**1.0. Business Understanding**

**1.1. Overview**

The world today is rife with information flowing from millions of users across different platforms based on a variety of topics including politics, celebrities, data science, and exercise to make my brain bigger. These opinions on the web garner more and more traffic and gain traction. At the same time, this information reaches a much larger audience who may also share the same information with their networks.

A technique known as Sentiment Analysis tackles the analysis of sentiments using Natural Language Processing.

Natural Language Processing is a machine learning technology that gives computers the ability to interpret, manipulate, and comprehend human language. This would be very useful in analysing human made opinions on the web.

These sentiments across the internet can be analysed using Natural Language Processing methodologies.

Every company/ business with an online presence, and even ones without, require some form of observing, recording, tracking and analysing of these online opinions of their products or services. Doing so will insure their business’ public image and ensure that opinions on the web do not burn the palettes of their users, and especially those of the potential users of their products or services, so to speak.

SentimentFlow leverages the power of cutting-edge NLP techniques to analyze sentiment in textual data, providing valuable insights for decision-making by the management of the vendor. This analysis would determine whether sentiments are positive, negative or neutral.

**1.2. Problem Statement**

With such a large volume of information shared by and / or received from many users and potential users, business would not be able to keep up with the information received if they attempt to track everything, everywhere all at once, manually.

Without fully comprehending the effects of the publics’ opinion, the businesses' public image would be tarnished. The poor public image could lead to potentially market share losses, loss of trust from it's repeat consumers, low credibility to its potential clients and also loss of investment/ partnership opportunities.

**1.3. Stakeholders**

1. **Companies (Apple and Google):** These organizations are directly impacted by public sentiment. They want to monitor how their products are perceived and identify areas for improvement.
2. **Marketing Teams:** Marketing teams can use sentiment analysis to adjust their campaigns, respond to negative feedback, and highlight positive aspects of their products.
3. **Decision-Makers:** Executives and managers need insights into public sentiment to make informed decisions about product development, customer support, and brand reputation.

**1.4. Proposed Solution**

Analysing the public opinion would help businesses monitor their brand and sentiments around their products and services coming in as customer feedback, and understand customer needs, while making them more conscious thus preventing poor public relations.

**1.5. Value Proposition**

By accurately classifying tweets, our NLP model can provide actionable insights to stakeholders. For example:

* Identifying negative sentiment can help companies address issues promptly.
* Recognizing positive sentiment can guide marketing efforts and reinforce successful strategies.
* Understanding neutral sentiment can provide context and balance.

**1.6. Objectives**

**Main Objective**

To create a NLP multiclass classification model that can analyse sentiments in either 3 categories - Positive, Negative or Neutral. This model targets to achieve a recall score of 85% and an accuracy of 90%.

**Specific Objectives**

* To identify the most common words used in the dataset using Word cloud.
* To confirm the most used words that are positively and negatively tagged.
* To recognize the products that have been opined by the users.
* To spot the distribution of the sentiments.
* To develop market strategy that improves the product positioning.

**2.0. Data Understanding**

**Data Sources**

The dataset originates from CrowdFlower via data.world. Contributors evaluated tweets related to various brands and products. Specifically:

* Each tweet was labeled as expressing positive, negative, or no emotion toward a brand or product.
* If emotion was expressed, contributors specified which brand or product was the target.

**Suitability of Data**

Here's why this dataset is suitable for our project:

1. **Relevance:** The data directly aligns with our business problem of understanding Twitter sentiment for Apple and Google products.
2. **Real-World Context:** The tweets represent actual user opinions, making the problem relevant in practice.
3. **Multiclass Labels:** We can build both binary (positive/negative) and multiclass (positive/negative/neutral) classifiers using this data.

**Dataset Size**

The dataset contains over 9,000 labeled tweets. We'll explore its features to gain insights.

**Descriptive Statistics**

* **tweet\_text:** The content of each tweet.
* **is\_there\_an\_emotion\_directed\_at\_a\_brand\_or\_product:** No emotion toward brand or product, Positive emotion, Negative emotion, I can't tell
* **emotion\_in\_tweet\_is\_directed\_at:** The brand or product mentioned in the tweet.

**Feature Inclusion**

Tweet text is the primary feature. The emotion label and target brand/product are essential for classification.

**Limitations**

* **Label Noise:** Human raters' subjectivity may introduce noise.
* **Imbalanced Classes:** We'll address class imbalance during modeling.
* **Contextual Challenges:** Tweets are often short and context-dependent.
* **Incomplete & Missing Data:** Could affect the overall performance of the models.

**2.2. Data**

SHAPE

Records in dataset are 9093 with 3 columns.

COLUMNS

Columns in the dataset are:

- tweet\_text

- emotion\_in\_tweet\_is\_directed\_at

- is\_there\_an\_emotion\_directed\_at\_a\_brand\_or\_product

UNIQUE VALUES

Column \*tweet\_text\* has 9065 unique values

Column \*emotion\_in\_tweet\_is\_directed\_at\* has 9 unique values

Top unique values in the \*emotion\_in\_tweet\_is\_directed\_at\* include:

- iPad

- Apple

- iPad or iPhone App

- Google

- iPhone

- Other Google product or service

- Android App

- Android

- Other Apple product or service

Column \*is\_there\_an\_emotion\_directed\_at\_a\_brand\_or\_product\* has 4 unique values

Top unique values in the \*is\_there\_an\_emotion\_directed\_at\_a\_brand\_or\_product\* include:

- No emotion toward brand or product

- Positive emotion

- Negative emotion

- I can't tell

MISSING VALUES

Column \*tweet\_text\* has 1 missing values.

Column \*emotion\_in\_tweet\_is\_directed\_at\* has 5802 missing values.

Column \*is\_there\_an\_emotion\_directed\_at\_a\_brand\_or\_product\* has 0 missing values.

DUPLICATE VALUES

The dataset has 22 duplicated records.

##### **Conclusions from the Data Understanding**:

1. All the columns are in the correct data types.
2. The columns will need to be renamed.
3. Features with missing values should be renamed from NaN.
4. Duplicate records should be dropped.
5. All records with the target as "I can't tell" should be dropped.
6. Corrupted records should be removed.
7. Rename values in the is\_there\_an\_emotion\_directed\_at\_a\_brand\_or\_product where the value is 'No emotion toward brand or product' to 'Neutral Emotion'

**3.0. Data Cleaning**

From the analysis, the intricate steps followed below cleaned the data before further analysis and modeling.

**Validity Checks:**

* All corrupted records were removed from the dataset,
* Removed all the sentiments that we would not account for.
* Streamlined the values in the third column.

**Completeness Checks:**

* Dropped any records with missing values in the first column.
* Filled in the missing values in the second column using signposts found in the tweet column.
* Streamlined the values in the emotions column

**Consistency Checks:**

* Dropped any duplicated records in the dataset

**Uniformity Checks:**

* Renamed the columns
* Reset the index of the data.