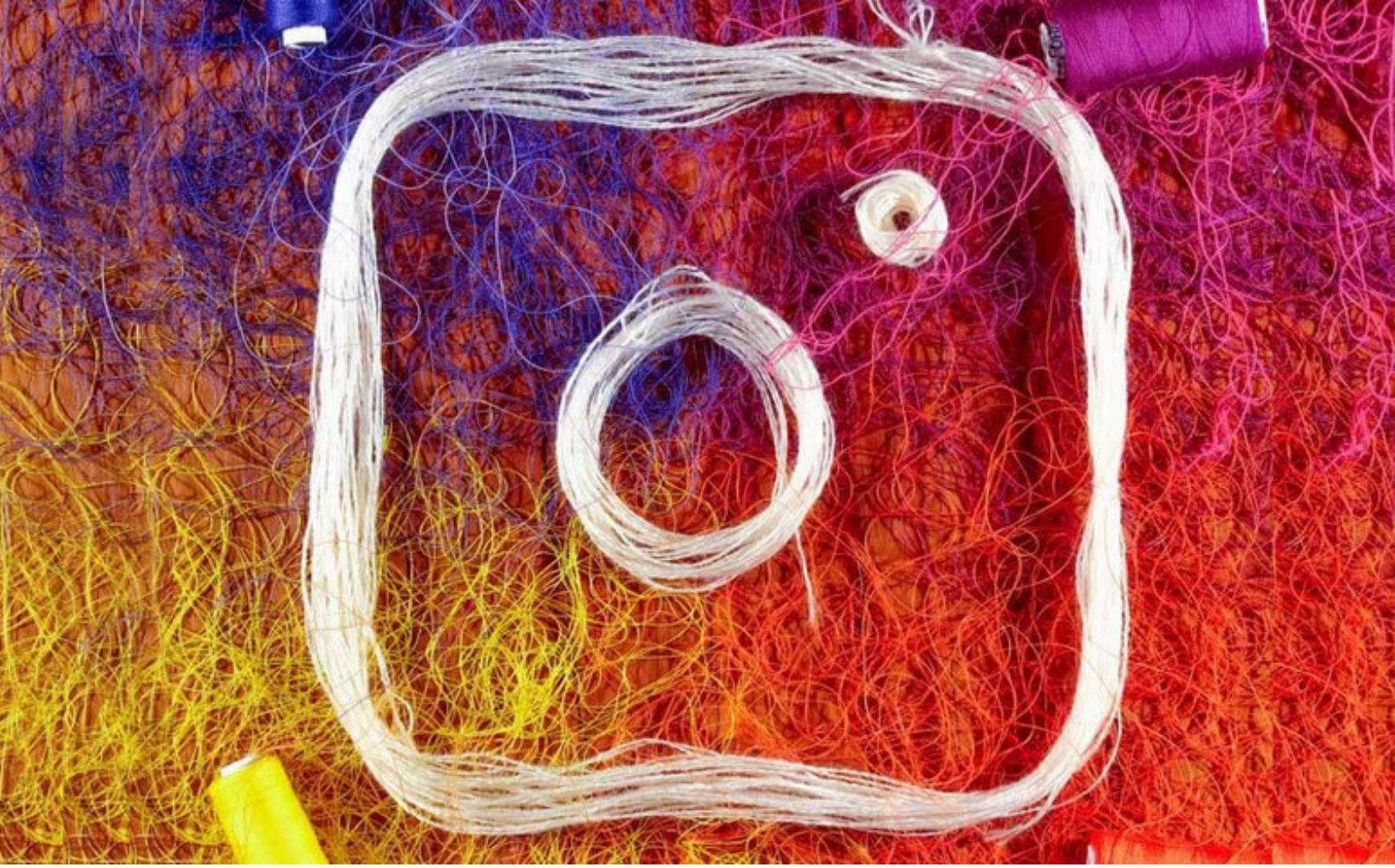


A Comprehensive Report on

Sentimental Analysis of Instagram Posts

April 2024

AUTHOR:
SAEED SHIRANI

About the Concept of Project

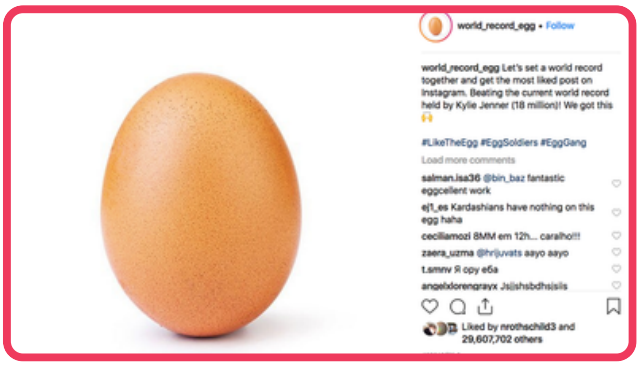
In today's world, Instagram stands out as one of the most impactful social media platforms. Its influence on people's minds and thoughts is unparalleled, making it challenging for other platforms to compete. With a diverse user base spanning the globe, Instagram offers a unique opportunity to intercept and analyze hashtags, activities, captions, likes, and comments, providing valuable insights into what's happening in different locations worldwide.

Businesses can leverage this wealth of information to guide their strategies effectively. By observing the reactions of influencers and users to specific concepts, they can tailor their approach to achieve optimal results.

Public Opinion is IMPORTANT!

Public opinion plays a crucial role in decision-making processes, especially in today's interconnected world where platforms like Instagram serve as virtual town squares. Egg's Instagram story exemplifies this significance by showcasing how a single post can catalyze widespread reactions and discussions among users. The engagement and responses generated provide valuable insights into the prevailing sentiments and preferences of the audience.

Decision-makers can leverage this information to gauge public perception, anticipate potential reactions, and tailor their strategies accordingly. Whether it's launching a new product, shaping a marketing campaign, or addressing a societal issue, understanding public opinion on platforms like Instagram empowers decision-makers to make informed choices that resonate with their target audience.



