DeepMMATE: Deep learning based MultiModal architecture for Audit Taxability classification with XAI

Y V S Harish
Amazon
India
yvsharis@amazon.com

ABSTRACT

Review of non-taxable products is an important internal audit which is carried out by majority of e-commerce stakeholders. This process usually cross checks the initial taxability assignments to avoid any unnecessary penalties incurred to the companies during the actual audits by the respective state compliance teams/tax departments. In order to handle millions of products sold online on e-commerce websites, we can adopt a machine learning solution to scale up the processing of products and make faster taxability predictions. However, a fine-grained classification cannot be achieved by visual analysis alone(product images). Often, the relevant information is present in the form of text on the product title, description & feature bullets etc. In this paper, we put forward a Multimodal Siamese based deep neural network which is capable of taking inputs from both product images and other textual content associated with it and predict the final output taxability. We show that this Multimodal architecture outperforms single modality networks which are only based on vision or language by a margin of atleast 5-6%. Furthermore, we reinforce confidence in our taxability outputs by incorporating an explainability wrapper around our model. This feature aids in establishing trust in the accuracy and reliability of our taxability predictions.

KEYWORDS

Explainable AI(XAI), Multimodality, E-Commerce Audit Taxability using Machine learning, Transfer learning for NLP, CV

ACM Reference Format:
Y V S Harish. 2024. DeepMMATE: Deep learning based MultiModal architecture for Audit Taxability classification with XAI. In *Proceedings of 3rd International Workshop on Interactive and Scalable Information Retrieval methods for eCommerce (ISIR-eCom) 2024 @ WSDM-2024, https://isir-ecom.github.io/ <https://isir-ecom.github.io/> (ISIR-eCom 2024 @ WSDM-2024).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Tax audit is a process which involves an independent body examining the financial accounts of an organization to authenticate fair dealings of a firm. A healthy audit feedback loop along with tax compliance leads to increased productivity for a better and efficient tax collection model. It aims to reduce the various problems faced by the tax authorities such as tax evasion, tax avoidance and other tax irregularities. The Tax department is responsible for paying tax obligations to governments and tax authorities around the world. In case of transactional tax, the e-commerce companies are responsible for collecting tax from consumers in accordance with the statues/regulation and pay it to Tax authorities. Every year, an external authority from respective state/country compliance team visits e-commerce companies to perform an audit where they cross check if the products sold are correctly taxed or not.

These e-commerce companies hosts millions of products coming from various vendors across the globe. Even when the products look similar or belong to the same category, their taxability(Taxable / Non-Taxable) might differ depending upon the manufacturing ingredients, description, type of audience its intended to, state/country in which they are sold, price bracket etc.

Usually all the e-businesses have an automated system in place to determine the product tax percentages. These systems are error prone considering they have rules/systems working to differentiate across minute product tax percentage slabs (e.g.: 5%, 8%, 15% etc.). As a result there exists an internal audit team which manually reviews the product taxability(just at a level of Taxable/Non-Taxable i.e., if any tax rate should be applicable or not in the first place). Since this process is manual, they mainly review the high-value products(based of net selling price) which are marked at 0%(Non-Taxable) by these automated systems to avoid penalty in the actual audit by respective state/country tax compliance teams. The goal of an internal audit is to detect products which are marked at 0% tax, but should be taxed. This process helps to save the tax these e-commerce companies have to pay to the authorities, on behalf of its customer and any additional penalties as a result of violation of tax laws.

In this paper, we explore the following question - is it possible to determine the taxability of an product across various states (refer: table 1) in the US(reason for selection mentioned in the section 3) by leveraging the different attributes of a given product(product Images, OCR text from images, product description, title & feature bullets etc.) using deep learning based methods. We finally introduce a multimodal siamese based deep network that learns from both the product images & other textual / language based attributes(title, description, feature bullets, OCR extracted text from images etc.). We design an end-to-end pipeline for feature extraction and fusion from image and textual inputs. We show that combining both the image(visual) and textual features to train the model outperforms single modality vision & language based baseline architectures in

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ISIR-eCom 2024 @ WSDM-2024, March 8, 2024, Mérida, Mexico

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

both accuracy and rate of convergence. Furthermore, with our explainability wrapper we output keywords from the product details which have led the model drive its decision towards a particular taxability. This thus ultimately helps the internal teams to justify the product taxability to external auditors.

2 RELATED WORK

Next, we discuss a quick overview of previous related works and reference topics. The classification problem was one of the first topics where modern deep learning was applied. There are several existing vision based deep networks for image classification which takes visual clues from images to predict the final output category. A major breakthrough in image classification domain came in 2012 with the introduction of AlexNet [7]. Some other noval architectures that were introduced after that include VGG [16] , ResNet [6], Inception [20], [21] and EfficientNet [22]. In this work we leverage both ResNet [6] & EfficientNet [22] as single modality vision based baseline architectures to solve the classification problem at hand. These models focus on extracting strong visual features from the images to classify the product based on their contrasts, objects and other visual feature clues present on the images.

