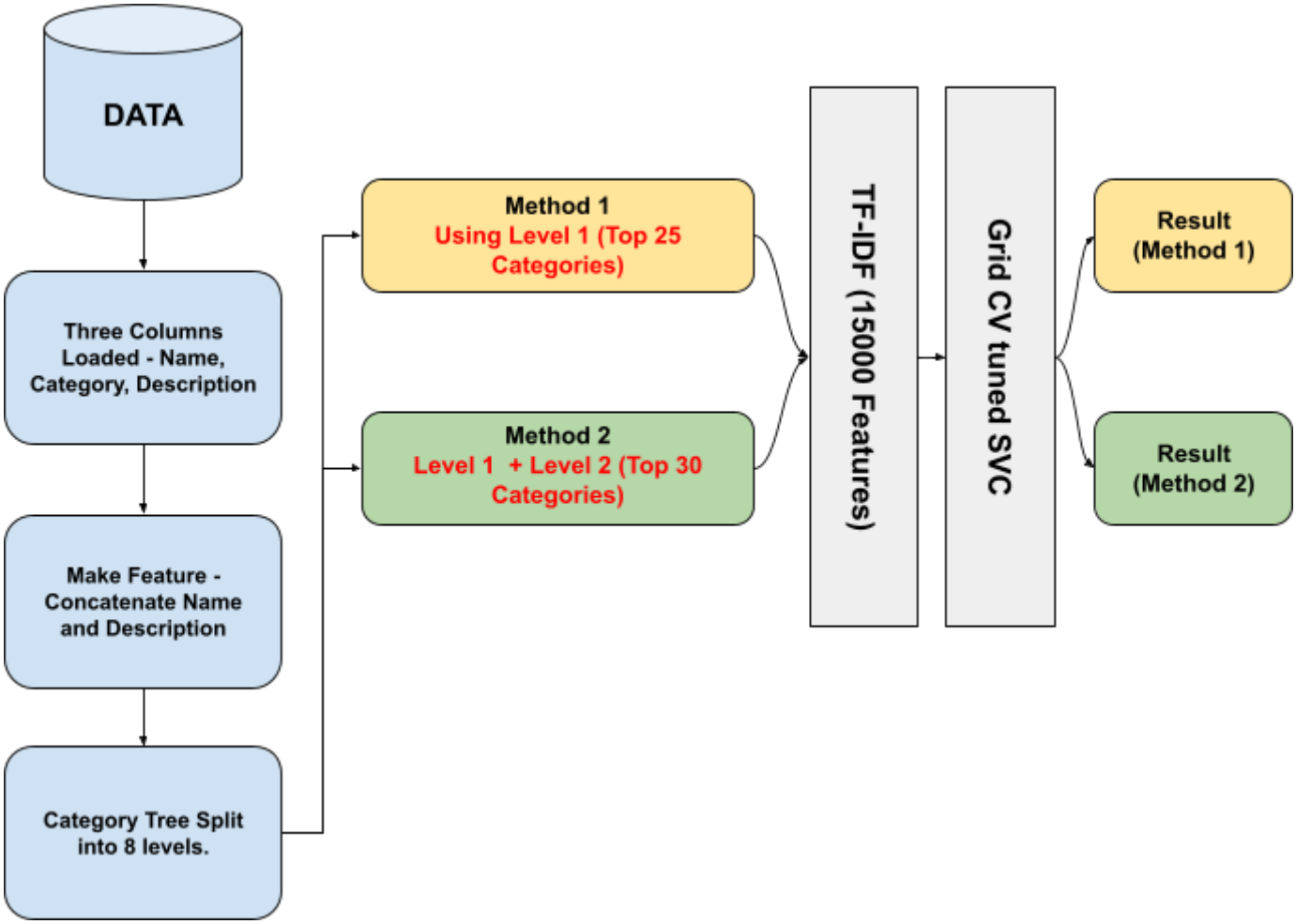
MIDAS - IIITD (Internship Task, 2021)
Task - 3 (NLP) - Detailed Explanation

