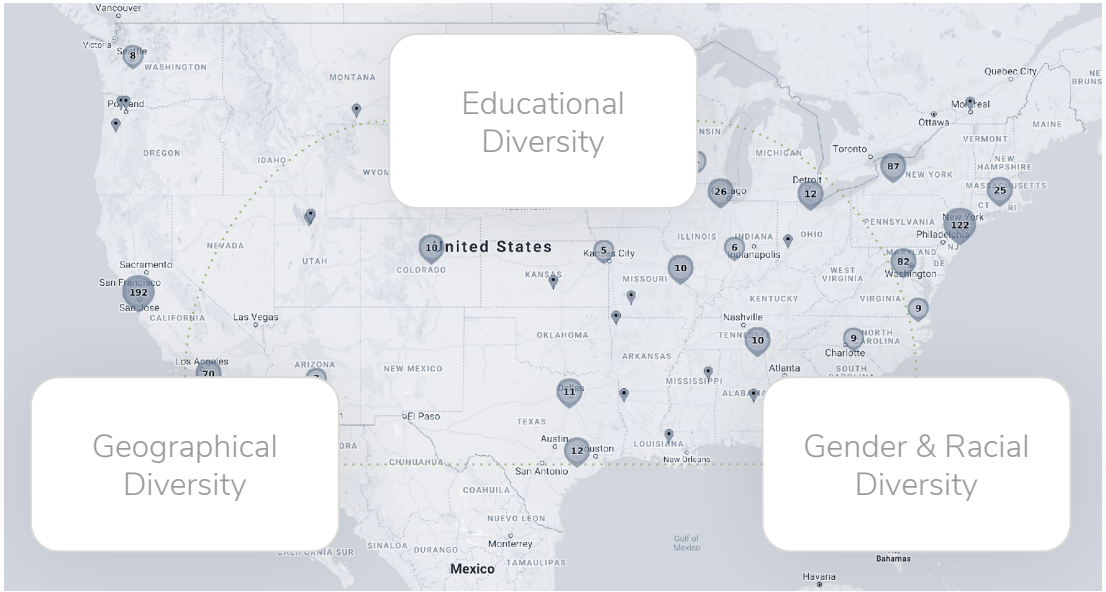
STEM-Away

STEM-Away®
Machine Learning Pathway

Rethinking internships as the way to expand the work-ready STEM talent pool beyond traditional demographics and colleges



STEM-Away Virtual-Internships are **organized as Mentor Chains® Projects**

STEM-Away Mentors (STEM Professionals)

Mentor Chains® Leads

Mentor Chains® Participants

Mentor Chains® Observers

STEM-Away® Virtual-Internships are organized as Mentor Chains® Projects:
Industry experts mentor student leads.
Leads mentor teams of 8-15 students.
Granularity of Mentor Chains® dependent on pathway.

Exponential increase in STEM internship opportunities with the Mentor Chains® structure.

Building the next generation of inclusive STEM leaders by giving students an early start on leadership & mentoring skills.

Project Goals are set by mentors and pathway leads.

Team-centric collaborative projects

End-to-end execution involving both lateral and strategic thinking

Safe environment to push boundaries

Stress on communication skills, organizational skills, and effective working with diverse teams

Periodic self-assessments

Free for students

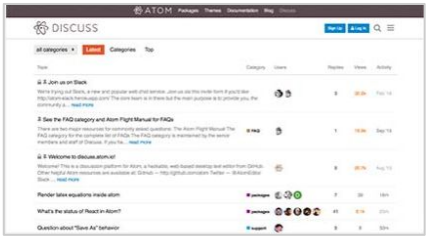
Discourse

The Open Source Forum Codebase with 1,500+ Customers

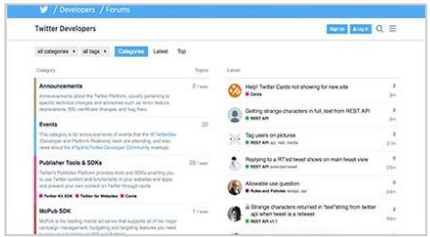
StackExchange Founder Vows to Reinvent Online Discourse

