IR Project Proposal
E-Commerce Competitive Analysis System
Group 22

Team Members: Akshat Saini (2020019), Saksham Bhupal (2020573), Aditya Nangia (2020168), Nakul Thureja (2020528), Jayan Pahuja (2020071)

Problem Statement:

Due to the highly competitive landscape of e-commerce marketplaces with numerous sellers offering similar products, sellers struggle to gather and analyse competitor data and customer feedback manually. This inefficiency hinders their ability to optimise product listings and gain a competitive edge.
Specifically, sellers face challenges in decision-making, particularly regarding identifying optimal pricing strategies. Without insights into competitor pricing trends and customer price sensitivity, it's difficult to determine the right price to maximise sales and profit margins. Sellers are unsure of what pricing strategy would work best and when or how much to discount products to attract customers without sacrificing profitability. An automated system that involves web scraping to automate data collection and unique feature extraction to perform analysis using deep learning models and various IR techniques such as TF-IDF can help streamline data collection, analysis, and interpretation across various platforms.
Such a system would empower sellers with actionable insights to optimise product listings, address customer needs, and ultimately enhance their competitiveness in the e-commerce market.

Motivation

With the advances in the field of Information Retrieval, there is a very high focus on consumer-oriented solutions; however, there is very little work being done for the producers/sellers. We aim to provide a solution for sellers working on platforms like Amazon, Flipkart, etc. Often, sellers list their products on multiple platforms, which makes manual analysis impractical and prone to errors. Moreover, it is the sheer volume of data available on e-commerce platforms which motivates us to work on the problem even further.
Understanding market dynamics, analysing competitor products, and identifying effective pricing and marketing strategies are essential for success in the highly competitive online landscape. Additionally, interpreting customer sentiments and preferences from product reviews poses another challenge for sellers seeking to enhance their product offerings. There is a need for an Information Retrieval based-system that can gather, analyse, and provide actionable insights from competitor data and customer feedback to help sellers make informed decisions and improve their competitiveness.

Literature Review

1. Increased information retrieval capabilities on e-commerce websites using scraping techniques [1].
   This study employs web crawling across three e-commerce websites, consolidating data into a database for streamlined retrieval. By scraping HTML tags and storing data systematically, the process ensures efficiency with a 100% recall rate and 93.9% precision rate, enabling rapid and accurate information retrieval. These techniques can be helpful and be further worked upon for our use case.
2. A Review on Web Scrapping and its Applications [2]
   This review dives into web scraping, exploring its applications and the technology behind it. It examines the reasons why web scraping is valuable, discusses its advantages and limitations,

and explores the tools and libraries commonly used for scraping tasks. Additionally, the review highlights the various applications of web scraping across different fields.

3. Opinion mining and sentiment analysis on online customer review [3].
   Opinion mining, crucial in e-commerce, sees rising importance with the surge in online transactions and user-generated content. Reviews on platforms like Amazon express customer sentiments, offering valuable insights. This study focuses on mining Amazon reviews, utilizing algorithms like Naïve Bayes, Logistic Regression, and SentiWordNet for sentiment analysis. The goal is to automate sentiment recognition and enhance understanding of user emotions.
4. Sentiment analysis: A literature review [4]
   This paper offers a survey of recent advancements in sentiment analysis, a technique that extracts emotional tones from text data. Due to the surge of subjective content online, particularly in reviews, sentiment analysis has become a hot research area. The paper delves into the core methods used in sentiment analysis research, including framework and lexicon development, feature extraction, and polarity classification (positive, negative, or neutral). It highlights the current methodologies, explores existing limitations, and provides an in-depth look at applications in business and blog analysis. Finally, the paper discusses potential future directions for sentiment analysis research.
5. Search engine optimization (SEO) for websites. [5]
   As the internet expands, search engines play a critical role in indexing and presenting relevant web content. However, many websites overlook the need for visibility, focusing solely on user experience and technical aspects. To address this, a web application is developed to analyze web pages and enhance their search engine friendliness. By providing actionable insights and recommendations, this application aims to improve website rankings and attract more visitors through Search Engine Optimization (SEO).
6. Classification of Customer Reviews based on Sentiment Analysis [6]
   The paper proposes a system that performs the classification of customer reviews of hotels by means of a sentiment analysis. They extract a domain-specific lexicon of semantically relevant words based on a given corpus, which backs the sentiment analysis for generating a classification of the reviews. The evaluation of the classification on test data shows that the proposed system performs better compared to a predefined baseline.
7. On Application of Learning to Rank for E-Commerce Search[7]
   E-commerce search is a burgeoning application of information retrieval, with Learning to Rank (LETOR) emerging as a pivotal strategy. While LETOR is extensively studied for web searches, its application to e-commerce searches remains unexplored. This paper addresses practical challenges in implementing LETOR for e-commerce search, including feature representation, obtaining reliable relevance judgments, and leveraging multiple user feedback signals. Experiments on industry datasets reveal insights: popularity-based features enhance relevance-based ones, reducing query attribute sparsity is beneficial, and order rate proves the most robust training objective, followed by click rate, while add-to-cart ratio is less reliable.

