# ===Analytics Startup Plan

**Synopsis: *This document provides a high-level walkthrough of the activities required to guide completion of the analysis.***

|  |  |
| --- | --- |
| **Project** | *Sentiment Analysis on Twitter* |
| **Requestor** | *Sudish Basnet, Omprakash Bista*  *301250603 , 301251726* |
| **Date of Request** | *17th July, 2023* |
| **Target Quarter for Delivery** | *14th August, 2023* |
| **Epic Link(s)** |  |
| **Business Impact** | *Enhance Brand Reputation, Drive Data-driven Decisions, Advance Competitive Analysis, Boost User Engagement* |

## 1.0 Business Opportunity Brief

|  |  |
| --- | --- |
|  | Clearly articulated business statement of the Ask, opportunity, or problem you are trying to solve for. An important step is to understand the nature of the business, system or process and the desired problems to be addressed. This will be communicated back to All stakeholders for alignment. |

Sentiment analysis is the process of extracting and understanding people's opinions, emotions, and attitudes from text data. Twitter, as one of the largest social media platforms, generates a massive amount of real-time data that can provide valuable insights into public sentiment. Sentiment analysis on Twitter’s tweets presents a valuable business opportunity to gain insights into customer perceptions, enhance brand reputation management, and drive data-driven decision-making. Real-time monitoring of sentiment allows for prompt response to customer feedback and crisis management.

## 1.1 Supporting Insights

|  |  |
| --- | --- |
|  | Define any supporting insights, trends and research findings. Where relevant, list key competitors in the market. What are their key messages, products & services? What is their share of market, nationally and regionally? |

Social media platforms have become crucial sources of information and insights for businesses. According to Statista, the global social media analytics market size is expected to reach $18.7 billion by 2027. Twitter, being one of the most influential platforms, offers a vast pool of data that can be harnessed for sentiment analysis. Therefore, the market potential for sentiment analysis on Twitter is significant and continues to grow. With over 330 million monthly active users, Twitter provides a wealth of data that can be leveraged by businesses, organizations, and individuals to gain insights into customer preferences, brand perception, public sentiment towards events or products, and more. Industries such as marketing, customer service, market research, and brand management can greatly benefit from accurate sentiment analysis on Twitter.

## 1.2 Project Gains

|  |  |
| --- | --- |
|  | *Describe any revenue gains, quality improvements, cost and time savings (as applicable). What will you do differently and why would our customers care. What are the implications if we do nothing? This section is particularly key for prioritization against company goals and KPI’s.* |

By doing sentiment analysis on Twitter, we can have the following project gains:

1. Improved brand reputation management through proactive identification and response to negative sentiment.
2. Enhanced customer satisfaction and loyalty by addressing concerns and improving products or services.
3. Deeper understanding of customer needs, preferences, and emerging trends.
4. Effective evaluation of social media campaigns and influencers.
5. Competitive analysis to identify areas for improvement and differentiate from competitors.

Failing to leverage sentiment analysis on Twitter may result in missed revenue opportunities, inadequate crisis management, inefficient resource allocation, and a lack of competitive advantage. Embracing sentiment analysis on Twitter is crucial for businesses to stay ahead, make data-driven decisions, and meet the evolving needs of their customers in a fast-paced digital landscape.

## 2.0 Analytics Objective

|  |  |
| --- | --- |
|  | List the key questions, assumptions and define the hypotheses. Often the deliverable may not just be an analysis output, however a recommended operating model or blueprint for a pilot etc.  Note: Asking the right questions and truly understanding the problem will lead to the right data, right mathematics, and right techniques to be employed. |

The primary objective of doing sentiment analysis on Twitter is to extract and understand the sentiment expressed in tweets. This involves classifying tweets as positive, negative, or neutral to gain insights into public opinion, customer sentiment, brand perception, and emerging trends.

The key questions for this analysis are:

1. What are the sentiments expressed in tweets related to a specific topic, brand, or event?
2. What are the factors influencing sentiment on Twitter? Are there specific keywords, hashtags, or user profiles associated with positive or negative sentiment?
3. How does sentiment on Twitter correlate with other business metrics, such as sales, customer satisfaction, or brand perception?
4. Can sentiment analysis on Twitter predict or provide insights into future consumer behavior or market trends?

Hypotheses:

1. There is a correlation between sentiment expressed on Twitter and real-world consumer behavior or market trends.

## 2.1 Other related questions and Assumptions:

|  |  |
| --- | --- |
|  | *List any assumptions that may affect the analysis* |

Questions:

1. What specific keywords, hashtags, or brand mentions should be tracked and analyzed?
2. Are there any regional or language-specific considerations in sentiment analysis?
3. How does sentiment vary over time and across different user segments?

Some assumptions for this analysis are:

1. The sentiment expressed in tweets can be accurately classified as positive, negative, or neutral.
2. Twitter users' expressed sentiments align with their true sentiments and reflect their opinions.

## 2.2 Success measures/metrics

|  |  |
| --- | --- |
|  | *What does success look like? Define the key performance indicators (success definition/indicators, drivers and key metrics) against which the objectives will be analyzed. These should be drawn from the interlock meeting with key stakeholders and will inform the approach and methodology for the analysis.* |

1. High accuracy of sentiment classification.
2. Customer satisfaction expressed on Twitter.
3. Impact and effectiveness of marketing campaigns.
4. Increase in customer engagement on Twitter.
5. Predictive power in anticipating consumer behavior and market trends.

