ML Team 4 Work

June 2020 Session

Forum post classification into a specific category

PLAN

A. Introduction
B. Project Outline
C. Data Collection & Processing
D. Basic Data Modeling
E. Advanced Data Modeling
F. Results Discussion
G. Conclusion

Introduction

ML team 4

Project Lead: Sara El-Ateif

PM Lead: Rohit Ganti

Participants: Kevin Dong, Gabriel Jai, @yeszoey, Khushali Verma, Ananya De, Mutaz Ahmed, Kitty Gu, @whuang19

Observers: Shalini Kumari, @WillP, @Merron_Tecleab, @fxc2000

14 interns : 2 lead, 8 participants, 4 observers

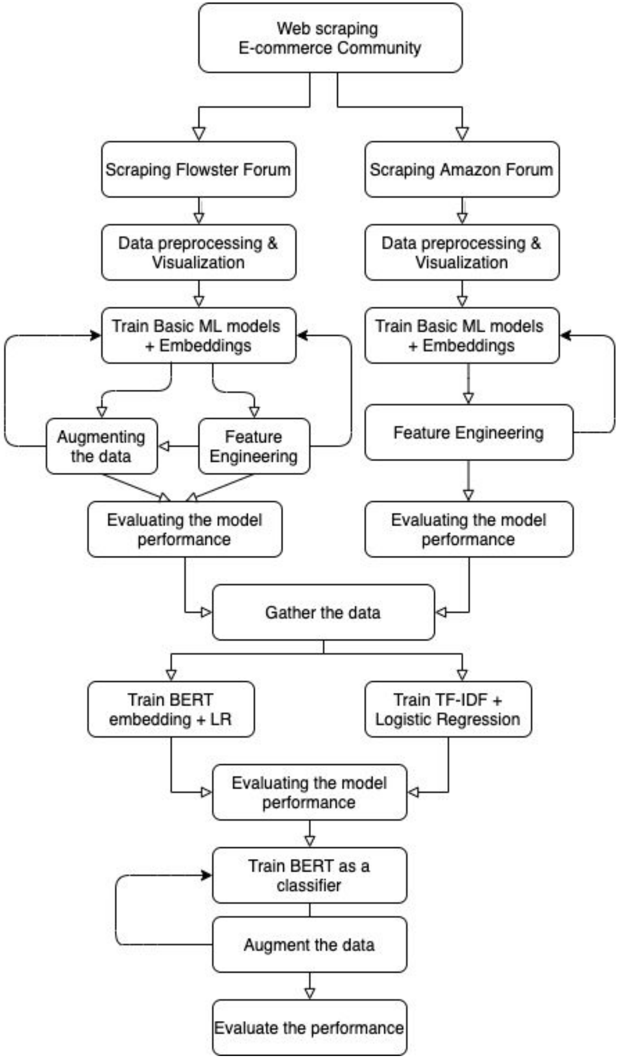
Please check our profiles in STEMAway Forum.

Goal of the project

Classify a given post from a specific forum from the E-commerce DiscourseHub Community into its appropriate category.

Project Outline

Project Workflow

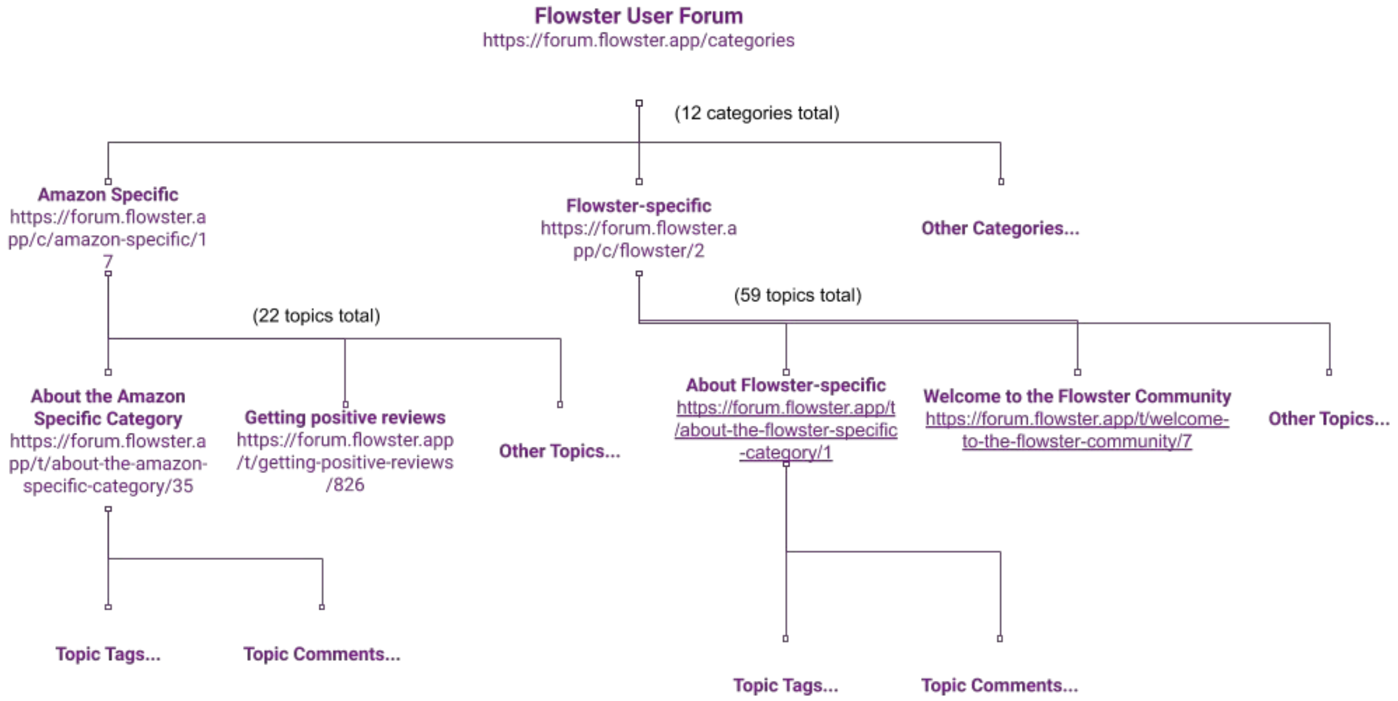


Web scraping
E-commerce Community
Scraping Flowster Forum
Scraping Amazon Forum
Data preprocessing & Visualization
Data preprocessing & Visualization
Train Basic ML models + Embeddings
Train Basic ML models + Embeddings
Augmenting the data
Feature Engineering
Feature Engineering
Evaluating the model performance
Evaluating the model performance
Gather the data
Train BERT embedding + LR
Train TF-IDF + Logistic Regression
Evaluating the model performance
Train BERT as a classifier
Augment the data
Evaluate the performance