On the other hand, text-based classification has long been investigated. We now look into various related works which used language based architecture to train models for category classification as down stream tasks. Several models like MPNET [17], BERT [2], XLNET [24], RoBERTa [10] which are pre-trained on on large-scale datasets(over 200 GB text corpa) have been leveraged in several down-streaming tasks by fine tuning these models with custom layers depeding upon the task in hand. We picked architectures like BERT & MPNET as single modality language base baseline networks. We take textual input coming from the product title, description, feature bullets & OCR text extracted from the product images and train these network on a down stream classification task to predict the final output categories(refer: table 2).

Some works tried to adapt the text and image based approaches to exploit both sources of information. Two such works which formulated generic visual-linguistic representation learning are VisualBERT [8] and Visual-Linguistic BERT (VL- BERT) [19]. In contrast to the above architectures, we propose a siamese based deep network where the images and text are processed in parallel towers of deep convolutional networks. The initial layers extract features specific to the data type. These features are flattened and concatenated into a single feature vector, grouping image features and text features separately. Finally, a Dense Neural Network (DNN) predicts the final product category from the combined feature vector. Similar works which try to exploit the Siamese base network include [3], [23], [9], [18] which exploit these networks for Alphabet predictions, Image Retrieval and Pattern Spotting, Object Tracking, other text Similarity Tasks for Multiple Domains and Languages etc.

Explainability in machine learning refers to the ability to understand and interpret the decisions or predictions made by a machine learning model. As machine learning models, particularly complex ones such as deep neural networks, become increasingly sophisticated & being used as an abstraction, making it challenging for humans to comprehend the reasoning behind their outputs. Understanding why a model makes a specific prediction is crucial for various reasons, including building trust in the model, ensuring regulatory compliance, uncovering biases, and facilitating the improvement of models. Explainability is especially important in scenarios where the impact of model decisions is significant, such as in healthcare, finance, and criminal justice.

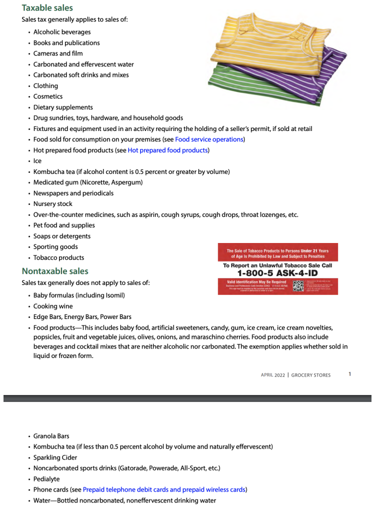
There are several methods and techniques for achieving explainability in machine learning:

1. Interpretable Models:
   - Use simpler models that are inherently more interpretable, such as decision trees [15] or linear models.
   - While these models might not always match the predictive performance of more complex models, they are easier to understand.
2. Local Interpretable Model-agnostic Explanations (LIME) [14]:
   - LIME [14] is a technique that approximates the decision boundary of a complex model using a simpler, interpretable model in a local region around a specific data point.
   - It helps provide insights into the model's decision-making process for individual instances.
3. SHapley Additive exPlanations (SHAP) [11]:
   - SHAP [11] values allocate the contribution of each feature to the prediction, providing a comprehensive understanding of feature importance.
   - SHAP [11] values can be used to explain the output of any machine learning model.
4. Partial Dependence Plots (PDP) [4] and Individual Conditional Expectation (ICE) [5]:
   - PDPs [4] illustrate the relationship between a feature and the model's prediction while keeping other features constant.
   - ICE [5] plots extend this concept to show the individual predictions for different instances.
5. Model-specific Techniques:
   - Some models have built-in mechanisms for explainability. For example, decision trees inherently provide a transparent decision-making process.
6. Rule-based Explanations:
   - Express the model's decision logic in the form of human-understandable rules.
7. Attention Mechanisms:
   - In the context of deep learning, attention mechanisms can highlight the input features that are crucial for a particular prediction.

Having said the above, Explainability is not a one-size-fits-all concept, and the choice of method depends on the type of model, the problem domain, and the level of detail required for explanation. Striking a balance between model complexity and interpretability is crucial for successfully incorporating machine learning models into real-world applications.