Moreover, Egg's Instagram story underscores the democratization of influence in the digital age. Unlike traditional media where only select voices could shape public discourse, platforms like Instagram empower individuals and influencers of all backgrounds to amplify their messages and perspectives. Decision-makers must recognize the diverse voices and opinions reflected in these online conversations. By engaging with and considering a broad spectrum of viewpoints, they can foster inclusivity, authenticity, and trust in their decision-making processes. Ultimately, Egg's Instagram story highlights the importance of embracing and leveraging public opinion as a dynamic and influential force in shaping decisions across various domains.

About The Dataset

In this Project i used Dataset of instagrame post's, This dataset was available on kaggle, but it is deleted now.
if you need the dataset, you can download it from my github: [Instagram Data](#).

this Dataset is include these columns:
'owner_id', 'owner_username', 'shortcode', 'is_video', 'caption', 'comments', 'likes', 'created_at', 'location', 'imageUrl', 'multiple_images', 'username', 'followers', 'following'.

owner_id	owner_username	shortcode	is_video	caption	comments	likes	created_at	location	imageUrl	multiple_images	username	followers	following
36063641	christendominique	C3_GS1ASeWI	False	I'm a brunch & Iced Coffee girlie	268	16382	1.709327e+09	NaN	https://instagram.flba2-	True	christendominique	2144626.0	1021.0

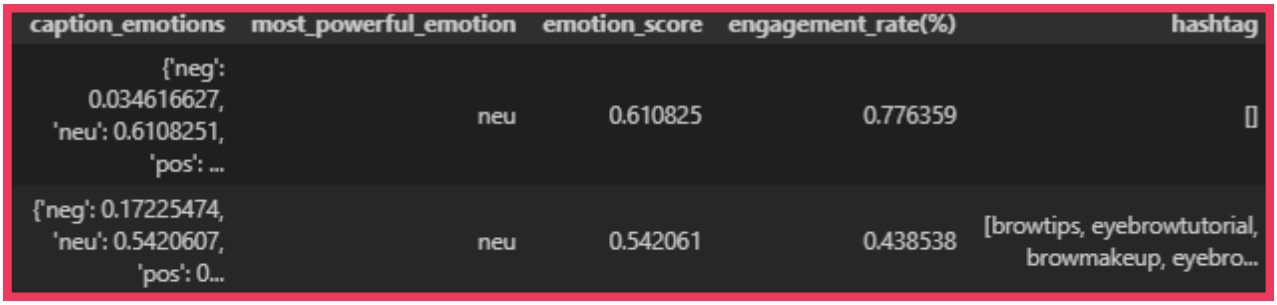
1.ETL and Data Wrangling

My initial task involved handling missing values, converting data types to ensure compatibility, and addressing duplicates within the dataset. Managing missing values presented a significant challenge for two primary reasons:
Firstly, each row represents a distinct post, reflecting real-world data essential for our analysis. Retaining these entries is crucial to ensuring that our analysis reflects real-world phenomena accurately. However, numerous rows contained missing values, necessitating their removal to ensure the integrity and reliability of our results.