Data Collection & Processing

Discourse Forum Site Map Example

The Flowster and Amazon Seller forums both use the same structure



Flowster User Forum
https://forum.flowster.app/categories

(12 categories total)

Amazon Specific
https://forum.flowster.app/c/amazon-specific/17

Flowster-specific
https://forum.flowster.app/c/flowster/2

Other Categories...

(22 topics total)

About the Amazon Specific Category
https://forum.flowster.app/t/about-the-amazon-specific-category/35

Getting positive reviews
https://forum.flowster.app/t/getting-positive-reviews/826

Other Topics...

Topic Tags...

Topic Comments...

(59 topics total)

About Flowster-specific
https://forum.flowster.app/t/about-the-flowster-specific-category/1

Welcome to the Flowster Community
https://forum.flowster.app/t/welcome-to-the-flowster-community/7

Other Topics...

Topic Tags...

Topic Comments...

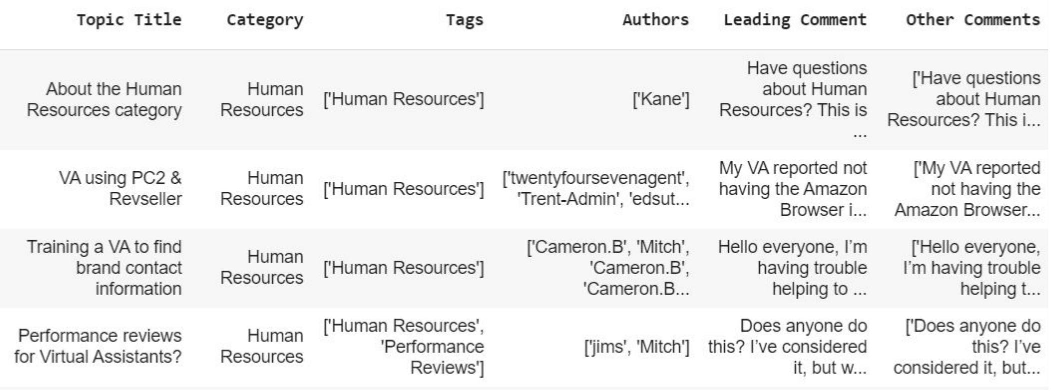
Data Collection from Amazon and Flowster Forums

Web Scraping Libraries:

Beautiful Soup

Selenium

Scraped Data:



Topic Title	Category	Tags	Authors	Leading Comment	Other Comments
About the Human Resources category	Human Resources	['Human Resources']	['Kane']	Have questions about Human Resources? This is ...	['Have questions about Human Resources? This i...
VA using PC2 & Revseller	Human Resources	['Human Resources']	['twentyfoursevenagent', 'Trent-Admin', 'edsut...	My VA reported not having the Amazon Browser i...	['My VA reported not having the Amazon Browser...
Training a VA to find brand contact information	Human Resources	['Human Resources']	['Cameron.B', 'Mitch', 'Cameron.B', 'Cameron.B...	Hello everyone, I'm having trouble helping to ...	['Hello everyone, I'm having trouble helping t...
Performance reviews for Virtual Assistants?	Human Resources	['Human Resources', 'Performance Reviews']	['jims', 'Mitch']	Does anyone do this? I've considered it, but w...	['Does anyone do this? I've considered it, but...

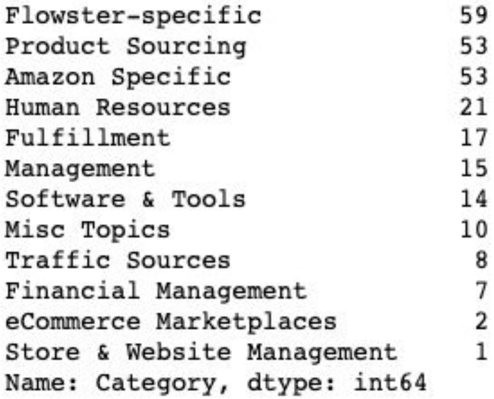
Flowster Data
Exploratory Data Analysis(EDA)

