PROJECT REPORT: HACK THE FEED - INSIGHTS FROM SOCIAL MEDIA DATA

Introduction

The goal of this project is to analyse Stanbic IBTC's social media data and turn it into valuable insights that can drive positive changes. The dataset under analysis covers about 10 years of Stanbic IBTC's social media data. This includes big platforms like Facebook, Instagram, LinkedIn, and Twitter.

The aim of this project is to extract meaningful insights regarding the core drivers of engagement within Stanbic IBTC's social media network. These insights will be transformed into informed recommendations, ultimately optimizing Stanbic IBTC's digital engagement strategies.

Methodology

The project's methodology was meticulously designed to conduct an in-depth analysis and extract meaningful insights from Stanbic IBTC's social media data. The approach involved the following pivotal steps:

1. Data Familiarization:

The project began by thoroughly understanding the provided datasets. Each dataset comprised 147 columns, yet it was evident that a significant number of values were missing. The Twitter dataset alone had 993,636 missing values, while the LinkedIn dataset had 997,785, the Instagram dataset had 1,277,690, and the Facebook dataset contained 887,357 missing values. Additionally, certain columns in the datasets were entirely populated with null entries.

Upon scrutinizing the data, it became apparent that content types accessible to users on Twitter and Facebook included Photo, Text, Video, and Links. LinkedIn expanded these options to encompass Polls, in addition to the four mentioned types. Conversely, Instagram offered Photo, Video, and Carousel as the available content types.

2. Data cleaning and processing:

In the initial phase of data processing, integration of the four datasets was undertaken to create a comprehensive master dataset. This merged dataset, which was made up of 36,092 rows and

147 columns, formed the foundation for subsequent analysis. The data cleaning stage comprised the following crucial steps:

a. **‘Post Type’ column update:** The 'Post Type' column underwent a comprehensive update to ensure precise representation of each post's type within the network. In cases where the 'Post Type' was not 'Tweet', it was systematically prefixed with the appropriate network identifier to enhance clarity. For instance, posts in the Facebook, LinkedIn, and Instagram datasets were modified to 'Facebook post', 'LinkedIn post', and 'Instagram post', respectively. This standardization contributed to a more structured and informative dataset.
b. **‘Post Id’ column update**: Here, the ‘Post Id’ column was updated for Facebook posts that has an underscore sign, which should be removed.
c. **Numeric columns cleaning**: Numeric columns, including but not limited to Impression, Organic Impression, Engagement, and Click-through rate, were meticulously cleaned. Extraneous characters such as commas and percentage signs were diligently removed, preparing the columns for a subsequent change in datatype. This step ensured uniformity and accuracy in the numeric data representation.
d. **Handling nulls**: Within the merged dataset, specific columns were found to be populated with null values, indicating an absence of data across all four social media channels. Consequently, these columns were removed from the merged dataset. Subsequently, a thorough analysis was conducted to ascertain the percentage of null values in each remaining column. Employing statistical analysis, a threshold of 50 percent was established, and columns exceeding this threshold were removed from the dataset to streamline the analysis process. After this curation, the resulting dataset comprised 36,092 rows and 18 columns.
e. **Handling Data Types:** In preparation for imputing null values in the remaining columns, a critical step involved converting the columns to their appropriate data types, ensuring data uniformity and accuracy for subsequent analyses.
f. Handling Null Values (2): Following the conversion of each column to its suitable data type, all the remaining null values were addressed through imputation, resulting in a dataset devoid of null entries.
g. Removing Duplicates: A thorough scan of the dataset was conducted to identify and eliminate any duplicate rows. However, this meticulous analysis revealed the absence of any duplicate entries within the dataset.
h. Cleaning the 'Date' Column: Within the dataset, the 'Date' column encompassed both the date and time of each post. To enhance data organization and clarity, the 'Date' column was

meticulously cleaned, segregating the date and time into distinct columns for improved structure and analysis.

i. **Engineering the 'Time of Day' Column:** A new column titled 'Time of Day' was created to signify the period during which a post was made. Posts made between '12:00:00' and '18:00:00' were labelled as 'Afternoon', those between '18:00:01' and '23:59:59' as 'Evening', and the rest as 'Morning', offering a clear representation of post timings.

j. **Engineering the 'Engagement Score' Column:** To comprehensively gauge post engagement, a novel metric known as 'Engagement Score' was introduced. This metric is a weighted average of engagement components such as likes, comments, impression, and more. It provides an independent measure for thorough analysis of a post's overall engagement. Notably, weights were strategically assigned to each feature, prioritizing human interactions such as Reactions, Likes, Comments, and Shares.