Introduction to Methodology

It is recommended to go through the ReadMe.md file for insights into directory structure and how to run the code on a local platform. For the convenience of the reader, the methodology hasbeen condensed into a flowchart below. **The chart only shows a high level overview and steps like cleaning and preprocessing are not shown.**



Data Preprocessing and Setup

Loading Data and Basic Inferences:

1. Basic EDA was carried out on the dataset. Data contained 15 columns and 20003 rows.
2. For the code, only 3 columns were loaded -> Product Name, Product Description, Product Category Tree. Others were ignored as they did not contain any data relevant to the prediction task.
3. While loading the dataset, rows with NaN type objects were not loaded and dropped. Product Name and Product Description were concatenated into a new feature column. Simple analysis after this, showed the following results:
4. Furthermore, the dataset was higly imbalanced due to the category tree being unique for most products.

The features of the dataset and a screenshot are shown below:

1. Number of Rows - 20000
2. Number of Columns - product_name, product_description, product_category_tree (3)
3. Unique Categories (Considering each tree as unique category) - 6466

	product_name	product_category_tree	description
0	Alisha Solid Women's Cycling Shorts	["Clothing >> Women's Clothing >> Lingerie, Sl...	Key Features of Alisha Solid Women's Cycling S...
1	FabHomeDecor Fabric Double Sofa Bed	["Furniture >> Living Room Furniture >> Sofa B...	FabHomeDecor Fabric Double Sofa Bed (Finish Co...
2	AW Bellies	["Footwear >> Women's Footwear >> Ballerinas >...	Key Features of AW Bellies Sandals Wedges Heel...
3	Alisha Solid Women's Cycling Shorts	["Clothing >> Women's Clothing >> Lingerie, Sl...	Key Features of Alisha Solid Women's Cycling S...
4	Sicons All Purpose Arnica Dog Shampoo	["Pet Supplies >> Grooming >> Skin & Coat Care...	Specifications of Sicons All Purpose Arnica Do...

Loaded Dataset

Splitting Category Tree into Levels and Selecting Targets:

1. Each category tree was split into levels. The max number of levels for any tree was seen to be 8.
2. **On average, only the first three levels were seen to contain usefull data throughout the dataset.**
3. A screenshot is shown below.

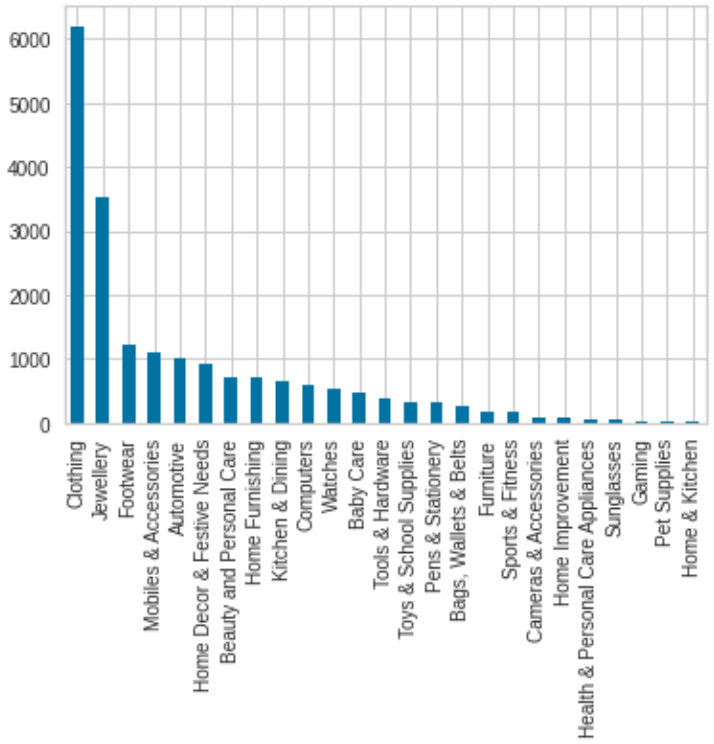
	Level1	Level2	Level3	Level4	Level5	Level6	Level7	Level8	NameDesc
0	Clothing	Women's Clothing	Lingerie, Sleep & Swimwear	Shorts	Alisha Shorts	Alisha Solid Women's Cycling Shorts	None	None	Alisha Solid Women's Cycling ShortsKey Feature...
1	Furniture	Living Room Furniture	Sofa Beds & Futons	FabHomeDecor Fabric Double Sofa Bed (Finish Co...	None	None	None	None	FabHomeDecor Fabric Double Sofa BedFabHomeDeco...
2	Footwear	Women's Footwear	Ballerinas	AW Bellies	None	None	None	None	AW BelliesKey Features of AW Bellies Sandals W...
3	Clothing	Women's Clothing	Lingerie, Sleep & Swimwear	Shorts	Alisha Shorts	Alisha Solid Women's Cycling Shorts	None	None	Alisha Solid Women's Cycling ShortsKey Feature...
4	Pet Supplies	Grooming	Skin & Coat Care	Shampoo	Sicons All Purpose Arnica Dog Shampoo (500 ml)	None	None	None	Sicons All Purpose Arnica Dog ShampooSpecifica...