Data Cleaning Techniques

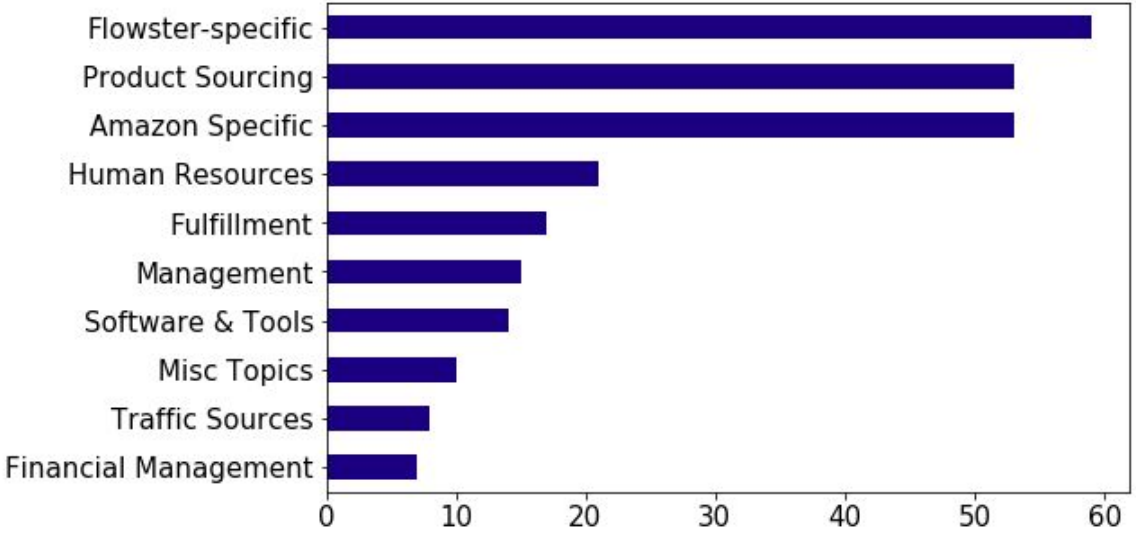
- Data augmentation by turning the topic reply comments into “new” topics within the same category
- Lowercase normalization
- Punctuation removal
- Stopword removal
- Rare word removal
- Numerics removal
- Non-ASCII character removal
- Numeric and cost replacements using unique identifiers “########” and “$$$$$$$$”

Class Imbalance

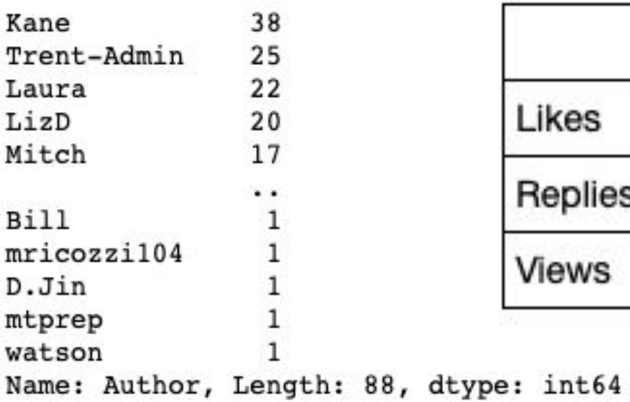
Need to be taken into account when training classifiers.



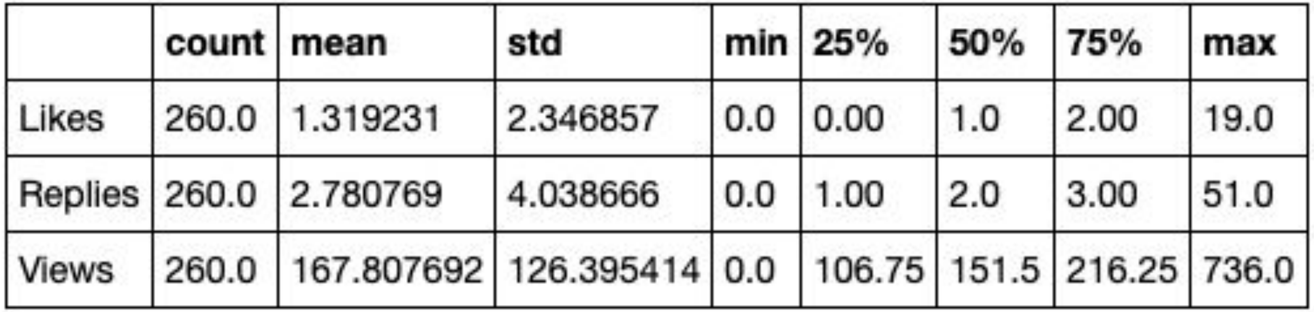
```
Flowster-specific             59
Product Sourcing              53
Amazon Specific               53
Human Resources               21
Fulfillment                   17
Management                    15
Software & Tools              14
Misc Topics                   10
Traffic Sources                8
Financial Management           7
eCommerce Marketplaces         2
Store & Website Management     1
Name: Category, dtype: int64
```



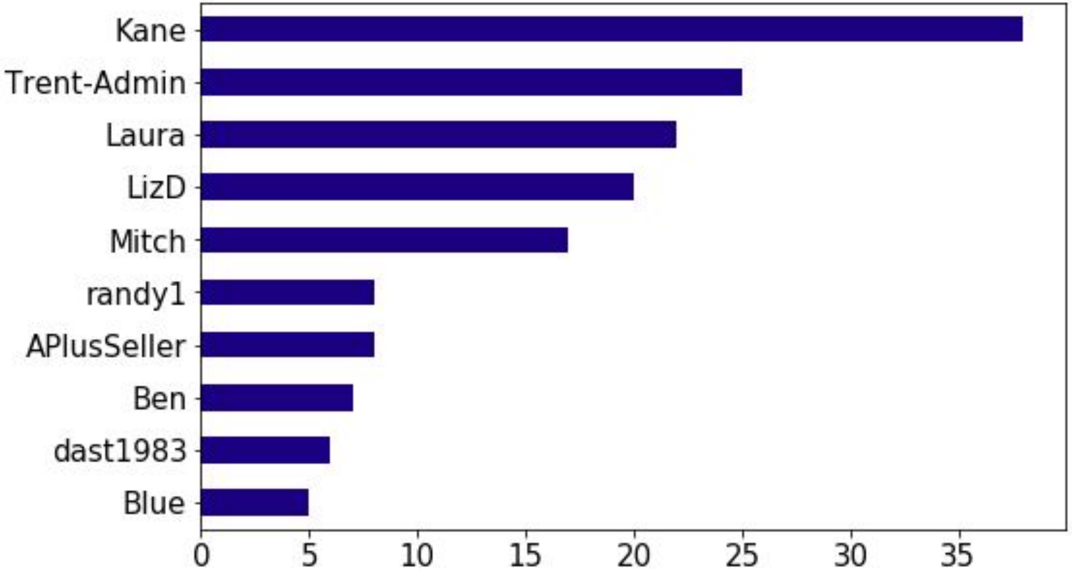
Relatively Low Activity of the Flowster Forum



```
Kane           38
Trent-Admin    25
Laura          22
LizD           20
Mitch          17
               ..
Bill            1
mricozzi104     1
D.Jin           1
mtprep          1
watson          1
Name: Author, Length: 88, dtype: int64
```

