Samantha Traversi
MDS 564 Final Project
Dr. Jim Kulich

Advanced Data Mining Analytics
Multilevel Marketing on Twitter

Business Understanding and Background

The principal user activities on social media platforms like Twitter, Facebook, Instagram, and LinkedIn have expanded beyond connecting socially to marketing and sales for small businesses. Some of these small money making ventures include legitimate small businesses like the sale of handmade items on platforms like Etsy and Minted, or secondhand sales of clothing and jewelry on platforms like Poshmark and RetailMeNot. A large number of these small businesses are actually rungs within the pyramid of a multilevel marketing company.

Multilevel marketing is problematic in that the companies are considered legitimate businesses. These companies sell products, like Lularoe sells patterned activewear for women, or services, like Primerica sells insurance policies; however, the majority of the individual seller's income is generated by recruiting new salespeople and receiving a cut of their recruits' sales. These businesses easily oversaturate the market and only early participants will receive significant profit, while people who join later rarely recuperate their sign-on fees.

One of the top strategies of multilevel marketing companies is participants posting overly enthusiastic commentary on social media platforms to hook new recruits. With all the noise within social media platforms, it can be difficult to sniff out a multilevel marketing company from a legitimate small business or side hustle. Since these companies are inherently predatory to the consumer, it is important to identify potential filters that can be put into place that might protect social media users from exposure to posts about these scams. Performing text analysis to determine what language is more frequently used in posts about multilevel marketing companies versus posts about legitimate small businesses is a crucial first step in the process.

Data Source and Attribute Definitions

Between all the social media platforms that are widely used, Twitter was selected as the focus. Twitter has a limit of 280 characters per post, which means users need to get to the point quite quickly. This platform was the origin of the current use of the hashtag, which is a way of marking words or phrases to make post content easily searchable. Twitter also has an Application Programming Interface (API) that makes it easy to scrape data.

For this project, there were time constraints that made it challenging to effectively scrape Twitter through a self-built application. Instead, we used a bot to scrape Twitter for a series of hashtags, specifically, eleven known multilevel marketing companies and eleven known legitimate small businesses, or companies that offer a similar product or service as the multilevel marketing companies. The first few rows of data are pictured in the image below:

	A	B	C	D	E	F	G	H
1	Hashtag	Username	User handle	Date of posting	Text	Retweet count	Like count	MLM
2	#etsy	Pinup Artist • sfw	@__ryusart	Mon Mar 28 22:1	Someone sugge https://t.co/ayeY	0	3	0
3	#lularoe		@_aaashleey	Mon Mar 28 03:3	So good I had to	0	0	1
4	#minted	Julz 🦃	@_AusJulz	Mon Mar 28 01:1	We got an agent 🤝 @basedfishmafi #BFM #NFTs #N Can't wait to see	10	51	0
5	#poshmark	Candice Couture	@_candicecoutu	Tue Mar 29 03:2	So good I had to	2	2	0
6	#minted	DoodleFriends.s	@_DoodleFrienc	Tue Mar 22 03:1	Great news! #20 https://t.co/30mV #NFTCommunity	0	2	0
7	#herbalife	MaFifi	@_Fifinky	Mon Mar 21 19:3	#herbalife #skinc I'm placing order	0	0	1
8	#avon	Keshia Owen	@_GoldenKe	Mon Mar 28 00:2	So good I had to	0	0	1
9	#gofundme	Heather Buckley	@_HeatherBuck	Tue Mar 29 03:0	This post alerted	1	2	0
10	#gofundme	Heather Buckley	@_HeatherBuck	Mon Mar 28 18:1	:D Asking for sto	3	4	0
11	#gofundme	Heather Buckley	@_HeatherBuck	Mon Mar 28 00:4	Asking for some	1	1	0
12	#gofundme	Heather Buckley	@_HeatherBuck	Sun Mar 27 17:2	Asking for some	7	7	0
13	#gofundme	Di Devl	@_idksophie_	Mon Mar 28 13:3	JUST GOT ACC	10	18	0

