Project Report: Decoding and Classifying Vanity License Plates

Team Captain: Mahan Agha Zahedi
Department of Computer Science
Stony Brook University

1 Introduction

Vanity license plates are one such unique way of personalizing one's vehicle, which can communicate quirkiness, humour, or individuality through a restrained number of letters, puns, and other symbolic devices. Many employ creative spelling, homophones, and abbreviations; how their intended meaning is extracted can greatly depend on context. It is this same mysterious element, though, that can sometimes result in a confusing interpretation or one that is ambiguous, contentious, or troublesome. This review process, carried out by the state DMV[1], for approval or rejection of these license plate requests, is extremely manual, cumbersome, and labour-intensive besides being quite subjective.
The given research explores the feasibility of applying large language models to address both understanding and classification issues related to vanity plates. In particular, two objectives have been set in the following manner:

1. **Decoding:** The objective is to uncover the concealed significance of a specific vanity license plate through the application of sequence-to-sequence learning models.
2. **Classification:** The aim is to assess if a suggested vanity plate complies with legal and societal criteria through the utilization of text classification methods.

To tackle these issues, we put forward a framework that harnesses the potential of fine-tuned LLMs. For this, a fine-tuned T5-small model[2] for sequence-to-sequence applications is employed to decode the cryptic text on license plates with their intended meaning. Subsequently, a classifier using BERT[3] will be deployed to check if any given plate meets legal and social norms by finding out if there is obscenity or possible legal issues. Overall, the system can automate an efficient scalable solution to assist DMV during the review process.

This is an important initiative because it reduces manual assessments, hence increasing efficiency and consistency in the approval process. At the same time, through the automation of this task, we reduce subjectivity connected with human interpretations of texts. This report describes how this system was developed based on data collection, initial analysis, methodological approaches, evaluation, and further development to make it more relevant and accurate. With this project, we aim to provide a robust and workable tool for the interpretation and classification of vanity license plates in real-world applications.

2 Background

Currently, interpreting vanity license plates is mostly a manual process, with state Departments of Motor Vehicles (DMVs) relying on rule-based systems. These systems focus mainly on the surface patterns in the alphanumeric combinations, but they struggle with deeper, more nuanced meanings. As a result, human reviewers are often needed, which introduces subjectivity and inefficiency.

Recently, there have been major advances in natural language processing (NLP) that tackle challenges similar to those posed by vanity plates. For instance, microtext normalization has helped transform informal text into standard language [4], and Seq2Seq models with attention mechanisms have proven effective at decoding informal or cryptic text [5]. Additionally, [6] introduced dynamic word vectors, which help capture the meaning of short, ambiguous texts—something that can be directly applied to understanding vanity plates.

Another important area of research is offensive language detection, where deep learning methods like Long Short-Term Memory (LSTM) networks have outperformed traditional approaches. Yi et al. (2021) [7] developed an LSTM-based system with word embeddings to detect profanity, which provides a strong framework for evaluating vanity plates. Large language models (LLMs) have further improved the performance of text classification, making them useful for tasks like checking legal compliance.

Models like BERT [3] and GPT [8] have shown great promise in both zero-shot learning and fine-tuning. Zero-shot learning allows these models to classify text even without specific training examples, which is especially useful for tasks like vanity plate classification, where labelled data might be scarce [9]. On the other hand, fine-tuning these models on specific datasets enhances their ability to accurately classify vanity plates based on their meaning and legality [9].

The goal of this project is to take advantage of these developments and create a system that will automatically decode vanity plates and classify them as acceptable or offensive. The system will be fine-tuning LLMs like T5-small[2] for decoding and BERT [3] for classification, hence providing DMVs with a more efficient automated solution, reducing the load on human reviewers, and increasing consistency in decision-making. This methodology furthers the broad domain of automated text interpretation, in general, and in contexts that predominantly use unconventional language, like vanity plates.

