HP Solve 2023

Knowledge Graph using Social Media Posts

Name:SRIKANTH THOLETI

Abstract:

This problem statement focuses on the importance of understanding customer sentiments about products, especially in the case of HP. With consumers expressing their feedback and experiences on social media platforms, it has become essential for HP to have a one-stop knowledge store that can store reviews, suggestions, complaints, and sentiments for all HP consumer printers, PC, and laptops. This knowledge store will help HP to understand their customers better, improve brand value, and Net Promoter Score (NPS). The objective is to enhance user experience and resolve customer issues faster by gaining insights from customer feedback.

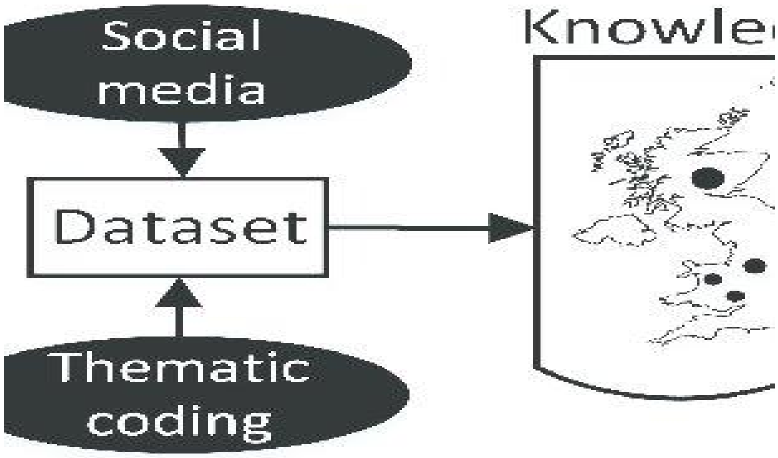

Fig1.1

KNOWLEDGE GRAPH: WHAT AND WHY

A knowledge graph consists of a collated group of interconnected entities with rich attributes (Singhal 2012). For example, attributes associated with an entity such as a social media user's post may include demographic data such as age, nationality, current location in addition to more rich data such as followers, friends and likes. Knowledge is derived by establishing facts from meaningful relations amongst entities which are highlighted in the knowledge graph. The inferred knowledge graph produced from the users' thematic analysis of the sample may include concepts from existing social media analytics frameworks. For example, the Honeycomb framework defines seven components of social media. These components include presence, relationships, reputation, groups, conversations, sharing and identity (Kietzmann et al. 2011). These may be used to define relations within a sample that has been annotated or assigned to various themes by a social scientist. Deriving knowledge from social media in this manner has the advantage of allowing social scientists to apply their own interpretation and analysis through thematic coding and further reflect upon these findings within a wider context such as those specified by the Honeycomb framework (Kietzmann et al. 2011). For example, consider the scenario whereby a social scientist (user) aims to investigate how narratives are developed by a political party through content (Tweets) posted on Twitter. Using our tool, the user collects a dataset which consists of all Tweets posted by a political party. The dataset also includes the number of followers for the political party's Twitter account and for each Tweet in the dataset, the number of retweets, favourites (likes), shares and what hashtags, mentions, links and geolocation tags were included. The user analyses the Tweets through a thematic coding approach. For example, the following tweet: "Politician X has been involved in corrupt and illicit acts!", is assigned the following themes: "accusation; corruption; rumours; opponent" by the user. Upon completion of the analysis, a subsequent knowledge graph produced examines potential relationships between the thematic analysis of the Tweets by the user and associated attributes for each coded Tweet and retweet. The knowledge graph indicates that Tweets coded with the theme Corruption are significantly popular amongst users located within location L. This is inferred by examining the volume of retweets, likes and shares from other locations for Tweets assigned with the code Corruption and comparing these to those assigned with the same theme, retweeted, liked and shared at location L. The user may conclude that location L will become a focal point for the political party and potentially other opponents, given the influence of narratives posted through Twitter and their impact at location L (see Figure 1).

Components,Interactions and Technologies used:

Social Media Data Collector: This component will collect and store social media data from various platforms, such as Twitter, Facebook, Instagram, and so on.
Data Preprocessor: This component will preprocess the collected data, such as cleaning, filtering, and transforming it into a suitable format for storage and analysis.
Sentiment Analysis Engine: This component will analyze the preprocessed data using Natural Language Processing (NLP) techniques to extract customer sentiments, such as positive, negative, or neutral, towards HP products.
Knowledge Store: This component will store the analyzed data, including reviews, suggestions, complaints, and sentiments for all HP consumer printers, PC, and laptops.
Analytics and Reporting: This component will generate insights from the stored data and generate reports to help HP understand customers better and improve brand value and NPS score.
Interactions:

Social Media Data Collector will interact with various social media platforms to collect customer data.
Data Preprocessor will interact with the Social Media Data Collector to preprocess the collected data.
Sentiment Analysis Engine will interact with the Data Preprocessor to analyze the preprocessed data and extract customer sentiments.
Knowledge Store will interact with the Sentiment Analysis Engine to store the analyzed data.
Analytics and Reporting will interact with the Knowledge Store to generate insights and reports.
Technologies used:

Social Media APIs: Used to collect data from various social media platforms.
Python: Used for data preprocessing, sentiment analysis, and analytics.
Natural Language Processing (NLP) Libraries: Used to extract sentiments from customer data.
Database Management Systems: Used to store and manage the data in the Knowledge Store component.
Reporting and Analytics Tools: Used to generate insights and reports.