3. **Exploratory Data Analysis:**

For the exploratory data analysis step, data wrangling was done to address specific questions and gain valuable insights. Some of the questions include:

i. **Which platform yielded the highest engagement?**

As seen in figure 1 below, the Facebook platform generated the highest engagement for the client, while LinkedIn generated the least engagement.

ii. **Which of the content types was engaged the most?**

As shown in figure 2 above, the content type with the most engagement was Poll, followed by Photo, and the least is Document.

iii. **What time during the day experienced the highest frequency of posts?**

As shown in figure 3 above, a significant portion of the posts across all social media platforms were published in the afternoon, specifically between 12 PM and 6 PM.

iv. **During which period do posts tend to receive the highest engagement?**

As shown in figure 4 below, on an average, posts made in the evening received the highest engagement across all social media platforms.

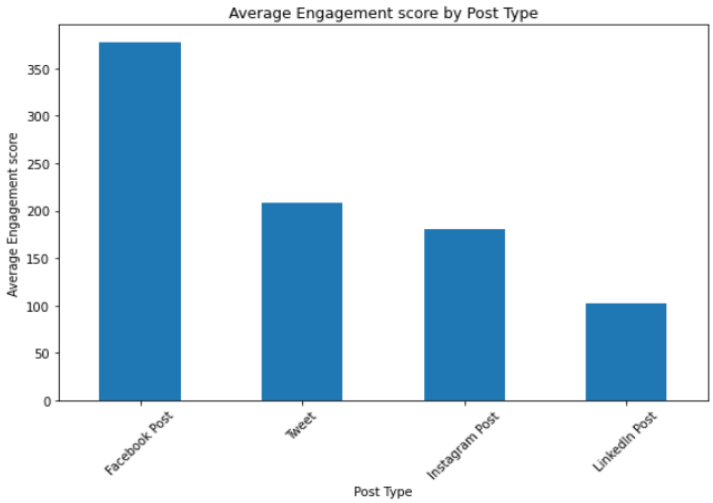

Figure 1: A plot of Average Engagement score by Post Type

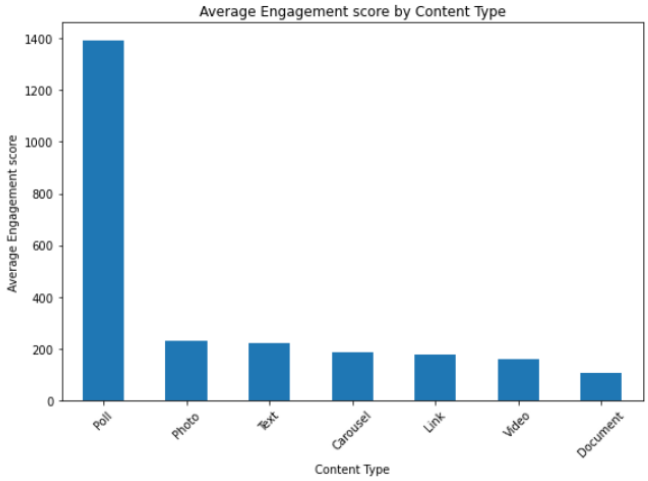

Figure 2: A plot of Average Engagement score by Content Type.

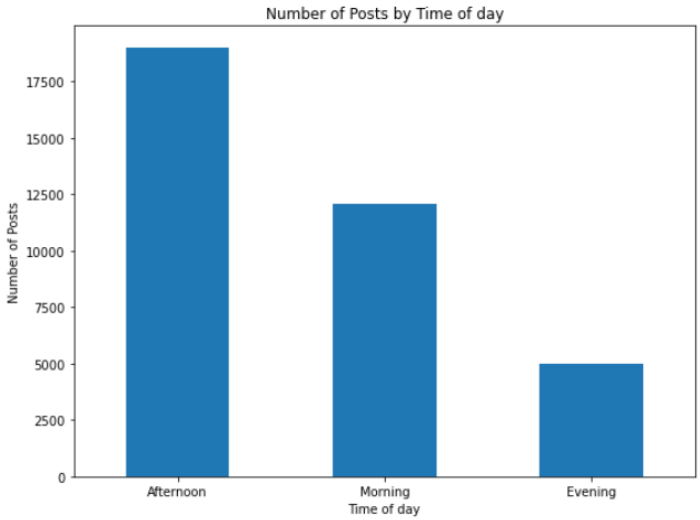

Figure 3: A plot of Number of posts by Time of day.