3 Dataset

The main dataset used in this project is based on more than 23,000 records from the California DMV, which include all vanity license plate applications from 2015 to 2016 [10]. Compiled by Veltman [11] in response to a public request by Muckrock [12], the dataset contains values that will be useful in both decoding and classification tasks, and is structured as follows:

1. **License Plate Number**: The alphanumeric sequence proposed by the applicant.
2. **Customer Meaning**: The applicant's explanation of the plate's intended meaning.
3. **Reviewer Comments**: DMV reviewers' interpretations and notes, serving as semantic cues for decoding.
4. **Status (Y/N)**: A binary indicator of whether the plate was accepted or rejected.

index ▲	plate	customer_meaning	reviewer_comments	status
0	AZIZ714	LAST NAME	714 AREA CODE	N
1	BATBOX1	BATMOBILE (BATMAN) PLUS SHAPE OF VEHICLE (SCION XB)	BOX	N
2	BBOMBS	NO MICRO AVAILABLE	BOMBS	N
3	BEACHY1	LOVE THE BEACH	BEACHY LOOKS LIKE BITCHY 1	N
4	BLK PWR5	STRENGTH OF FAMILY	BLACK POWER	N
5	BOT TAK	THIS IS IT	CAN NOT TRANSLATE	N
6	CHERIPI	CHERRY PIE	CHERRY PIE	N
7	CIO FTW	NO MICRO AVAILABLE	FUCK THE WORLD	N
8	DAVES88	DAVES 1988 TOYOTA	88 HITLER REFRENCE	N
9	DMOBGFY	GREEK SLANG FOR I LOVE YOU	MOB	N
10	DOITFKR	DO IT	DO IT FUCKER	N
11	EGGPUTT	NO MICRO AVAILABLE	SLANG FOR FAT GUY PER URBAN DICTIONARY	Y
12	F DIABDZ	IT MEANS FOREVER DIABETES. ITS ARE DIABETES FOUNDATION WE WALK EVERY YEAR.	FUCK DIABETES	N
13	FJ 666	FJ IS MY BORN CITY.666IS MY LUCKLY NUMBER	666 EVIL	Y
14	FKK OFF	FREIKÖRPER-KULTUR (GERMAN)	FUCK OFF	N
15	FKN BLAK	FKN IS INITIALS FOR MY 3 GREAT GRANDMAS FRANCINE, KAREN AND "N" FOR NANNY. AND THE WORD BLAK "BLACK" BECAUSE MY CAR APPEARS BLACK	FUCKING BLACK	N
16	FLT ATCK	NO MICRO AVAILABLE	FLAT ATTACK	N
17	F LUPUS	FIGHT LUPUS	FUCK LUPUS	N
18	HVNNHEL	ALBUM TITLE	HEAVEN AND HELL	Y
19	H8DES	FAVORITE GREEK MYTH	HATE DES	N
20	ILL NKA	"ILL" POPULAR TERM IN THE 90'S MEANING COOL. "NKA" SHORTER FOR MY NICKNAME 'NEIKA'.	ILL NIGGA	N

Figure 1: Dataset Representation

Figure 1 provides a valuable resource for understanding the intended meaning of vanity plates and assessing their legal and ethical compliance. The **Reviewer Comments** field is particularly important, offering insights

into how DMV officials interpret the plates. For example, the plate BLUGNG was interpreted as "blue gang," suggesting a potential gang affiliation, which led to its rejection [10].
Besides the main dataset, additional data was included from the California DMV 2013 [13] and other states such as New York [12], providing a wide variation in plate style and rejection criteria. Where the California dataset was mostly comprised of short and often cryptic strings, the New York dataset included much longer phrases or even complete sentences; this added variety to the analysis across the different jurisdictions.

The preprocessing of these data sets was necessary to prepare the data for machine learning model training. First, standardization of formats and removal of stop words were made to highlight key phrases. In this respect, experiments with expansion of text using LLM did not show improved performance, since the meanings decoded often deviated from ground truth.

Ultimately, a lightweight pre-processing pipeline was most effective with the removal of stop words and normalization of plate strings. These datasets form the bedrock for the proposed machine-learning pipeline that will allow for the creation of a robust system to decode and classify vanity license plates within a wide range of legal and cultural contexts.

4 Exploratory Data Analysis

The first step in our analysis was preprocessing the dataset to remove invalid entries, such as those with missing data or incorrect formats. We then examined the distribution of the dataset's `Status` column, which indicates whether a license plate was accepted (`Y`) or rejected (`N`). The dataset shows a significant imbalance, with a rejection-to-acceptance ratio of approximately 4:1, reflecting the stringent DMV review process.

