Student Declaration:

By submitting this assignment, I confirm that I am aware of DBS's policy regarding cheating, plagiarism and all other forms of academic impropriety. The coursework submitted is my own or my group's work, and all other sources consulted have been appropriately acknowledged. I am aware that in the case of doubt an investigation will be held.

Title: Critical Analysis of Text Classification & Prediction

Student Name	Mohammad Minhazul Amin
ID	20005267
Module Code	B9BA103
Module Leader	Kunwar Madan

Date of Submission: 11/12/2023

Table of Contents

Use of ChatGPT

Chatgpt is used for data generation for this project as per the requirement where 1000 rows of Toothpaste reviews are made with two columns ‘Text’ consisting of 1000 customer reviews with 500 positive reviews and 500 negative reviews and ‘Label’ contains ‘Positive’ and ‘Negative’**.**

Conditions used in Prompts for Data Generation

- **Mislabel Reviews:** Label some positive reviewed as negative and vice versa
- **Uniqueness:** Ensure each review is unique in wording, structure, and usage purpose.
- **Expression Variety:** Articulate similar reviews in different styles
- **Real-World Reflection:** Ensure the dataset reflects real-world communication patterns
- **Non-Alphabetic Characters:** Include characters like "@", "http://", etc., in each review
- **Contextual Variability:** Include cultural and linguistic diversity in content.

Advantages

1. The Future of Balanced Datasets: Many datasets are naturally imbalanced can lead to biased prediction by favoring the majority classes. Chatgpt can mitigate this challenge through creating diverse and representative samples for underrepresented categories, thus achieving a more equitable distribution of classes.

2. Efficiently Generating Dataset: Chtagpt can assist significantly in the dataset creation process. The generation of such a large dataset would typically be conducted through multiple iterations or prompts, with each prompt producing a manageable number of reviews. However, traditional methods of dataset collection, such as surveys, user feedback, or web scraping, can be significantly slower and more resource-intensive.

3. Regular Expression: Chatgpt can incorporate real-world noise “[A-Za-z0-9._%+-]+” which can train the model in a complex way and the model can provide better performance in processing and understanding diverse and noisy real-world data and enhance flexibility and adaptability to perform better in unseen data.

Review: Recently switched to toothpAste and I'm quite impressed!. They've 0-9got amazing resources on oral health at their blog https://blog.herbal-dent.com. Highly recommend it!"

4. Tone Diversity: Human made datasets consist of diversity of expression which reflect a range of human communication. Since Chatgpt is trained on its training data based on large text datasets it can provide different communication styles to differentiate each review of toothpaste.

Enthusiastic	Absolutely love this toothpaste! The way it leaves my mouth feeling so fresh and clean is just unbeatable. And the mint flavor? Simply the best!"
Analytical	"After using this toothpaste for a month, I've noticed some subtle improvements in teeth whiteness. The fluoride content aligns with dental recommendations, and the texture is pleasantly smooth."

Limitation

- **Duplicate use of words:** ChatGPT can sometimes repeat or use synonyms in close proximity, which might not always be stylistically ideal. This can happen due to the way the model processes and generates language, where it aims to ensure clarity or emphasis but might end up being redundant.

Prompt: Generate 50 unique toothpaste reviews each labeled as half positive and negative.
This toothpaste doesn't just battle cavities; it unleashes a burst of minty freshness that leaves your mouth invigorated and ready to take on the day.
This toothpaste doesn't simply shield against cavities; it delivers an exhilarating wave of minty freshness that leaves your mouth feeling revitalized and ready to conquer the world

- **Generalization from Training Data:** If the user does not provide a specific usage for a variant of toothpaste, Chatgpt will provide generic reviews without emphasizing the specific purpose of the usage in the context of a variant. Similar situation is applicable for the diverse usability of toothpaste. Additionally, if the user mentions the variant ‘Whitening’ it will generate reviews on toothpaste. Users have to mention the toothpaste variant and variety of usage for diversified data generation. This may happen because chatgpt generates feedback based on common attributes associated with toothpaste in general, such as freshness, cleaning effectiveness, or flavor, rather than focusing on specific variants unless prompted to do so. However, this could be alarming for generating large datasets based on the real-world.