3 AUDIT OVERVIEW & CHALLENGES

In this section we formulate the business scenario for which we show various quantitative results in 6. In section 1, we mentioned that we will determine the taxability of an product across various



Taxable sales

Sales tax generally applies to sales of:

- Alcoholic beverages
- Books and publications
- Cameras and film
- Carbonated and effervescent water
- Carbonated soft drinks and mixes
- Clothing
- Cosmetics
- Dietary supplements
- Drug sundries, toys, hardware, and household goods
- Fixtures and equipment used in an activity requiring the holding of a seller's permit, if sold at retail
- Food sold for consumption on your premises (see Food service operations)
- Hot prepared food products (see Hot prepared food products)
- Ice
- Kombucha tea (if alcohol content is 0.5 percent or greater by volume)
- Medicated gum (Nicorette, Aspergum)
- Newspapers and periodicals
- Nursery stock
- Over-the-counter medicines, such as aspirin, cough syrups, cough drops, throat lozenges, etc.
- Pet food and supplies
- Soaps or detergents
- Sporting goods
- Tobacco products

The Sale of Tobacco Products to Persons Under 21 Years of Age Is Prohibited by Law and Subject to Penalties

To Report an Unlawful Tobacco Sale Call

1-800-5 ASK-4-ID

Valid Identification May Be Required

Nontaxable sales

Sales tax generally does not apply to sales of:

- Baby formulas (including Isomil)
- Cooking wine
- Edge Bars, Energy Bars, Power Bars
- Food products—This includes baby food, artificial sweeteners, candy, gum, ice cream, ice cream novelties, popsicles, fruit and vegetable juices, olives, onions, and maraschino cherries. Food products also include beverages and cocktail mixes that are neither alcoholic nor carbonated. The exemption applies whether sold in liquid or frozen form.

APRIL 2022 | GROCERY STORES 1

- Granola Bars
- Kombucha tea (if less than 0.5 percent alcohol by volume and naturally effervescent)
- Sparkling Cider
- Noncarbonated sports drinks (Gatorade, Powerade, All-Sport, etc.)
- Pedialyte
- Phone cards (see Prepaid telephone debit cards and prepaid wireless cards)
- Water—Bottled noncarbonated, noneffervescent drinking water

Figure 1: Example image from California Sales and Use Tax Laws

states in the US. The reason we chose US was that each state has its own set of taxability rules defined i.e., the taxability (taxable / non-taxable) of an product in the mentioned states(refer: table 1) differ depending upon various factors(manufacturing ingredients, description, keywords etc). For e.g.: California has the following rules and regulations defined for taxing Grocery store items - California Sales and Use Tax Law. The above document is designed for owners, managers, and other operators of grocery stores and provides basic information on the application of the California Sales and Use Tax Law to grocery store sales and purchases. In Figure:1, we show Examples of taxable and nontaxable sales(not limited to these) in California. Considering the diverse set of product categories which are sold on e-commerce websites, we have come up with categories (ref: table 2) which any product can be classified into, once we categorise the product in a state to its respective category we then derive its taxability using a static map i.e., the category to taxability for a state is a 1:1 mapping. Some of the categories in the mentioned list are self explanatory i.e., if a product in a state is mapped to an Exempt Clothing category then it is non-taxable.

Usually, the manual audit process can cover only upto 10% of the total non-taxable products for re-verification. Also, if the audit team wants to increase the regions/states it is difficult to scale considering all the manual rule formulations etc. The purpose of this paper is to provide the ability to completely automate the Audit process and cover 100% of the products in as many regions as required.

The following steps briefly describes the process followed by the audit teams in an e-commerce setting:

- The Audit team fetches product list which are labelled as Non-taxable either by vendors/internal rule engines.
- Shortlists the high-valued products based on net selling price.
- Goes over the product images and information on the websites.
- Fetches details about the product like: Manufacturing ingredients, description, feature bullets & other keywords.
- Refer the tax laws of respective region/state of audit & derive product taxability manually based on taxability rules.

State List		
Arkansas	Minnesota	Texas
Connecticut	Utah	Florida
Nevada	Vermont	New Jersey
Washington	Indiana	New York
Iowa	North Carolina	Wisconsin
Kansas	Wyoming	Kentucky
Rhode Island	Oklahoma	South Dakota
Tennessee	Michigan	Nebraska
Georgia	West Virginia	North Dakota
Ohio		

Table 1: List of US states

4 APPROACH

We have decided to adopt a Parent → Child architecture[2 stage process] as shown below (refer fig: 2). We pass the input features / data source to the Model Architectures. In this case, first through the Parent Model and then depending upon the output parent category we select the appropriate child model and pass the product though it and make the final prediction.

In the below diagram (refer fig: 3) you can see the parent categories filled in black and the corresponding child categories above them. So basically as a first level, what we have done is grouped categories with similar semantical overlap. E.g: on the left most you can see the parent category as Food and there are many child categories inside this like : soft drink, candy, Dietary supplement etc. And similarly we have a parent category Medication and Hygiene and various other child categories inside it.

