Final Project Report

McGill University - Desautels Faculty of Management

Introduction & Problem Statement

Black Friday, an event of immense economic significance, sees consumer spending surge to approximately $60 billion as shoppers hunt for the best deals. Understanding customer sentiment towards this day is vital for businesses to strategize effectively, aiming to meet consumer expectations and capitalize on this peak shopping period. This analysis aims to examine the evolution of sentiment scores regarding Black Friday before and after the COVID-19 pandemic. By utilizing Natural Language Processing (NLP) and regression techniques, it seeks to provide insights into consumer attitudes and behaviors.

Data Retrieval &Pre-Processing

To conduct the analysis data from amazonprime, blackfriday, bestbuy, anticonsumption and walmart subreddits were used. The text data extracted from Reddit threads underwent a comprehensive preprocessing and Natural Language Processing (NLP) pipeline. Initially, the data from various threads were merged into a single DataFrame, enabling a unified approach to analysis. The timestamp within each post was utilized to extract the year, facilitating temporal analysis and categorization of the data. Additionally, a `var_year` column was introduced as a categorical variable, simplifying the process of grouping, and comparing data across different periods. The NLP steps initiated with tokenization, which involved decomposing complex sentences into individual words, laying the groundwork for further analysis. This was followed by lemmatization, where words were reduced to their base or dictionary forms (lemmas), ensuring consistency in the dataset. Part-of-speech (POS) tagging was then applied, assigning grammatical categories to each token based on its definition and context, enhancing the accuracy of subsequent processing steps. N-Gram modeling was employed to capture textual context using unigrams (1-gram) for individual words, bigrams (2-gram) for pairs of words, and trigrams (3-gram) for sequences of three words. This approach allowed for the analysis of word patterns and phrase usage within the text. The final step in the preprocessing was the application of Term Frequency-Inverse Document Frequency (TF-IDF) to evaluate the significance of words across the dataset for all monograms, bigrams, and trigrams, both in the lemmatized and non-lemmatized datasets. This method quantified the importance of each word within documents relative to the entire corpus, highlighting terms that offer unique insights into the content being analyzed.

Sentiment Analysis

To conduct a comprehensive sentiment analysis, both VADER sentiment analysis and K-Means clustering methods were utilized. Although VADER is well-known for its robustness with social media language, we wanted to ensure its suitability for our Reddit posts. This is why we sought to compare its performance against another algorithm. In the VADER approach, each post was analyzed to assign both a categorical sentiment label and a numerical sentiment score, leveraging a compound score to categorize sentiments. Specifically, a compound score of 0.1 or higher was indicative of a positive sentiment, a score of -0.1 or lower suggested a negative sentiment, and scores between -0.1 and 0.1 were classified as neutral. Analysis of the total posts revealed that 26,002 were positive, 14,746 were neutral, and 8,561 were negative. Subsequently, a K-Means clustering approach was applied to the TF-IDF vectorized counts for monograms, bigrams, and trigrams. This analysis highlighted significant dissimilarities between the sentiment categorizations derived from the VADER and clustering approaches, particularly noting a 50% dissimilarity rate in the classification of negative sentiments between the two models. To further investigate these discrepancies, a manual review was conducted on 10 posts classified as negative by the K-Means model, comparing these classifications against those derived from VADER. This review concluded that the VADER classifications were generally more aligned with the actual sentiment expressed in the comments. Consequently, VADER-derived sentiments were favored for the subsequent phases of the analysis. This strategic approach to sentiment analysis, incorporating both automated sentiment analysis via VADER and exploratory clustering via K-Means, offered an understanding of sentiment distribution across the dataset. The decision to rely on VADER's sentiment classifications for further analysis steps was informed by the method's closer alignment with the perceived sentiment of the textual data, underscoring the importance of method selection in sentiment analysis tasks.

Latent Dirichlet Allocation (LDA)

A Latent Dirichlet Allocation (LDA) model was used to discern latent topics within the textual data. Each post in the dataset was represented as a combination of topics, with probabilities indicating the extent of relevance to each topic. Each post is assigned probability scores for two topics. After experimenting with different combinations of preprocessing techniques and number of topics, it was discerned that two topics yielded the most coherence. With more topics, the topic distribution was skewed, with almost 100% of the posts being grouped into just two topics. With two topics, the distribution was more balanced at 47% and 53%. Words like 'online', 'Amazon', 'link', 'code', and 'deal' suggested that the theme of one of the topics was online shopping, possibly indicating people hunting for deals online. Conversely, the theme of the second topic was in-store shopping, with words like 'Walmart', 'store', 'people', 'work', and 'Thanksgiving', possibly reflecting people shopping with their families during the holidays.

