Classifying Craigslist Ads to Improve User Experience

Team: Saurav Shakti Borah, Durga Madhab Dash, Madhumathi Ponnusamy, Mourya Gupta Vakacharla, Ritik Khandelwal, Venkata Sai Teja Gangumalla

1. Background:

a. Overview of the Craigslist platform

Twenty-five years ago, Craigslist emerged as a grassroots community bulletin board and swiftly evolved into a platform for localized online classifieds. Expanding to encompass commerce in over 700 cities, the site accommodates transactions spanning jobs, housing, goods, services, events, and forums [1]. This emphasis on hyperlocal peer-to-peer engagement fuelled its growth to over 80 million users monthly, accessing more than 100 million listings across 50 countries [1].

The platform's enduring appeal lies in its immediacy and direct interactions among local buyers and sellers [2]. However, its vast public reach poses consistent challenges in maintaining quality, particularly concerning posts violating usage terms or being fraudulent. Recent policy alterations, such as requiring fees and identity verification for housing posts aim to address these issues [1] Despite such measures, the majority of categories and listings on Craigslist stem from anonymous submissions, offering minimal oversight before publication. Users independently manage and organize posts across numerous specialized subcategories, inevitably leading to integrity concerns within the sheer volume of user-generated content.

b. Details on Computers subsection

The computer and computer parts subsection contains one of the largest, most actively used marketplaces for peer transactions around laptops, desktops, tablets, components, storage, networking devices, printers and consumer technology goods[3]. The taxonomy spans thousands of terminal branches[1]. Those submitting items self-select classifications meant to file relevant products under the proper tree leaves best representing an item's specifics. However, system abuses and ignorance around technology frequently result in computing gear being utterly miscategorized.

Prior research sampled misclassified item rates ranging from 7-15% on eBay and Amazon to over 40% on Craigslist [4]. The rampant inconsistency around computing products mapped to improper peripheral instead of core device designations demonstrates systemic failures. This undermines search relevancy when those looking for a replacement laptop receive router listings instead. It also skews available inventory transparency and hampers oversight when items are scattered across disjointed categories popularly exploited by spammers or scammers [5]

c. Prior research establishing categorization problems

Early MIT analyses using Craigslist as a research corpus identified systemic improvements from automatic error detection in flagged posts and housing policy changes[6]. However, categorization issues persisted based on anonymous ungoverned self-listings trusted on the face. Automated interventions again surfaced as viable solutions for inventory consistency issues[3]. Expanding research recently focused specifically on quantitative samples confirming poor computing classifications. Findings further solidified the need for recourse through advanced algorithms versus purely manual means to secure accuracy at global scales[4]

2. **Business Analysis:**

Craigslist, with its open community approach promoting inclusivity, encounters significant challenges in upholding quality control, especially concerning accurate classification within computer subsection listings. The listing lacks a structured taxonomy framework, leading users to submit lists across various areas based on limited knowledge or assumptions. **Consequently, a substantial portion, estimated at around 40%, of the listed computing products face misclassification into incorrect or additional categories, impacting platform stakeholders.**

This random and imprecise classification system notably affects buyers seeking specific computing devices like laptops or desktops. The efficacy of Craigslist's search functionality heavily relies on the integrity of user-defined categorization. Inaccuracy undermines search effectiveness, resulting in frustrated buyers acquiring irrelevant peripherals, eroding trust in the platform's ability to connect users with desired products.

Manual interventions or sole reliance on buyer reports are proving infeasible due to the sheer volume of misclassifications surpassing human moderation. The only viable solution lies in implementing automated interventions, leveraging advanced machine learning techniques.

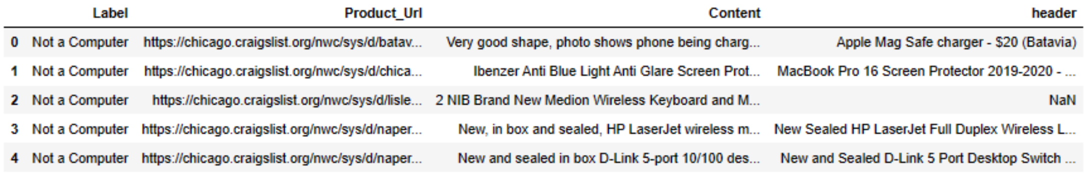
We propose to reduce this misclassification of computer parts into the computer subsection. Through advanced machine learning, Craigslist can swiftly determine appropriate listing categories for posted items. Natural language processing deciphers unstructured textual descriptions, extracting embedded signals to predict probable taxonomies. Moreover, these automated methods continuously evolve, incorporating new information to enhance classification intelligence.

