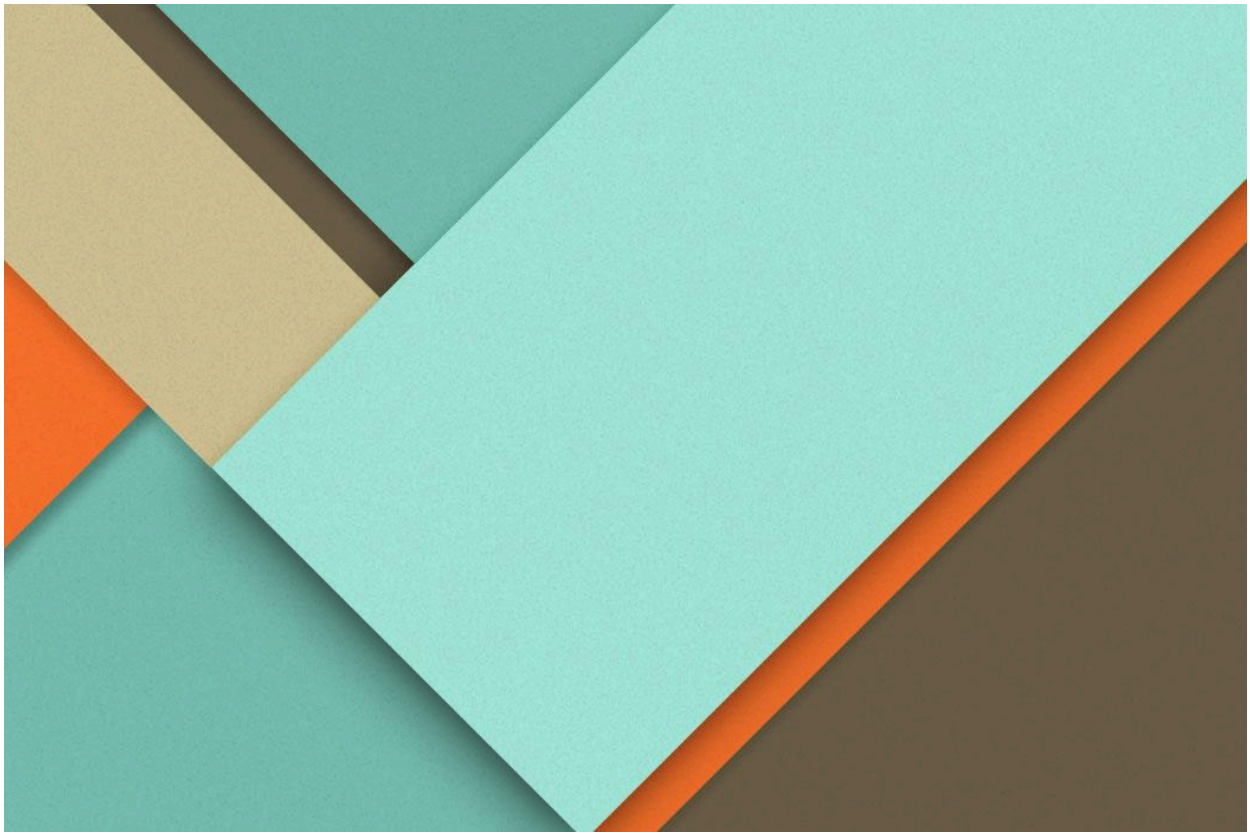


Aspect-Based Sentiment Analysis

12.03.2024

—

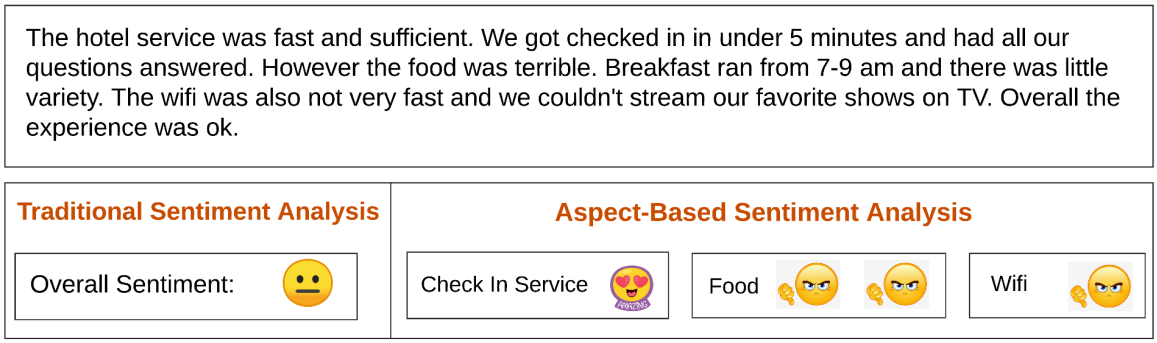
Team: Sifael Sebastian Ndandala, Piumi Abeyrathne, Protogen Hahirwabayo

Topic: Natural Language Processing: Aspect-Based Sentiment Analysis

Github: https://github.com/DataMining-CU-Spring2024/Final-Project

Overview and Description

Our project proposal for the Data Mining course focuses on the implementation of Aspect-Based Sentiment Analysis (ABSA) using the popular Amazon products review datasets. Aspect-Based Sentiment Analysis (ABSA) is an advanced technique in natural language processing (NLP) that goes beyond determining the overall sentiment of a text. Instead, it identifies sentiments related to specific aspects of a product or service mentioned in the text. The example below demonstrates the difference between **Traditional Sentiment Analysis** and **Aspect-Based Sentiment Analysis** using a sample Hotel Review.



