I do not provide tracking numbers unless there is a problem with the shipment. Please do not request a tracking number unless it has taken more than 5 days (excluding Sundays) from the day your order was marked shipped. If you do have to request a tracking number, be sure to include the sale number in the message with your tracking request.

I am usually able to ship the specific strain requested. However, due to high sales volumes I sometimes run out of particular strains before updating the listings. On those rare occasions I reserve the right to substitute an alternative strain so as to fulfill your order in a timely manner.

B. DREAM MARKET PRODUCT #362297

Available strains listed below.

****

$22 ADD-ON ITEMS!!!

For an extra $22 you can get a CO2 cartridge or edible item added to your order (regular price $30). Select “USPS Plus Extra Item” shipping option.

****

Currently available strains (select first and second choices. No mixted [*sic*] strain orders):

Sunset Sherbet

Superglue

Platinum

Girlscout Cookies

Kosher Kush

Boss OG Kush

-------SHIPPING-------

Please make sure your shipping info is accurate and correctly formatted.

SAMPLE ADDRESS:

JOE DRAKE

585 MAIN ST

APT 284

CHICAGO, IL 48392

Due to security issues and high order volumes, I do not provide tracking numbers unless there is a problem with the shipment. Please do not request a tracking number unless it has taken more than 5 days (excluding Sundays) from the day your order was marked shipped. If you do have to request a tracking number, be sure to include the sale number in the message with your tracking request.

LIST OF REFERENCES

Alsayra. (2012, May 1). *AZSecure-data intelligence and security informatics data sets*. AZSecure-data.org. http://azsecure-data.org/dark-net-markets.html

Banik, R. (2018). *Hands-on recommendation systems with Python: Start building powerful and personalized, recommendation engines with Python*. Packt Publishing Ltd. https://learning.oreilly.com/library/view/hands-on-recommendation-systems/9781788993753/

Bearman, J. (2017, May 1). Silk Road: The untold story. *Wired*. https://www.wired.com/2015/05/silk-road-untold-story/

Bitcoin Project. (2020). *How does bitcoin work?* https://bitcoin.org/en/how-it-works

Branwen, G., Christin, N., Décary-Hétu, D., Munksgaard Andersen, R., StExo, El Presidente… Goode, S. (2019, May 22). *Darknet market archives (2013–2015)*. https://www.gwern.net/DNM-archives

Brownlee, J. (2019, August 7). *How to prepare text data for machine learning with scikit-learn*. https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/

Chen, H. (2008, June). Sentiment and affect analysis of dark web forums: Measuring radicalization on the internet. In *2008 IEEE International Conference on Intelligence and Security Informatics* (pp. 104–109).

Crooks, J. (2020, April 27). *Dream market review and URL*. Deep websites Links. https://www.deepwebsiteslinks.com/dream-market-url/

DarknetStats. (2019). *Vendor review—The grass company*. https://www.darknetstats.com/vendor-review-the-grass-company/

eBay Inc. (2020). *Electronics, cars, fashion, collectibles & more*. https://www.ebay.com/sch/i.html?_sofindtype=0&_byseller=1&_nkw=&_in_kw=1&_ex_kw=&_sacat=0&_udlo=&_udhi=&_ftrt=901&_ftrv=1&_sabdlo=&_sabdhi=&_samilow=&_samihi=&_sadis=15&_stpos=93924&_sargn=-1%26saslc%3D1&_salic=1&_fss=1&_fsradio=%26LH_SpecificSeller%3D1&_saslop=1&_sasl=debbme_0&_sop=12&_dmd=1&_ipg=50&_fosrp=1

Elendner, H., Trimborn, S., Ong, B., & Lee, T. M. (2016). *The cross-section of crypto-currencies as financial assets: An overview* (No. 2016–038). SFB 649 Discussion paper. https://www.econstor.eu/bitstream/10419/148874/1/870165852.pdf

Empire Marketplace. (2020). *Empire market: Empire marketplace link*. https://empiremarketlink.com/

Finklea, K. M. (2017). *Dark web* (CRS Report No. R44101). Congressional Research Service. https://fas.org/sgp/crs/misc/R44101.pdf

Graczyk, M., & Kinningham, K. (2015). *Automatic product categorization for anonymous marketplaces*. Technical report, Stanford University. http://cs229.stanford.edu/proj2015/184_report.pdf

Graham, K. (2017, October 11). Dark web ransomware economy growing at 2,500 percent annually. *Digital Journal*. http://www.digitaljournal.com/tech-and-science/technology/dark-web-ransomware-economy-growing-at-2-500-percent-annually/article/504795

Gupta, S. (2007). *Modelling deception detection in text* [Master’s thesis, Queen’s University]. https://qspace.library.queensu.ca/handle/1974/922

Hancock, J. T., Curry, L. E., Goorha, S., & Woodworth, M. (2007). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45(1), 1–23. Taylor & Francis Group, LLC. https://sml.stanford.edu/ml/2008/01/hancock-dp-on-lying.pdf

Holt, T. J., Smirnova, O., & Hutchings, A. (2016). Examining signals of trust in criminal markets online. *Journal of Cybersecurity*, *2*(2), 137–145. https://academic.oup.com/cybersecurity/article/2/2/137/2525525

Hunter, J. D. (2007, May-June). Matplotlib: A 2D graphics environment. In *Computing in Science & Engineering*, *9*(3), 90–95. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4160265