Levels split into columns. Only the first three are considered.

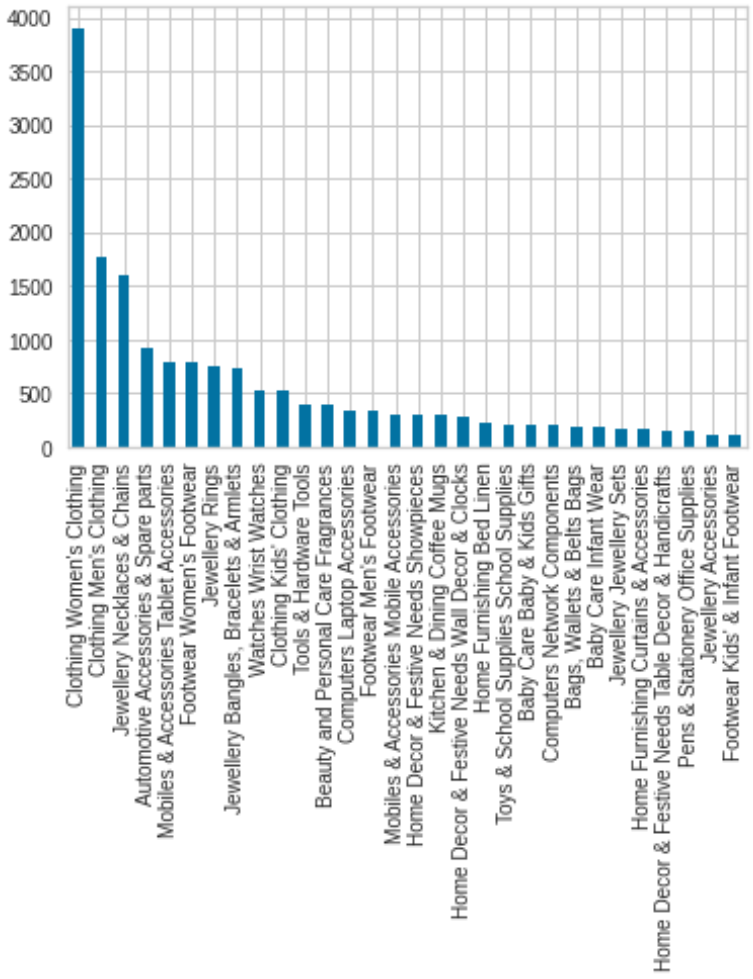
4. Level 1 and Level 2 contained 265 and 218 unique categories repectively. A brief description of both levels is given below.
5. Level 1 - 265 unique categories. **Mean frequency of each category was only 75. 98% of the data was concentrated in the first 25 categories.**
6. Level 2 - 218 unique categories. **Mean frequency of each category was only 91. 82% of the data was concentrated in the first 25 categories.**
7. Finally, **Level 1 and Level 2 were concatenated to form a new product category target.** This contained 436 unique categories but the advantage was that the data was less imbalanced. **84% of the data was in the top 30 categories of the data.**
8. **In conclusion, 25 categories of level 1 were selected as one target set, while 30 categories of level 1+2 were selected as the other target set.**