Goals and Objectives:

- Utilise IR techniques to provide sellers with insights into competitor products and market trends.
- Optimise product listings by analyzing pricing, reviews, descriptions, and other relevant data points.
- Enhance seller competitiveness by identifying opportunities for improvement.
- Improve decision-making processes based on data-driven insights.

Baseline Solution/Results:

The following step-by-step solution serves as our baseline model. We have been able to implement HTML scraping to collect competitor data automatically. This is followed by feature extraction and identifying our seller's product as well as the top three competitors. Finally we have also implemented a basic sentiment analysis on the seller's reviews.

1. **Web Scraping with Ethical Considerations:**
    a. Technology: Leverages libraries like Beautiful Soup to extract data from product pages on e-commerce websites.
    b. Challenge Addressed: Automates data collection, eliminating manual work and ensuring access to the latest competitor information.
    c. Ethical Considerations:
        i. Respect robots.txt files that instruct web crawlers on restricted areas.
        ii. Adhere to website terms of service to avoid copyright infringement.
        iii. Be mindful of scraping frequency to avoid overloading website servers.
2. **Feature Extraction for Detailed Competitor Comparison:**
    a. Technology:
        i. Image Features: Uses a pre-trained deep learning model like ResNet to extract visual features from competitor product images (e.g., color, shape, texture).
        ii. Text Features: Employs TF-IDF (Term Frequency-Inverse Document Frequency) to analyse product descriptions and identify keywords that highlight product features and benefits.
    b. Challenge Addressed: Enables a more nuanced comparison than just product titles. This allows sellers to identify similar products even with slight variations in names or descriptions.
3. **Identifying Key Competitors Through Feature Matching:**
    a. Technology: Utilizes a similarity metric (e.g., cosine similarity) based on the extracted image and text features from the seller's product and competitor's products.
    b. Challenge Addressed: Provides a targeted set of the most relevant competitors for analysis rather than an overwhelming list from the entire marketplace. This allows sellers to focus their energy on the most impactful competition.
4. **Sentiment Analysis of Customer Reviews:**
    a. Technology: Employs Natural Language Processing (NLP) tools such as VADER Sentiment Analysis, NLTK, and Beautiful Soup to analyse customer reviews for the seller's product and identify positive, negative, or neutral sentiment.
    b. Challenge Addressed: Provides sellers with insights into customer satisfaction and helps them understand areas for improvement in their own product or listing.

[Update] Challenge Addressed: Limited data scope due to scraping from a single e-commerce website.

Improvement Implemented:

- **Multi-Website Scraping:** The system now gathers data from multiple e-commerce websites, expanding the competitive landscape and potentially providing a more comprehensive understanding of the market.
- **Data Consolidation:** The scraped data from various websites is combined and processed to ensure consistency and facilitate comparative analysis. This allows you to compare products and sentiment even if the websites present information differently.