Decoupling the categories into parent and child sections will help the models to perform better as the child models will now only concentrate on the subset categories rather than worrying about other miscellaneous categories. In section 5 we explain the architectures we leveraged for the parent and child models.

5 MODEL ARCHITECTURAL OVERVIEW

In the above section we explained the 2 stage process in order to predict the final product category. In this work, we explore similar architectural designs for both the parent and child models. The discussions in this sections are thus common for both the parent and child models.

5.1 Vision only model

Deep convolutional networks have established themselves as the state of the art on many tasks like Object detection, recognition / classification, segmentation etc. There exist wide range of novel visual architectures which can be leveraged for image classification tasks as well. Deep features extracted by pre-trained or fine tuned deep CNNs constitute a strong baseline for image classification tasks [13]. Considering how diverse various product images present on e-commerce sites are, we planned to fine-tune CNNs pretrained on ImageNet in order to extract visual features on our

Category List	
1. Exempt Food, Food Ingredients & Beverages	2. Taxable Food, Food Ingredients & Beverages
3. Soft drinks	4. Candy
5. Dietary Supplements	6. Bundles
7. Exempt Clothing	8. Taxable Clothing and Accessories
9. Taxable Digital products	10. Exempt Digital products
11. Taxable TPP	12. Nontaxable TPP
13. OTC Drugs & Medicines	14. Feminine Hygiene Products
15. Cosmetics, Grooming & Hygiene Products	16. Durable Medical Equipment
17. Prosthetic Devices	18. Mobility Enhancing Equipments
19. Exempt Medical Supplies	20. Taxable Medical Supplies
21. Pet Products	22. Protection Plans(Warranties)/ Maintenance Contracts
23. Taxable Services	24. Exempt Services
25. Coins/ Bullions	26. Gift Cards

Table 2: List of categories a product is classified into before predicting the final taxability

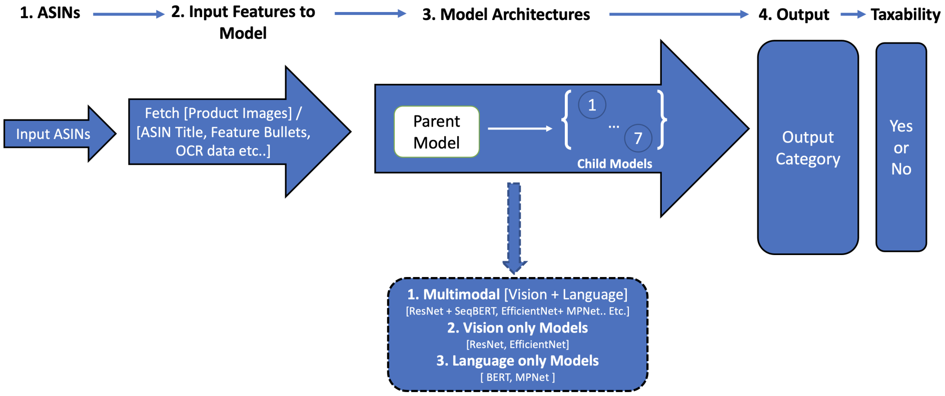


Figure 2: Tax Audit Machine Learning Work Flow

images. We wanted the networks to learn both low and high level features efficiently and due to this the depth of the network was pretty important to us. Scaling the depth of the neural network corresponds to adding more layers to the original network. Mostly this type of scaling has helped achieve better performance in standard datasets. However in the conventional scaling techniques, scaling was done in random fashion and required a lot of human intervention and expertise and often the corresponding increase in performance was not significant. Considering various vision deep convnets we decided to use 1. **ResNets**: These networks were mainly designed to mitigate gradient loss in very deep architectures [6]. The identity mapping helps learns residual mappings instead of learning the entire function map. The input (X) is passed through a two-layer path (approximating F(X)) and a skip connection path. The outputs of the two paths are summed at the output of the ResNet. 2. **EfficientNet**: The EfficientNet architecture proposes scaling using compound coefficient, a simple, efficient technique [22]. This technique scales the dimensions of depth, width and resolution in a uniform manner using scaling coefficients. The balanced scaling improves overall performance rather than just increasing accuracy. The compound scaling balances the scaling dimensions using a constant ratio. The intuition for the networks is, if the input image is bigger, then the network needs more layers to increase the receptive field and more channels to capture more fine-grained patterns on the bigger image.