Source: Wired

ATOM

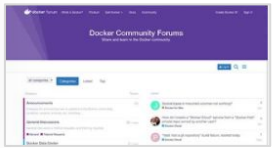


TWITTER DEVELOPERS



RUST

DOCKER

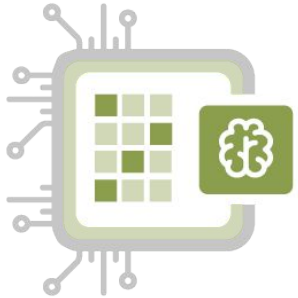


EMBER.JS

Exploratory Data Analysis

Sub optimal properties of collected data

How pre-processing affects performance of different models



Sub optimal properties of collected data

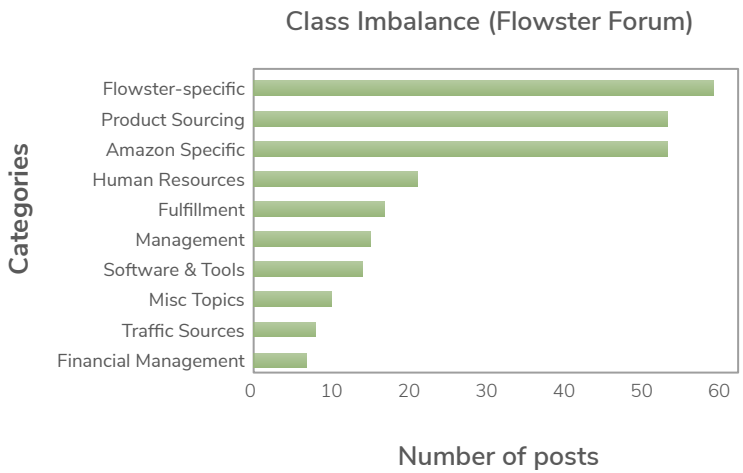


Figure 1: Number of posts per category (Flowster Forum)

High count of Stop Words (Flowster Forum)

	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

Table 1: Word count per sentence (or row)

TF-IDF: Term Frequency - Inverse Document Frequency

$$W_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term **x** within document **y**

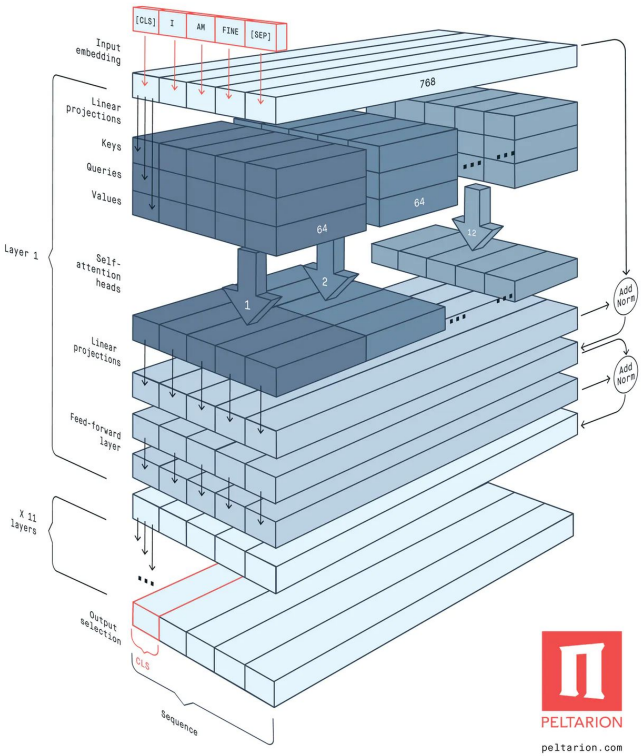
$tf_{x,y}$ = Frequency of **x** in **y**

df_x = number of documents containing **x**

N = total number of documents

Source: https://www.quentinfily.fr/tf-idf-pertinence-lexicale/

BERT: Bidirectional Representation for Transformers



Source: https://peltarion.com/blog/data-science/illustration-3d-bert

Evaluation metrics

Confusion Matrix

Predict \ Actual	0	1
0	TN	FN
1	FP	TP

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

TF-IDF+Linear SVM with various data preprocessing

Data Preprocessing (Flowster forum)	Augmentation	Accuracy (Linear SVM)
Lemmatization + Remove digits, words containing digits, extra spaces, punctuations, rare words, common words, stop words lemmatization	NO	51%
Remove digits, words that contain digits, extra spaces, punctuations	NO	56%
Remove punctuations and stop words	YES	**78%**

Fine-Tuning the BERT model

Parameters (batch_size=8)	Accuracy (BERT + Simple FFN)
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 = 4.0	**70%**

Results of BERT

Despite the model yielding decent results for most categories, it had difficulty of identifying specific characteristic of categories with less data

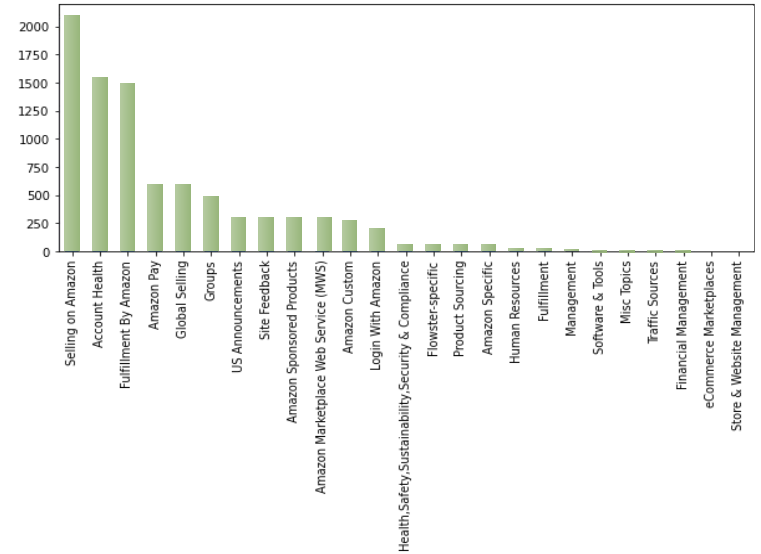
In order to address this issue and improve the performance of the model, data augmentation would seem to be needed