The attributes of this data set are as follows:

- Hashtag - this is how the tweets were found within Twitter
- Username and User handle - these simply identify the posters
- Date of posting - this is both the date and time of posting
- Text - herein lies the corpus of documents, this is the content of the tweets
- Retweet count - this is the number of other users that re-posted this tweet to their personal account
- Like count - this is the number of other users that liked the tweet
- MLM - this is a binary target variable that is used to categorize the content of the tweets as being about a multilevel marketing company or a legitimate business

Data Cleaning

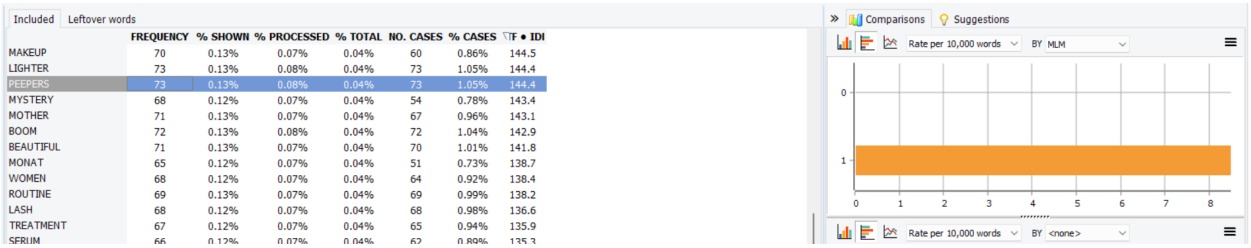
The first steps that were taken to clean this dataset were to remove the emojis, html entities, html tags, and non-printing characters. Twitter users tend to add a lot of extraneous symbols and spacing to call attention to their tweets, so removing this noise was a necessary step.

Exploratory Analysis

The initial exploration of the text was performed in Wordstat as a content categorization with the tweet being the content and the binary classes of multilevel marketing company or legitimate business being the categories. As one might imagine, there was an enormous amount of words and phrases selected through the auto-settings on the application. There were some obvious

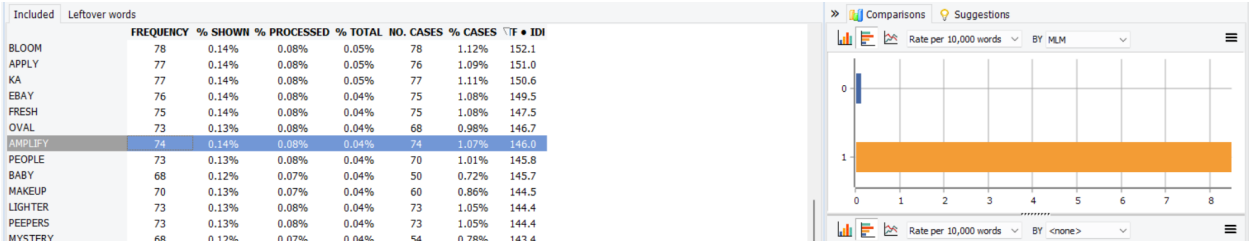
classifiers, like “Etsy” and “Etsy shop” which clearly belong in one of the binary categorizations since “Etsy” was one of the hashtags used to scrape Twitter. Setting aside those n-grams, there were several strong classifiers.

Peepers (synonym for eyes) is a strong indicator of a multilevel marketing company.