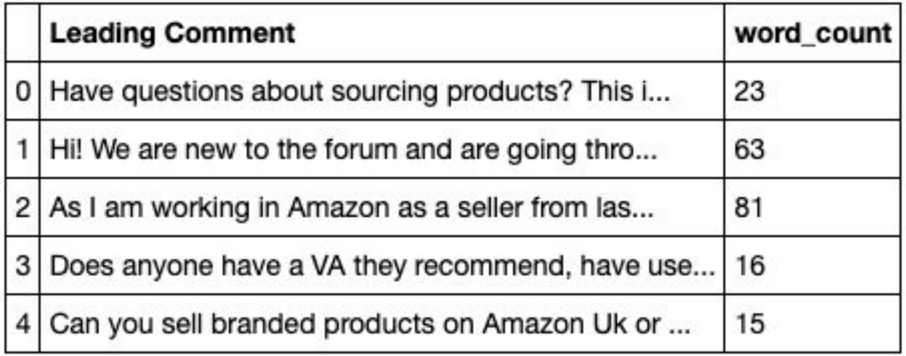


	count	mean	std	min	25%	50%	75%	max
Likes	260.0	1.319231	2.346857	0.0	0.00	1.0	2.00	19.0
Replies	260.0	2.780769	4.038666	0.0	1.00	2.0	3.00	51.0
Views	260.0	167.807692	126.395414	0.0	106.75	151.5	216.25	736.0

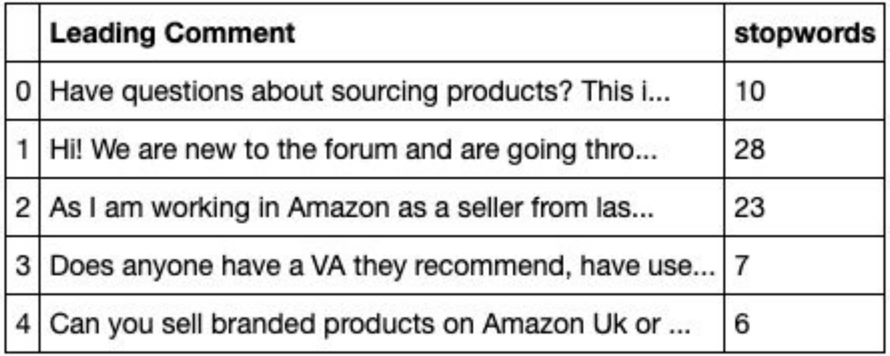


Many Stop Words

Implies that removing stop words may hurt the accuracy of our analysis



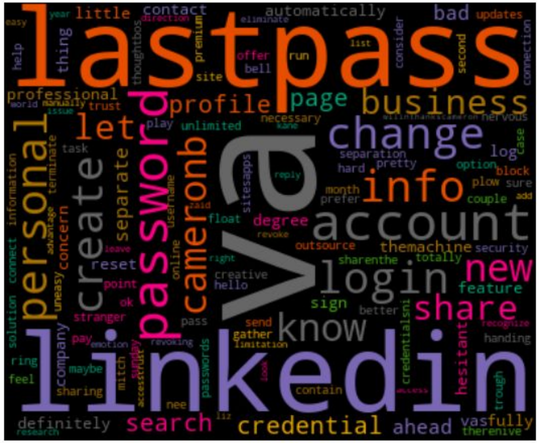
	Leading Comment	word_count
0	Have questions about sourcing products? This i...	23
1	Hi! We are new to the forum and are going thro...	63
2	As I am working in Amazon as a seller from las...	81
3	Does anyone have a VA they recommend, have use...	16
4	Can you sell branded products on Amazon Uk or ...	15



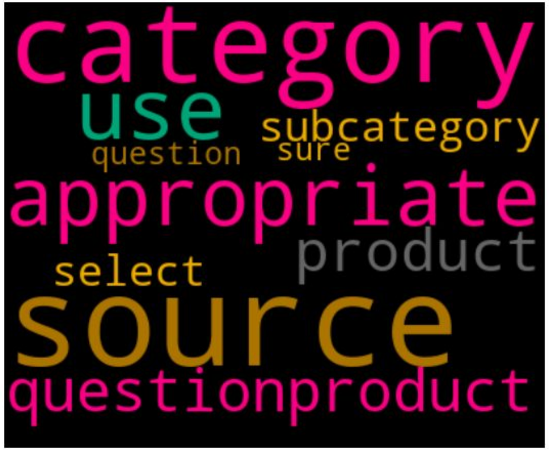
	Leading Comment	stopwords
0	Have questions about sourcing products? This i...	10
1	Hi! We are new to the forum and are going thro...	28
2	As I am working in Amazon as a seller from las...	23
3	Does anyone have a VA they recommend, have use...	7
4	Can you sell branded products on Amazon Uk or ...	6