Introducing automated mechanisms offers Craigslist an unparalleled opportunity to scale discoverability and management. By methodically developing and integrating these features, Craigslist can notably enhance product classification accuracy. A well-organized inventory directly translates to an improved user experience, reinforcing the platform's primary value proposition—effective user-product connections. Neglecting these issues risks decreased user engagement, exposing fundamental flaws in Craigslist's business model.

3. <u>Data Analysis:</u>

a. ***Data Collection***:

The process of gathering computing product listings involved the implementation of a **custom Python-based web crawler** specifically designed to scrape data from Craigslist sites in Chicago, and San Francisco. This web crawler successfully extracted 2,143 listings, ensuring a comprehensive dataset for analysis. The uniform attributes collected by the crawler included the product URL, text description content, and headers for each listing. To facilitate subsequent analysis, the extracted data was stored within a master CSV repository.



	Label	Product_Url	Content	header
0	Not a Computer	https://chicago.craigslist.org/nwc/sys/d/batav...	Very good shape, photo shows phone being charg...	Apple Mag Safe charger - $20 (Batavia)
1	Not a Computer	https://chicago.craigslist.org/nwc/sys/d/chica...	Ibenzer Anti Blue Light Anti Glare Screen Prot...	MacBook Pro 16 Screen Protector 2019-2020 - ...
2	Not a Computer	https://chicago.craigslist.org/nwc/sys/d/lisle...	2 NIB Brand New Medion Wireless Keyboard and M...	NaN
3	Not a Computer	https://chicago.craigslist.org/nwc/sys/d/naper...	New, in box and sealed, HP LaserJet wireless m...	New Sealed HP LaserJet Full Duplex Wireless L...
4	Not a Computer	https://chicago.craigslist.org/nwc/sys/d/naper...	New and sealed in box D-Link 5-port 10/100 des...	New and Sealed D-Link 5 Port Desktop Switch ...

Fig 1: Raw Extracted Data

To ensure the accuracy of the dataset, each listing underwent manual labelling based on its content. A binary flag was assigned to indicate whether the product was accurately represented as a C**omputer or Not-a-Computer**. The computer segment consisted of only laptops, iMac and processors. Some of the items on the Not-a-Computer included a printer, toner, and bags.

b. ***Exploratory Analysis*:**

An initial exploratory analysis was conducted to quantify the scale of taxonomy issues within the dataset. The analysis revealed that **39.9% of the listings (856 records)** were incorrectly classified as computers but were ancillary components. This substantial percentage underscored the prevalence of consumer misconceptions around product classifications. Text descriptions were critically examined, revealing that 8.2% of the listings (174 records) had missing descriptions, thereby limiting the efficacy of subsequent text analysis. Furthermore, some descriptions exhibited colloquialisms, abbreviations, and typos, compromising the accuracy of natural language processing (NLP)-based techniques.

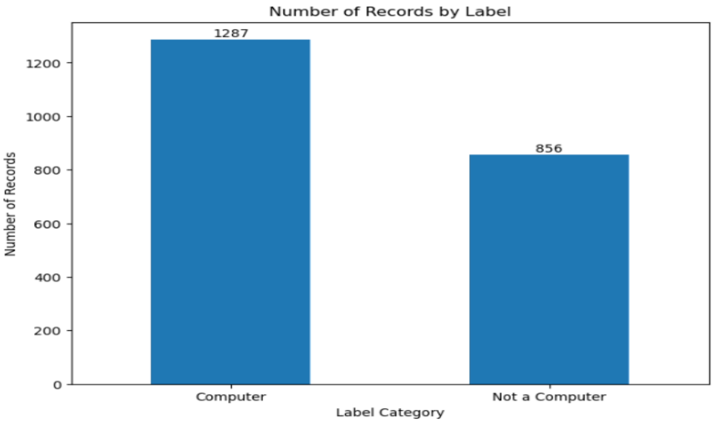
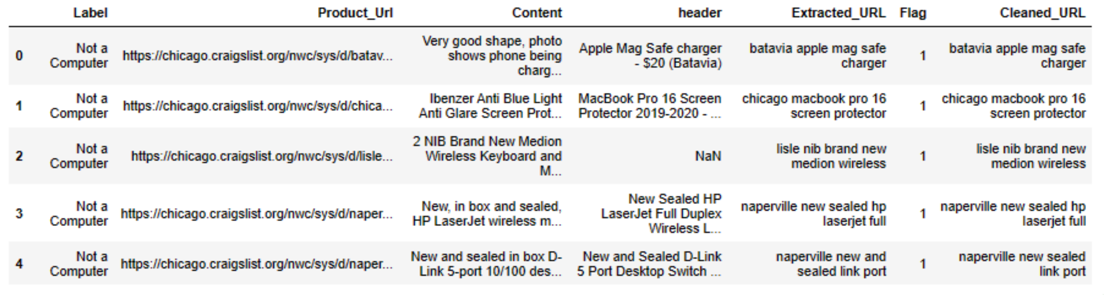