Secondly, during the data processing, it became evident that certain data points within rows were missing. To obtain accurate and meaningful results, it was imperative to address these missing values appropriately. Consequently, dropping rows with several missing values became a necessary step in our data preparation process.

1.1.Additional Columns

After Cleaning and Preparing Data, I Calculate and find all related values of data frame. these are added columns:



caption_emotions	most_powerful_emotion	emotion_score	engagement_rate(%)	hashtag
{'neg': 0.034616627, 'neu': 0.6108251, 'pos': ...	neu	0.610825	0.776359	[]
{'neg': 0.17225474, 'neu': 0.5420607, 'pos': 0...	neu	0.542061	0.438538	[browtips, eyebrowtutorial, browmakeup, eyebro...

Here is a list of the added columns:

1.1.1. Caption Emotions:

For this column, I utilized the Roberta model to conduct sentiment analysis on the caption text. I implemented a Python function and applied it row by row to the data frame using pandas' apply function.

1.1.2. Most Powerful Emotions:

This column indicates the predominant emotion expressed in the caption. Additionally, in the subsequent column, I included the score associated with this emotion.

1.1.3. Instagram Post Engagement Rate:

The Instagram Post Engagement Rate serves as a pivotal metric, gauging the degree of interaction users exhibit with content on Instagram. It is computed by summing the total number of likes and comments garnered by a post. This sum is then divided by the number of followers the account possesses. Finally, the result is multiplied by 100 to derive a percentage. The formula can be expressed as:

$$\text{Engagement Rate} = \frac{\text{Post's Likes} + \text{Post's Comments}}{\text{Followers}} \times 100\%$$

Here's how we can interpret the score:

- **Below 1%**: Considered low. This might suggest that your content is not effectively engaging your audience or that your follower base is inactive.

- **1% to 3.5%**: Average to good. A rate within this range is typical for most profiles and shows a healthy level of interaction.

- **3.5% to 6%**: High. This suggests that your content is very engaging and is generating a lot of interest and interaction.

- **Above 6%**: Very high. This is typically seen with highly compelling content or profiles with a smaller, more engaged audience.

1.2.Sentimental Analysis

Sentiment analysis is indeed crucial for assessing the performance of an account, team, or product. In many cases, models are employed to analyze the emotions conveyed in text. However, this project faced several challenges. The captions were multilingual, and dealing with a large number of rows, especially in terms of translating or labeling them using REST APIs like Google Translate API, proved to be extremely challenging and time-consuming.

After numerous unsuccessful attempts with different strategies, a breakthrough was achieved with the discovery of a model called Roberta. Built upon the BERT Model, it is capable of handling multilingual text effectively.

The implementation codes for this can be found in the GitHub repository of the project.

1.3.Changed Columns

I did not initially add some of the columns, but I made modifications to them, such as the "**location**" column. Locations were either represented as **NaN** or in **Dictionary format**, like this:

"{'id': '34537076', 'has_public_page': True,
'name': 'Salem, Massachusetts',
'slug': 'salem-massachusetts'}"

I created a function to replace these dictionaries with the value of the ‘name’ key. For instance, in the example mentioned above, the final result is formatted as “salem-massachusetts”, with all spaces removed and all letters converted to lowercase.

1.4.Exporting Processed Data

In conclusion, I segregated the dataset into two distinct subsets: one containing **location-related data** and the other comprising **non-location-related data**. These subsets were then stored in a folder named "**Prepared Data**". The entire data processing workflow, including handling missing values, type conversion, and duplicate removal, is documented in the **ETL** notebook.

2. Explanatory Data Analysis

In this section of the report, I will discuss the process outlined in the EDA Notebook, which is accessible on GitHub under the name "EDA Notebook".

My analysis is divided into the following stages:

1. Analysis of User Behavior. (for a single account)
2. Sentiment Analysis Based on Time and Location.
3. Sentiment Analysis Based on Time without considering Location.
4. Hashtag Analysis (including Sentiments Behind Most Popular Hashtags).