Prompt: Generate 50 unique reviews of toothpaste labeled as Positive and Negative
good value refreshing breath highly recommend improves oral health pleasant aroma eco-friendly packaging perfect for daily use gentle on gums good value refreshing
Prompt: Generate 50 unique reviews of whitening toothpaste labeled as Positive and Negative
improves oral health gentle on gums whitens teeth refreshing breath perfect for daily use whitens teeth good value pleasant taste perfect for daily use pleasant

- **Complexity of Multiple Conditions:** Prompt with multiple conditions considering different aspects enhance difficulty for toothpaste to generate reviews. They require balancing accuracy, creativity, linguistic diversity, and cultural sensitivity, which can be challenging for any AI language model, including ChatGPT.

- **Real World Reflection Challenge:** Even though chatgpt can capture sound reviews from its training data, it can not capture individual nuances, detailed personal experiences, and specific preferences often found in authentic user-generated reviews. However, this gap becomes more apparent when generating large datasets.

Reality	I've always had sensitive gums, so I'm careful about the toothpaste I use. I decided to try this toothpaste because my usual one was out of stock. Surprisingly, it didn't irritate my gums, and the mint flavor is just the perfect intensity for me. It reminds me of the spearmint gum I used to have as a kid. I also noticed that my teeth look a bit brighter. It's a bit pricier than my usual brand, but I think it's worth it.
Chatgpt	I recently tried this new toothpaste and absolutely loved it! The mint flavor is just right - not too strong but still refreshing. It leaves my mouth feeling clean and fresh all day. Definitely going to buy it again!

- **Uniqueness & Variety Challenges:** Though one of strengths of chatgpt is to generate creative response with clarity but ensuring uniqueness and review reflecting diversified styles over a large dataset might lead to potential repetition of reviews in different sentences and stylistic inconsistencies.

AI Generated Toothpaste Reviews
eco-friendly packaging perfect for daily use gentle on gingiva pleasant taste good value quality product highly recommend quality product perfect for daily use long-lasting effect
eco-friendly packaging highly recommend long-lasting effect improves oral health pleasant aroma highly recommend improves oral health long-lasting effect highly recommend highly recommend perfect for daily

Text-to-Structured-Data Conversion

In Text Mining, Term Frequency-Inverse Document Frequency(TFI-DF) is used to evaluate the prominent usage of a word in each review. A cleaned review column created after removing regular expression. The TF-IDF with specific parameters, min_df = .03 parameter means that a term must be in at least 3% of the documents to be considered. The ngram_range=(1,3) parameter indicates that unigrams, bigrams, and trigrams will be used. Afterwards, 576 rows created consist of vocabulary created for implementing the text mining algorithms.

- **Missing Entries:** TF_IDF is not able to identify missing values within the dataset. Usually, missing values for text classification are handled through imputation technique. However, TF-IDF operates based on the presence and frequency of words or terms within a document but does not have the capability to infer missing context or information.

Synthetic Review	Review text missing for entry 684 Review text missing for entry 684 Review text missing for entry 684 Review text missing for entry 684 Review ðŸ ƒ

- **Difficulty with Rare Terms:** In the context of Text classification dataset, if a toothpaste review contains a unique review ‘hydroxyapatite’. TF-IDF might weight this term higher due to its rarity. This approach potentially overestimates its importance throughout the vocabulary which is however more common and contains equally important ingredients such as ‘fluoride’.