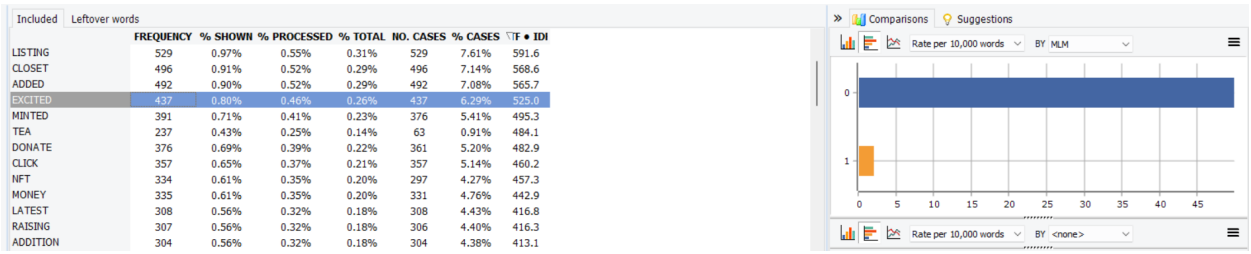
	FREQUENCY	% SHOWN	% PROCESSED	% TOTAL	NO. CASES	% CASES	TF • IDI
MAKEUP	70	0.13%	0.07%	0.04%	60	0.86%	144.5
LIGHTER	73	0.13%	0.08%	0.04%	73	1.05%	144.4
PEEPERS	73	0.13%	0.08%	0.04%	73	1.05%	144.4
MYSTERY	68	0.12%	0.07%	0.04%	54	0.78%	143.4
MOTHER	71	0.13%	0.07%	0.04%	67	0.96%	143.1
BOOM	72	0.13%	0.08%	0.04%	72	1.04%	142.9
BEAUTIFUL	71	0.13%	0.07%	0.04%	70	1.01%	141.8
MONAT	65	0.12%	0.07%	0.04%	51	0.73%	138.7
WOMEN	68	0.12%	0.07%	0.04%	64	0.92%	138.4
ROUTINE	69	0.13%	0.07%	0.04%	69	0.99%	138.2
LASH	68	0.12%	0.07%	0.04%	68	0.98%	136.6
TREATMENT	67	0.12%	0.07%	0.04%	65	0.94%	135.9
SERUM	66	0.12%	0.07%	0.04%	62	0.89%	135.3

Amplify is another strong indicator of a multilevel marketing company.



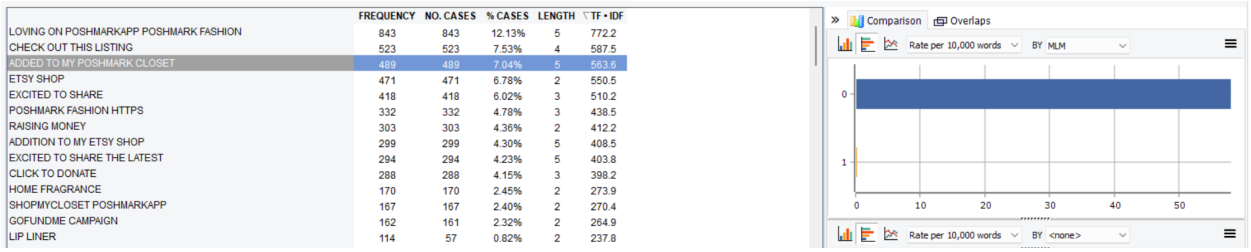
	FREQUENCY	% SHOWN	% PROCESSED	% TOTAL	NO. CASES	% CASES	TF • IDI
BLOOM	78	0.14%	0.08%	0.05%	78	1.12%	152.1
APPLY	77	0.14%	0.08%	0.05%	76	1.09%	151.0
KA	77	0.14%	0.08%	0.05%	77	1.11%	150.6
EBAY	76	0.14%	0.08%	0.04%	75	1.08%	149.5
FRESH	75	0.14%	0.08%	0.04%	75	1.08%	147.5
OVAL	73	0.13%	0.08%	0.04%	68	0.98%	146.7
AMPLIFY	74	0.14%	0.08%	0.04%	74	1.07%	146.0
PEOPLE	73	0.13%	0.08%	0.04%	70	1.01%	145.8
BABY	68	0.12%	0.07%	0.04%	50	0.72%	145.7
MAKEUP	70	0.13%	0.07%	0.04%	60	0.86%	144.5
LIGHTER	73	0.13%	0.08%	0.04%	73	1.05%	144.4
PEEPERS	73	0.13%	0.08%	0.04%	73	1.05%	144.4
MYSTERY	68	0.12%	0.07%	0.04%	54	0.78%	143.4

Excited is a strong indicator of a legitimate business, which is a surprise, since so much of the common language used by direct sellers of multilevel marketing companies is biased toward high energy and high enthusiasm.