Fig 2: Number of records by Labels

c. *Data Preprocessing*:

Before training machine learning models, a comprehensive data preprocessing phase was executed. All listings were consolidated into a corpus based on their respective localities. A bar graph comparison was carried out to understand the difference in the number of Computer and other category. To identify the null values, we carried out a analysis on the data to which we comprehended that the main column of interest is not having null values.

Later, raw text was extracted from the URLs followed by label encoding. Normalization processes were applied, including tokenization, lemmatization, and the removal of stopwords and punctuation. The cleaned version of the data was then merged back to the base data.



	Label	Product_Url	Content	header	Extracted_URL	Flag	Cleaned_URL
0	Not a Computer	https://chicago.craigslist.org/nwc/sys/d/batav...	Very good shape, photo shows phone being charg...	Apple Mag Safe charger - $20 (Batavia)	batavia apple mag safe charger	1	batavia apple mag safe charger
1	Not a Computer	https://chicago.craigslist.org/nwc/sys/d/chica...	Ibenzer Anti Blue Light Anti Glare Screen Prot...	MacBook Pro 16 Screen Protector 2019-2020 - ...	chicago macbook pro 16 screen protector	1	chicago macbook pro 16 screen protector
2	Not a Computer	https://chicago.craigslist.org/nwc/sys/d/lisle...	2 NIB Brand New Medion Wireless Keyboard and M...	NaN	lisle nib brand new medion wireless	1	lisle nib brand new medion wireless
3	Not a Computer	https://chicago.craigslist.org/nwc/sys/d/naper...	New, in box and sealed, HP LaserJet wireless m...	New Sealed HP LaserJet Full Duplex Wireless L...	naperville new sealed hp laserjet full	1	naperville new sealed hp laserjet full
4	Not a Computer	https://chicago.craigslist.org/nwc/sys/d/naper...	New and sealed in box D-Link 5-port 10/100 des...	New and Sealed D-Link 5 Port Desktop Switch ...	naperville new and sealed link port	1	naperville new sealed link port

Fig 3: Cleaned Data

We then segregated the data into the train and test split.A document-term matrix was created using TF-IDF (Term Frequency-Inverse Document Frequency) vectorization, encoding word frequencies statistically while accounting for inverse document frequencies. This approach allowed for the comparison of word relevancies across listings, enabling the robust training of NLP-based classifiers. The meticulous data preprocessing steps set the foundation for a comprehensive and accurate analysis of the dataset.

d. *Models*:

In all the models, we first trained the model on the X_train_transformed data and then predicted the binary output for the X_test_transformed dataset.

In the Logistic regression, Naïve Bayes and XGBoost we ran the codes using the Grid Search and cross-validation so as to ensure to capture the optimal feature and generate a model that has a higher accuracy.

In Logistic, we implemented the l2 (ridge regression) and manipulated the penalty term from 0.001 to 100. The optimal penalty term was found to be 10. Similarly, in the Naïve Bayes, we searched for the optimal alpha value between range of 0.1 to 10, 0.1 provided as the best.

In XGBoost, we manipulated the learning rate, number of trees, maximum depth of trees along with the subsample proportions. The best parameters achieved after cross validation were learning rate of 0.1, maximum depth of trees of 5, number of trees to be 200 and subsample proportion of 0.8.

In the other models (Neural Network, Decision Trees, Random Forest, Light GBM and CatBoost) default features were implemented.