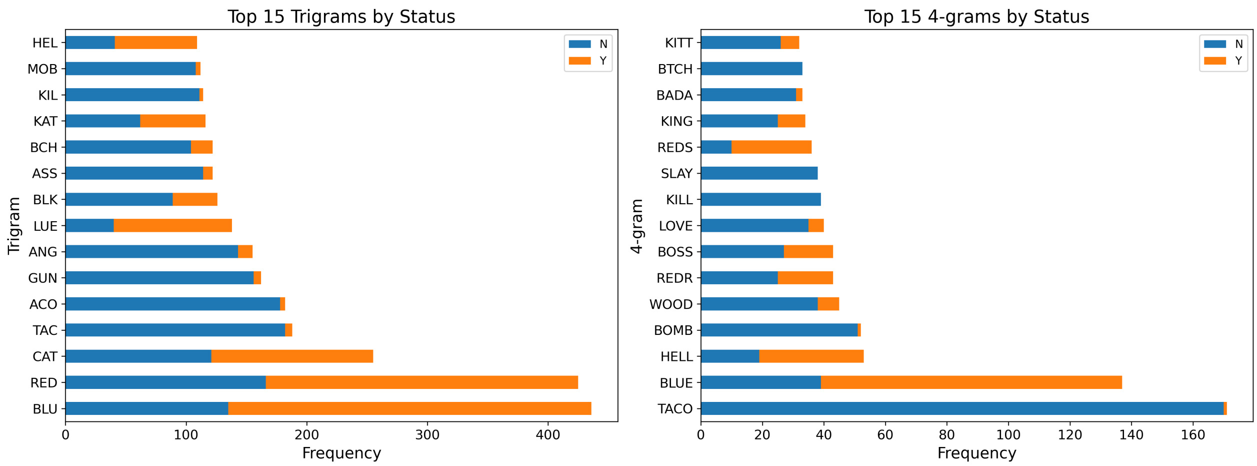

Figure 2: Top 15 Most Common Trigrams and 4-Grams in Vanity Plates

Next, we conducted a frequency analysis of n-grams in the license plates to identify common patterns. Figure 2 shows the top 15 most frequent trigrams and 4-grams. A notable observation is the frequent occurrence of the 4-gram `BOMB`, which has a near-max rejection rate due to its violent associations. In contrast, the 4-gram `BLUE` is more context-dependent, often appearing in acceptable plates related to colour or mood. Similarly, trigrams like `BLU` and `RED` appear frequently and have varied interpretations but may also be flagged for potential gang affiliations, depending on customer comments. These observations underscore how contextual nuances influence the acceptance or rejection of vanity plates, which will be essential for our decoding and classification models.

To further understand the dataset, we analyzed the top 15 bigrams in `reviewer comments` compared to `customer meanings`. This comparison highlights critical inconsistencies in the dataset and provides insight into how different stakeholders interpret vanity plates.

Figure 3 shows the stark contrast between the most frequent bigrams in `reviewer comments` and `customer meanings`. Reviewer comments often contain genuine reasons for rejection, such as `"area code"` and `"gang reference"`, which align with DMV regulations and rejection policies. On the other hand, customer meanings are frequently fabricated to pass plates through the DMV. These include creative or deceptive interpretations that downplay the plate's controversial or inappropriate nature.