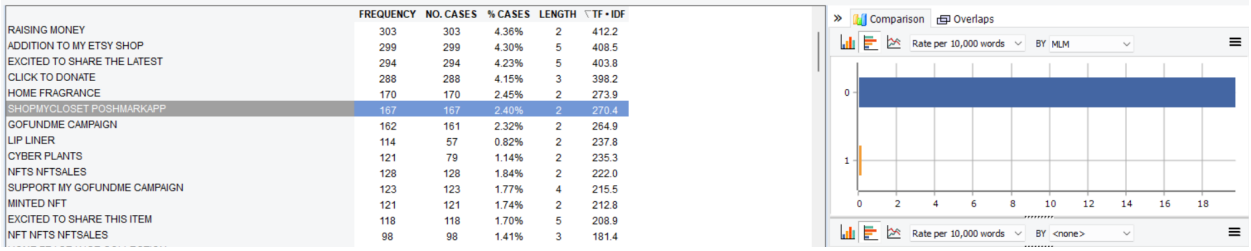


	FREQUENCY	% SHOWN	% PROCESSED	% TOTAL	NO. CASES	% CASES	TF • IDI
LISTING	529	0.97%	0.55%	0.31%	529	7.61%	591.6
CLOSET	496	0.91%	0.52%	0.29%	496	7.14%	568.6
ADDED	492	0.90%	0.52%	0.29%	492	7.08%	565.7
EXCITED	437	0.80%	0.46%	0.26%	437	6.29%	525.0
MINTED	391	0.71%	0.41%	0.23%	376	5.41%	495.3
TEA	237	0.43%	0.25%	0.14%	63	0.91%	484.1
DONATE	376	0.69%	0.39%	0.22%	361	5.20%	482.9
CLICK	357	0.65%	0.37%	0.21%	357	5.14%	460.2
NFT	334	0.61%	0.35%	0.20%	297	4.27%	457.3
MONEY	335	0.61%	0.35%	0.20%	331	4.76%	442.9
LATEST	308	0.56%	0.32%	0.18%	308	4.43%	416.8
RAISING	307	0.56%	0.32%	0.18%	306	4.40%	416.3
ADDITION	304	0.56%	0.32%	0.18%	304	4.38%	413.1

Pairing “Poshmark” with “Closet” is a strong indicator of a legitimate business. This is not initially a surprise since Poshmark was one of the hashtags for legitimate business.