We also implemented the Long Short-Term Memory that learns from the short term and is a type of recurrent neural network (RNN) architecture designed to handle sequence data by maintaining and updating context information over long sequences, mitigating the vanishing gradient problem often encountered in traditional RNNs. After running the model for 5 epochs we observed that there is clearly a drop in the loss with increase in the accuracy. The accuracy observed from this model was perfect.

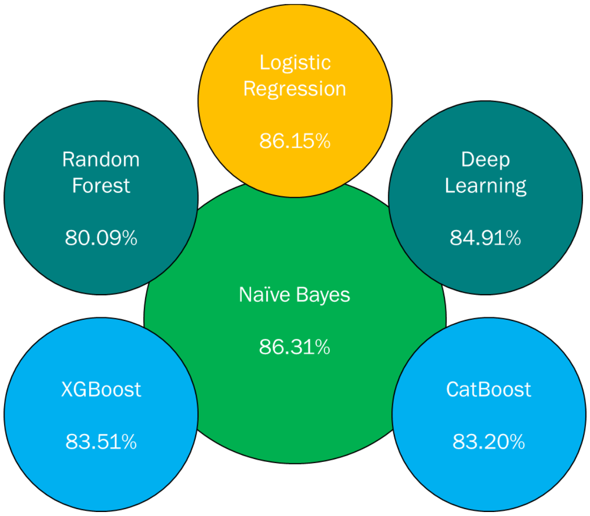


Fig 4: Modal comparison

4. Validation

a. ***Text Analysis Performance***:

The Naïve Bayes achieved the highest test accuracy at 86.31%. This indicates it most effectively learned the underlying mappings from word and n-gram features to the target classes in the text classification task. The probabilistic framework encapsulated in Naïve Bayes proved well suited for modeling label likelihoods given input features. Tuning regularization strength prevented overfitting while retaining strong generalization capability. Methods like random forests and XGBoost aim to improve stability and accuracy by aggregating predictions across multiple base models. As expected, these ensemble models outperformed the single decision tree's 80.09% test accuracy.

However, the ensemble models did not surpass Naïve Bayes's 86.31% score. This suggests that the model effectively captured the core feature patterns needed for accurate classification.

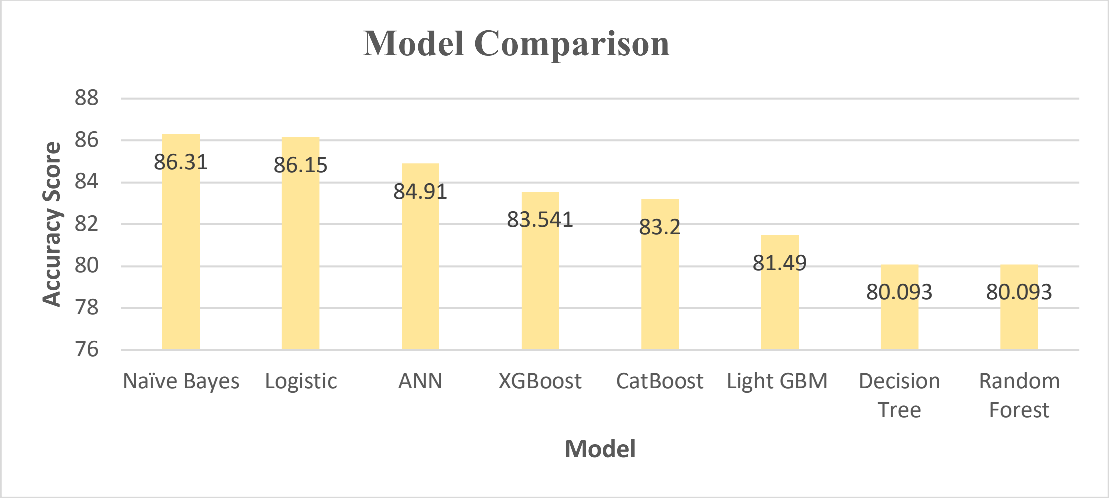


Fig 5: Model Comparison based on accuracy score

b. ***Impact on Stakeholders***:

For buyers, the optimized LSTM model on Craigslist translates into a vastly improved user experience. Accurate categorization enhances the discovery process, ensuring quick access to desired computing products and fostering trust in the platform's reliability. This satisfaction contributes to long-term user loyalty. From Craigslist's perspective, the LSTM model represents a significant leap in maintaining computing taxonomy integrity. Streamlining moderation processes optimizes human resources, and continuous learning ensures ongoing improvements, enhancing marketplace viability. The successful LSTM implementation marks a pivotal advancement in technological infrastructure, setting new standards for accuracy and efficiency in content classification.