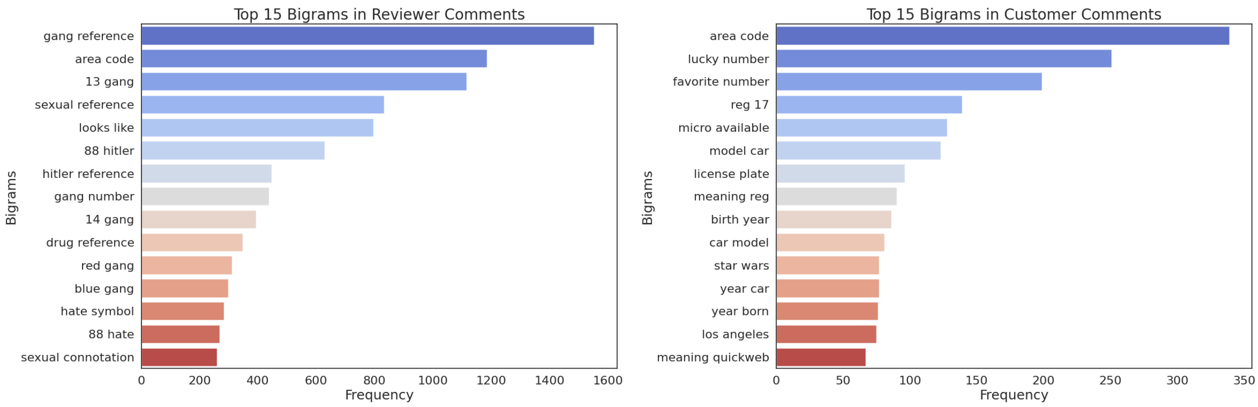

Figure 3: Top 15 Bigrams in Reviewer Comments vs. Customer Meanings

The frequency disparity between the two categories is striking. In `reviewer comments`, the most common bigram exceeds 5000 occurrences, while in `customer meanings`, the most common bigram barely reaches 340 occurrences. This disparity underscores the inconsistency and lack of reliability in customer-provided meanings. Consequently, customer meanings were excluded from our pipeline as they introduce significant noise and undermine the validity of the decoding and classification processes.
These findings further support the importance of leveraging reviewer comments as the primary source of contextual guidance for decoding models and highlight the challenges of working with highly imbalanced and inconsistent data.

5 Methodology

The methodology explores three approaches for classifying vanity plates: direct plate classification, classification using reviewer comments, and a hybrid pipeline that decodes plates into meaningful strings before classification. This section details the models, strategies, and rationale for each approach.

5.1 Pipeline Overview

Our research implemented a two-stage pipeline for decoding and analyzing vanity license plates:

1. **Decoding Vanity License Plates:** We used zero-shot, few-shot and fine-tuning of Seq2Seq[5] models and Large Language Models (LLMs) to interpret vanity license plates, using reviewer comments to provide contextual guidance and enhance interpretability.

2. **Classification of Plates:** We applied Bert Classification model[3] and Machine Learning algorithms such as random forest, SVMs and Logistic Regression for classification models to evaluate the legality of the decoded strings, categorizing them as either `accepted` or `rejected`.

5.2 Decoding Vanity License Plates

1. **No Preprocessing:** This approach leveraged the full text of vanity plates without filtering or modification, preserving all original information. While effective for maintaining raw context, it occasionally included irrelevant or noisy details, which could dilute interpretability.

2. **Stop Word Removal:** This preprocessing step was used to enhance the focus on meaningful words, reducing the noise in decoded strings. This could possibly lead to better and more concise decoding of vanity plates with less noise, which could eventually help in better classification.

3. **Keyword and Obscene Word Extraction with LLMs:** Considering the effectiveness of instruction-based tasks in LLMs and Transformer models, we opted to utilize them to preprocess and extract useful information from reviewer comments. Implemented zero-shot and few-shot prompting for Keywords and Obscene word extraction using zero-shot and few-shot prompting to identify critical terms effectively.