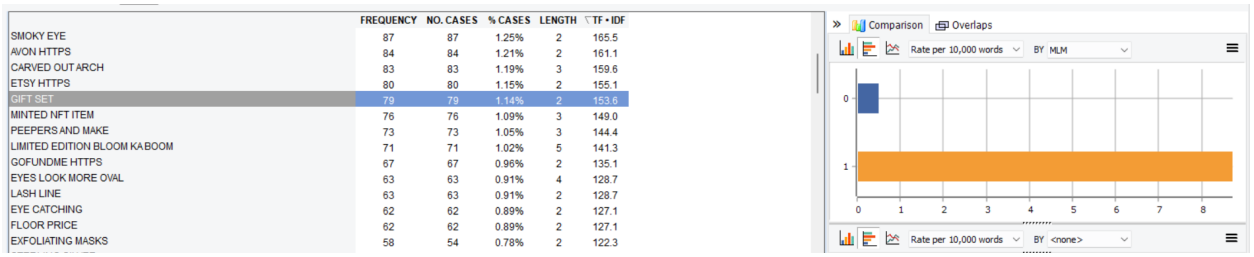


	FREQUENCY	NO. CASES	% CASES	LENGTH	TF • IDF
LOVING ON POSHMARKAPP POSHMARK FASHION	843	843	12.13%	5	772.2
CHECK OUT THIS LISTING	523	523	7.53%	4	587.5
ADDED TO MY POSHMARK CLOSET	489	489	7.04%	5	563.6
ETSY SHOP	471	471	6.78%	2	550.5
EXCITED TO SHARE	418	418	6.02%	3	510.2
POSHMARK FASHION HTTPS	332	332	4.78%	3	438.5
RAISING MONEY	303	303	4.36%	2	412.2
ADDITION TO MY ETSY SHOP	299	299	4.30%	5	408.5
EXCITED TO SHARE THE LATEST	294	294	4.23%	5	403.8
CLICK TO DONATE	288	288	4.15%	3	398.2
HOME FRAGRANCE	170	170	2.45%	2	273.9
SHOPMYCLOSET POSHMARKAPP	167	167	2.40%	2	270.4
GOFUNDME CAMPAIGN	162	161	2.32%	2	264.9
LIP LINER	114	57	0.82%	2	237.8



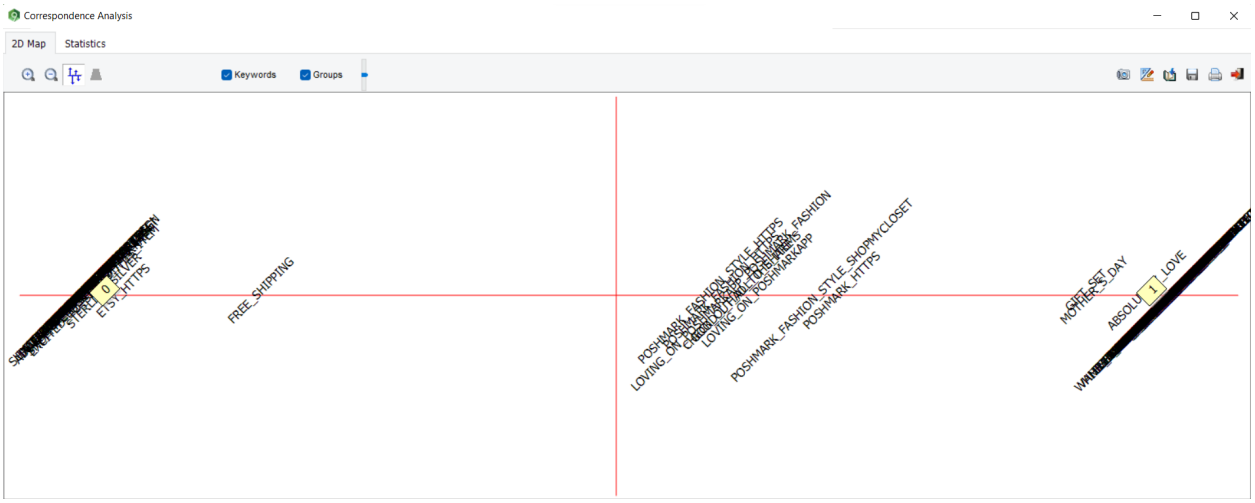
	FREQUENCY	NO. CASES	% CASES	LENGTH	TF • IDF
RAISING MONEY	303	303	4.36%	2	412.2
ADDITION TO MY ETSY SHOP	299	299	4.30%	5	408.5
EXCITED TO SHARE THE LATEST	294	294	4.23%	5	403.8
CLICK TO DONATE	288	288	4.15%	3	398.2
HOME FRAGRANCE	170	170	2.45%	2	273.9
SHOPMYCLOSET POSHMARKAPP	167	167	2.40%	2	270.4
GOFUNDME CAMPAIGN	162	161	2.32%	2	264.9
LIP LINER	114	57	0.82%	2	237.8
CYBER PLANTS	121	79	1.14%	2	235.3
NFTS NFTSALES	128	128	1.84%	2	222.0
SUPPORT MY GOFUNDME CAMPAIGN	123	123	1.77%	4	215.5
MINTED NFT	121	121	1.74%	2	212.8
EXCITED TO SHARE THIS ITEM	118	118	1.70%	5	208.9
NFT NFTS NFTSALES	98	98	1.41%	3	181.4

“Gift set” is a strong indicator of a multilevel marketing company.

