IT & Computer Science Department

Pak-Austria Fachhochschule: Institute of Applied Sciences and Technology, Haripur, Pakistan

Project Part-1:

DATA UNDERSTANDING AND PROBLEM DESCRIPTION

Introduction:

In this initial phase, our objective is to lay the groundwork for an effective data mining process. Suppose working as a data scientist, the client expects you to carefully select a dataset, thoroughly understand the problem it addresses and provide insights into its business importance. Additionally, gaining a comprehensive understanding of the dataset's features is crucial.

1.1 TASKS:

1. Select Dataset (5 marks):

Your task is to choose a dataset from reputable sources such as Kaggle, UCI Machine Learning Repository, or any other platform. Ensure that the dataset aligns with the client's objectives.

Answer:

I've selected a dataset from **Kaggle** for sentiment analysis, and it perfectly aligns with the client's objectives. Clients aim to understand customer sentiments regarding their product (medicines and vaccinations), service, or brand.

Data set named "**Sentimental Analysis**".

2. Problem Description (10 marks):

Write detailed description of the problem that the dataset aims to address. Clearly articulate the primary goal of mining this data and how it aligns with the client's business objectives.

Answer:

The sentiment analysis dataset aims to assess public perceptions and emotions surrounding medicines and vaccinations. It comprises text-based data from various sources, reflecting sentiments either positive, negative, or neutral regarding vaccinations and medicines. It includes data of different time periods to capture evolving sentiments, considering that public perception towards vaccinations and medications can change over time due to various factors like outbreaks, misinformation, or new research findings.
In the current environment, it is critical to understand public opinion about vaccinations and medications. This dataset aims to solve the problem of evaluating the public thoughts, attitudes, and feelings around medications and vaccinations.

Client primary goal is "To develop models to categorize sentiments expressed in text regarding vaccines and medicines". It also perfectly aligns with the clients objective of model development, to assess sentence tones about vaccines accurately and using sentiment analysis to guide evidence-based policy development and public health communication initiatives.

Importance of data mining in this dataset involves understanding of the public sentiment how vaccines are perceived by the general population and for the enhanced decision making.

3. <u>Business Importance Analysis (10 marks):</u>

Provide a thorough analysis of the business importance of solving the identified problem. Discuss why addressing this problem is crucial for the client's business strategy.

<u>Answer:</u>

Analyzing the sentiment of customer opinions through the chosen dataset holds vast business importance for the client. Understanding customer sentiments regarding their product, service, or brand is essential in shaping effective business strategies.

Customer sentiment directly impacts brand reputation. Positive sentiments can strengthen brand image and loyalty, while negative sentiments can lead to a damaged reputation. Addressing issues highlighted by negative sentiments allows for proactive management of the brand's image and enhances customer satisfaction.

Datasets analysis aids in product/service improvement. Identifying areas that receive negative feedback helps in pinpointing specific aspects that need attention or modification. This iterative process ensures continual enhancement of offerings, aligning them better with customer needs and preferences.

Sentiment analysis contributes significantly to marketing strategies. Understanding the sentiment behind customer reviews or feedback helps in crafting targeted and personalized marketing campaigns. It allows for better communication strategies, resonating more effectively with the intended audience. Sentiment analysis can also provide insights into competitor analysis.

Understanding how customers perceive competitors products/services aids in benchmarking and positioning strategies. It assists in identifying competitive advantages or weaknesses, allowing the client to capitalize on market opportunities.

Sentiment analysis from the dataset aligns with the clients objectives by enabling them to make data-driven decisions. This process of analyzing customer sentiment via the dataset serves as a powerful tool that not only shapes business strategies but also cultivates brand loyalty, fosters product/service improvement, refines marketing approaches, aids in competitive positioning, and most importantly, empowers the client with data-backed decision-making capabilities in the dynamic landscape of customer-centric business environments.

4. <u>Feature Understanding (10 marks):</u>

Investigate the features present in the dataset. Write a detailed description of each feature, emphasizing their relevance to the problem at hand. This will help for subsequent analyses.

<u>Answer:</u>

Each of the below features contributes uniquely to the dataset, providing context, user-related information, and metadata that can aid in comprehensive sentiment analysis regarding vaccines and medications.

The following below features are present in the dataset:

1. **id:** This feature likely represents a unique identifier or code assigned to each entry in the dataset. While it may not directly contribute to sentiment analysis, it serves as a primary key for referencing and organizing individual data points.
2. **user_name**: This feature indicates the username or handle of the individual expressing opinions or sentiments. While not directly influencing sentiment, it could be valuable in understanding if certain users consistently express specific sentiments, potentially indicating influencers or recurring opinions within the dataset.

3. **user_location:** This feature denotes the geographical location provided by users in their profiles. Understanding the geographic distribution of sentiments can offer insights into regional variations in perceptions about vaccines and medications, contributing to a more nuanced analysis.

4. **user_description:** This feature likely contains a brief bio or description provided by users in their profiles. While not directly influencing sentiment, it might provide context or background about the users expressing opinions, aiding in understanding their perspectives and potential biases.

5. **user_created:** This feature represents the date when the user account was created. It might not directly impact sentiment analysis but could be used to filter or analyze sentiments based on the account's longevity or age, potentially identifying patterns related to user behavior over time.

6. **user_followers:** This feature indicates the number of followers a user has. Understanding the influence of users with varying follower counts might be relevant in identifying sentiments that gain traction or visibility within the platform.

7. **user_friends:** This feature likely represents the number of users a particular user is following. While not directly related to sentiment, it might offer insights into user behavior or social connections that could influence the spread of opinions.

8. **user_favourites:** This feature indicates the number of tweets or content a user has marked as favorites. While not directly influencing sentiment, it might indicate the type of content a user resonates with, providing context for their opinions.

9. **user_verified:** This binary feature might indicate whether the user account is verified by the platform. While not directly impacting sentiment, verified accounts might carry more credibility, potentially influencing the spread or impact of expressed sentiments.

10. **date:** This feature denotes the timestamp when the tweet or opinion was posted. Analyzing sentiments over time can reveal trends, spikes, or shifts in public opinion related to vaccines and medications, aiding in understanding evolving perceptions.

11. **text:** This feature contains the actual text or content of the tweet expressing opinions about vaccines and medications. This is the primary feature relevant to sentiment analysis, serving as the textual data on which sentiment analysis models will be applied.

12. **Hashtags:** This feature includes any hashtags used in the tweet. It might indicate trending topics or themes related to vaccines and medications, offering context or themes associated with expressed sentiments.

13. **source:** This feature denotes the platform or application used to post the tweet. Understanding the source might provide insights into user behavior or preferences on specific platforms regarding discussions about vaccines and medications.

14. **retweets:** This feature represents the number of times the tweet has been shared or retweeted. Higher retweet counts might signify sentiments that resonate widely among users, potentially indicating influential opinions.

15. **favorites:** This feature denotes the number of times the tweet has been marked as a favorite by other users. Similar to retweets, higher favorite counts might indicate popular sentiments among the audience.

16. **is_retweet:** This binary feature likely indicates whether the tweet is a retweet or an original post. Understanding retweet behavior might help distinguish original opinions from echoed sentiments, providing clarity on the spread of sentiments.