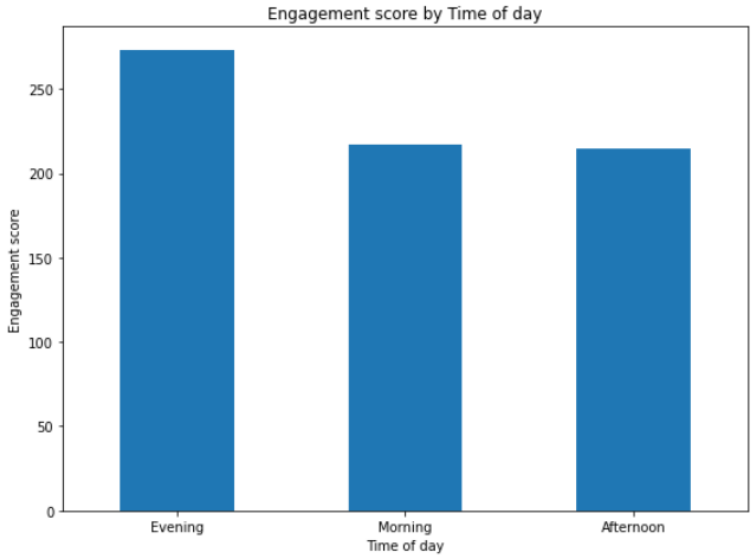

Figure 4: A plot of Average Engagement score by Time of day.

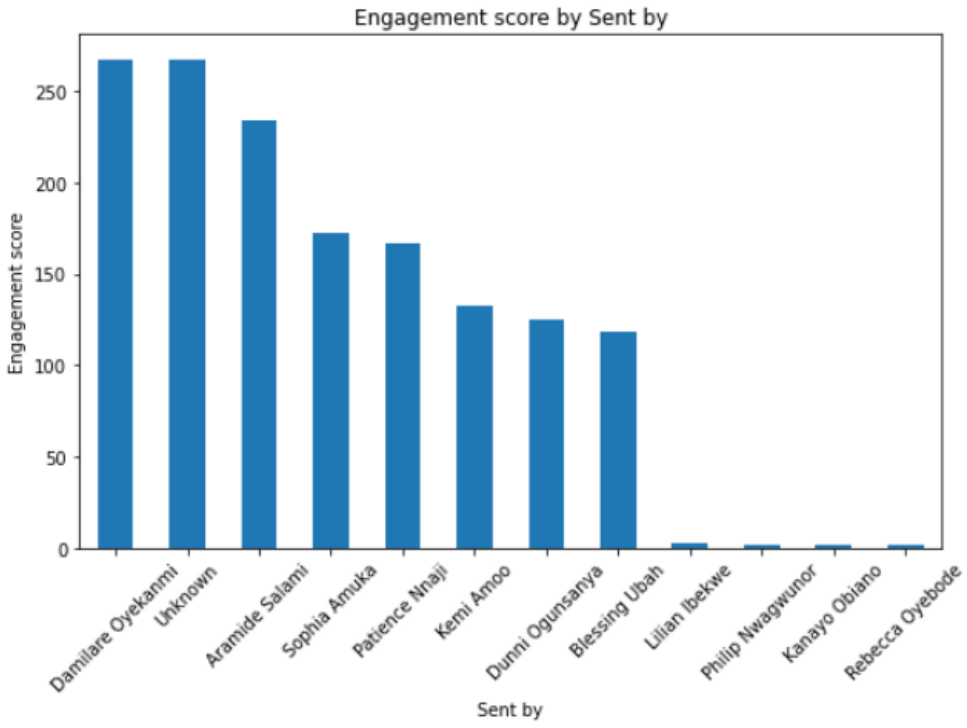

Figure 5: A plot of Average Engagement score by sent by.

v. **Whose post had the highest engagements?**

Majority of the posts were contributed by individuals whose identities are unknown. A good number of the post were made by Blessing Ubah, Sophia Amuka, Aramide Salami and Damilare Oyekanmi, while Rebecca Oyebode, Kanayo Obiano and Philip Nwagwunor made just one post.

However, as shown in figure 5 above, Damilare Oyekanmi's post were more engaged while that of Rebecca Oyebode were least engaged. This is quite interesting because Damilare Oyekanmi made fewer post (2991) compare with Blessing Ubah (5712) and Sophia Amuka (5276) and Damilare had a highest average engagement score.

vi. **How does Damilare's posts compare with his colleague Blessing Unah's?**

The analysis reveals that Damilare stands out as a prominent contributor to Facebook posts, particularly favoring afternoon timings. A notable characteristic of Damilare's posts is their predominantly textual nature. Similarly, Blessing exhibits a substantial presence on Facebook, also emphasizing textual content. However, Blessing's posting frequency is notably higher during the evening hours.

Despite analyzing three key features, we did not unearth sufficient insights to explain Damilare's superior engagement levels compared to his/her colleagues. Further investigation into the post content is warranted, as it may unveil the underlying reasons for this disparity.

vii. **How does the engagement metrics compare with each other?**

As shown in figure 6 below, higher impressions tend to occur when metrics such as reactions, likes, and comments are high. This suggests a correlation between user engagement metrics and the reach or visibility of the post, wherein increased engagement is associated with a larger post impression.