5.2 Language only model

Since our use case focuses on product category classification, we can leverage various means of textual data present on the product title, product description, feature bullets and OCR extracted data on the product images. In order to extract the OCR data on images we used the DetectDocument API of the AWS Textract's OCR engine [1]. Recent literature in NLP suggests that pretrained word embeddings offer a strong baseline which surpasses traditional shallow learning approaches. The only prior these word embeddings assumes is a good tokenisation of words, i.e. most embeddings remove noisy and meaning less data, ignore punctuation and do not deal with out-of-vocabulary (OOV) words or are mapped to closest in-vocabulary word based on the Levenshtein distance. For this problem, we have selected BERT [2] and MPNET [17](since

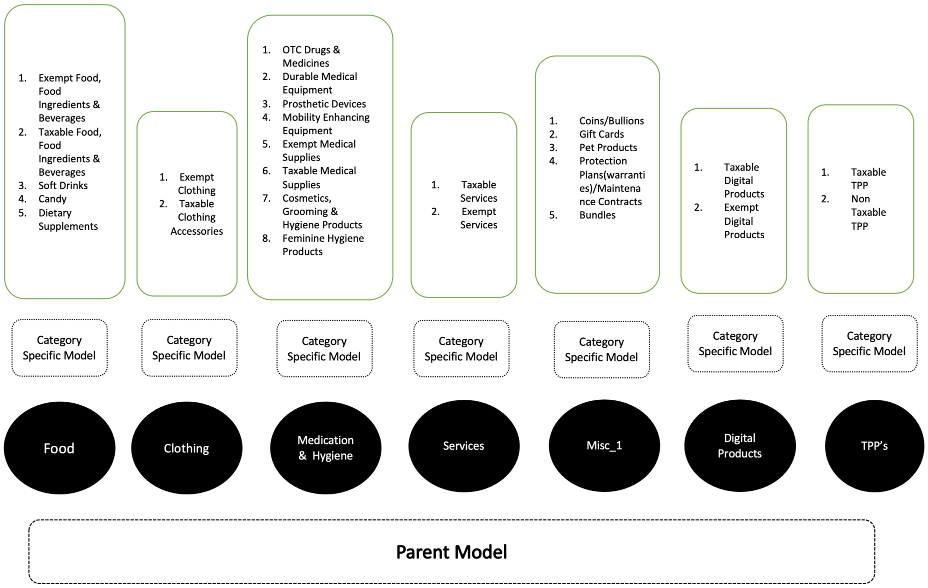


Figure 3: Parent - Child Categorization

MPNET overcomes some of the problem prevalent in the classical BERT pre-trained model) . 1. **BERT**: BERT, short for Bidirectional Encoder Representations and Transformers is used as a pre-trained sentence encoder. BERT adopts masked language model (MLM) enabling biredirectional learning from text by masking a word in a sentence and forcing it to use the words on the either side of the covered word to predict the masked word. 2. **MPNet**: MPNet adopts a novel pre-training method, named masked and permuted language modeling, to inherit the advantages of masked language modeling and permuted language modeling for natural language understanding. We leverage the above 2 architectures by fine tuning them by adding one additional layer on top to create model for the category prediction task.

5.3 Multimodal Architecture

Multi-Modal Modeling of images and text combines semantic knowledge extracted from text with knowledge of spatial structures extracted from images. Models of this type learn joint representations of images and text. These joint representations have been used to relate images and text to improve search-and-retrieval, classification, and self-supervised learning. In sections 5.1 & 5.2, we explained various off the shelf models we used as a single modalities to predict the output categories. For the multimodal architecture, once the text and image features have been extracted we feed them to a final classification layers. Inorder to carry out this we need to fuse the incoming feature vectors into one. This could be done in 2 ways: 1. Averaging the feature representations by normalizing the vector sizes 2. Concatenating the feature representations. The theoretical down sides of method 1 is that the two incoming feature vectors have different dimensional meanings, their vector spaces are different and once clubbed they will loose their individuality as the averaged out vector might not make any semantic sense. We observe this in our experiments where the second method mentioned above performed better than the first. Therefore, in section 6, we report the metrics by using the fusion technique mentioned in method 2 above.

After testing out the singular modality vision and language models, we trained an end-to-end siamese based multimodal architecture using the vector fusing mechanism discussed above. The multimodal architecture is modular enough to plug and play various feature extractors mentioned in section 5.2 & 5.1. We tried out the off the shelf base networks and figure out the impact of such collaboration for the tax audit use case. We have seen that the multimodal architecture completely outperforms the single modality networks by huge margins reported in section 6. The complete Siamese based multimodal architecture can be seen in fig 4.

