**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, wordle, 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.

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 to 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.

A major mobile vendor who has been collecting sentiments across brands, products and services reached out to us at **SentimentFlow** to address the business problem above.

SentimentFlow leverages cutting-edge NLP techniques to analyze sentiment in textual data, providing valuable insights for decision-making by the management of the vendor.The analysis would be used to determine whether data is 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 could 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. 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.

**1.4. 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.

**1.5. Constraints**

The following potential constraints were identified as:

* Data Quality - Incomplete data, imbalanced classes and missing Data could affect the overall performance of the models.

**2.0. Data Understanding**

**2.1. Sources**

The data was sourced from Data World. It contained 9093 records and 3 columns.

**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.

##### Comments:

* All the columns are in the correct format
* The columns names will need to be changed
* Features with missing values should be renamed from NaN
* Duplicate records should be dropped
* All records with the target as "I can't tell" should be dropped