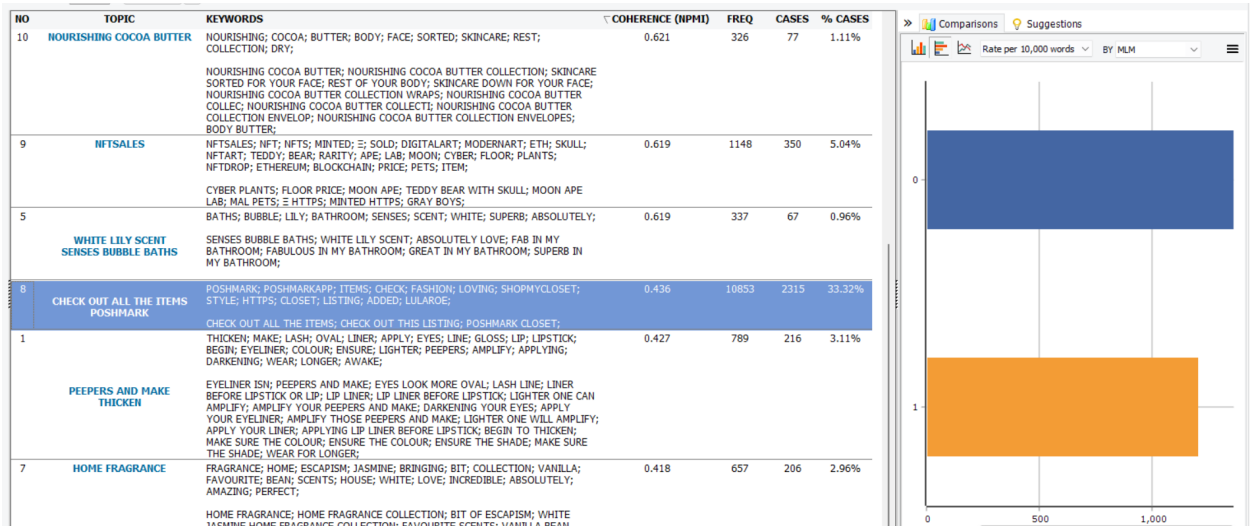


	FREQUENCY	NO. CASES	% CASES	LENGTH	TF • IDF
SMOKY EYE	87	87	1.25%	2	165.5
AVON HTTPS	84	84	1.21%	2	161.1
CARVED OUT ARCH	83	83	1.19%	3	159.6
ETSY HTTPS	80	80	1.15%	2	155.1
GIFT SET	79	79	1.14%	2	153.6
MINTED NFT ITEM	76	76	1.09%	3	149.0
PEEPERS AND MAKE	73	73	1.05%	3	144.4
LIMITED EDITION BLOOM KA BOOM	71	71	1.02%	5	141.3
GOFUNDME HTTPS	67	67	0.96%	2	135.1
EYES LOOK MORE OVAL	63	63	0.91%	4	128.7
LASH LINE	63	63	0.91%	2	128.7
EYE CATCHING	62	62	0.89%	2	127.1
FLOOR PRICE	62	62	0.89%	2	127.1
EXFOLIATING MASKS	58	54	0.78%	2	122.3

Putting the phrases into a spectrum between the two categories shows interesting results. For example, “Free shipping” tends to appear in more content about legitimate businesses, while, again, “Gift set” is an indicator of a multilevel marketing company, as is the phrase, “Absolutely love”



An interesting result from this exploration is that Poshmark was originally listed as a legitimate business, which it is, but apparently content about Poshmark appears in tweets about both legitimate businesses and multilevel marketing companies. At this point, I looked back into Poshmark to find out that some of the direct sales on Poshmark are for products and services offered by multilevel marketing companies.