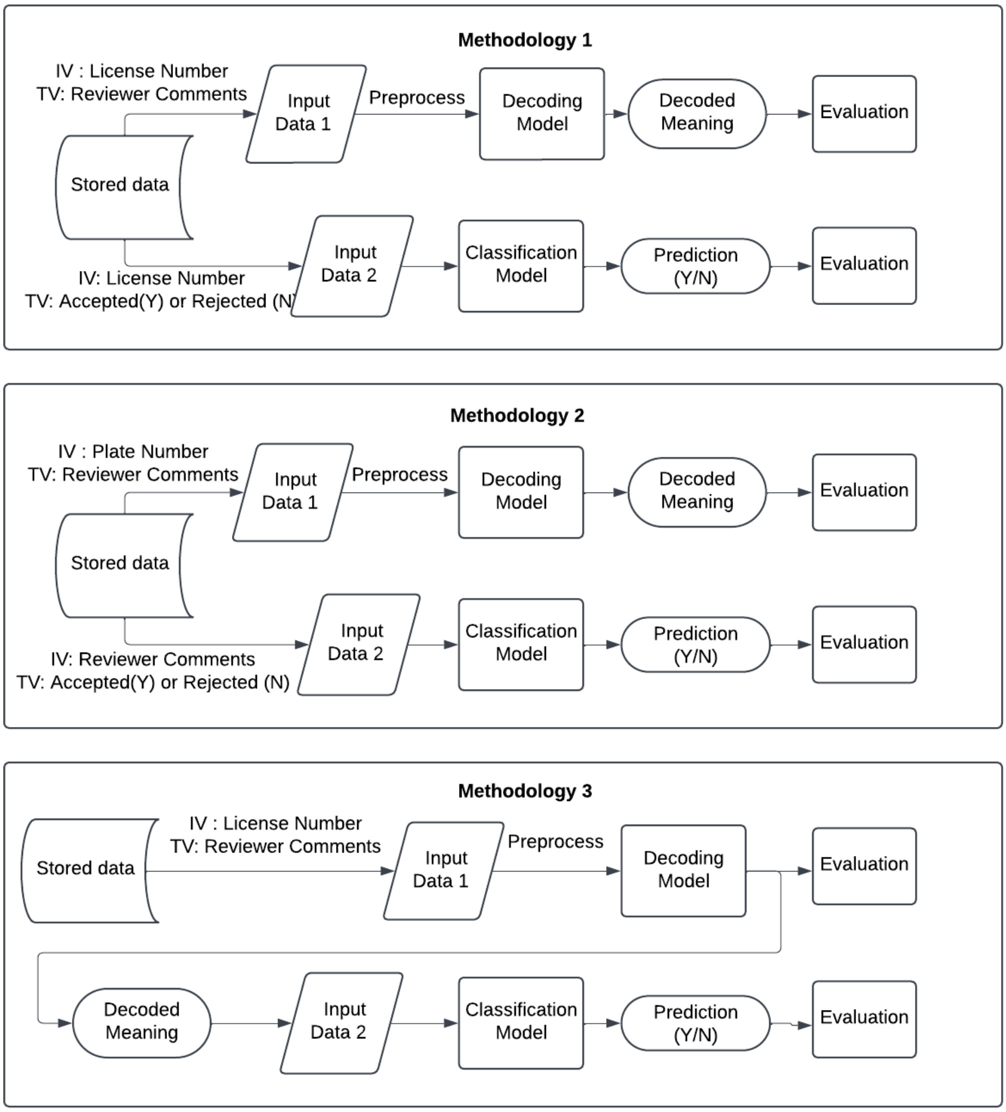

Figure 4: Three different approaches opted for task completion

5.2.1 Model Selection and Training

We evaluated a range of transformer-based models and large language models to transform vanity license plates into meaningful strings:

- **T5 (small and base):**[2] A versatile seq2seq model trained to perform diverse text-to-text transformations.
- **LLaMA 3.2:**[14] A high-efficiency and adaptable large language model.
- **Falcon 7B:**[15] A state-of-the-art model renowned for its large-scale pretraining and fluent language generation capabilities.
- **GPT-2:**[8] A widely used autoregressive model capable of generating coherent text, though it requires larger input contexts.

The model selection was done by carefully balancing performance with computational constraints. Although larger LLMs such as Falcon 7B[15] and Llama[14] provide more detailed and accurate results, their substantial computational requirements pose significant challenges when processing our vast dataset of license plate numbers. On the other hand, smaller models such as T5-transformer-based LLMs[2] provided a more computationally efficient solution without much compromising decoding quality.

We ultimately prioritized the T5 models, particularly their small and base variants, due to their comprehensive pretraining across multiple natural language understanding tasks. This broad training enabled the models to comprehend text structure and nuances, effectively transforming short, cryptic license plate strings into coherent and contextually appropriate decoded representations. The text-to-text architecture of T5 models made them particularly well-suited for our decoding objective.

Model training was completed by running over 20 epochs and with rigorous parameter fine-tuning. Analysis of cross-entropy loss was performed over every few epochs with an aim to prevent overfiting to training data. Through this methodical approach, we trained the models to generate meaningful decoded strings from vanity plates, creating a robust input for the subsequent classification stage. This critical step ensured that the decoded strings contained sufficient context for accurately determining the legality of each plate.