6 EXPERIMENTS

In this section we present quantitative results of the parent and child models on the following architectures: 1. Multimodel architecture (Vision + Language) 2. Vision Only architecture 3. Language only architectures reported on the Amazon Review Data [12]. The values reported in each of these architectural design belong to models mentioned in section 5, i.e. for the vision only architecture we report metrics from the best performing model out of ResNet and EfficientNet(B7 version), for the language only model it is between BERT and MPNet & for the Multimodal architecture we report the best performing model from various possible combinations of the above two[in this case a combination of ResNet-152 and SeqBERT was best performing multimodal]. Table 3 reports metrics(F1 score and final accuracy) on the parent level stage. Table 4 reports the averaged out metrics for the child level models across states. We have seen that the rate of convergence of the multimodal architecture is close to (1.5-2) times faster as compared to the single modality

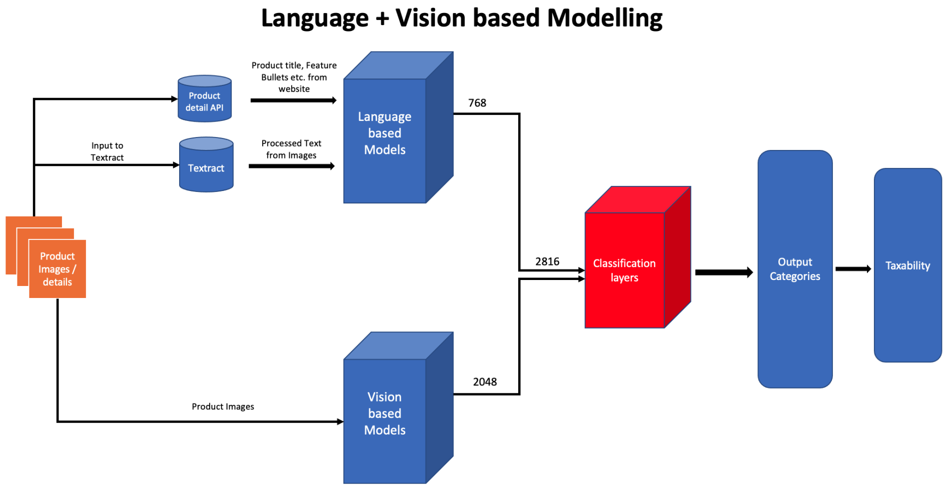


Figure 4: Siames based Multimodal Architecture for Audit taxability prediction

models. We can add this comparison table in the final version if required.

6.1 Parent level model metrics

We can clearly see in table 3 that the Siamese based Multimodal architecture outperforms the singular modality models.

6.2 Child level model metrics

In the above section 6.1, we have seen that the Multimodal architectures have performed better than the single modality networks. We have observed similar performance even at the child level classification. We therefore, report precision, recall and F1 metrics at the child level only for the Multimodal architecture[ResNet-152 + SeqBERT]. From table 4, we see that the accuries for the child level classes for the parent categories of Food and Medication & hygiene are a bit on the lower side due to the products belonging to these child categories are semantically closer which makes it difficult for the model to differentiate the products inside these categories.

7 EXPLAINABILITY

In actual audits, just outputting the category / Taxability of the product won't suffice. We also need to tell the auditors what ingredients/features/words of the product led to this taxability/category prediction. Hence, inorder to bring in transparency into this taxability process we bring in the concept of XAI(Explainable AI) which will tell us the reasons why we categorised the product into a particular category and hence taxability bucket.

In order to show the effectiveness of our explainability wrapper around the classification model leveraging LIME [14](most effective of all in this case), we take the following example from an e-commerce website where the product title was the following - "Spry Xylitol Peppermint Sugar Free Candy - Breath Mints That Promote Oral Health, Dry Mouth Mints That Increase Saliva Production, Stop Bad Breath." On passing this product to our siamese model, the parent class was predicted as Food and the child class was Candy. Next we run the explainability module to get the keywords that drive the model towards the decision. Figure 5 & 6, gives the explainability output at the parent level & child level respectively. In fig. 6 on the left you can see the probability score for each category. The words on the right of this straight line are the reason why the model is inclined towards the candy & since Candy in Texas(state considered in this case) is taxable, we output the taxability as Yes. The numbers shown Indicate the importance of the words helping the model to come to the category conclusion in descending order.

We can see words like peppermint, mints, Xylitol, oral, mouth, sugar as the reasons that drives the model towards the candy category. At the same time on the extreme right, we have words like health, production etc which drive the model towards Dietary supplements. Similarly, we have other categories where the model lacks the confidence from the input text to drive itself to these categories. Here Xylitol is an important Ingredients in candies, it is a natural sugar alcohol found in plants. The model during its training phase might have picked up this insight where it saw product titles containing the word Xylitol were mostly candies & hence, we can see the word driving the model towards candy. These insights help the auditors gain more confidence on the models performances and improves reliability.