NO	TOPIC	KEYWORDS	COHERENCE (NPMI)	FREQ	CASES	% CASES
10	NOURISHING COCOA BUTTER	NOURISHING; COCOA; BUTTER; BODY; FACE; SORTED; SKINCARE; REST; COLLECTION; DRY; NOURISHING COCOA BUTTER; NOURISHING COCOA BUTTER COLLECTION; SKINCARE SORTED FOR YOUR FACE; REST OF YOUR BODY; SKINCARE DOWN FOR YOUR FACE; NOURISHING COCOA BUTTER COLLECTION WRAPS; NOURISHING COCOA BUTTER COLLEC; NOURISHING COCOA BUTTER COLLECTI; NOURISHING COCOA BUTTER COLLECTION ENVELOP; NOURISHING COCOA BUTTER COLLECTION ENVELOPES; BODY BUTTER;	0.621	326	77	1.11%
9	NFTSALES	NFTSALES; NFT; NFTS; MINTED; Ξ; SOLD; DIGITALART; MODERNART; ETH; SKULL; NFTART; TEDDY; BEAR; RARITY; APE; LAB; MOON; CYBER; FLOOR; PLANTS; NFTDROP; ETHEREUM; BLOCKCHAIN; PRICE; PETS; ITEM; CYBER PLANTS; FLOOR PRICE; MOON APE; TEDDY BEAR WITH SKULL; MOON APE LAB; MAL PETS; Ξ HTTPS; MINTED HTTPS; GRAY BOYS;	0.619	1148	350	5.04%
5	WHITE LILY SCENT SENSES BUBBLE BATHS	BATHS; BUBBLE; LILY; BATHROOM; SENSES; SCENT; WHITE; SUPERB; ABSOLUTELY; SENSES BUBBLE BATHS; WHITE LILY SCENT; ABSOLUTELY LOVE; FAB IN MY BATHROOM; FABULOUS IN MY BATHROOM; GREAT IN MY BATHROOM; SUPERB IN MY BATHROOM;	0.619	337	67	0.96%
8	CHECK OUT ALL THE ITEMS POSHMARK	POSHMARK; POSHMARKAPP; ITEMS; CHECK; FASHION; LOVING; SHOPMYCLOSET; STYLE; HTTPS; CLOSET; LISTING; ADDED; LULAROE; CHECK OUT ALL THE ITEMS; CHECK OUT THIS LISTING; POSHMARK CLOSET;	0.436	10853	2315	33.32%
1	PEEPERS AND MAKE THICKEN	THICKEN; MAKE; LASH; OVAL; LINER; APPLY; EYES; LINE; GLOSS; LIP; LIPSTICK; BEGIN; EYELINER; COLOUR; ENSURE; LIGHTER; PEEPERS; AMPLIFY; APPLYING; DARKENING; WEAR; LONGER; AWAKE; EYELINER ISN; PEEPERS AND MAKE; EYES LOOK MORE OVAL; LASH LINE; LINER BEFORE LIPSTICK OR LIP; LIP LINER; LIP LINER BEFORE LIPSTICK; LIGHTER ONE CAN AMPLIFY; AMPLIFY YOUR PEEPERS AND MAKE; DARKENING YOUR EYES; APPLY YOUR EYELINER; AMPLIFY THOSE PEEPERS AND MAKE; LIGHTER ONE WILL AMPLIFY; APPLY YOUR LINER; APPLYING LIP LINER BEFORE LIPSTICK; BEGIN TO THICKEN; MAKE SURE THE COLOUR; ENSURE THE COLOUR; ENSURE THE SHADE; MAKE SURE THE SHADE; WEAR FOR LONGER;	0.427	789	216	3.11%
7	HOME FRAGRANCE	FRAGRANCE; HOME; ESCAPISM; JASMINE; BRINGING; BIT; COLLECTION; VANILLA; FAVOURITE; BEAN; SCENTS; HOUSE; WHITE; LOVE; INCREDIBLE; ABSOLUTELY; AMAZING; PERFECT; HOME FRAGRANCE; HOME FRAGRANCE COLLECTION; BIT OF ESCAPISM; WHITE	0.418	657	206	2.96%

Modeling
The model that was selected for this project was to use a Passive Aggressive Classifier with a TFIDF Vectorizer.

