Mid Term BTP Evaluation

Data Analysis using Queryable Knowledge Graphs

SEMESTER 7 , 2023

Netaji Subhas University of Technology

New Delhi-1100784

Supervisor:- Dr. Preeti Kaur

Aibhinav Upadhayay-2020UCO1535

Rohit Lahori-2020UCO1530

Mahika Kushwaha-2020UCO1659

INDEX

Introduction

In today's fiercely competitive technology landscape, the success of any company hinges not only on innovative products and services but also on its ability to understand and cater to the ever-evolving preferences and sentiments of its customers. In light of this, the motivation behind this research project is to harness the formidable potential of social media data, sentiment analysis, and knowledge graph technology. By doing so, we aim to gain deeper insights into customer sentiments, identify user communities, and utilize this wealth of information for strategic decision-making that is poised to enhance the performance of a specific product within the market.

This project embarks on a journey to position our company as a paragon of customer-centricity and data-driven excellence.
In the pages that follow, we will delve deeper into the methodologies and strategies that underpin our mission. We will explore the technical aspects of sentiment analysis and knowledge graph construction while elucidating their practical applications within the context of a customer-centric and data-driven organization. Through this research, we aim to equip our company with the tools and insights needed to thrive in a rapidly evolving marketplace, making decisions that resonate with our customers and drive success in the technology industry.

Motivation

The motivation behind this project is to use data, sentiment analysis, and knowledge graph technology to gain deeper insights into customer sentiments, identify user communities, and use this information for strategic decision-making aimed at improving a specific product's performance in the market. The project aims to position itself as a customer-centric and data-driven company in the competitive technology industry.This allows

A. Enhancing Customer-Centricity:
By analyzing sentiment, we aim to put the customer at the center of its decision-making processes, ensuring that product development and service improvements align with customer preferences and needs.

B. Interdisciplinary Insights:
The generated knowledge graph will be integrated with other tools and datasets, enabling interdisciplinary analysis to uncover deeper insights and patterns in customer behavior and market trends.

C. Market and Consumer Analysis:
The project's primary goal is to harness the knowledge graph for market and consumer analysis, enabling organizations to make data-driven decisions regarding product offerings, marketing strategies, and customer engagement.

D. Predictive Analytics:
By identifying communities and analyzing data,we aim to develop predictive models to anticipate future trends, product demand, and potential issues.

E. Improving Company Performance:
This project seeks to enhance the overall performance as a company by staying attuned to customer feedback and market dynamics, thereby ensuring competitiveness and sustained customer satisfaction.

Literature Survey

1. **(R) Social media data extraction based on knowledge graph and LDA models**

   Authors: Ma You, Yue Kun
   Main Contribution: This paper presents a method for social media data extraction and topic analysis using the Latent Dirichlet Allocation (LDA) model. The LDA model is employed to mine hidden topics in social media data and extract features.
   Dataset Used: Data extracted from "weibo.com."
   Strengths:
- Effective use of the LDA model for topic analysis.
- Provides a foundation for feature extraction from social media data.

   Limitations:
- Limited dataset variety, lacking multi-platform and multi-dimensional data.
- Challenges in accurately dividing topics using unsupervised learning.
- Incompleteness of domain knowledge for knowledge graph construction.

2. **(R) Keyphrase Extraction Using Knowledge Graphs**

   Authors: Wei Shi, Weiguo Zheng
   Main Contribution: This paper focuses on improving keyphrase extraction by integrating semantic document key term graphs from knowledge graphs with traditional phrase features.
   Dataset Used: Contains 308 news articles.
   Strengths:
- Enhances keyphrase extraction performance.

   Limitations:
- Cannot return phrases when the number of keyphrases exceeds a certain threshold.

3. **Sentiment Analysis in Twitter Based on Knowledge Graph and Deep Learning Classification**

   Author: Fernando Andres
   Main Contribution: This paper proposes a sentiment analysis method using knowledge graphs and deep learning. It employs graph similarity metrics and LSTM/Bi-LSTM neural networks for sentiment prediction.
   Dataset Used: Twitter tweets or micro-blogging texts.
   Strengths:

- Knowledge graphs improve sentiment analysis accuracy.
- Interpretability using LIME.

  Limitations:

- Challenges with unrecognized or incorrectly tagged entities.
- Performance on longer texts or different domains not discussed.