- **Syntactics Text**

Original Review	absolutely love this toothpaste because it's gentle on my sensitive gums, unlike other brands that often cause irritation and discomfort
Syntactic Review	Unlike other brands that often cause irritation and discomfort, this toothpaste is gentle on my sensitive gums, I absolutely love because

TF-IDF Implication: TF (Term Frequency) usually takes the frequency of the word into consideration to evaluate how frequently each word appeared within the 3% of the document.

Drawback: This approach may lead to misinterpretation because the original review reflects the satisfaction whereas the rearranged one is negative review. This approach of detecting the frequent usage of words within the corpus leads to inaccurate sentiment analysis in data procession.

- **Semantics Text**

Original Review	The new whitening toothpaste I tried is quite abrasive, which I found effective for removing stains, but it may be too harsh for some

TF-IDF Implication: Term Frequency (TF) for "abrasive" = 1 occurrence in the review / 19 total words and Inverse Document Frequency (IDF) for "abrasive" (assuming it appears in 10 out of 1000 documents). Total score would be 0.2424 which may reflect high due to the high score of IDF 4.605 and consider this word as a significant word in the document compared to other words.

Drawback: However, it might interpret "abrasive" as purely negative, missing its dual implications. It wouldn't capture the nuanced usage of "abrasive" as both a potentially negative attribute (being harsh) and a positive one (effective in removing stains). This could lead to skewed product feedback analysis, with potential repercussions for marketing or product development. So, the TFI-DF has its limitation to understand the semantic depth of a word.

Model Deployment Performance

Business Scenario

A renowned toothpaste brand is receiving a high volume of feedback after their new launch. So, the company is looking forward to aiming for these feedbacks in order to improve overall customer experience. The feedback is diverse and collected from various channels from social media and e-commerce platforms.

Business Objective

The primary objective here is to categorize these feedback into positive and negative labels so that the company identifies the areas of improvement through this classification. Here the positive label is “Negative” because it is crucial for the company to currently identify negative customers which are about to leave the company due to their misinterpreted feedback.

Cost of Acquisition/Loss

If a customer decides to leave because they feel their negative feedback is being misinterpreted,

- A customer spends an average cost of €100 per year
- Typically stays with the company for 5 years
- Lifetime Time Value = €100 x 5 years = €500
- Cost of Acquiring a New Customer = €300

- Negative Word-of-Mouth Impact = 15 people
- Lost Opportunity Cost = 20% of LTV

Total Cost = LTV + CAC + (Negative Word-of-Mouth Impact x LTV) + Lost Opportunity Cost

= €500 + €300 + (15 x €500) + (€500 x 20%) = €8,400 ((False Negative)

If a customer decides to leave because they feel their positive feedback is being misinterpreted,

- Lifetime Time Value = €100 x 5 years = €500
- Cost of Overcompensation = €200
- Negative Word-of-Mouth Impact = 10 people
- Resource Misallocation = 5% of LTV

Total Cost = €500 + €200 + (10 x €500) + (€500 x 5%) = €5,725 (False Positive)

Since, True Positives and True Negatives are correct classifications, they ideally should not involve additional costs .

Performance of the Model

Since the cost of False Negative is higher than False Positive. Recall would be an ideal metric that will minimize the cost of False Negative. Among two algorithms Fasttext scored slightly higher recall than Nearest Centroid so Fasttext will be applicable for the toothpaste company.

Algorithm	Precision	Recall
Nearest Centroid	0.9076734693877553	0.8826692133422641
Fasttext	0.8999999999999999	0.8847904128533587

Why does the Model Perform Well?

Term Frequency (TF): Certain terms like whitening, flavor, sensitivity are frequently used in toothpaste reviews TF weighted each of these words higher to reflect their importance in labeling positive and negative.

Inverse Document Frequency (IDF): However, some words are common across many reviews but not relevant to decide the sentiment of the review such as toothpaste, teeth, product. IDF reduces the weight of these common terms across the dataset, focusing instead on words that are unique or less common but more informative.