4. **Feature Engineering:**

In this step, more features that are crucial to the analysis were engineered. This includes:

i. Day of the week column
ii. Month of the year column and
iii. Day type column (Weekday or weekend).

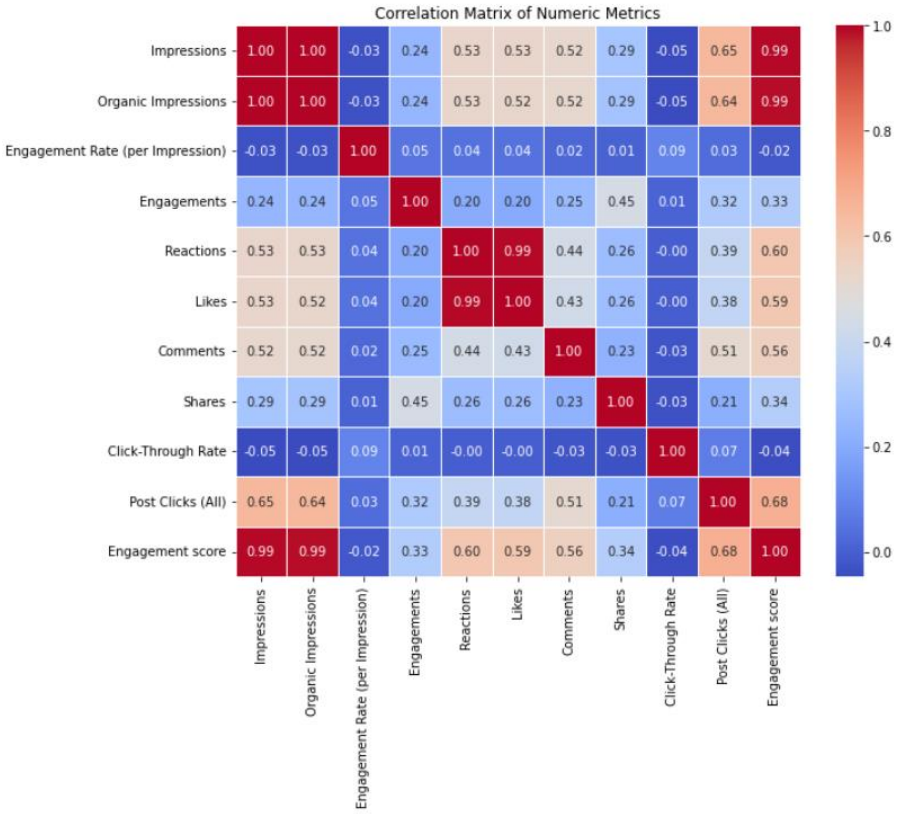

Figure 6: Corelation Matrix of numeric features.

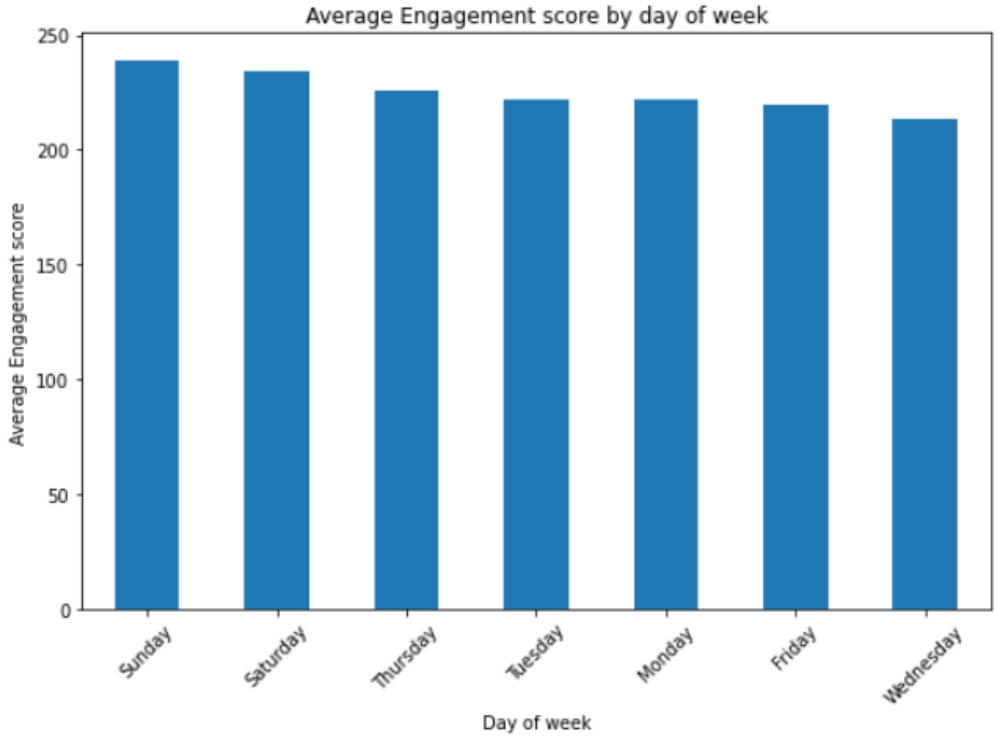

Figure 7: A plot of Average Engagement score by Day of the week.