5.3 Classification of Decoded Strings

Three distinct approaches were evaluated to classify vanity plates into the binary categories of `accepted` or `rejected`:

5.3.1 Direct Plate Classification (Methodology 1)

The first approach involved directly classifying the vanity plates themselves without any additional decoding or contextual information. In this case, the raw plate string was fed directly into a sequence classification model **(BERT)**[3], which was trained to predict the legality of the plate. This approach aimed to determine if the plate alone could provide enough context for a reliable classification. The model was fine-tuned on a labelled dataset consisting of plates and their associated legal status.

5.3.2 Classification Based on Reviewer Comments (Methodology 2)

In the second approach, the classification model was trained using only the reviewer's comments as input. The idea was to see if the detailed context and the explanations provided by reviewers could offer sufficient information to classify the plates. This approach ignored the vanity plates themselves, instead relying solely on the textual descriptions provided by the reviewers to make a decision. This method benefits from the clarity and accuracy of the reviewer's insights.

5.3.3 Proposed Pipeline: Using Decoded Strings for Classification (Methodology 3)

The third approach used a more intricate pipeline, combining both decoding and classification steps. In this pipeline, the vanity plates were first decoded into meaningful strings using the T5-based decoder model described earlier. The decoded strings, enriched with the contextual guidance from the reviewer's comments, were then fed into a sequence classification model (BERT[3]) to predict the legality of the plate. This approach aimed to leverage the strengths of both the vanity plates and the reviewer comments, with the decoding step acting as an intermediary process to consolidate and clarify the information. The importance of this methodology arises due to the lack of reviewer comments in other datasets. This methodology aims at bypassing the requirement of a reviewer overall.

6 Evaluation and Experimental Results

The evaluation examines the performance of decoding strategies and classification methodologies used to determine the legality of vanity plates. The results highlight the effectiveness of each method in terms of weighted F1 scores, human evaluation feedback, and preprocessing strategies. The results are highlighted in Figure 5.

6.1 Decoding Vanity Plates Analysis

Beyond initial human evaluation of results, the **Metrics used** to perform this analysis was the BERT score. The main reason for using this metric over other metrics, such as BLEU and CHRF, was due to its ability to compare two texts not only based on the literal characters or words but also on their meaning and context. BLEU and CHRF mainly depend on comparing n-grams without considering the context and meaning of the sentences. Since our reviewer comments were in the form of sentences rather than just full forms of abbreviations, the BERT score was more applicable to our use case. Few-shot and zero-shot prompting using examples from the dataset gave subpar results, with mostly gibberish and completely unrelated meanings. We tested various prompts across multiple Open-Source models such as Llama, Falcon, GPT2 and T5 without

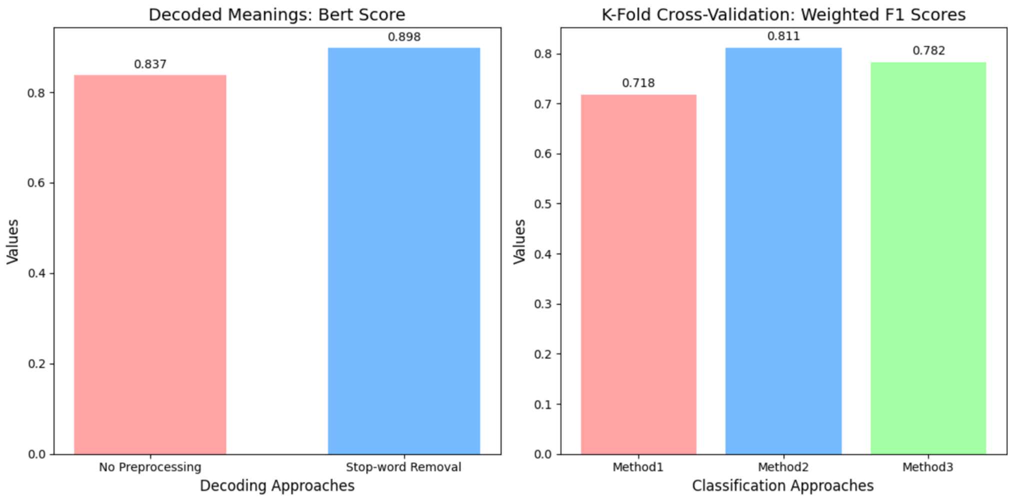

Figure 5: Decoding and Classification Testing Data results (M1, M2 and M3) are methodologies as explained in 4

any quantifiable improvement even with the more complex and sophisticated models, whereas fine-tuning gave promising results. Due to computational constraints, priority was given to t5 models as explained in section 5.2.1. Decoding without any preprocessing yielded a **BERT score of 0.83**, establishing a strong baseline. Removing stop words improved the decoding process, achieving a **BERT score of nearly 0.9**. The results have been visualized in Figure 5. After using LLMs to extract Keywords and Obscene word extraction from Reviewer comments, it was highly effective at extracting relevant terms which were analysed using human evaluation. But, it failed to maintain overall sentence meaning which led to inaccurate decoding when trained for t5 models. As a result, this approach was discarded following human evaluation, despite its technical efficiency in identifying key elements.