8 CONCULUSION

Constant improvements in the Computer Vision and Natual Language processing domains have led to better tackling the category classification problems. The semantic representations of the language and visual structures of the product are well captured using the siamese based model. This work asserts that transfer learning between these two modalities provides a robust solution to noise & improving the overall performance accuracy of classification tasks. We compared the single modality baseline model performances

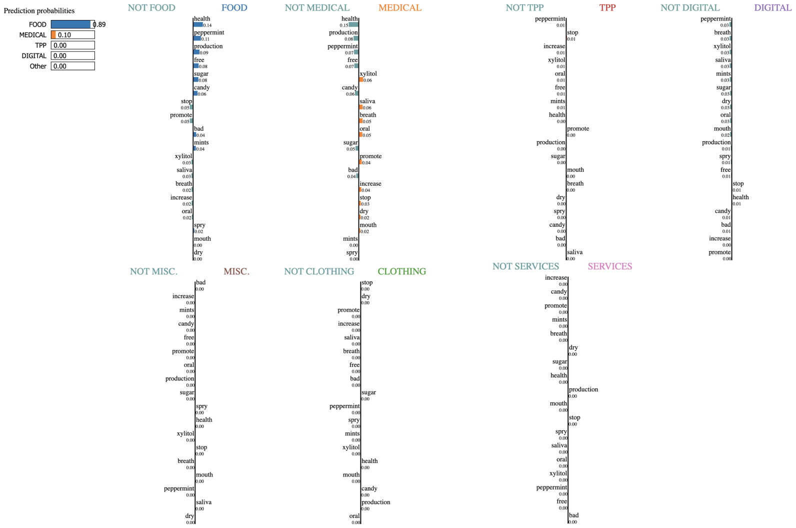


Figure 5: Explainability wrapper output for parent level classification for example product in Texas state

```
*** Taxability of the Product ***:  TAXABLE
```

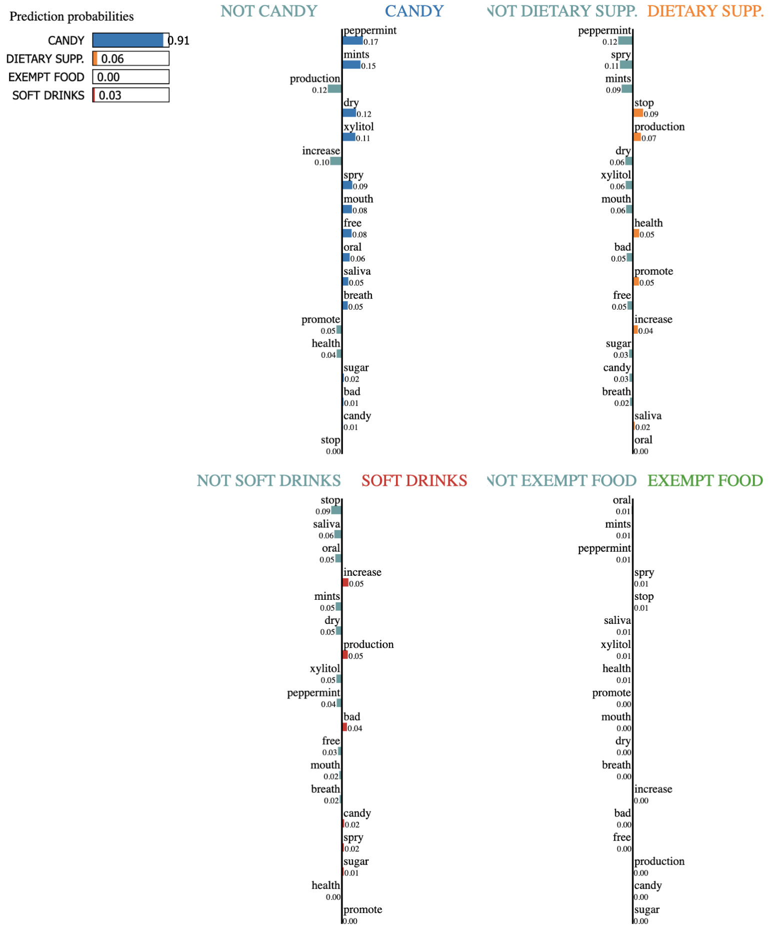


Figure 6: Explainability wrapper output for child level classification for example product in Texas state

F1 scores of various model architectures - Testing set - Parent Model				
	Sample Size	Multimodal architeture	Vision only	Language only
Parent Category				
Medication & Hygiene	6509	**0.89**	0.85	0.82
TPPs	1590	**0.45**	0.34	0.37
Digital Products	1482	**0.93**	0.92	0.91
Clothing	1342	**0.77**	0.71	0.73
Misc 1	2225	**0.90**	**0.90**	0.81
Food	4273	**0.98**	0.94	**0.98**
Services	846	**0.58**	0.57	0.32
Final Accuracies	**18267**	**0.89**	0.83	0.80

Table 3: The above reported values are the best F1 scores belonging to the model combinations mentioned in section 5