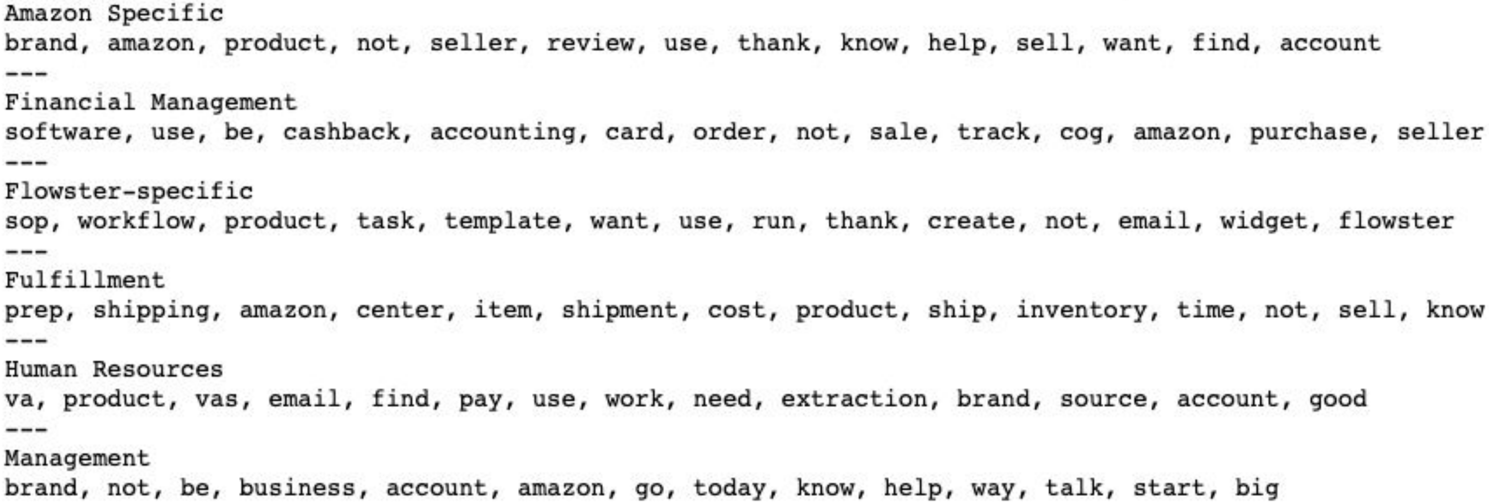
Generated Word Clou and lists of frequent words better understand the data

Popular Post



Unpopular Post





```
Amazon Specific
brand, amazon, product, not, seller, review, use, thank, know, help, sell, want, find, account
---
Financial Management
software, use, be, cashback, accounting, card, order, not, sale, track, cog, amazon, purchase, seller
---
Flowster-specific
sop, workflow, product, task, template, want, use, run, thank, create, not, email, widget, flowster
---
Fulfillment
prep, shipping, amazon, center, item, shipment, cost, product, ship, inventory, time, not, sell, know
---
Human Resources
va, product, vas, email, find, pay, use, work, need, extraction, brand, source, account, good
---
Management
brand, not, be, business, account, amazon, go, today, know, help, way, talk, start, big
```

Amazon Data
Exploratory Data Analysis(EDA)

Data Cleaning Techniques

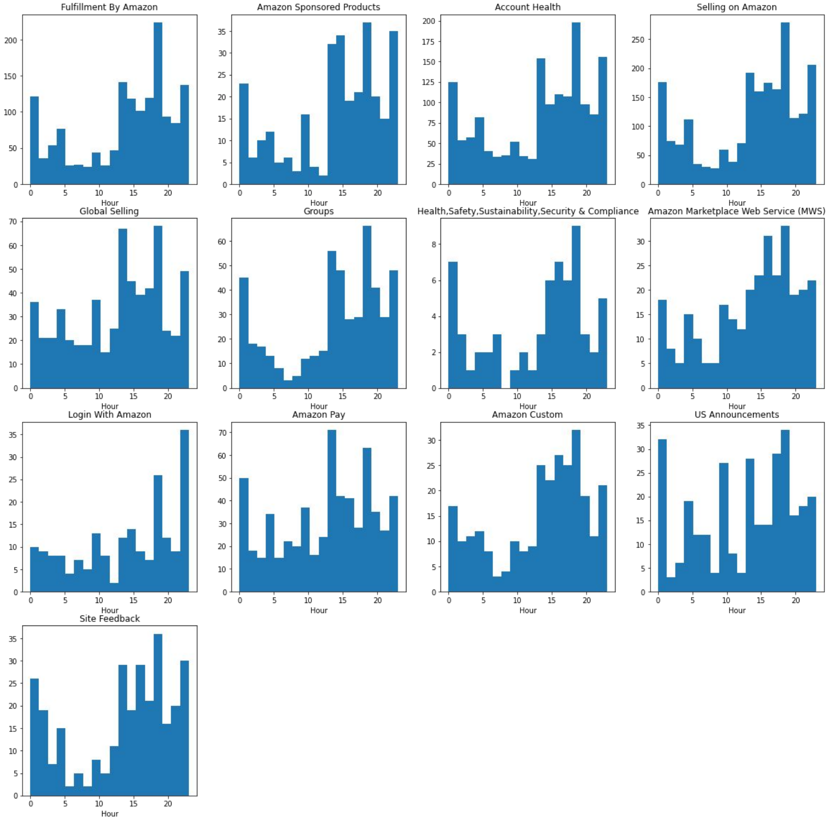
- Lowercase normalization
- Punctuation removal
- Stopword removal
- Non-ASCII character removal
- URL removal

Numeric features do not seem helpful in distinguishing categories

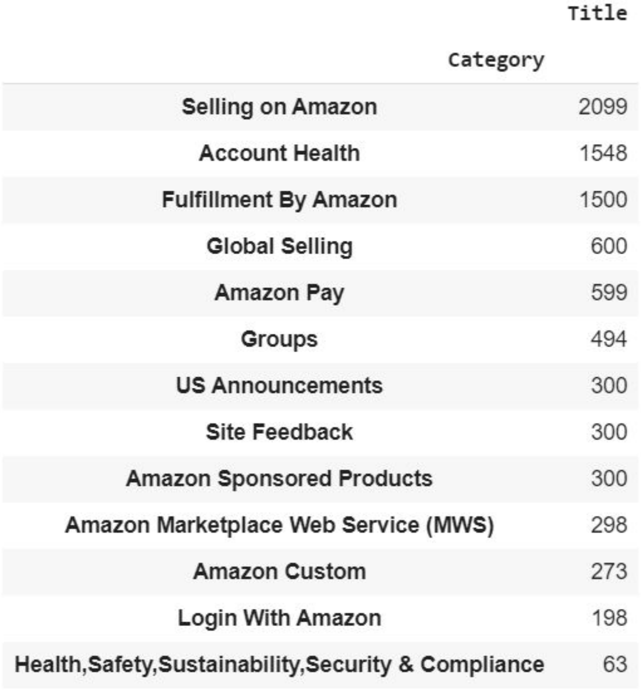
Most of the posts in every category possessed a similar pattern in numeric features.