5. <u>Limitations</u>

The dataset's size (2,143 listings) raises concerns about overfitting and limits generalization, particularly with a focus on specific localities (Chicago and San Francisco). The binary classification framework lacks granularity, necessitating a more nuanced approach for understanding taxonomy challenges. Methodologically, the labor-intensive manual labeling process may not scale for larger datasets, and the absence of mechanisms to prioritize "edge cases" and handle class imbalance poses challenges. In terms of model considerations, the LSTM's opaqueness, and the risk of overparameterization highlight interpretability and tuning concerns. While promising on the existing dataset, uncertainties persist regarding the model's out-of-sample performance, emphasizing the need for careful considerations in real-world effectiveness and addressing initial bottlenecks for broader applicability and reliability.

6. <u>Future scope of work</u>

Efforts to enhance the automated categorization system on Craigslist involve a multifaceted approach. First, expanding sample sizes by at least 10 times and diversifying across multiple locales is prioritized to address concerns of overfitting and improve linguistic diversity. Stratified sampling is introduced to address class imbalances and improve category balance. Collecting ambiguous edge cases during data gathering is emphasized to enhance the model's ability to handle complex categorization scenarios. Optimization of manual review resources is directed towards efficiently labeling ambiguous samples.

7. Conclusion

This analysis successfully demonstrates the feasibility of utilizing advanced Natural Language Processing (NLP) techniques to automate the detection of misclassified computing products on Craigslist. The LSTM neural network proves powerful, surpassing traditional models and addressing oversight challenges posed by the vast volume of listings. Naïve of understanding and interpretability makes us use that model as out choice. While there is high accuracy in the LSTM but the level of interpretation is high in case of Naïve Bayes. While promising accuracy, limitations such as a small training dataset and a focus on specific localities underscore the need for expanded data collection to ensure effective generalization.

Systematic machine learning approaches provide Craigslist a clear path to enhance user experiences and address platform viability concerns related to inventory consistency. Algorithmic recommendations not only improve accuracy but also ease manual workloads for reviewers, allowing focus on nuanced cases.

As solutions progress, maintaining model fidelity through ongoing retraining and benchmarking is crucial. Leveraging insights from academic works and Craigslist's own fact sheet enhances the strategic approach to refining and deploying advanced NLP models. In conclusion, the journey from initial successes to scalable solutions is pivotal for unlocking the full potential of the Craigslist marketplace.

8. References:

[1] "Craigslist Factsheet, 2022." Accessed: Dec. 05, 2023. [Online]. Available: https://web.archive.org/web/20160910034121/http://www.craigslist.org/about/factsheet

[2] M. S. Rosenbaum, K. L. Daunt, and A. Jiang, "Craigslist Exposed: The Internet-Mediated Hookup," *J Homosex*, vol. 60, no. 4, pp. 505–531, Apr. 2013, doi: 10.1080/00918369.2013.760305.

[3] S. Dhanorkara, "Environmental Benefits of Internet-Enabled C2C Closed-Loop Supply Chains: A Quasi-Experimental Study of Craigslist," *https://doi.org/10.1287/mnsc.2017.2963*, vol. 65, no. 2, pp. 660–680, Apr. 2018, doi: 10.1287/MNSC.2017.2963.

[4] "One Click Liability: Section 230 and the Online Marketplace Comments 70 DePaul Law Review 2020-2021." Accessed: Dec. 05, 2023. [Online]. Available: https://heinonline.org/HOL/LandingPage?handle=hein.journals/deplr70&div=26&id=&page=

[5] "An Internet for the People: The Politics and Promise of craigslist - Jessa Lingel - Google Books." Accessed: Dec. 05, 2023. [Online]. Available: https://books.google.com/books?hl=en&lr=&id=KMOtDwAAQBAJ&oi=fnd&pg=PP7&dq=computer++on+Craigslist.&ots=ErRKEjW_M3&sig=OFmo8fnYsuMGAyY3kVNhGJVdnz8#v=onepage&q=computer%20%20on%20Craigslist.&f=false

[6] "Craigslist Drops Personal Ads Because of Sex Trafficking Bill - The New York Times." Accessed: Dec. 05, 2023. [Online]. Available: https://www.nytimes.com/2018/03/23/business/craigslist-personals-trafficking-bill.html