	precision	recall	f1-score	support
0	0.81	0.88	0.84	293
1	0.85	0.40	0.54	58
2	0.62	0.68	0.65	65
3	0.65	0.69	0.67	118
4	0.00	0.00	0.00	13
5	0.64	0.79	0.71	61
6	0.00	0.00	0.00	3
7	0.00	0.00	0.00	12
8	0.00	0.00	0.00	1
9	0.70	0.74	0.72	313
10	0.55	0.47	0.50	120
11	0.71	0.64	0.67	88
12	0.00	0.00	0.00	12
13	0.00	0.00	0.00	3
14	0.69	0.73	0.71	37
15	0.00	0.00	0.00	2
17	0.00	0.00	0.00	11
18	0.56	0.64	0.59	440
19	0.28	0.13	0.18	52
20	0.00	0.00	0.00	5
22	0.75	0.84	0.79	62
accuracy			0.66	1769

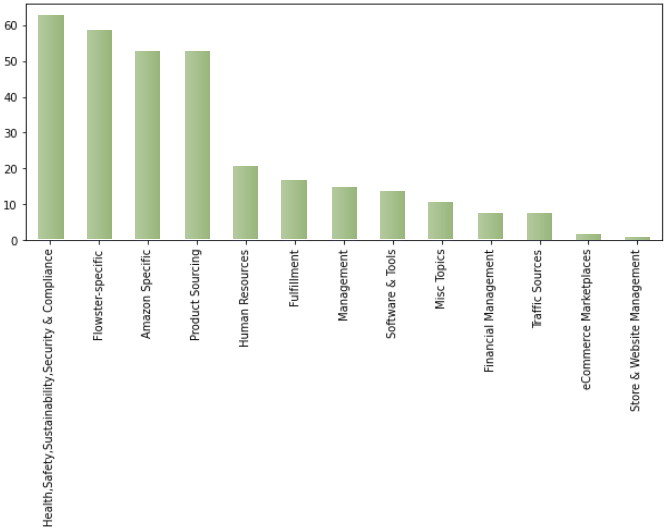
Data Augmentation

Class Imbalance

All Categories



Categories with <250 Samples



Augmentation Methods

Method 1: Omit categories that have a small sample size

Method 2: Augment (by substitution) categories with small sample sizes using TF-IDF, Roberta, BERT, DistillBert, WordNet and GPT-2

Method 3: Combine leading comments and reply comments before augmenting (by substitution) this new dataset using BERT, DistillBERT and WordNet

Method 4: Augment categories with small sample sizes using multiple rounds of round-trip translation (RTT)

Data Augmentation Results

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

Augmenting using TF-IDF, Roberta, BERT, DistillBert, WordNet, GPT-2

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 repackaged 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 .

Combining leading and reply comments, followed by augmenting the new data by using BERT, DIstilBERT and WordNet

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 ! ! !

Augmenting using RTT

Original

['About the Sales Channels & Marketplaces Category', 'Have questions about sourcing the various sales channels available to you? This is the category to use. Please be sure to select the most appropriate sub-category for your questions.']

['Information on the category of sales channels and retail space', 'Do you have questions about finding different sales channels? This is the category to use. Please remember to select the most appropriate subcategory for your questions.']

['About sales channels and marketplaces', 'Do you have questions about obtaining the various sales channels available? This is the category to use. Please make sure you select the most appropriate subcategory for your questions.']

Data Augmentation Results Discussion

	precision	recall	f1-score	support
0	0.89	0.90	0.90	134
1	0.80	0.64	0.71	25
2	0.83	0.75	0.79	32
3	0.80	0.76	0.78	49
4	0.89	1.00	0.94	25
5	0.84	0.87	0.86	31
6	1.00	1.00	1.00	14
7	1.00	1.00	1.00	17
8	0.95	1.00	0.98	20
9	0.77	0.81	0.79	142
10	0.64	0.71	0.67	55
11	0.93	0.74	0.82	53
12	0.92	0.92	0.92	24
13	1.00	1.00	1.00	15
14	0.79	0.69	0.73	16
15	1.00	1.00	1.00	15
16	1.00	1.00	1.00	9
17	1.00	1.00	1.00	19
18	0.65	0.70	0.67	168
19	0.56	0.33	0.42	27
20	1.00	1.00	1.00	21
21	1.00	1.00	1.00	13
22	0.83	0.86	0.85	35
23	1.00	1.00	1.00	10
accuracy			0.81	969

Panel Discussion Topics

1) Differences between working on these types of ML projects as individual research vs working in a team to deploy a product
    a) Collaboration
    b) Strategic thinking
    c) End-to-end execution
2) Challenges with remote teamwork
    a) Project management and planning
    b) Task assignment and accountability
    c) Team bonding and motivation
3) How success is evaluated in the context of these STEM-Away internship
4) Key takeaways and future plans