For example, the majority of posts was published in the afternoon and evening.

Because they were very similar, it would not be very helpful to use the numeric feature as one of the predictors.



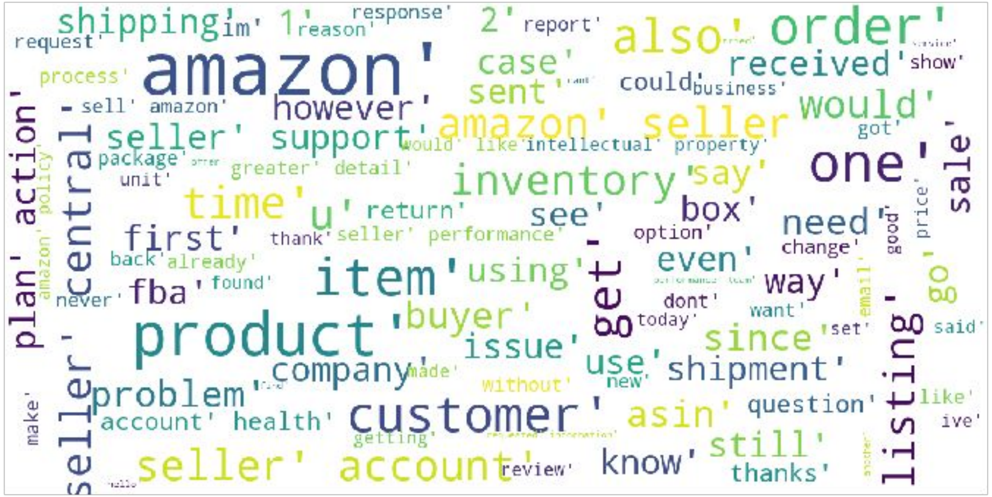
Class Imbalance



Category	Title
Selling on Amazon	2099
Account Health	1548
Fulfillment By Amazon	1500
Global Selling	600
Amazon Pay	599
Groups	494
US Announcements	300
Site Feedback	300
Amazon Sponsored Products	300
Amazon Marketplace Web Service (MWS)	298
Amazon Custom	273
Login With Amazon	198
Health,Safety,Sustainability,Security & Compliance	63

Class imbalance for the amazon data is less drastic than that of the flowster's data.

Generated Word Map:



Basic Data Modeling

Models & Embeddings Experimented

Basic ML models

Naive Bayes

Linear SVM

Logistic Regression

XGBoost

Word embeddings

Bag of Words (Count Vectorizer)

TF-IDF

Word2Vec

Doc2Vec

Flowster Data
Basic Data Modeling

ML Models vs Text Embeddings Performance (Flowster Dataset)

	Bag of Words	TF-IDF	Word2Vec	Doc2Vec
Linear SVM	--	0.55	--	0.6
Logistic Regression	0.85	0.51, 0.83	0.19	0.68, 0.55
Naive Bayes	0.59	0.59, 0.48	0.11	0.29

TF-IDF + Linear SVM

Data Preprocessing	**Augmentation**	**Accuracy**
Lowercase all words + Lemmatization + Remove digits, words containing digits, extra spaces, punctuations, rare words, common words, stop words lemmatization	NO	51%
Lowercase all words + Remove digits, words that contain digits, extra spaces, punctuations	NO	56%
Remove punctuations and stop words + Lowercase all words	YES	Got different results from different code source: 78% 53%

Amazon Data
Basic Data Modeling

ML Models vs Text Embeddings Performance (Amazon Dataset)

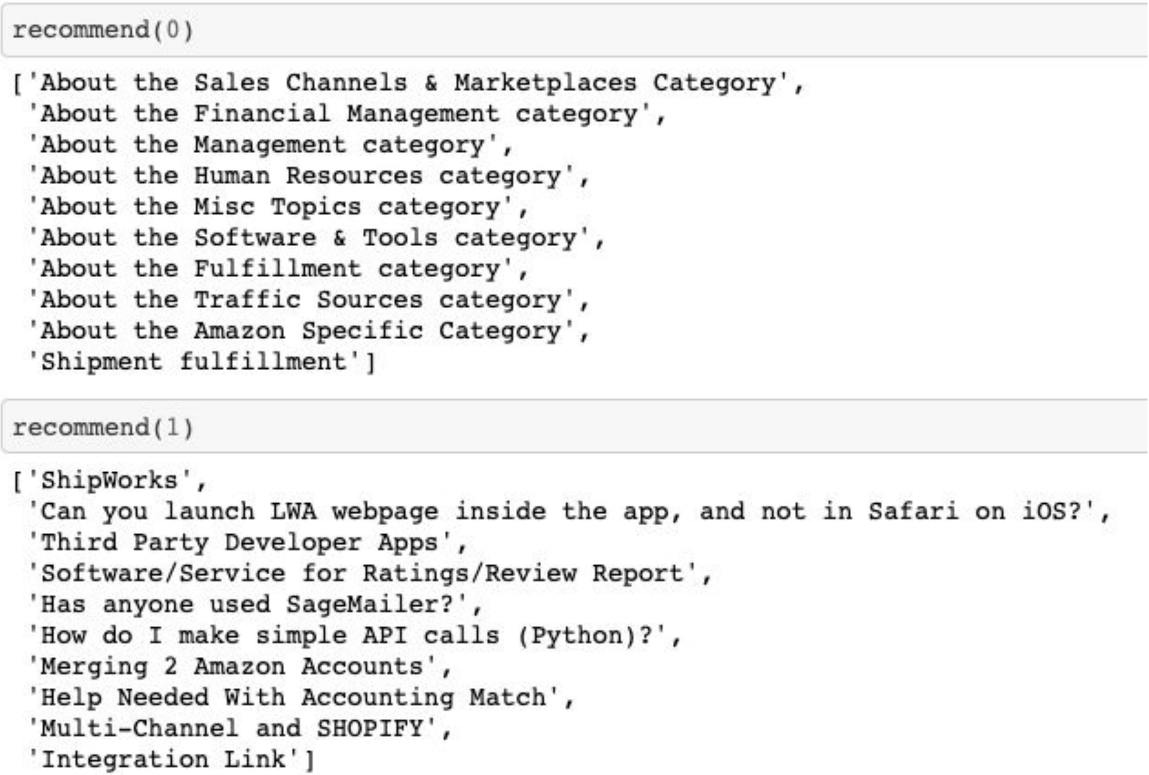
	CountVectorizer	**TF-IDF**	**Doc2Vec**
Logistic regression	0.68	0.72	0.29
Random forests	--	0.49	0.22
XG Boost	--	0.64	0.27