This is shown graphically below:

LEVEL 1 - Top 25



LEVEL 1 + 2 - Top 30



The final data with the new targets -> **LEVEL 1 and NEWCAT** are shown below:

	Level1	Level2	NameDesc	NewCat
0	Clothing	Women's Clothing	Alisha Solid Women's Cycling ShortsKey Feature...	Clothing Women's Clothing
1	Furniture	Living Room Furniture	FabHomeDecor Fabric Double Sofa BedFabHomeDeco...	Furniture Living Room Furniture
2	Footwear	Women's Footwear	AW BelliesKey Features of AW Bellies Sandals W...	Footwear Women's Footwear
3	Clothing	Women's Clothing	Alisha Solid Women's Cycling ShortsKey Feature...	Clothing Women's Clothing
4	Pet Supplies	Grooming	Sicons All Purpose Arnica Dog ShampooSpecifica...	Pet Supplies Grooming

Statistical data of both is given below:

Dataset 1 (Level 1, Top 25) (19620, 3)
Dataset 2 (Level 1 + Level 2, Top 30) (16996, 3)
Train - Test Split was 0.85 - 0.15. Furthermore, label encoder was used to make the labels in digitized format.

Data Cleaning

The following are the steps for data cleaning:

1. Removal of blank rows.
2. Convert all text to lowercase.
3. Remove punctuations using NLTK.
4. Tokenization of each description.
5. Removal of stop words based on NLTK.
6. Application of WordNet Lemmatizer based on three POS Tags:
    a. Adjective
    b. Verb
    c. Adverb

Vectorization and Training

For training purposes, a TF-IDF vectorizer with 15000 features was used. The reasons for this are:

1. A word2vec model was also tested. It provided very similar scores but was discarded due to the large training time and computational expenses. The TF-IDF model performed commensurate to it, in a much less time.
2. A deep learning based model was also tested (GLoVE). A similar trend to the above point was observed.
3. Furthermore, since this is a relatively simpler predictive task, not involving need for context and also based majorly on words related to the category present in description, TF-IDF was a much better fit.

Training:
For training, both Naive Bayes and SVM based classifiers were tested. Finally, SVC was selected due to more options in hyperparameter tuning and better relative performance. The same model was used for both datasets. Grid CV (Cross Validation) was used for hyperparameter tuning. The details are given below:

Parameters Tuned:

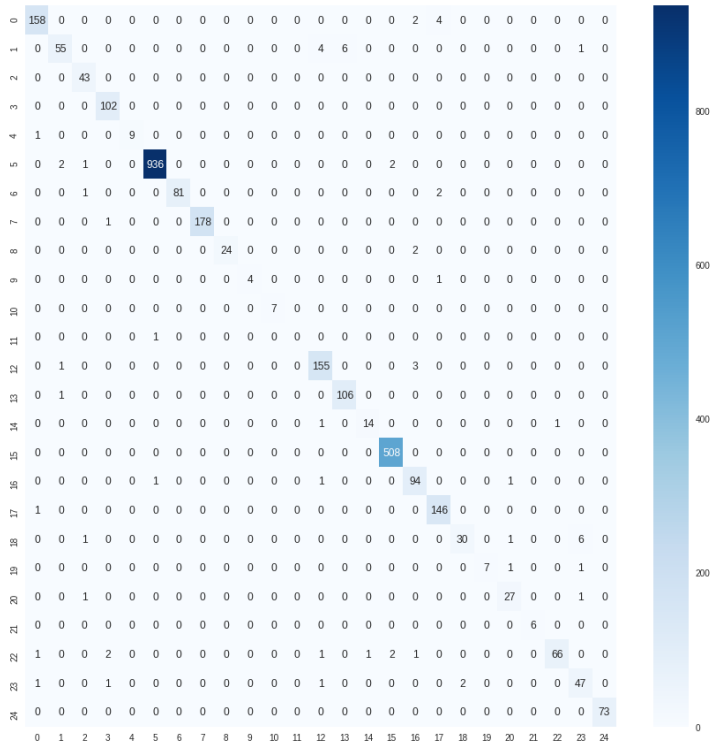
C: 1, 10
Gamma: 0, 0.1
Kernel: Linear, RBF