4. **Knowledge-Based Sentiment Analysis and Visualization on Social Networks**

   Author: Joulio Kouji
   Main Contribution: Proposes a knowledge-based methodology for sentiment analysis on social networks using knowledge graphs, semantic processing, and graph theory algorithms.
   Datasets Used: Amazon Reviews (11,702 comments) and Twitter Comments (500 comments from CNN).
   Strengths:

- Focus on knowledge graphs and disambiguation for precise sentiment identification.
- Includes a visualization system for better understanding of sentiment and social interactions.

  Limitations:

- Ethical considerations not elaborated.
- Limited adaptability to emerging language and expressions.
- Complex sentiments not considered.

5. **Building and Using Personal Knowledge Graph to Improve Suicidal Ideation Detection on Social Media**

   Authors: Lei Cao, Huijun Zhang, Ling Feng
   Main Contribution: Introduces a framework using personalized knowledge graphs to detect suicidal ideation. Incorporates property and neighbor attention mechanisms.
   Datasets Used: Sina Weibo (users with and without suicidal ideation) and Reddit (categorized suicide risk).
   Strengths:

- Improved accuracy in suicidal ideation detection.
- Property and neighbor attention mechanisms for identifying risk factors.

   Limitations:

- Insufficient exploration of personality-related factors.
- May require more domain-specific expertise for accurate risk identification.

6. **Knowledge Graphs: Opportunities and Challenges**

   Authors: Ciuan Peng, Feng Xia
   Main Contribution: Offers an analysis of opportunities and challenges related to knowledge graphs in various fields.
   Strengths:

- Provides insights into knowledge graph applications in healthcare, education, research, and social networks.

   Limitations:

- Could benefit from specific examples or case studies to illustrate applications.

7. **(A) Multimodal knowledge graph construction of Chinese traditional operas and sentiment and genre recognition**

   Authors: Tao Fan, Hao Wang, Tobias Hodel

Main Contribution: Constructs a multimodal knowledge graph for Chinese traditional operas and introduces sentiment and genre recognition.
Dataset Used: Chinese national-level ICH category Traditional Opera.
Strengths:

- First of its kind knowledge-based database for Chinese Opera.

Limitations:

- Challenges in visualizing various sentiments.

8. **(A) Tracing and analyzing COVID-19 dissemination using knowledge graphs**

   Authors: Gabriel H.A. Medeiros, Lina F. Soualmia, Cecilia Zanni-Merk, Ramiz Hagverdiyev
   Main Contribution: Addresses complex COVID-19 data management using knowledge graphs for analysis and proposes measures for control.
   Dataset Used: OpenSky's Dataset of Corona Virus.
   Strengths:

- Uses graphs and visualizations to analyze virus spread.

Limitations:

- Needs improvement in measuring additional parameters.

9. **(A) Knowledge graph analysis of artificial intelligence application research in nursing field based on visualization technology**

   Authors: Siyu Duan, Yang Zhao
   Main Contribution: Quantitative analysis of AI application research in the nursing field using knowledge graphs and visualization.
   Datasets Used: Multiple databases and CiteSpace for visualization.
   Strengths:

- Provides insights into research trends and collaborations.

Limitations:

- Complex knowledge graph clusters may be challenging to decipher.

Problem Statement

To address the aforementioned challenges, this research project seeks to leverage the power of social media data, sentiment analysis, and knowledge graph technology to create a robust framework for customer-centric decision-making in the technology industry. By doing so, we aim to empower companies with the ability to:

- Understand and adapt to customer sentiments in real-time.
- Identify and engage with user communities effectively.
- Utilize interdisciplinary insights to drive innovation and marketing strategies.
- Conduct comprehensive market and consumer analysis.
- Develop predictive models for anticipating market trends and customer demands.
- Elevate overall company performance by ensuring sustained customer satisfaction and competitiveness.

By addressing the problem of obtaining timely customer insights and integrating data into actionable knowledge, we aim to transform companies into customer-centric, data-driven organizations that thrive in an ever-changing marketplace.