Logistic Regression Analysis

Initially we load the results of the sentiment analysis (this will be our labels), the LDA topics and the (1:3)-grams TFIDF word vector into the same dataset in order to run a logistic regression (at first we tried to predict the sentiment scores using linear regression but the results were poor so we simplified the approach with classification). The negative comments were upsampled for each model in order to have a balanced dataset.

Logistic regression models are then trained on the balanced datasets to predict sentiment based on features including subreddit IDs, sentiment (Label), comment score, year information and LDA results. After training, model performance is evaluated using accuracy, precision, recall, and F1 score metrics, and the coefficients for the last nine features are extracted and displayed to understand their impact (coefficients other than the 5000 word vector features).

Furthermore, the code segments the data by year (before and after a specific year, excluding year 3 to have a clear before and after covid delimitation) and performs the same preprocessing, upsampling, and logistic regression analysis for each segmented dataset to compare the effect of time on sentiment.

Finally, causal inference analysis is conducted using the causalml library to estimate the Average Treatment Effect (ATE) of the time period (treated as a binary treatment variable) on sentiment scores, employing both T-learner (using gradient boosting) and S-learner (using linear regression) models. This part aims to quantify how sentiment scores are influenced by the transition from before to after the specified event (potentially COVID-19, given the reference to changes after COVID). The ATE estimates from both models suggest a slight decrease in sentiment scores following the event, indicating a potential negative impact of the event on public sentiment as expressed in the analyzed text data.See Appendix 1

Expected Impact and Key Takeaways

Individual elements of the approach highlighted in this report, such as sentiment analysis using VADER or topic modeling with LDA, are widely used in analyzing consumer sentiment across various domains including retail, e-commerce, and social media discussions, the integration of these methods to specifically analyze Black Friday sentiment over time, especially in relation to pre- and post-pandemic sentiments, may not be as common.

The findings from this type of analysis can have significant implications for businesses, the economy, and society. For businesses, understanding shifts in consumer sentiment towards Black Friday can help in tailoring marketing strategies, optimizing inventory and pricing models, and enhancing customer experiences to meet evolving preferences. Economically, insights into how sentiments and shopping behaviors have changed, especially post-pandemic, can inform predictions about future consumer spending patterns, potentially guiding policy, and decision-making to stimulate economic activity during key shopping periods.

Societally, examining the sentiment towards events like Black Friday can reveal broader trends in consumer culture, such as the growing emphasis on sustainability, the shift towards online shopping, or changing attitudes towards consumption and materialism, which can influence public discourse, policy, and individual behaviors.

The topic is of importance for several reasons. Firstly, Black Friday is a significant event in the retail calendar, representing a peak shopping period that can account for a substantial proportion of annual sales for many businesses. Analyzing the evolution of sentiment towards Black Friday, especially in the context of the COVID-19 pandemic, offers insights into how major global events can impact consumer behavior and market dynamics. This can help businesses and policymakers to better navigate future challenges. Lastly, the topic sheds light on the societal attitudes towards consumption and the sustainability of such shopping practices, contributing to broader discussions about consumerism, economic sustainability, and the environmental impact of major retail events.

Appendix 1

Feature	**Coefficient Overall**	Coefficient **BEFORE** COVID-19	Interpretation Set 1	Coefficient **AFTER** COVID-19	Interpretation Set 2
score	-0.324609	0.01974	Increase	-0.75727	Decrease
in_store_shopping_prob	-0.138068	-0.25425	Decrease	-0.23051	Decrease
online_shopping_prob	0.141088	0.24963	Increase	0.24718	Increase
Year_var	0.021753	0.00000	No effect	0.00000	No effect
subreddit_Anticonsumption	-0.105235	-0.26945	Decrease	-0.05509	Decrease
subreddit_amazonprime	0.017945	0.24879	Increase	-0.21033	Decrease
subreddit_bestbuy	0.040561	0.06539	Increase	0.09920	Increase
subreddit_blackfriday	0.162523	0.15258	Increase	0.32097	Increase
subreddit_walmart	-0.112774	-0.20013	Decrease	-0.13808	Decrease