5. Further Exploratory Data Analysis

In this step, with the newly engineered features, further questions were posed to gain more insight into the analysis.

i. Which day of the week recorded the highest engagement?

As shown in figure 7 above, post made on Sunday recorded a little more engagement than other days, while Wednesday recorded the least.

ii. Which month of the year recorded the highest engagement?

As shown in plot 8 below, post made in March recorded a little more engagement than other months, while November recorded the least.

iii. Did post made on weekends record more engagements than that of weekdays?

Yes. As shown in figure 9 below, posts made on weekends accrued more engagements.

6. Deep Dive Analysis: Natural Language Processing

In this phase, we conducted an in-depth exploration leveraging Natural Language Processing (NLP) techniques. Our primary objective was to delve into the post column, analyzing the content of posts across all social media platforms. The goal was to discern patterns and correlations to understand if post content plays a significant role in driving engagement levels.

The dataset was split into two distinct subsets: one comprising posts with high engagement (Engagement Score > 1000), and the other containing posts with lower engagement (Engagement Score <= 1000). This segregation enabled a more focused and nuanced analysis.

i. Frequently occurring words in posts with high engagement

In Figure 10a and 10b below, the chart and word cloud visualizations highlight the prevalent words within the dataset. Notably, terms such as 'Stanbicibtc' and '909' (the bank's USSD code) stand out, representing the bank's prominent identity. Moreover, words such as 'Money,' 'Fund,' 'Join,' 'Winner,' and 'Win' are notably displayed, implying enticing offers from the bank. Additionally, the prominence of terms like 'http,' 'com,' and 'click' suggests the frequent inclusion of links in the posts.

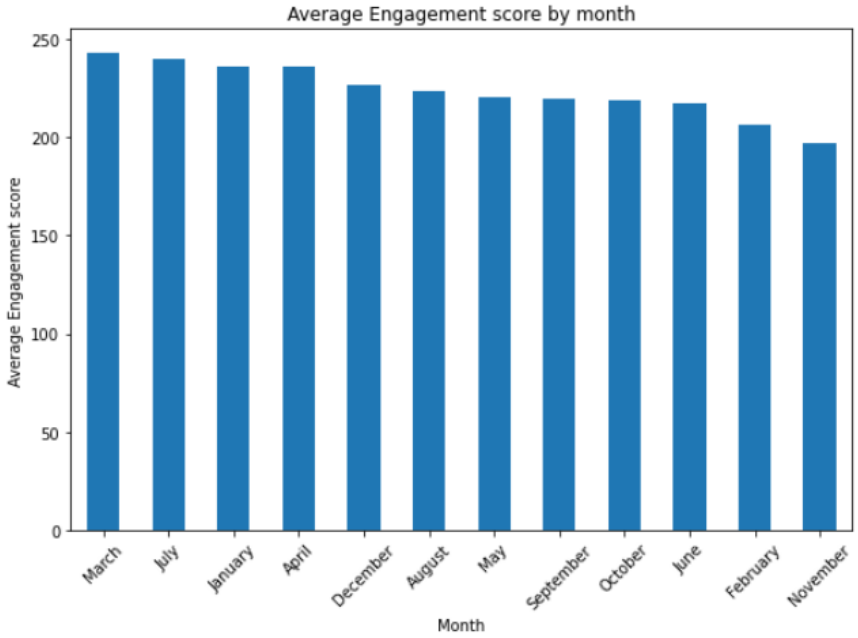

Figure 8: A plot of Average Engagement score by Month.