Objectives

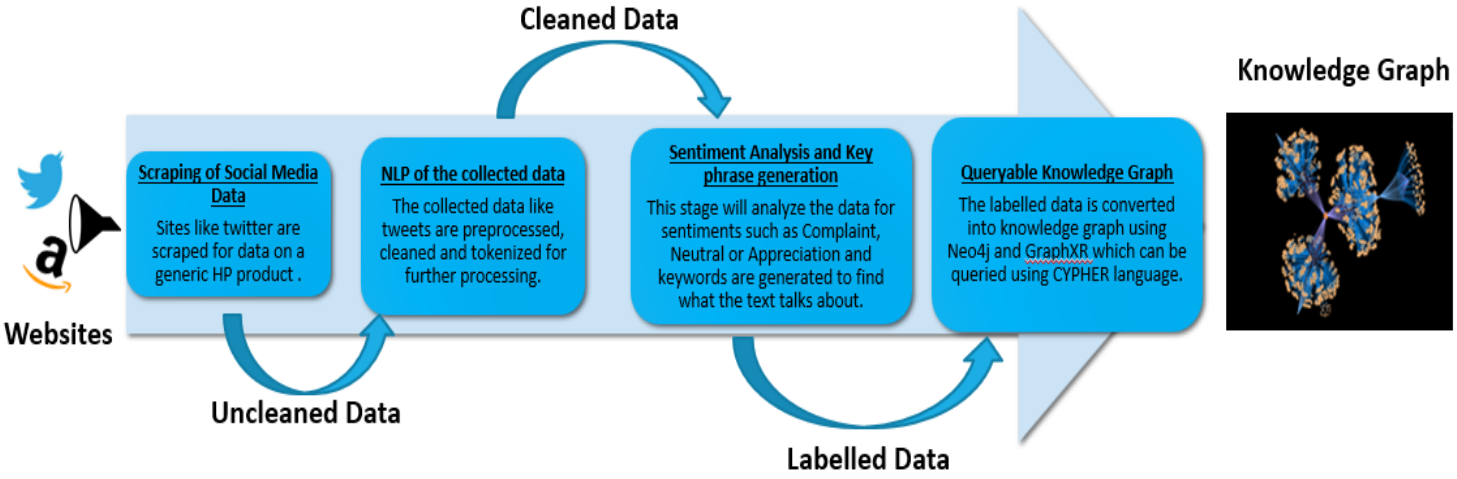
The following problem statement consists of 3 objectives: -

1. Collecting data from various sources (Twitter, Open Datasets, Instagram, Facebook, etc.) on market products like PC and printers.

2. Analyzing the data collected for the products, i.e., segregate the posts into complaints, appreciation, and suggestion after sentiment analysis and understand the content.

3. Store the gained understanding of the data and convert it into a knowledge graph which can be worked upon with other tools.

Methodology



The various major steps can thus be divided into the following segments:

1. Collection of data on the internet by suitable tools for Web scraping by checking hashtags, tags, etc.
2. Conversion of the data into an analysable format such as CSV, Relational Database, etc.
3. Tokenization of the strings and Sentiment analysis of the posts.
4. Segregation of posts into various types based on the content.
5. Conversion of the segregated data into a knowledge graph using rules and Graphs.

Tools Required

- Web API for Data Mining / Web Driver Automation
- Natural Language Processing
- Machine Learning
- Rule Generator
- Graphing Tools

Implementation

1. Data Sourcing/Web Scraping

- The data can be taken from an open source database that contains reviews of products regarding a single product or a multitude of products of one or many companies.
- We can also do so using Web Scraping.For the first step, we import the libraries into the Jupyter notebook and initialize a web driver using selenium. The web driver is then directed towards the intended page of the website (twitter in this case) and the user is asked to login with his/her credentials. This step is necessary as most social media websites have blocked or have paid incoming API requests.
- Once completed, the search bar is passed with the product for which data is needed and the scraping begins. The driver is automatically scrolled for the tweets that are saved to a list using the sleep function and data is then extracted from each tweet. For this we use the XML path of the HTML tags containing the content of the post.
- We are then presented with the Username, Handle, likes, retweets, number of comments, and text of the tweet which can be stored to a .csv file.

2. Sentiment Analysis

- In the next step, the uncleaned data is converted into cleaned data. The NA values are converted to zero using pandas and the text is converted to lexicon and tokens with the stop words being removed.
- Emoji Analyzer, a custom built model which performs sentiment analysis on a given string by assigning a score to every lexicon, is used. Upon its operation we are presented with a label positive, negative, neutral and compound with values ranging from -1 to 1. This way seems fit as we have an unsupervised model to be clustered.

- By rule forming with the scores, we can assign a label to every tweet, and it is saved to the tweets dataframe with the help of pandas. The number of sentiments can be increased and decreased and cross checked on a testing data with hyperparameter tuning and the net accuracy can be increased.