Advanced Data Modeling

Understanding the advanced models embeddings

DistilBERT (embedding) + Similarity Calculation

Built a recommend function (trained with DistilBERT embedding) that takes in the index of the topic and gives a list of 10 most similar topics as recommendations.

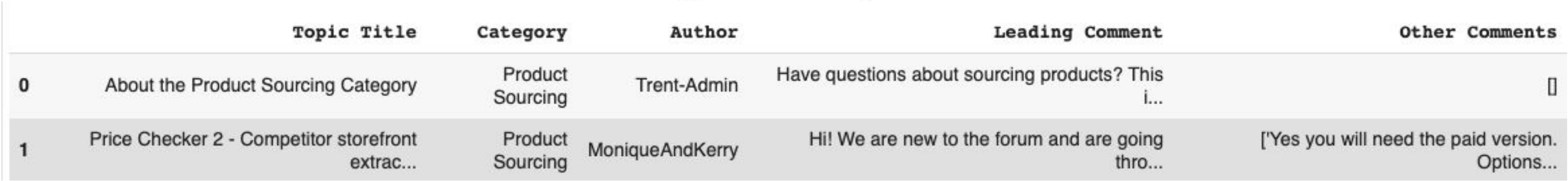


```
recommend(0)
```

```
['About the Sales Channels & Marketplaces Category',
 'About the Financial Management category',
 'About the Management category',
 'About the Human Resources category',
 'About the Misc Topics category',
 'About the Software & Tools category',
 'About the Fulfillment category',
 'About the Traffic Sources category',
 'About the Amazon Specific Category',
 'Shipment fulfillment']
```

```
recommend(1)
```

```
['ShipWorks',
 'Can you launch LWA webpage inside the app, and not in Safari on iOS?',
 'Third Party Developer Apps',
 'Software/Service for Ratings/Review Report',
 'Has anyone used SageMailer?',
 'How do I make simple API calls (Python)?',
 'Merging 2 Amazon Accounts',
 'Help Needed With Accounting Match',
 'Multi-Channel and SHOPIFY',
 'Integration Link']
```



	Topic Title	Category	Author	Leading Comment	Other Comments
0	About the Product Sourcing Category	Product Sourcing	Trent-Admin	Have questions about sourcing products? This i...	[]
1	Price Checker 2 - Competitor storefront extrac...	Product Sourcing	MoniqueAndKerry	Hi! We are new to the forum and are going thro...	['Yes you will need the paid version. Options...

Training with different data tuning

Method 1: Training without categories that have small amount of samples

Method 2: Training with data augmented (by substitution) on categories had small sample size (<250) using TF-IDF, Roberta, BERT, DistillBert, WordNet, GPT-2 using nlpaug library

Method 3: Training by first transforming and incorporating the reply comments into leading comments, and then augmenting (by substitution) this new dataset by using BERT, DistillBERT and WordNet using nlpaug library

Method 4: Training with data augmented by round-trip translation (RTT). Multiple rounds of RTT were performed on leading and reply comments to increase the sample numbers of the unbalanced categories.

Training TF-IDF + Logistic Regression (LR) on merged data

Fine-Tuned TF-IDF + Logistic Regression Results

Strategy	**Accuracy**
Training **without data augmentation**	68%
Augmented Data using **method 1** formula	69%
Augmented Data using **method 2** formula	**88%**

Training BERT classifier without data augmentation

Fine tuning the BERT model

Parameters (batch_size=8)	Accuracy

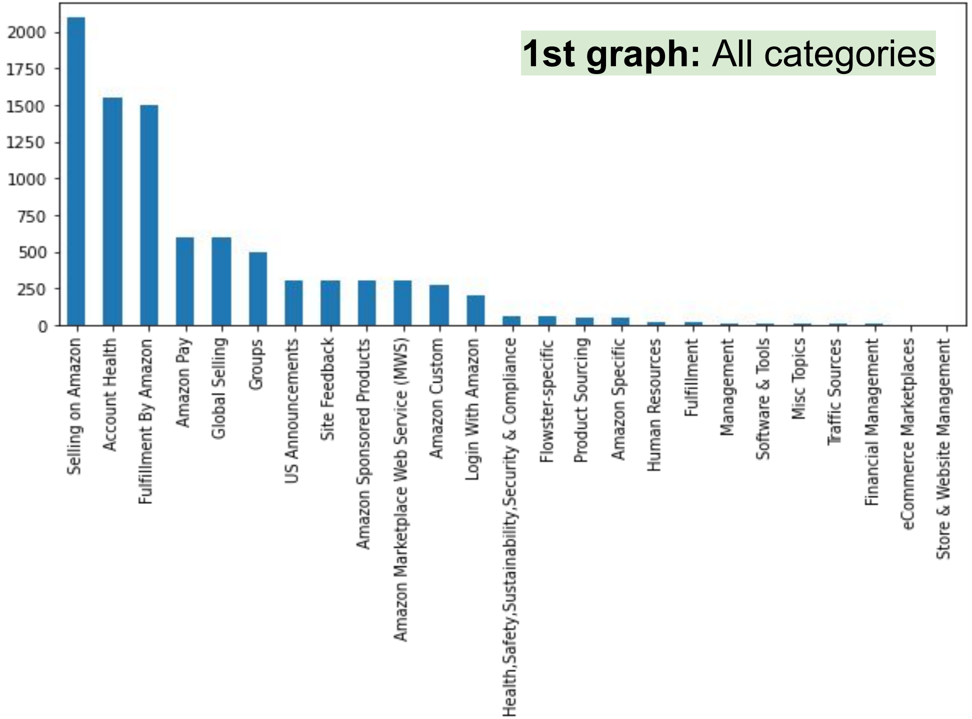
Max_seq_length = 128
Num_train_epochs = 4.0
	67%