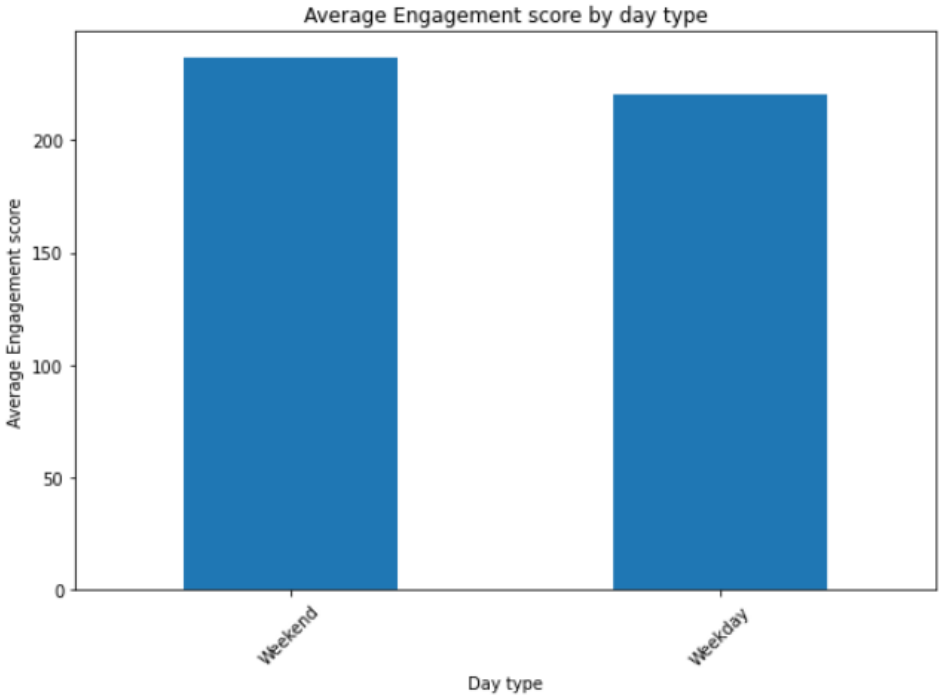

Figure 9: A plot of Average Engagement score by Day type.

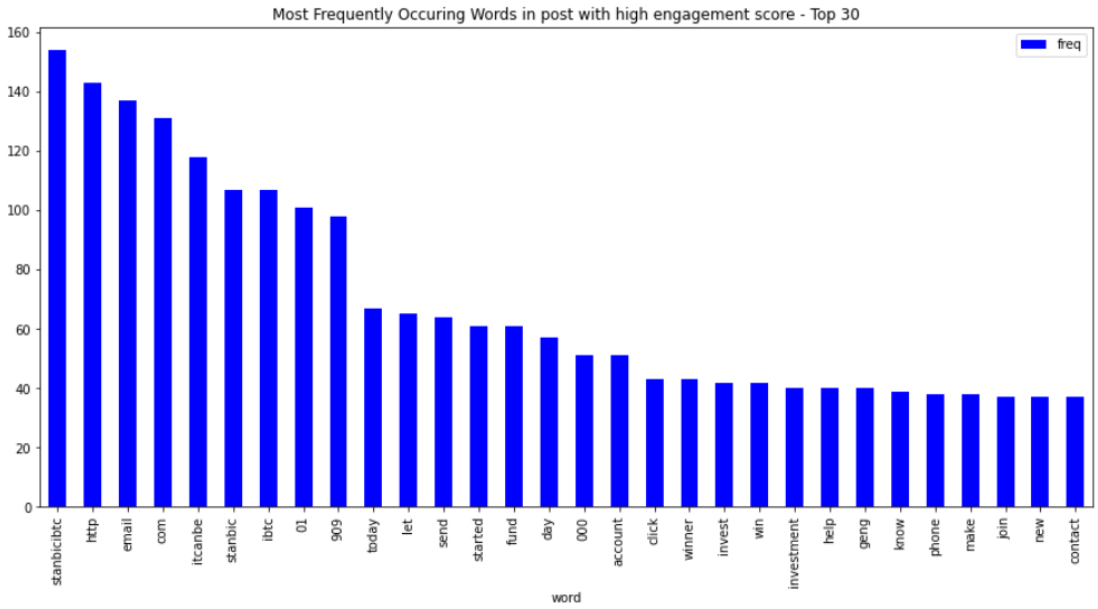

Figure 10a: A plot of the top 30 frequently occurring words in post with high engagement.