3. Key Phrase Generation

- For generating which product the post talks about and what features it mentions, we can use keyword/keyphrase generation. For this we shall be using yake library to perform both the tasks in one go.
- In this each extracted keyphrase is assigned a score based on how important it is deemed by the unsupervised model. Thus, as long as keyphrases exist in the text they can be extracted. For this model we shall be extracting the top 20 phrases, but this can be increased as deemed necessary by the developer to generate more information from the post by sacrificing on the fact that some keyphrases may not be useful.
- Finally, as we have a list of keyphrases, storing it into the data frame can be tricky, for which we first assign a column of list and then insert the lists of keyphrases generated. Now we have our completely labeled data with their keyphrases and we can move to creation of a knowledge graph.

4. Knowledge graph Conversion

- The dataframe with the data is now converted to a knowledge graph. For this we shall be using neo4j, a free and native graph database. The desktop version is used and is linked to the jupyter notebook using its associated library.
- The nodes are loaded into the database using cypher queries which are also used to create nodes of types – replies and original, and of

the sentiments to which the data has been categorized with their associated relationships.
- The GraphXR plugin is installed and is used in the Neo4j desktop to extract the list of keyphrases associated with the individual nodes into separate nodes and link them to their associated post.
- The built knowledge graph is visualized using GraphXR and the nodes are changed to appropriate color and placement in a 3d space. The complete data of nodes and relationships is stored into .csv files also.

Current Work

So far, we've thoroughly reviewed various research papers, identifying their strengths, the metrics they've employed, and their shortcomings. This process has provided valuable insights into the areas we should focus on for enhancement compared to existing models. It has also guided us in incorporating specific features into our model to address the limitations observed in other models.

To begin, we gather information from the user about the product for which they require Twitter scraping. Following that, when the user provides a product as input, we proceed to conduct web scraping on that specific product utilizing Selenium. We're utilizing web scraping techniques to extract data from the Twitter search bar's latest tab. This process involves using XPath to gather information, including content that requires scrolling to be visible. Once we've collected the tweets, we save them in a CSV file. Subsequently, we perform data cleaning and categorize the information based on sentiment analysis.

Afterward, we evaluated the performance metrics of our sentiment analysis model and conducted a comparison with other existing sentiment analysis models.
The table below shows the different metrics for the the custom model built and other pre existing models :-

Model Used	Precision	Recall	F1-Score
Emoji Analyzer	.8928	.8571	.8745
PolaritySemRel Manual	.6280	.5890	.6350
PolaritySemRel Auto	.6690	.6180	.6220
SSSemRel Manual	.8560	.5790	.7080

Future Work

In the near future we aim on delivering the final knowledge graph that will be queryable and can be easily used to make product based predictions and interpretations. After the sentiment analysis segregation, the next step would be key phrase generation using the yake library. Here, a score will be assigned to phrases based on their importance with the help of an unsupervised model and the top 20 phrases will be extracted and stored in a list called labeled tweets, in a dataframe using pandas.

The dataframe will then be converted into a queryable knowledge graph using Neo4j and will be linked to the jupyter notebook. The graph will be made of Subject-Verb-Object relations and/or RDF triples. The built knowledge graph will be visualized using GrapghXR. Finally, the complete data of nodes and relationships will be stored into .csv files as well.

References

1. https://xblk.ecnu.edu.cn/CN/html/20180516.htm
2. https://link.springer.com/article/10.1007/s41019-017-0055-z
3. https://www.sciencedirect.com/science/article/pii/S2590291122000912
4. https://www.mdpi.com/2079-9292/10/22/2739

5. https://www.researchgate.net/publication/343981459_Knowledge-Based_Sentiment_Analysis_and_Visualization_on_Social_Networks
6. https://www.researchgate.net/publication/337606032_Knowledge_Graph_Construction_for_Intelligent_Analysis_of_Social_Networking_User_Opinion
7. https://arxiv.org/pdf/2012.09123.pdf
8. https://www.researchgate.net/publication/301335561_Sentiment_Analysis_of_Twitter_Data_A_Survey_of_Techniques
9. https://link.springer.com/article/10.1007/s10462-023-10465-9
10. https://ceur-ws.org/Vol-2918/paper6.pdf
11. http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.679.pdf
12. https://www.sciencedirect.com/science/article/pii/S2772662223000838
13. https://www.sciencedirect.com/science/article/abs/pii/S1296207423000584
14. https://www.sciencedirect.com/science/article/pii/S1877050922011632
15. https://www.sciencedirect.com/science/article/pii/S1877050922011632
16. https://www.researchgate.net/publication/343981459_Knowledge-Based_Sentiment_Analysis_and_Visualization_on_Social_Networks