GROUP FINAL PROJECT (PHASE 2)

NLP & RECOMMENDER SYSTEMS

ANALYSIS REPORT

GROUP MEMBERS:

DHRUVIL PATEL

JENISHKUMAR JETANI

MAULIK KHOULADIYA

PARITA SHAH

NILKANTH PATEL

Contents

[OBJECTIVE 3](#_Toc163844241)

[Data loading and exploration: 3](#_Toc163844242)

[Text representation: 3](#_Toc163844243)

[Text classification using Naïve bayes and SVM. 4](#_Toc163844244)

[TEXT CLASSIFICATION WITH LEXICON MODELS 4](#_Toc163844245)

[ENHANCING THE RATING VALUES. 5](#_Toc163844246)

[CALCULATED WEIGHTED RATING. 5](#_Toc163844247)

[SUMMARY GENERATION 6](#_Toc163844248)

[USE OF LLM MODEL TO AUTOMATICALLY CREATE A RESPONSE. 8](#_Toc163844249)

[MACHINE LEARNING MODELS VS ANALYSIS MODELS 9](#_Toc163844250)

OBJECTIVE**:**

The first phase involves uploading the data, cleaning it up, pre-processing the data to create a textual representation, and finally, building and testing the Lexicon classifier. In the second phase, the team needs to construct the same procedure using a Machine learning approach and compare the results of each approach. Lastly, a study of how to utilize the same review data to construct a recommender system is required.

# Data loading and exploration:

* Data Sampling: Selected 2000 reviews from the original dataset.
* Data exploration:
  + Distribution of rating
  + Counting Unique Reviewers and Products.
  + Visualizations of sentiment, review length and rating distribution.
  + Average rating by sentiments.
  + Review length vs sentiments.

# Text representation:

* Word Embeddings (Word2Vec): Word embeddings capture semantic relationships between words by representing them as dense vectors in a continuous vector space. Word2Vec is a widely used technique for learning word embeddings, known for its ability to capture context and meaning from large text corpora.
* **Actions:**
  + **Tokenization:** Data is tokenized to break down the reviews into individual words.
  + **Word2Vec Model Training:** Using the Gensim library the model is trained on the tokens.

# Text classification using Naïve bayes and SVM.

* **Purpose:** To perform sentiment analysis on text data using two classifiers: Naïve bayes and SVM.
* **Actions:**
  + Data splitting: splitting into train test to evaluate model performance.
  + Pipeline Definition: Used to streamline the workflow by combining TF-IDF and model training.
  + Grid Search Cross- Validation: Used it to find the best hyperparameters for both the models.
  + Model evaluation: Classification reports and accuracy to measure the model performance.

|  |  |  |
| --- | --- | --- |
|  | SVM | NAÏVE BAYES |
| BEST PARAMS | 'clf\_\_C': 1,  'tfidf\_\_max\_df': 0.5,  'tfidf\_\_min\_df': 1, 'tfidf\_\_ngram\_range': (1, 2) | 'clf\_\_alpha': 0.5 |
| BEST SCORE | 0.9806142183027772 | 0.9246228823392899 |
| ACCURACY | 0.9882943143812709 | 0.9431438127090301 |
| REPORT |  |  |

# TEXT CLASSIFICATION WITH LEXICON MODELS

* **Purpose:** Comparison of the two machine learning models on sentiment analysis using the two analysis models. (Using the same data used for the analysis models)
* **Actions:**
  + Data preparation and splitting: splitting the input and target variables into train and test.
  + Naïve bayes model (Textblob and Vader):
    - The best performing Naïve bayes model from previous grid search (trained on the two analysis models) is applied to the test data to make predictions.
  + SVM model (TextBlob and Vader):
    - The best performing SVM model from previous grid search (trained on the two analysis models) is applied to the test data to make predictions.
* **COMPARISON:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Accuracy |
| NB (TEXTBLOB) | 0.932576 | 0.926667 | 0.917558 | 0.926667 |
| NB (VADER) | 0.892512 | 0.876667 | 0.843460 | 0.876667 |
| SVM(TEXTBLOB) | 0.964873 | 0.963333 | 0.961247 | 0.963333 |
| SVM(VADER) | 0.956956 | 0.956667 | 0.954008 | 0.956667 |

# ENHANCING THE RATING VALUES.

In the article “Recommender systems based on user reviews: the state of the art” it suggests many ways to enhance rating values using the data. Here are some of the methods:

1. Review Helpfulness: The reviews are based on how important the readers find them and they are measured by the number of votes they receive. More helpful reviews are seen as more important and trustworthy.
2. Weighted Rating with Review Topics: By comparing these topics with what the user is interested in, ratings can be weighted accordingly.
3. Incorporating Review Contexts: The context in which reviews are written (for example: when, where, and whom) can be important in predicting how useful an item will be to the user.
4. Utilizing Review Emotions: Emotions expressed like happiness or anger, can give insights into how much a user might like an item.