Diagram:

```
+-------------------+
|Social Media       |
|Data Collector     |
+-------------------+
          |
          |
          v
+-------------------+
|Data Preprocessor  |
+-------------------+
          |
          |
          v
+-------------------+
|Sentiment Analysis |
|Engine             |
+-------------------+
          |
          |
          v
+-------------------+
|Knowledge Store    |
+-------------------+
          |
          |
          v
+-------------------+
|Analytics and      |
|Reporting          |
+-------------------+
```

Data Collection and PreProcessing

Data Collection:
The first step is to collect customer data from various social media platforms such as Twitter, Facebook, Instagram, etc. The social media data collector component of the system will perform this task. The collector will use the APIs provided by these platforms to extract data such as customer reviews, complaints, and feedback on HP products. The collected data will include text, images, and videos.
Pre-processing:
The collected data may contain irrelevant information, such as spam, duplicate data, or data in different languages. The next step is to preprocess the data to remove such information and convert it into a suitable format for storage and analysis. The data preprocessor component of the system will perform this task. The preprocessor will perform the following tasks:
Cleaning: The preprocessor will remove irrelevant information such as spam, duplicate data, or irrelevant data.
Filtering: The preprocessor will filter out data that is not related to HP products such as data related to competitors.
Transformation: The preprocessor will transform the data into a suitable format for analysis. This may include converting data to lowercase, removing punctuation, and removing stop words.
The preprocessed data will then be passed on to the Sentiment Analysis Engine component of the system for further analysis. The data collected and preprocessed will be stored in a knowledge store component to allow for analysis and reporting of insights.

NLP and ML algorithms used

Tokenization: This technique involves breaking down a sentence into individual words or tokens. It is an essential step in NLP because it makes it possible to analyze individual words or tokens' sentiments.

Stop Words Removal: Stop words are common words in a language that do not carry significant meaning, such as "the," "and," or "is." Removing these words can improve the efficiency and accuracy of the sentiment analysis process.

Lemmatization/Stemming: These techniques involve reducing words to their base form to simplify the analysis process. For example, "running" and "ran" can be reduced to their base form "run."

Part of Speech Tagging: This technique involves identifying the part of speech of each word in a sentence, such as noun, verb, adjective, or adverb. This can help in determining the context and sentiment of the words in the sentence.

Naive Bayes Classifier: This is a popular machine learning algorithm used for sentiment analysis. It is a probabilistic algorithm that uses Bayes' theorem to classify the sentiment of text into positive, negative, or neutral categories.

Support Vector Machines (SVM): This is another machine learning algorithm that can be used for sentiment analysis. SVM tries to find the best boundary between different classes by maximizing the margin between them.

Recurrent Neural Networks (RNNs): RNNs are a type of deep learning algorithm that can analyze the sentiment of text. RNNs can capture the context and sequence of words in a sentence, making them particularly useful for sentiment analysis.

In summary, the Sentiment Analysis Engine component of the system can use a combination of these NLP techniques and ML algorithms to analyze customer sentiments and provide insights for improving HP products and customer experience.

Knowledge Graph Creation

Creating a knowledge graph is an important component of the system described in the problem statement. A knowledge graph is a structured representation of information that captures relationships and dependencies between different entities. It can be used to store and query data efficiently, making it an ideal solution for analyzing large volumes of customer data.

The knowledge graph can be created by following these steps:

Entity Recognition: The first step is to identify the entities in the customer data, such as product names, customer names, and features of the product. This can be done using NLP techniques such as named entity recognition (NER).

Relationship Extraction: Once the entities are identified, the next step is to extract the relationships between them. This can be done using NLP techniques such as dependency parsing, which can help identify the subject, object, and verb in a sentence.

Graph Creation: After identifying the entities and relationships, the knowledge graph can be created using a graph database such as Neo4j or RDF store. In the knowledge graph, each entity will be represented as a node, and each relationship will be represented as an edge connecting the nodes.

Knowledge Store Integration: The knowledge graph can be integrated with the knowledge store to enable efficient storage and querying of

customer data. The knowledge store can also be used to store additional information such as customer feedback and product specifications.
Querying the Knowledge Graph: Once the knowledge graph is created and integrated with the knowledge store, it can be queried to extract insights about customer sentiments and preferences. Queries can be made using a query language such as SPARQL or Cypher, which allows for complex queries that can capture relationships between entities.

Reference Links:

Batrinca, B. and P. C. Treleaven (2015). Social media analytics: a survey of techniques, tools and platforms. AI & SOCIETY 30(1), 89–116.
Cao, N., L. Lu, Y.-R. Lin, F. Wang, and Z. Wen (2015). Socialhelix: visual analysis of sentiment divergence in social media. Journal of Visualization 18(2), 221–235.

Golbeck, J., C. Robles, and K. Turner (2011). Predicting personality with social media. In CHI'11 Extended Abstracts on Human Factors in Computing Systems, pp. 253–262.

ACM. Guille, A., C. Favre, H. Hacid, and D. A. Zighed (2013). Sondy: An open source platform for social dynamics mining and analysis. In Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, pp. 1005–1008.