The TFIDF Vectorizer consists of two parts: term frequency and inverse document frequency. This is a measure of the terms that make a big impact when they appear in the documents within the corpus. To measure term frequency alone means that across the corpus, a term appears often, which may be significant, but may not be. To measure only inverse document frequency means to give weight to rarer words in documents. TFIDF measures a combination of the two, both the important terms, and the documents in which they appear.

The Passive Aggressive Classifier is an iterative algorithm that learns incrementally. The algorithm responds passively to correct classifications (it lets the correct classifications remain as they are), and aggressively to incorrect classifications (the adjustment to the classification method occurs upon incorrect classification results).

The training set is 20% of the total dataset and the test set is the remaining 80%.

```
[12] x_train,x_test,y_train,y_test = train_test_split(t_df['Text'], labels, test_size=0.2, random_state=7)

[13] tfidf_vectorizer = TfidfVectorizer(stop_words='english', max_df=0.7)

[14] tfidf_train = tfidf_vectorizer.fit_transform(x_train)
     tfidf_test = tfidf_vectorizer.transform(x_test)

[15] pac = PassiveAggressiveClassifier(max_iter = 50)
     pac.fit(tfidf_train,y_train)

     PassiveAggressiveClassifier(max_iter=50)

[16] y_pred = pac.predict(tfidf_test)

[17] score = accuracy_score(y_test,y_pred)
     print(f'Accuracy: {round(score*100,2)}%')

     Accuracy: 91.8%
```

The accuracy of the model is 91.8%, which is a very high level of accuracy for such a small dataset and with minimal feature engineering.

Model Evaluation
One of the model evaluation methods employed for this project was a confusion matrix:

```
[19] confusion_matrix(y_test,y_pred, labels=[1,0])
0s

     array([[657,  18],
            [ 96, 619]])
```

We can see in the array that out of all the tweets about multilevel marketing companies in the test set, 657 of them were correctly categorized as multilevel marketing tweets. Of all the tweets about legitimate companies, 619 were correctly categorized as legitimate companies. Ninety-six of the legitimate companies were categorized as multilevel marketing companies, and 18 of the multilevel marketing companies were categorized as legitimate companies.

If our purpose is to prevent vulnerable individuals from being exposed to and potentially scammed by content about multilevel marketing, this skew towards over-predicting twitter posts as multilevel marketing content is a positive.

Recommendations for Future Steps

If further analysis is performed, the first step would be to scrape other social media platforms like Facebook, Instagram, and LinkedIn. These platforms do not restrict users to the character limit that Twitter enforces, so the text analysis will likely uncover more patterns within the text that can be used to identify strategies commonly used by multilevel marketing companies.

For a more in-depth analysis and higher prediction accuracy, I recommend using a dictionary method, such as dictionary-based sentiment analysis. These analytical methods can identify brand personality which can be used to more immediately recognize these types of posts.

Finally, if a much larger dataset was collected, the posting behavior of the content creators could potentially be used in conjunction with the text analysis as a predictor of multilevel marketing content. From my own experience as a social media user, I recognize that multilevel marketing posts are published on a schedule and with a high frequency of post output. This pattern of behavior could be analyzed and incorporated into the model to create highly accurate predictions of multilevel marketing content.

Works Cited

Botester Team. “Twitter Hashtag Scraper.” *botster*, 2022, https://botster.io/bots/twitter-hashtag-scraper.