## 2.3 Methodology and Approach

|  |  |
| --- | --- |
|  | *Now that you have a good understanding of the Ask and deliverable, detail the recommended approach/methodology.* |

The approach for our analysis will follow the following steps:

1. Collect and preprocess Twitter data using appropriate tools and APIs.
2. Apply natural language processing techniques for tokenization, cleaning, and transformation of text data.
3. Train a sentiment analysis model using machine learning algorithms.
4. Evaluate and fine-tune the model's performance using relevant evaluation metrics.
5. Extract insights from sentiment analysis results using statistical analysis and visualization techniques.

**Type of Analysis:** The initial approach will be to use NLP to determine the sentiment of a tweet and then followed by ML model to train and predict the given dataset while simultaneously testing with new input dataset. To check the accuracy of our classification model, we will be using confusion matrix on our testing dataset.

**Methodology:** Key questions from ‘Analytics objective’ will be tackled in ascending order as outlined in ‘5.0 Timelines and deliverable section’.

We will use the Twitter API to retrieve tweets and do an exploratory data analysis to perform some dimension reduction as we need just one feature which will be tweets, but we will try to analyze most of the classes initially so that we can use them later if needed to justify specific problem domain to be investigated. And then we will preprocess the raw text data (tweets) to clean and prepare it for analysis. This will include removing URLs, mentions, hashtags, punctuation, emoji, and special characters. Afterwards, we will split the preprocessed text into individual words or tokens to create a structured representation of the text data. Within the text, we will advance text cleaning techniques like stop word removal, stemming, and lemmatization to reduce noise in the data. As ML algorithms need numeric data to process, we will do feature extraction by converting the preprocessed text data into numerical features using any relevant approaches (BoW, TF-IDF, Word2Vec or GloVe).

After all these steps, our features will be ready to be trained. We will be using either Naive Bayes, Random Forests, or Recurrent Neural Networks as per the need. After training, we will evaluate our model using classification matrix.

Furthermore, we will apply the trained model to predict sentiment on new, unseen tweets or text data. The model then needs to assign sentiment labels (positive, negative, or neutral) to each input based on the learned patterns and features. Finally, we can analyze the sentiment predictions and perform any necessary post-processing steps like aggregating sentiment scores, calculating sentiment percentages, or visualizing sentiment trends using Matplotlib or Seaborn. If time allows then we will be conducting market trend on different segments based on different hot topics.

**Output:** The output will be in the form of categorical variable, either Positive, Negative, or Neutral. This will help to determine the sentiment of the tweets.

## 3.0 Population, Variable Selection, considerations

|  |  |
| --- | --- |
|  | Capture learning about the data available today location, structure, and reliability; this would include data in operational systems including dealer sourced, data warehouse and any CRM or email marketing systems available today. |

**Audience/population selection: Twitter Users**

**Observation window: Consideration of latest tweets**

**Inclusions: Tweets, User Location, Source, User Verified, Date, Re-Tweets, and any other relevant attribute. (We will use almost every attribute for EDA for future use but mainly we will be working on tweets)**

**Exclusions: Unnamed 0, ID, Username, Length, and Likes**

**Data Sources:** https://www.kaggle.com/datasets/sudishbasnet/twitter-dataset

**Audience Level:** Not Applicable

**Variable Selection: Tweets, (Target variable will be created manually after analyzing the tweets)**

**Derived Variables: None**

**Assumptions and data limitations:** None

## 4.0 Dependencies and Risks

|  |  |
| --- | --- |
|  | Identification of key factors that may influence the outcome of the project and likelihood of it happening: |

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk** | **Likelihood (based on historical data)** | **Delay (based on historical data)** | **Impact** |
| Dependency on Twitter's verification for data collection as a twitter developer, subject to any limitations or changes.  Risks include potential biases or noise in the data.  Challenges in accurately classifying sentiment due to language nuances, sarcasm, or contextual understanding.  Risks of misinterpretation or misrepresentation of sentiment analysis results. | *Low*  *Medium*  *High*  *Low* | *1-2 days*  *1-2 days*  *1-3 days*  *1-2 days* | Data sources should be changed and need to look for data from other open sources.  Impact sentiment analysis accuracy.  Impact sentiment analysis accuracy.  Actual motive of this analysis will not be meet. |
|  |  |  |  |

## 5.0 Deliverable Timelines

|  |  |
| --- | --- |
|  | List key dates and timelines as a work-back schedule. Activate line items based on complexity and line-of-sight required. Will set the stakeholder expectations for the process. |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **SN** | **Major Events / Milestones** | **Description** | **Scope** | **Days** | **Date** |
| 1. | Kick-off / Formal Request | *Analysis plan discussion and approval of analysis.* | *Finalization of blueprint for the analysis.* | *7* | *July 17,2023* |
| 2. | Exploratory Data  Analysis (EDA) | *Summarizing the main characteristics to analyze the data.* | *Data quality descriptive statics* | 3 | *July 24,2023* |
| 3. | Data Preparation   * Issues with duplicates, missing, incomplete, and noisy data. | *Handling missing values, removing duplicates, reducing dimension, extracting the important keywords from noisy data and convert them to numerical form.* | *Use of domain scope and variable scope with integration of different imputation and reduction techniques.* | 3 | *July 24, 2023* |
| 4. | Modeling | *Building a model using appropriate ML model and NLP.* | *Interpret tweets with NLP and initialize the ML model to fit, train, and test the twitter data.* | *13* | *August 7,2023* |
| 5. | Governance | *Setting internal standard for model.* | *Properly governed model.* | *5* | *August 12,2023* |
| 6. | Documentation | *Documenting the final report covering all aspects of analysis.* | *Analysis Report* | *2* | *August 14, 2023* |
| 7. | Presentation | *Delivering presentation by providing all the necessary inputs while working in the projects.* | *Final presentation on Twitter analysis.* | *2* | *August 14,2023* |