Figure 6: Generated Wordclouds for Actual Meanings and Decoded Meanings

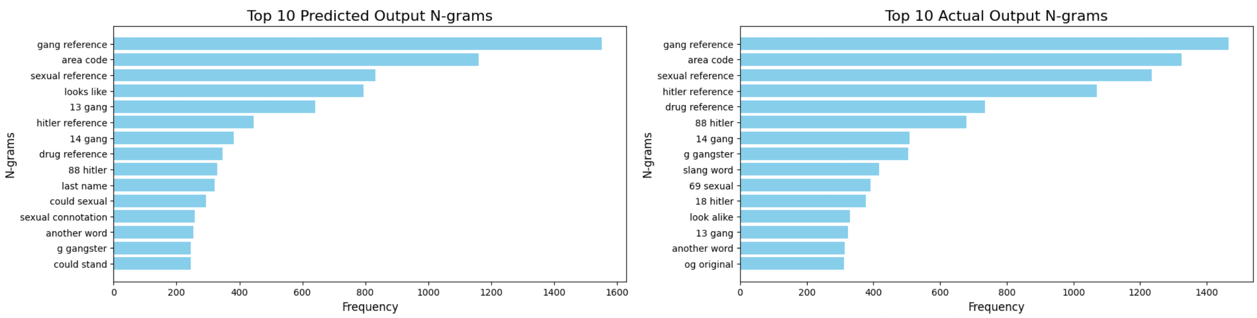

Figure 7: Most Frequently N-grams of Actual Outputs (reviewer's comments) and Predicted Outputs

Along with BERT scores, we also generated word clouds as in Figure 6 and most commonly occurring N-grams as referenced in Figure 7 to further verify the model's performance. These graphs show key similarities between the generated outputs and actual reviewer comments with commonly occurring bigrams like "gang reference", "area code" and "sexual reference" being the most frequent cases.

6.2 Classification Performance

Metrics and Evaluation methodology opted due to a relatively smaller dataset and the class imbalance was a k-fold cross-validation process and weighted F1-Score. This helps avoid overfitting and a more robust evaluation across different folds of data for training and testing. a weighted F1-score is optimal as it accounts for both, precision and recall as well as taking into consideration the less frequent class.

In Direct Plate Classification *(Methodology 1)*, raw vanity plate strings were directly classified using a BERT model, achieving a **weighted F1 score of 0.71**. This method relied solely on the cryptic plate text, which often lacked sufficient context for precise classification. Classification based solely on reviewer comments*(Methodology 2)* yielded the best results, with a **weighted F1 score of 0.81**. The model benefited from detailed contextual insights provided by reviewers, enabling it to make more informed predictions. However, the models were unable to achieve a very high accuracy given the high variance in sentence formation and structure of reviewer comments.

The hybrid pipeline combined decoded plate strings with reviewer comments *(Methodology 3)*, resulting in a **weighted F1 score of 0.78**. While it slightly underperformed compared to Method 2, it effectively integrated insights from both plates and comments, reducing ambiguity. This method highlighted the potential of combining multiple data sources to enhance interpretability and accuracy. The results have been visualized in Figure 5.

7 Conclusion and Future Scope

In conclusion, this study demonstrates that our hybrid approaches can help decode and classify vanity license plates without the need for manual human evaluation. The decoding task, which already yields respectable results as indicated by the BERT scores 5, could be further refined by extended fine-tuning of more sophisticated large language models. The T5-small model is highly promising for decoding, considering its computational efficiency and specialization in sequence-to-sequence tasks. In classification, the comparison of BERT-based classification using both reviewer comments and decoded strings shows that decoding indeed assists classification but still falls slightly short of being as effective as reviewer comments as it relies on the effectiveness of Decoded meanings of vanity license plates. The hybrid pipeline, which integrated the decoded plate strings with the reviewer comments, performed well, showing that a combination of raw plate data with the insight provided by the reviewers reduces ambiguity in classification. Most importantly, this removes human evaluation from the loop for the classification of plates altogether, going full circle to complete automation.