This project aims to take traditional sentiment analysis a step further, to dissect customer reviews into specific sentiment details thereby providing granular and actionable insights into what aspects customers liked or disliked about the product and service.

Purpose of the Project

We decided to focus on Aspect-Based Sentiment Analysis (ABSA) for several compelling reasons that align with our educational and professional goals. Firstly, the project offers us an opportunity to apply and expand upon the knowledge and skill we've acquired from our Data Analytics course. The course laid a solid foundation in various analytical methodologies, including classification, clustering, and deep learning. By undertaking this ABSA project, we aim to leverage these insights and techniques to not only analyze data more effectively but also to devise predictive algorithms/models capable of handling new, unseen datasets.

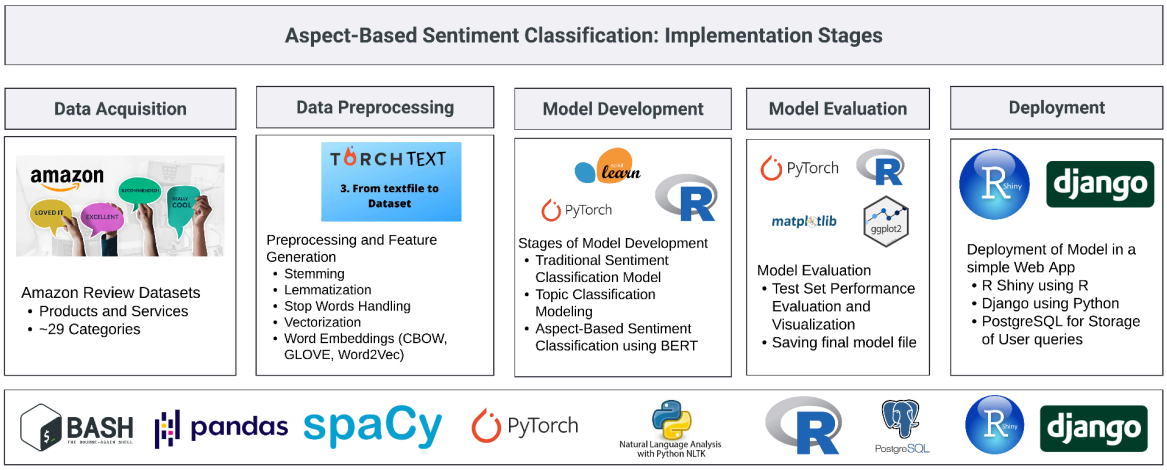
Secondly, our choice is driven by an interest in delving into the field of Natural Language Processing (NLP), a domain that holds significant relevance in both industry and academic

research. Through this project, we aspire to familiarize ourselves with the core concepts underpinning some of the most sophisticated NLP solutions, such as ChatGPT. Furthermore, through this project, we expect to navigate the complexities of NLP and understand its inherent challenges which will prepare for more advanced studies such as Text Generation.

Finally, the ABSA project serves as a platform for us to explore and master various technologies and software tools, thereby enhancing our proficiency in programming, data processing, and machine learning. We expect to explore toolkits such as **NLTK** and **Spacy** for NLP processing, **Pytorch** for deep learning and even leverage pre-trained models such as **pre-trained Word Embeddings** and **BERT Models**. Furthermore, we will explore tools to develop user applications such as **Django** and **R-shiny** to package the final product into a user-friendly live application. By integrating these elements, our project not only consolidates our current understanding but also prepares us for more complex challenges ahead, ensuring we develop comprehensive competencies in these critical areas.

Project Implementation Stages and Technology Stack

The implementation process for the Aspect-Based Sentiment Analysis (ABSA) project involves several key stages: Data Collection, Data Preprocessing, Model Development (Several Stages), Evaluation and Testing, and Deployment. This section outlines the technology stack, models, algorithms, and steps necessary to achieve the project's objectives.



Data Collection

The project's initial phase involves collecting and preprocessing Amazon product reviews, accessible from a well-known repository. This extensive dataset encompasses a wide range of product categories, from books and arts to beauty and fashion, offering a rich diversity for analysis. The variety in categories challenges us to develop specialized techniques that go beyond basic sentiment classification, aiming to extract specific insights related to various aspects mentioned in the reviews. Implementing this step, we will employ scripts to automate the download and initial preprocessing of these datasets, ensuring a broad and representative sample for our analysis.