Benefits:

- **Richer competitor insights:** By scraping data from multiple websites, you can identify a wider range of competitors, even those operating on different platforms.
- **More comprehensive market analysis:** This broader data set allows you to gain a more complete picture of customer sentiment and product offerings within the market.
- **Website-Specific Insights:** We can identify trends and patterns within customer reviews and competitor data on each platform. This provides a deeper understanding of customer preferences and buying behaviors specific to each website's audience.
- **Tailored Product Listing Recommendations:** By analyzing website-specific data, we can generate targeted recommendations for improving your product listings on each platform. This could involve optimizing titles, descriptions, or visuals to better resonate with the audience of each website.
- **Maximized Sales Across Channels:** With these tailored recommendations, your customer can adjust their product listings to cater to the specific needs and preferences of each e-commerce platform's audience. This has the potential to significantly improve conversion rates and maximize sales across all sales channels.

Work Remaining:

Consolidated Competitor Analysis: We've implemented website-specific competitor analysis, but there's room for improvement. We can develop a method to consolidate competitor data across all websites. This could involve:

- **Feature Normalization:** Ensure features extracted from different websites (e.g., product descriptions) are on a comparable scale before combining them.
- **Weighted Scoring:** Assign weights to data from different websites based on factors like website popularity or relevance to our customer's target audience.
- **Combined Similarity Scores:** Develop a method to combine similarity scores obtained through website-specific analysis to create a single, consolidated competitor strength score.
- **State-of-the-Art Models:** Upgrading our existing models is a great step forward. We're exploring advancements in areas like:
    - **Image Feature Extraction:** Utilize cutting-edge deep learning models like CLIP (Contrastive Language-Image Pre-training) for richer image understanding beyond color, shape, and texture.
    - **Text Feature Extraction:** Explore advancements in natural language processing (NLP) like BERT (Bidirectional Encoder Representations from Transformers) for a more nuanced understanding of product descriptions and competitor offerings.
    - **Sentiment Analysis:** Look into advancements in sentiment analysis tools like FLAIR (Fasteners for Language Representation) that can provide more granular sentiment analysis beyond just positive, negative, and neutral.

AI Chatbot for Review Analysis: Developing an AI chatbot to answer customer questions based on reviews is an ambitious but valuable addition. Here are some key aspects to consider:

- **Question Classification:** Train the chatbot to categorize user questions into relevant topics (e.g., product features, competitor comparisons, customer satisfaction).

- **Answer Retrieval:** Implement a method for the chatbot to retrieve relevant information from processed reviews based on the classified question.
- **Answer Generation:** Train the chatbot to generate clear, concise, and informative answers to user questions using the retrieved information.