Child Model metrics - Testing set			
	Precision	Recall	F1-score
Child Categories			
Taxable Clothing & Accessories	0.92	0.94	0.93
Exempt Clothing	0.93	0.91	0.92
Taxable Services	0.873	1.00	0.93
Exempt Services	1.00	0.33	0.50
Gift Cards	1.00	1.00	1.00
Coins/ Bullions	0.98	0.97	0.98
Pet Products	0.99	0.98	0.98
Bundles	0.95	0.97	0.96
Protection Plans (Warranties)/ Maintenance Contracts	1.00	1.00	1.00
Nontaxable TPP	0.95	0.92	0.93
Taxable TPP	0.96	0.97	0.97
Taxable Digital products	0.87	0.96	0.91
Exempt Digital products	0.95	0.85	0.90
OTC Drugs & Medicines	0.86	0.82	0.84
Prosthetic Devices	0.78	0.82	0.80
Taxable Medical Supplies	0.74	0.80	0.77
Durable Medical Equipment	0.74	0.58	0.65
Cosmetics, Grooming & Hygiene Products	0.89	0.94	0.91
Feminine Hygiene Products	0.93	0.98	0.96
Mobility Enhancing Equipments	0.70	0.86	0.77
Exempt Medical Supplies	0.74	0.66	0.70
Dietary Supplements	0.80	0.89	0.85
Candy	0.69	0.88	0.78
Taxable Food, Food Ingredients & Beverages	0.46	0.15	0.22
Exempt Food, Food Ingredients & Beverages	0.67	0.64	0.65
Soft Drinks	0.71	0.92	0.80

Table 4: The above reported values are the best F1 scores belonging to the model combinations mentioned in section 5

where the category classification tasks are performed separately for each of the vision and language streams with the performance of the Multimodal siamese based network. The latter incorporates features from both the vision and language streams. The proposed Multimodal architecture (ResNet-152 & SeqBERT) outperform the single modality models giving us an accuracy boost of atleast 5-6 % both at the parent and child level stages. Additionally, with our explainability wrapper we reinforce how we can enable wider adoption of such ML solutions to end customers(auditors in this case).

REFERENCES

[1] [n. d.]. Amazon Textract - Detecting Text. https://docs.aws.amazon.com/textract/latest/dg/how-it-works-detecting.html.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

[3] Atoany Fierro and Karina Perez-Daniel. 2020. Siamese Convolutional Neural Network for ASL Alphabet Recognition. *Computación y Sistemas* 24 (09 2020). https://doi.org/10.13053/cys-24-3-3481

[4] Jerome Friedman. 2000. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* 29 (11 2000). https://doi.org/10.1214/aos/1013203451

[5] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. 2014. Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. arXiv:1309.6392 [stat.AP]

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1* (Lake Tahoe, Nevada) *(NIPS'12)*. Curran Associates Inc., Red Hook, NY, USA, 1097–1105.

[8] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. arXiv:1908.03557 [cs.CV]

[9] Su-Chang Lim, Jun-Ho Huh, and Jong-Chan Kim. 2022. Deep Feature Based Siamese Network for Visual Object Tracking. *Energies* 15, 17 (2022). https://doi.org/10.3390/en15176388

[10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL]

[11] Scott Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. arXiv:1705.07874 [cs.AI]

[12] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. 188–197. https://doi.org/10.18653/v1/D19-1018

[13] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN Features off-the-shelf: an Astounding Baseline for Recognition. arXiv:1403.6382 [cs.CV]

[14] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv:1602.04938 [cs.LG]

[15] Lior Rokach and Oded Maimon. 2005. *Decision Trees*. Vol. 6. 165–192. https://doi.org/10.1007/0-387-25465-X_9

[16] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs.CV]

[17] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and Permuted Pre-Training for Language Understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS'20)*. Curran Associates Inc., Red Hook, NY, USA, Article 1414, 11 pages.

[18] João Souza, Lucas Oliveira, Yohan Gumiel, Deborah Carvalho, and Claudia Moro. 2020. *Exploiting Siamese Neural Networks on Short Text Similarity Tasks for Multiple Domains and Languages*. 357–367. https://doi.org/10.1007/978-3-030-41505-1_34

[19] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. arXiv:1908.08530 [cs.CV]

[20] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2014. Going Deeper with Convolutions. arXiv:1409.4842 [cs.CV]

[21] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2818–2826. https://doi.org/10.1109/CVPR.2016.308

[22] Mingxing Tan and Quoc V. Le. 2020. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv:1905.11946 [cs.LG]

[23] Kelly L. Wiggers, Alceu S. Britto, Laurent Heutte, Alessandro L. Koerich, and Luiz S. Oliveira. 2019. Image Retrieval and Pattern Spotting using Siamese Neural Network. In *2019 International Joint Conference on Neural Networks (IJCNN)*. 1–8. https://doi.org/10.1109/IJCNN.2019.8852197

[24] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv:1906.08237 [cs.CL]