# CALCULATED WEIGHTED RATING.

* **Purpose**: To provide a more personalized assessment of reviews for a user by considering the topics tht are relevant to their interests. By assigning weights to the reviews based on the intersection of topic and user’s topic profile, the weighted ratings highlight reviews that align closely with user’s preference.
* **Actions:**
  + Topic Extraction: Identifying topics allows for assessment of review content in relation to the user’s interests.
  + Weight Calculation: Calculate weight for each review based on the intersection of review topics and user’s topic profile.
  + Weighted rating calculations: Integrates the user’s topic preference with the review ratings, producing personalized weighted rating.
* **Results**:

A screenshot of a computer

Description automatically generated

A grey background with white text

Description automatically generated

# SUMMARY GENERATION

* **Purpose:** To generate summaries for a set of reviews using pre-trained sequence-to-sequence language model(T5).
* **Actions:**
  + Loading pre-trained model and tokenizer
  + Data Preprocessing: Drop rows with missing review text and select a subset with a minimum of 100 words. This is to ensure that the input data is clean.
  + Summary generation: Iterate through each review, generates a summary, and stores the original review text, summary.
* **Results**:

|  |  |
| --- | --- |
| Initial Review | Generated Summary |
| i bought a pair from DSW for 50$ and they are very comfortable but i bought a size 6.5 when I normally where a size 7. and even the 6.5 feels a little roomy. I have wide flat feet btw. im not sure if I will keep them even though i love their look and comfort. I've been having feet problems while taking a gym class and I think they won't give me enough support when running. they're great for wearing at home or gym and doing exercises in one place- extremely light weight and flexible, but i don't know if they'll last long if you use them everywhere else. :( i really like them but i can’t just buy 2 different shoes.  also, i really like the strings inside the lining on the side of the shoes to adjust and make the shoe fit snug. it makes me feel secure :D | i bought a pair from dSW for 50$ and they are very comfortable. they are great for wearing at home or gym and doing exercises in one place. i really like them but i cant just buy |
| Super light-weight, decent arch support (mine is higher than average). I was pleasantly surprised by the sole/tread on these for my circuit training classes, as I couldn't really tell by online pics. Also pleasantly surprised that these ran ever so slightly bigger than Nike running shoes I use to wear 15+years ago - I always wore one to 1.5 sizes larger in their "Hurache's"/running shoes, so I was really worried my size 6 would be too small - there is actually more toe room in these shoes than any other size 6 shoe I own (athletic, heel, boot). But it is the correct amount of room from big toe to edge of shoe for me (an inch or less, I don't want my big toe being jammed into edge of shoe when lunging, jumping, etc.). Shoes arrived with an odd yellowish color on side of one shoe, but no damage to box, paper lining etc - I was happy that a clorox wipe easily cleaned the shoe and did not discolor any further. I am very happy and highly recommend, although I have only worked out it once! | shoes arrived with an odd yellowish color on side of one shoe, but no damage to box, paper lining etc. shoes arrived with an odd yellowish color on side of one shoe. |

A screenshot of a black and white table

Description automatically generated

# USE OF LLM MODEL TO AUTOMATICALLY CREATE A RESPONSE.

* **Purpose:** Generate a customer service response to a review that contains a question. By using a pre-trained model (GPT-2), the code aims to provide automated responses.
* **Actions:**
  + Loading pre-trained model and tokenizer.
  + Identifying review with question nature: Focuses on reviews where customers have expressed inquiries or doubts.
  + Response generation: Enables the model to generate contextually appropriate responses for the specific customer question.
* **Results**:

|  |  |
| --- | --- |
| Question | Generated Answer |
| It's hard to find a lightweight athletic shoe with a wide toe box. This shoe has that. I am happy with the support for walking and everyday wear. I wouldn't run in them. Now I just have to stock up on as many pair as I can afford before Nike discontinues them :). Don't you hate when that happens? | I'm sorry, but I can't find a shoe that is as good as this. I'm sorry, but I can't |

# MACHINE LEARNING MODELS VS ANALYSIS MODELS

|  |  |  |
| --- | --- | --- |
|  | ANALYSIS MODELS | Machine Learning Models |
| TYPE OF MODEL | Rule-based | Data-driven |
| LEARNING APPROACH | Based on predefined rules | Learn patterns from data |
| COMPLEXITY | Simpler, often interpretable | Can be highly complex |
| FEATURE ENGINEERING | Relies on handcrafted features | Can learn feature representation. |
| TRAINING DATA SIZE |  | Requires large datasets |
| SCALABILITY | Limited scalability | Scalable to large datasets |
| EASE OF IMPLEMENTATION | Relatively easy to implement | Can be complex to implement |
| PERFORMANCE | May perform well for simple tasks | Can achieve high accuracy on complex tasks |