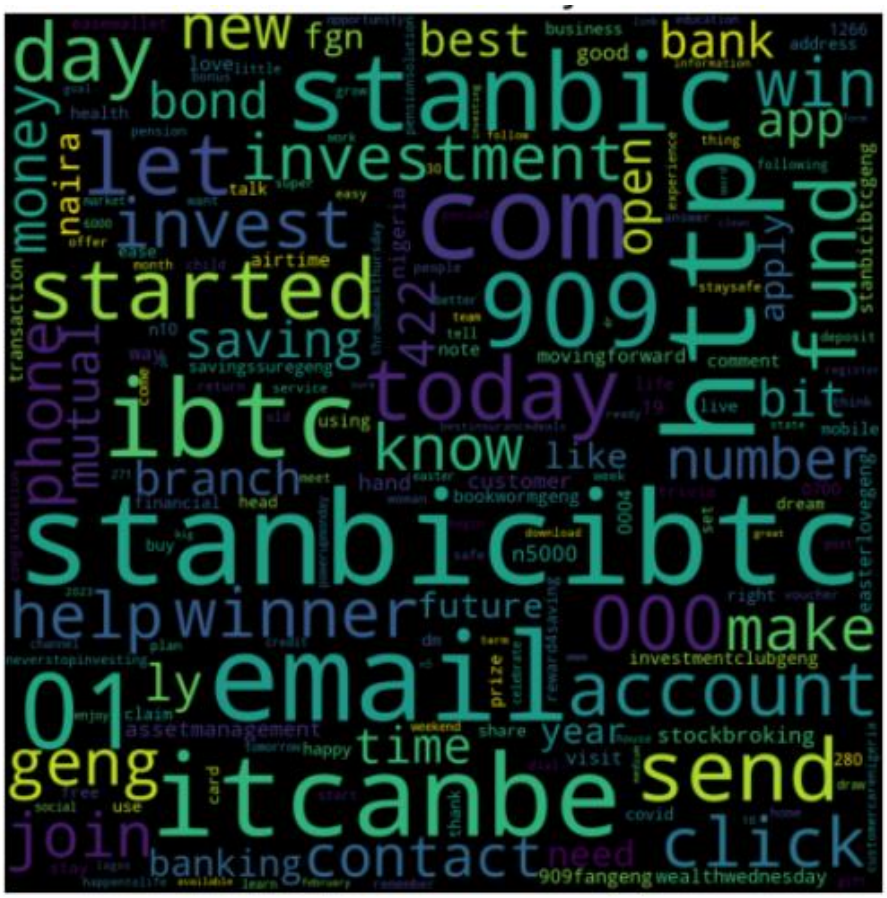

Figure 10b: A Word Cloud chart of the top 30 frequently occurring words in post with high engagement.

ii. Frequently occurring words in posts with low engagement

In Figure 11a and 11b below, the chart and word cloud visualizations highlight frequently occurring words, such as 'itcanbe' and 'Stanbicibtc,' representing the bank's brand. Additionally, 'visit and 'email, are also prominent. Words like 'App,' 'account', and 'Business' are also notable in the chart.

Insights

i. Facebook emerged as the primary engagement generator, while LinkedIn displayed the least engagement.
ii. Polls were found to be the most engaging content type; however, only a minimal number (2) of polls were posted across all social media platforms over the 10-year period.
iii. Posts made in the evening received higher engagement compared to other times of the day.
iv. Despite having a lower number of posts, Damilare Oyekanmi's posts were highly engaging, making him a notable contributor.
v. Blessing Ubah had a higher frequency of posts as compared with Damilare. Just like Damilare, she primarily utilized Facebook for posting, focusing mainly on text-based content, but she displayed a notable preference for evening posts. However, despite her frequent posts, she experienced lower engagement levels.
vi. Posts shared on Sundays witnessed heightened engagement, indicating a receptive audience during this day.
vii. March saw a surge in engagement, surpassing other months in terms of audience interaction.
viii. Engagement levels were notably higher during the weekends compared to weekdays, showcasing distinct engagement patterns based on the day of the week.
ix. Between 2013 and 2023, posts made in 2019 had the highest engagement.
x. Specific keywords such as 'Money,' 'Fund,' 'Join,' 'Winner,' and 'Win' appeared to capture audience attention, resulting in increased engagement levels.