Best Params Returned - {C:10, Gamma:0.1, RBF: 10}

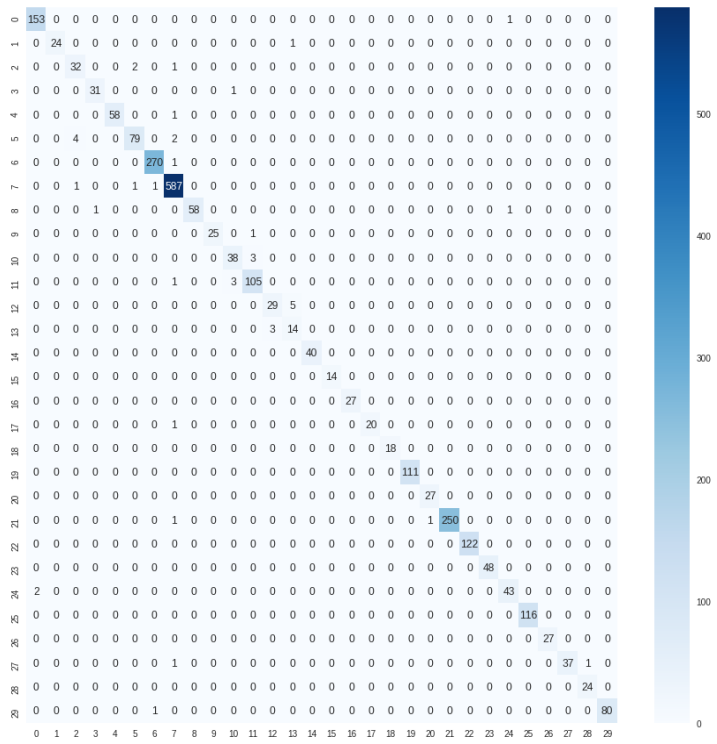
Time Taken To Apply GridCV - 17 minutes 26 seconds

Results and Discussion

Dataset	Method Used	Accuracy	Unweighted Recall	Weighted Recall
Prediction on Top 25 Level 1 Categories	TF-IDF with Hypertuned SVC	98%	91%	98%
Prediction on Top 30 Level 1 + 2 Categories	TF-IDF with Hypertuned SVC	98%	**96%**	98%



Heatmap For Level 1



Heatmap For Level 1 + 2

Discussion

We can see that the model obtains excellent performance on both the dataset configurations. On the model predicting only on Level 1, the model does face some imbalance in classes causing the recall to be lower than what is expected. The model tends to overfit on the first 10 categories. On the other hand, the second model performs much better since the dataset reduces in imbalance when 2 levels are combined. This is also seen clearly in heatmap 2, wherein, even with more classes, the model performs in a muchmore balanced manner.

Possible Improvements

1. Use of a biderectional deep learning model like ELMo for better encapsulation of lexical features.
2. One of the main factors in improving dataset recognition accuracy on this particular database is the right selection of target variables. Since there are 8 levels of variables, there could be many combinations of the target variables. A careful selection of those variables is essential for this problem. The same has been attempted above but could be further improved.
3. Use of oversampling to reduce imbalance in dataset.
4. Topic Modelling to improve recognition of keywords.

References

1. ML Powered Product Categorization, Abhimanyu Sundar
2. Multi-level Deep Learning based E-commerce Product Categorization, E- Com 2018, Yu et al.
3. Analysis of TF-IDF effectiveness, Pallavi Ahuja