Kang, R. (2018, January 23). *Think of the internet as an iceberg* [png]. IIAP. https://iapp.org/news/a/welcome-to-the-dark-web-a-plain-english-introduction/

Kirkpatrick, K. (2017). Financing the Dark Web. *Communications of the ACM*, 60(3), 21–22. https://dl.acm.org/doi/fullHtml/10.1145/3037386

Kumar, A., & Rosenbach, E. (2019, September). The truth about the Dark Web. IMF. https://www.imf.org/external/pubs/ft/fandd/2019/09/pdf/the-truth-about-the-dark-web-kumar.pdf

Lacson, W., & Jones, B. (2016). The 21st century Darknet Market: Lessons from the fall of Silk Road. *International Journal of Cyber Criminology*, *10*(1). https://www.cybercrimejournal.com/Lacson&Jonesvol10issue1IJCC2016.pdf

Leefeldt, E. (2017, May 17). Hacker’s paradise: Secrets of the “Dark Web.” *CBS news*. https://www.cbsnews.com/news/wannacry-ransomware-hackers-dark-web/

Lorenzo-Dus, N., & Di Cristofaro, M. (2018). ‘I know this whole market is based on the trust you put in me and I don’t take that lightly’: Trust, community and discourse in crypto-drug markets. *Discourse & Communication*, *12*(6), 608–626. Sage Journals. https://www.gwern.net/docs/sr/2018-lorenzodus.pdf

McKinney, W. (2010). Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference*. 445. https://pdfs.semanticscholar.org/ef4e/f7f38bb907e5d7b4df3e6ff1db269d4970f5.pdf?_ga=2.235264902.609808697.1592176667-650151641.1592176667

Meland, P. H., & Sindre, G. (2019, December). Cyber attacks for sale. In 2019 *International Conference on Computational Science and Computational Intelligence (CSCI)*. 54–59. https://ieeexplore.ieee.org/document/9070875

Microsoft. (2020). Create a PivotTable to analyze worksheet data. https://support.office.com/en-us/article/create-a-pivottable-to-analyze-worksheet-data-a9a84538-bfe9-40a9-a8e9-f99134456576

Neo4j, Inc. (2020). Relational databases vs. graph databases: A comparison. https://neo4j.com/developer/graph-db-vs-rdbms/

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). *Lying words: Predicting deception from linguistic styles*. *Personality and Social Psychology Bulletin*, 29(5), 665–675. https://pdfs.semanticscholar.org/2373/cfd1ceefb253294ac97c63522671adc26c2f.pdf?_ga=2.222895004.609808697.1592176667-650151641.1592176667

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. … Cournapeau, D. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(85), 2826–2830. https://dl.acm.org/doi/pdf/10.5555/1953048.2078195

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. University of Texas at Austin. https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf

Perone, S. (2013, December 9). Machine learning: Cosine similarity for vector space models (part III). *Terra Incognita*. http://blog.christianperone.com/2013/09/machine-learning-cosine-similarity-for-vector-space-models-part-iii/

Reitano, T., Oerting, T., & Hunter, M. (2015). Innovations in international cooperation to counter cybercrime: The joint cybercrime action taskforce. *The European Review of Organised Crime*, 2(2), 142–154.

SAS Institute, Inc. (2020). Make every voice heard. https://www.sas.com/en_us/offers/19q3/make-every-voice-heard.html?utm_source=google&utm_medium=cpc&utm_campaign=campaign-ai-ml-us&utm_content=GMS-119788&keyword=sas+nlp&matchtype=e&publisher=google&

Sedgwick, K. (2020, January 12). White House market wants to become the Darknet's toughest DNM. *The Bitcoin News*. https://news.bitcoin.com/white-house-market-wants-to-become-the-darknets-toughest-dnm/

Soska, K., & Christin, N. (2015). Measuring the longitudinal evolution of the online anonymous marketplace ecosystem. In *24th USENIX Security Symposium (USENIX Security '15),* 33–48. https://www.usenix.org/system/files/conference/usenixsecurity15/sec15-paper-soska-updated.pdf

Takaaki, S., Atsuo, Inomata (2019, March). Dark web content analysis and visualization. In *Proceedings of the ACM International Workshop on Security and Privacy Analytics,* 53–59. https://dl.acm.org/doi/pdf/10.1145/3309182.3309189

Toma, C. L., & Hancock, J. T. (2010, February). Reading between the lines: linguistic cues to deception in online dating profiles. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work* (pp. 5–8). https://dl.acm.org/doi/pdf/10.1145/1718918.1718921

The Tor Project, Inc. (2020). About Tor. http://www.torproject.org/about/overview.html.en

Van Buskirk, J., Naicker, S., Bruno, R. B., Breen, C., & Roxburgh, A. (2016). *Drugs and the internet*. National Drug & Alcohol Research Centre. https://ndarc.med.unsw.edu.au/sites/default/files/ndarc/resources/DNeT%20Mar%202016_0.pdf

van Wegberg, R., & Verburgh, T. (2018). Lost in the dream? Measuring the effects of operation bayonet on vendors migrating to dream market. In *Proceedings of the Evolution of the Darknet Workshop*, 1–5. https://pure.tudelft.nl/portal/files/46185682/Wegberg_Verburgh_Lost_in_the_Dream.pdf

Weimann, G. (2016). Terrorist migration to the dark web. *Perspectives on Terrorism, 10*(3), 40–44. www.jstor.org/stable/26297596

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California