Max_seq_length = 256
Num_train_epochs = 3.0
	68%

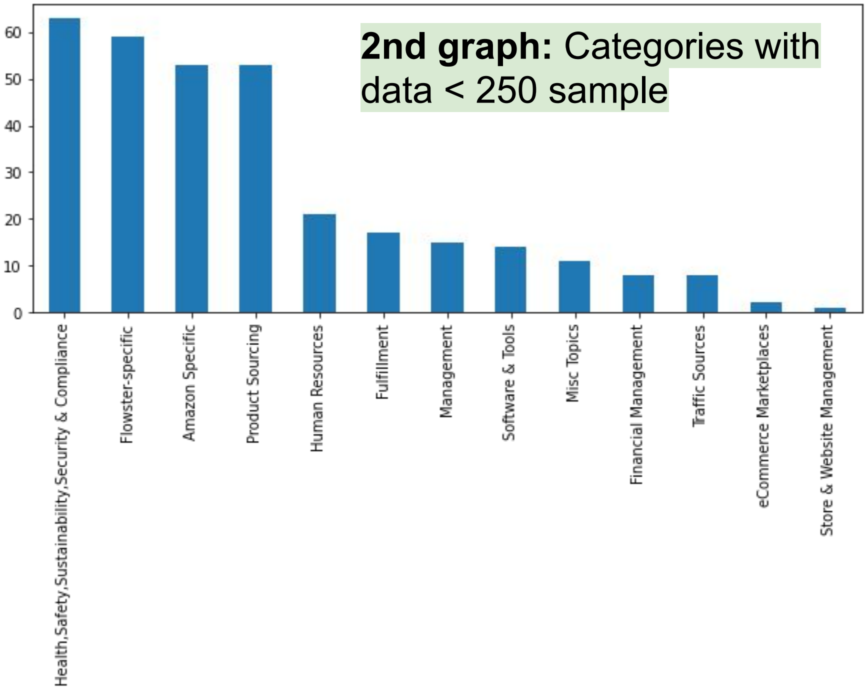
Max_seq_length = 512
(442 tokens from head, 70 tokens from tail)
Num_train_epochs = 3.0
	68%

Max_seq_length = 512
Num_train_epochs = 3.0
	68.7%

Max_seq_length = 512 **Num_train_epochs = 4.0**	**70%**



1st graph: All categories



2nd graph: Categories with data < 250 sample

Class Imbalance

Our dataset suffers from the class imbalance problem in a distinguished manner as shown in the 1st graph.

As the 2nd graph demonstrates we have 12 categories with data less than 250 samples.

Training BERT classifier with data augmentation

Data Augmentation or Dropping Effects

Parameters (batch_size=8, max_seq_length=512, epochs=4)	**Accuracy**
Method 1: Dropping categories with less data	72%
Method 2: Augmenting using TF-IDF, Roberta, BERT, DistillBert, WordNet, GPT-2	4%
Method 3: Incorporating the Reply Comments into Leading Comments, followed by augmenting the new data by using BERT, DIstilBERT and WordNet	78%
Method 4: Augmenting using RTT	**81%**

Results Discussion

Data augmentation results discussion

- Using TF-IDF, Roberta, BERT, DistilBert, WordNet, GPT-2 from nlpaug library dropped the accuracy to a worrying level and this could be due to the noise they introduced to the data.

Example: Using '**roberta-base**'

```
Original:
Have questions about Store & Website Management? This is the category to use. Please be sure to
select the most appropriate sub-category for your questions.
Augmented Text:
[' FDA is asking me to send them with a registration of my products I did submit them a schedule
but they said it � � d not the correct one . I buy the ingredient of the products from a man u af
act urer , then I rep ack aged them and sell . I have know idea on how they get on the FDA website
because they only allowed me to register a small facility .
```

Data augmentation results discussion

- Training by first transforming and adding the Reply Comments as Leading Comments and then augmented (by substitution) this new data by using BERT, DistilBERT and WordNet using nlpaug library.

  This approach was less noisy in comparison with the previous one as we can see from the following example.

  Example: Using **‘wordnet’**

```
Original:
['Amazon is asking me to provide them with a registration of my product. I did submit them a
registration but they said it’s not the correct one. I buy the ingredient of the capsules from a
manuafacturer, then I repackaged them and sell. I have know clue on how to register on the FDA
website because they only allowed me to register a food facility. Please help me!!!']
Augmented Text:
Amazon is asking me to provide them with a registration of my product . I did relegate them a
enrollment but they said it ’ s not the correct one . I corrupt the ingredient of the capsules
from a manuafacturer , and then 1 repackaged them and sell . I sustain jazz clue on how to
register on the FDA website because they only allowed me to register a food facility . Please help
me ! ! !
```

Conclusion

Shortcomings & Improvements

Shortcomings:

- The model still mis-classifies some classes
- Data Quality

Improvements:

- Playing around with the tokenization process
- Fine tuning BERT learning rate and batch size parameters
- Improve the data augmentation process
- Improve the recall score
- Improve data quality