Min-df (‘min_df: .03’): min_df=.03 means that each ngram (unigram, bigram, & trigram) must be present in at least 30 documents for it to be considered as a token (0.03 * 1000 = 30). This assisted the model to filter out the rare terms and emphasize its focus on relevant terms.

N-grams('ngram_range=(1,3)): 1 gram represent as unigrams capture individual keywords that are used frequently in toothpaste review labeling 'Positive' and 'Negative'. Additionally, integrated each common word to identify its contextual relationship and complex expression for understanding the nuances of the customer review for the final sentiment.

Fasttext Hyper Parameters

Learning Rate ('lr=0.5'): Learning rate basically determines how quickly or slowly a model learns. Toothpaste reviews contain recurring themes or phrases related to product features (taste, effectiveness, packaging). A higher learning rate quickly adapts to these patterns for accurate classification.

Number of epochs ('Epoch=25'): This parameter ensures the model does not learn too much from the noise or specificities in the training data. Ecoch iterates the model over 25 times to ensure that it sufficiently learns from all parts of the data.

Vector Dimension ('dim=100'): To understand the various ways of customers expressing their opinions, vector dimension provides 100 dimensional space to capture semantic relationships and understand these nuanced expressions which is not as large as to overfit the training data.

Minimum Count ('minCount=2'): This minimum count assists the model to focus on common words which are relevant to the overall sentiment instead of considering rarity or uncommon words.

Limitation of Fasttext Model

- **The Song beneath the Line:** Fasttext is designed to focus on capturing subword information such as prefixes and suffixes to understand meaning of the words. However, this approach is not adequate to grasp the full context of the sentence. For instance, 'strong flavor' in a toothpaste might have an equal chance of labeling into positive or negative which depends on the holistic context of the review.

- **Cultural and Linguistic Variability:** Phrase, idioms, local slang contain mixed usage of language and demonstrate culture diversity which is difficult for Fasttext to handle due to its subword approach. Such linguistic nuance requires better understanding beyond the context. Toothpaste reviews from rural areas may reflect such challenges for fasttext as these elements require a deeper understanding of cultural contexts and language use patterns that go beyond mere subword analysis.

- **Overemphasize on subwords:** "This toothpaste has a refreshing, non-traditional minty flavor that's quite invigorating." Fasttext will focus on each subwords "non-", "traditional", "minty" and "refresh". However, this focus on subwords could interpret negative sentiment without understanding the context which might lead to potential inaccuracies in understanding the sentiment of the review.

Conclusion

Creating dataset from the real-world can be costly and time-consuming, sometimes they don't have adequate information and variety of response to create a robust model whereas large language models like chatgpt generated dataset can be tailored according to specific needs and cost-effective and time-consuming approach.

This report is basically based on AI-generated dataset on toothpaste review and implication of machine algorithm (Nearest Centroid) and natural language processing (NLP) (fasttext) algorithm to train the model for the real-world for classifying customer feedback as 'Positive' and 'Negative' label.

According to the business context, the positive label is "Negative" which clearly indicates the correctness of predicting negative feedback as negative. Considering the higher cost of false negative and alarming concern of minimizing the cost, recall is the ideal metric since recall minimizes the cost of false negative.

From the model, fasttext scored the highest recall compared with the traditional machine learning algorithm Nearest Centroid. Fasttext harnesses the speciality of capturing subwords to understand the meaning of each word and connects its correlation with the label.

FastText might be a suitable model for classifying toothpaste reviews, given its capability to learn word representations from training data. However, it's practical to explore other models in real-world scenarios to identify the most effective solution. While FastText is designed specifically for learning word embeddings from its training dataset, experimenting with different NLP algorithms can help determine the best fit for this particular classification task.

Reference

1. Clearbox AI. (2023) Why ChatGPT isn’t your best bet to generate data [Online]. Available at: https://www.clearbox.ai/blog/2023-07-11-why-chatgpt-isnt-your-best-bet-to-generate-data (Accessed: [21st November, 2023]).