Preprocessing and Feature Generation

Following data collection, our focus shifts to NLP data preprocessing, an essential step to prepare the text for analysis. This process involves cleaning the text, tokenization, removing Stopwords, and lemmatization, employing NLP libraries such as NLTK and spaCy. These techniques are critical for refining the data, enabling the effective extraction of meaningful insights and ensuring our models can accurately interpret the nuances of language used in customer reviews. Our implementation strategy includes creating a preprocessing pipeline that standardizes the text data, making it suitable for detailed sentiment analysis.

Model Development

For model development, we plan to leverage advanced NLP models and machine learning frameworks, specifically focusing on fine-tuning pre-trained models like BERT for aspect-based sentiment analysis. This approach allows us to build upon the sophisticated understanding these models have of language, adapting them to our specific task of identifying and analyzing sentiments related to different aspects of products. We will implement this phase by selecting appropriate model architectures, adjusting parameters, and integrating our preprocessed data to train models capable of capturing the complexities of aspect-based sentiments.

Model Evaluation and Tuning

Once our models are developed, we will rigorously evaluate and tune them to ensure optimal performance. This includes using standard metrics such as accuracy, precision, recall, and F1-score, alongside more specific evaluations tailored to the nuances of aspect-based sentiment analysis. Our tuning process will involve iteratively adjusting model parameters, incorporating feedback from the evaluation phase, and potentially exploring ensemble methods to enhance accuracy and robustness. Implementing this involves setting up a systematic

evaluation framework that can accurately measure the model's effectiveness in real-world scenarios.

Deployment

The final stage involves deploying our model into a user-friendly application, making our ABSA tool accessible to businesses and researchers. This entails integrating the model with a web framework, such as Django or R-Shiny, to create an interactive interface where users can input text and receive detailed sentiment analysis. Additionally, we will ensure the deployment infrastructure is scalable and secure, allowing for efficient processing of large volumes of data. The implementation of this phase will focus on building a robust platform that showcases the practical utility of our ABSA model, demonstrating its value in extracting nuanced insights from textual data.

Learning Objectives

Our learning outcomes are closely aligned with the reasons for selecting Natural Language Processing, as noted in the Purpose of the Project section. Here, we dive into the specific outcomes of what we expect to achieve as Learning outcomes.

Deepening Understanding of NLP

A significant motivation behind our project choice is the shared keen interest in delving deeper into the realm of Natural Language Processing (NLP). This field's relevance in both academic research and the industrial sector cannot be overstated, and through this project, we anticipate gaining understanding of the core concepts that drive some of the most sophisticated NLP solutions, such as ChatGPT. Through the project, we hope to learn about the complexities and challenges in NLP and develop a solid foundation for more advanced applications such as Text Generation.

Developing Technical Competencies in Technologies and Software

This project presents us with an opportunity to explore and master a variety of technologies and software tools, significantly enhancing our competencies in programming, data processing, and machine learning. Our exploration will include utilizing NLP processing toolkits such as **NLTK** and **Spacy**, deep learning frameworks like **Pytorch** and **Tensorflow**, and leveraging powerful pre-trained models, including **Word Embeddings** and **BERT** for aspect based sentiment Analysis. Furthermore, we aim to harness tools like **Django** and **R-shiny** to develop user applications, packaging our findings into user-friendly live applications. This integrative

approach not only consolidates our existing knowledge but also prepares us to tackle more complex challenges beyond the scope of this course.

Data Preprocessing and Analysis Proficiency of Textual Data

A foundational element of our project involves developing a robust proficiency in data preprocessing and analysis techniques critical for NLP tasks. This includes mastering tokenization, lemmatization, and dependency parsing among others, to effectively prepare and refine data for analysis and model training. Such skills are indispensable for ensuring the accuracy and effectiveness of our NLP models, laying the groundwork for successful sentiment analysis and predictive modeling.

Mastering Aspect-Based Sentiment Analysis Techniques

Central to our project is the objective to master Aspect-Based Sentiment Analysis techniques, which will enable us to intricately dissect and interpret sentiments related to specific aspects within textual data. This skill is particularly valuable for extracting nuanced insights from varied sources such as customer reviews or feedback, providing a deeper understanding of public sentiment towards products, services, or topics. By achieving proficiency in ABSA, we aim to enhance our analytical precision and contribute meaningful insights to the field.

Integration of NLP into Real-World Applications

A key ambition of ours is to seamlessly integrate our NLP models into practical, real-world applications. Our goal is to create a Web App that deploys the model we have developed into the real-world. This endeavor challenges us to think beyond the technical aspects of model development and consider the practical implications and utility of our work, bridging the gap between theoretical NLP solutions and their real-world applications.

Collaboration and Version Control through Git

Finally, an essential component of our group project involves enhancing our technical collaboration skills through the use of version control systems like **Git** and **Github**. This tool is pivotal for efficient teamwork on coding projects, enabling us to maintain a well-organized and collaborative codebase. Mastering version control is crucial for our project's success, facilitating seamless integration of our collective efforts and ensuring the robust development of our ABSA project.

References:

ABSA Research Paper and assets:
https://paperswithcode.com/task/aspect-based-sentiment-analysis

Dataset: https://nijianmo.github.io/amazon/index.html