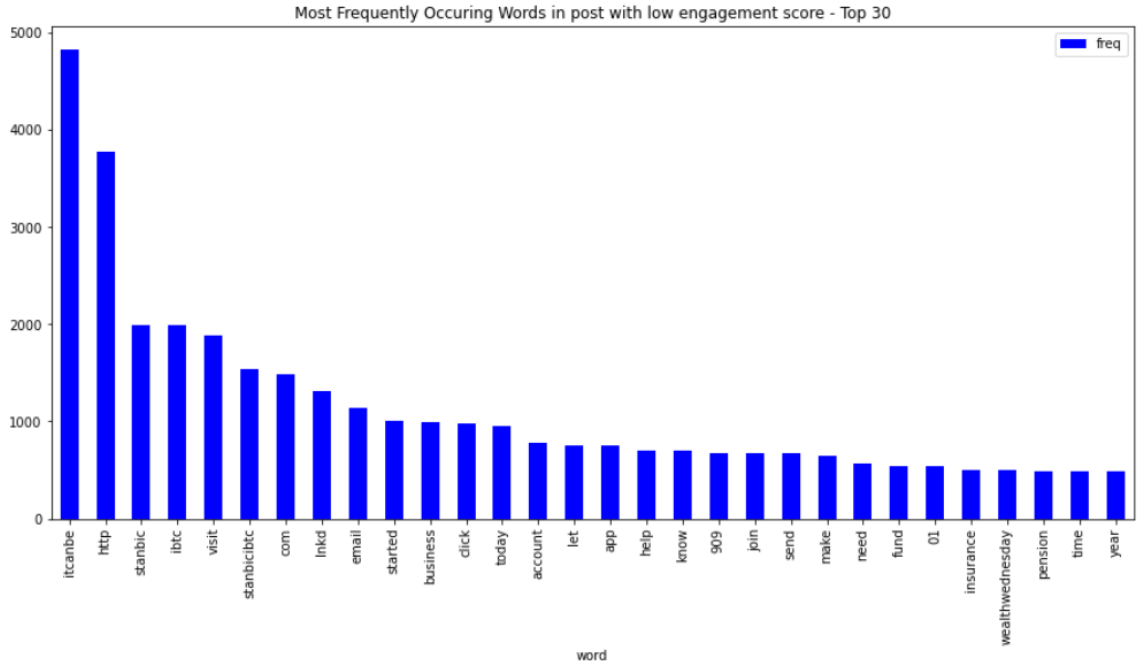

Figure 11a: A plot of the top 30 frequently occurring words in post with low engagement.

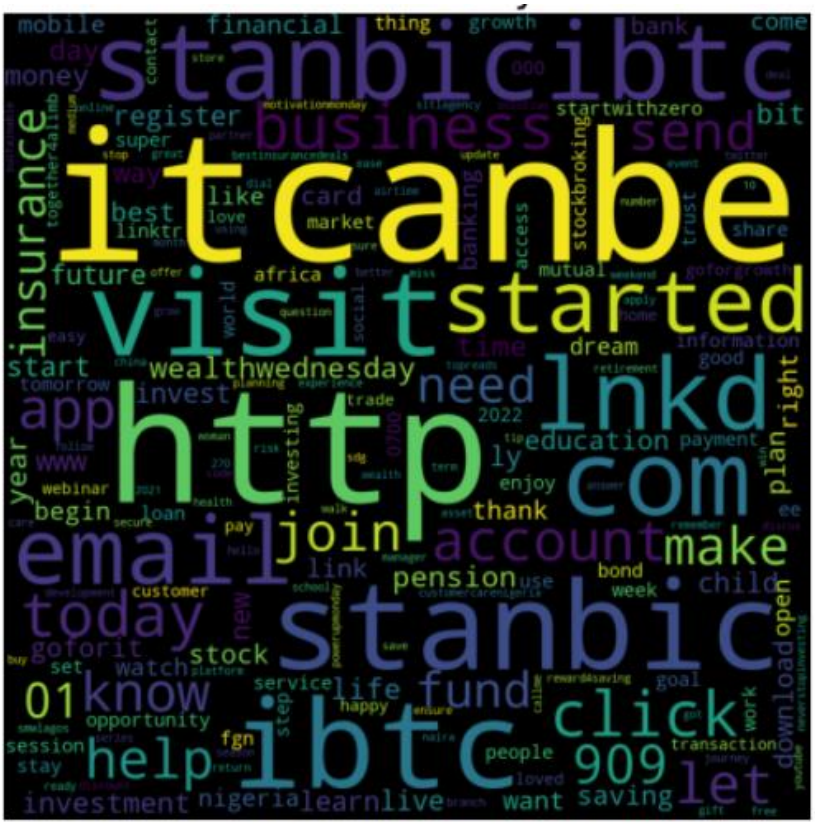

Figure 10b: A Word Cloud chart of the top 30 frequently occurring words in post with low engagement.

Recommendations

i. Given Facebook's significant engagement, this platform should be prioritized for content distribution. Contents should be tailor to suit Facebook's audience preferences and behaviour.
ii. Despite their infrequent use, polls can be a powerful tool to drive interactions and engagement. Engaging and relevant poll questions should be designed to boost audience participation.
iii. Focus should be placed on scheduling posts during the evening, as this timeframe consistently showed higher engagement.
iv. The influence of key contributors like Damilare Oyekanmi would be leveraged. He/She should be involved in impactful campaigns or collaborations, maximizing his/her engagement potential.
v. The considerable number of posts lacking author attribution in the datasets should be addressed. Mechanism addressing comprehensive data recording should be implemented allowing for more accurate insights and tailored engagement strategies.
vi. The surge during peak periods (such as Sundays and the month of March) should be capitalized on. Engaging and informative post should be scheduled for these peak periods.
vii. The type of content and engagement drivers employed in 2019 should be analysed, replicated, or modified for future campaigns.
viii. Engaging words such as 'Money,' 'Fund,' 'Join,' 'Winner,' 'Win,' etc. should be subtly and strategically integrated into future post to capture attention and drive higher engagement.

Conclusion

Implementing these recommendations will enhance engagement levels, effectively connecting with the audience and driving favourable outcomes for Stanbic IBTC's social media presence.