The scope of future work would be to first and foremost acquire more data including reviewer comments or the meaning of vanity license plates via online forums, crowdsourcing or other methods. This would give us the ability for better fine-tuning of more sophisticated LLMs like LLama, GPT and Claude, for longer epochs to yield much better decoding and also classification tasks. With more training, these models could capture the subtlety of vanity plates and what they mean, thus driving a more accurate, efficient, and scalable system for DMV applications. Our codes can be found here[16][17][18].

References

[1] C. D. of Motor Vehicles, "California dmv," https://www.dmv.ca.gov/, [Accessed: Nov. 14, 2024].

[2] C. Raffel, N. Shazeer, A. Roberts, K. Lee, T. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, pp. 1–67, 2020. [Online]. Available: https://arxiv.org/abs/1910.10683

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[4] S. K. R. S. Satapathy and S. R. B. R. Raju, "Microtext normalization using seq2seq models for informal text," *International Journal of Computer Science*, vol. 34, no. 5, pp. 1204–1211, May 2019. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8851895

[5] "Seq2seq: A framework for sequence-to-sequence learning," https://proceedings.neurips.cc/paper_files/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf, [Accessed: Nov. 14, 2024].

[6] R. K. M. P. Chaturvedi and S. Kumar, "Affectivespace: Dynamic word vectors for implicit meaning representation," *Journal of Natural Language Processing*, vol. 42, no. 1, pp. 85–94, Jan 2024.

[7] H. W. Z. Yi and Y. Li, "Profanity detection in social media using word embeddings and lstm networks," *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 56–67, July 2021.

[8] OpenAI, "Gpt: Generative pre-trained transformer," 2024, accessed: 2024-12-05. [Online]. Available: https://openai.com/gpt

[9] Y. Chae and T. Davidson, "Large language models for text classification: From zero-shot learning to instruction-tuning," SocArXiv Preprints, August 2023. [Online]. Available: https://osf.io/preprints/socarxiv/sthwk

[10] C. D. of Motor Vehicles, "Vanity license plate applications dataset (2015-2016)," https://github.com/veltman/ca-license-plates, [Accessed: Nov. 14, 2024].

[11] D. Veltman, "Ca license plates: A dataset for california vanity plates," 2024, accessed: 2024-12-05. [Online]. Available: https://github.com/veltman/ca-license-plates

[12] MuckRock, "Muckrock: A platform for government transparency," https://www.muckrock.com, n.d.

[13] C. D. of Motor Vehicles, "Vanity license plate applications dataset (2015-2016)," https://www.muckrock.com/foi/california-52/rejected-personalized-aka-vanity-license-plates-in-2013-department-of-motor-vehicles-11655/, [Accessed: Nov. 14, 2024].

[14] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, and et al., "Llama: Open and efficient foundation language models," arXiv preprint arXiv:2302.13971, 2023. [Online]. Available: https://arxiv.org/abs/2302.13971

[15] T. T. I. Institute), "Falcon 7b: A large-scale, open-access language model," 2024, accessed: 2024-12-05. [Online]. Available: https://huggingface.co/tiiuae/falcon-7b

[16] "Evaluation: License plate interpretation," 2024, accessed: 2024-12-05. [Online]. Available: https://colab.research.google.com/drive/1_ps_oL5k5OTvAmImMGJ_QUOJKkCVtzZZ?authuser=3#scrollTo=RhrEmRQxhehb

[17] "Fine tune llm: License plate decoding with llms," 2024, accessed: 2024-12-05. [Online]. Available: https://colab.research.google.com/drive/1NcHaBG4WFEG4HHxEgslySuMjjGwyRd0D?authuser=3

[18] "Classification: Vanity plate classification with bert," 2024, accessed: 2024-12-05. [Online]. Available: https://colab.research.google.com/drive/1m6iS7LFhkE-fJNks0_OTgx6GrEd2LE_i?authuser=3