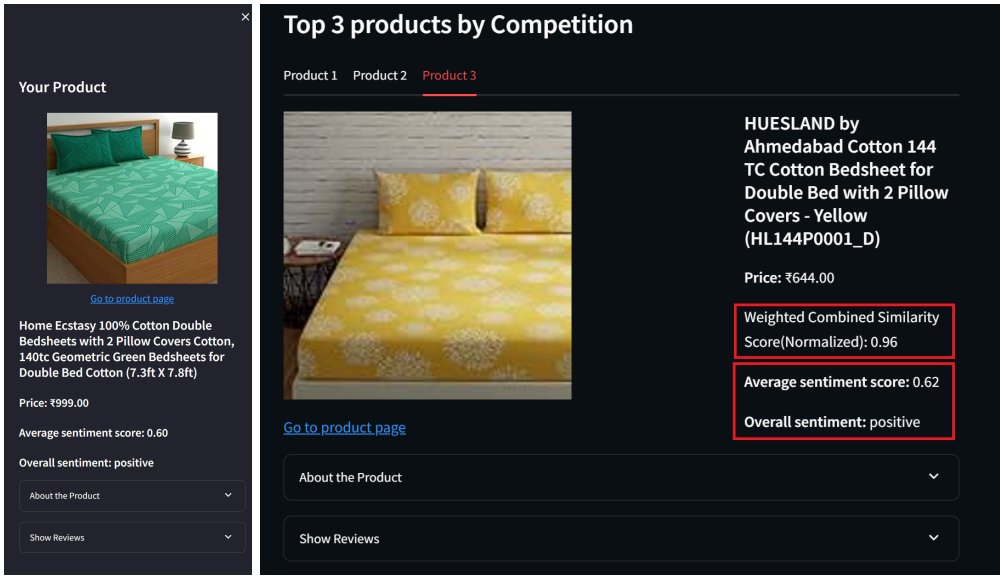
Lofty Aim: Listing Strength Prediction

The idea of a "listing strength predictor" similar to password strength meters is an innovative goal. Here are some key aspects to consider:

- **Data Integration:** Combine the insights from website scraping, competitor analysis, sentiment analysis, and potentially other relevant data sources.
- **Feature Engineering:** Develop features that comprehensively represent the strength of a listing. This could include factors like keyword usage, title clarity, image quality, competitor comparison, positive sentiment in reviews, etc.
- **Machine Learning Model:** Train a machine learning model (e.g., Random Forest) on historical data to predict listing strength based on the engineered features. The model will learn to associate specific features with successful listings and use that knowledge to predict the strength of new listings.
- **Visual Representation:** Implement a slider or visual indicator similar to password strength meters to represent the predicted listing strength (e.g., weak, okay, good, excellent).

Evaluation

We evaluated our baseline system's performance by comparing it to human manual data collection. Our approach achieved high accuracy in sentiment scoring, with results closely matching those obtained through human review of written reviews. Furthermore, the similarity scores were checked using image and text features. We intend to improve further upon the evaluation techniques and employ more ML techniques for better results.



References

[1] Kurniawati, D., & Triawan, D. (2017, November). Increased information retrieval capabilities on e-commerce websites using scraping techniques. In *2017 International Conference on Sustainable Information Engineering and Technology (SIET)* (pp. 226-229). IEEE https://ieeexplore.ieee.org/abstract/document/8304139

[2] V. Singrodia, A. Mitra and S. Paul, "A Review on Web Scrapping and its Applications," 2019 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2019, pp. 1-6, doi: 10.1109/ICCCI.2019.8821809. keywords: {Robots;Libraries;Tools;Data mining;Web pages;Java;Browsers;Web Scrapping;Internet;Big Data;Business Intelligence.}, https://ieeexplore.ieee.org/abstract/document/8821809

[3] Kumar, K. S., Desai, J., & Majumdar, J. (2016, December). Opinion mining and sentiment analysis on online customer review. In *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)* (pp. 1-4). IEEE. https://ieeexplore.ieee.org/abstract/document/7919584

[4] Z. Nanli, Z. Ping, L. Weiguo and C. Meng, "Sentiment analysis: A literature review," 2012 International Symposium on Management of Technology (ISMOT), Hangzhou, China, 2012, pp. 572-576, doi: 10.1109/ISMOT.2012.6679538. keywords: {Feature extraction;Blogs;Semantics;Computers;Algorithm design and analysis;Data mining;Educational institutions;Sentiment analysis;Information extraction}, https://ieeexplore.ieee.org/abstract/document/6679538

[5] Barbar, A., & Ismail, A. (2019, April). Search engine optimization (SEO) for websites. In *Proceedings of the 2019 5th international conference on computer and technology applications* (pp. 51-55). https://dl.acm.org/doi/abs/10.1145/3323933.3324072

[5] Gräbner, D., Zanker, M., Fliedl, G., & Fuchs, M. (2012). Classification of customer reviews based on sentiment analysis. In *Information and communication technologies in tourism 2012* (pp. 460-470). Springer, Vienna. https://link.springer.com/chapter/10.1007/978-3-7091-1142-0_40

[6] Karmaker Santu, S. K., Sondhi, P., & Zhai, C. (2017, August). On application of learning to rank for e-commerce search. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval* (pp. 475-484). https://dl.acm.org/doi/abs/10.1145/3077136.3080838