2.1.Analysis of User Behavior

For this purpose, it is suitable to identify a username with the highest number of posts. According to my research, the account with the username "**mensfashions**" holds the record for the highest number of posts in this dataset, with over 30 posts.

```
mensfashions = non_locational_data.loc[non_locational_data["owner_username"] == "mensfashions"]
mensfashions.head(3)
```

	owner_id	owner_username	shortcode	is_video	caption	comments	likes	created_at	imageUrl	multiple_images	username	followers	following
1381	184378318	mensfashions	C3U-uVbt0SL	False	Suit up Inspo 🔥 by @artworth_brothers	5	2307	1.707914e+09	https://instagram.flhr13-1.fna.fbcdn.net/v/t51...	True	mensfashions	1957620.0	11.0

Analyzing the activities of this account can provide insights into its potential future actions. This account is particularly well-suited for this objective due to its large follower base (**approximately 1.9 million**) and low following count (**just 11 accounts**).

I plotted the number of likes on posts from the “menfashions” account over time, and the output is quite interesting. It reveals a lot about the integrity of our dataset.

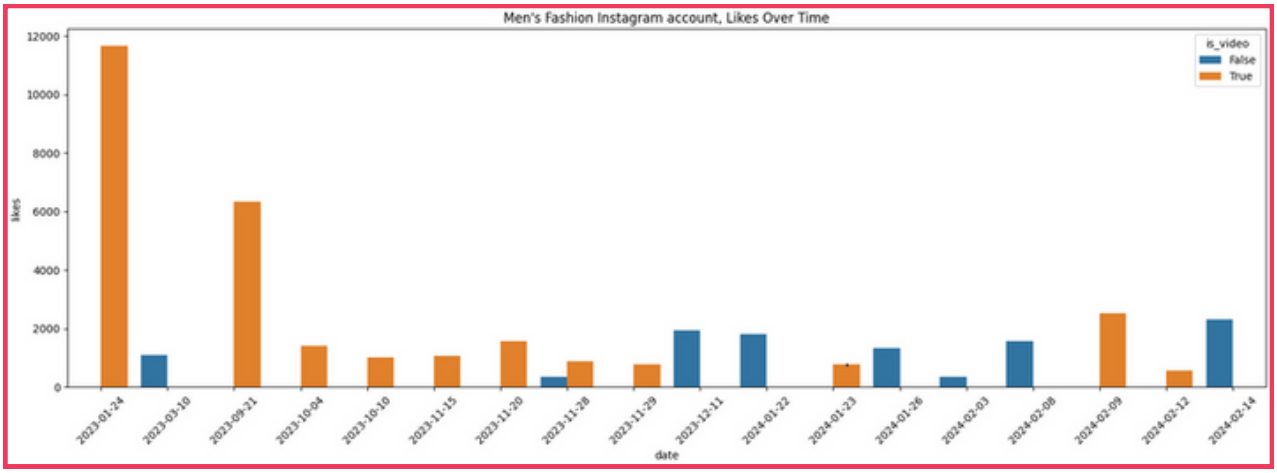


Diagram one

The first issue is the long time gaps between posts from this Instagram account. Upon checking the account directly, I noticed that some posts were missing from our dataset. This discrepancy was also observed when examining other accounts. Such discrepancies are impacting our ability to accurately identify trends and viral posts from this account.

The second issue concerns the low number of posts for each account. It's worth noting that the highest number of posts for any account in this dataset is 37, with the next highest number being 24 posts. In addition to the time gaps between posts, this dataset does not encompass an extensive timeframe for any single account. Furthermore, even within these time periods, not all posts from an account are included.

This incompleteness significantly impacts the data integrity of the dataset.

Wait a minute!

Let’s forget the Defects of this dataset and analyze what we saw in **diagram one**:

As you can see, the most liked content of this account is videos.

What are these posts about?

2023-01-24:
this is the link to the video: see thee video
The video is a humorous one, having gone viral prior to being reposted by this account. It features a man assisting a young, beautiful woman. Meanwhile, his own wife, who is also young and attractive, expresses jealousy and amusement at the situation involving the stranger woman.

2023-09-21:

this is the link to the video: see the video

The video depicts a young, flexible woman with blonde hair executing acrobatic moves to gracefully enter a car with closed doors and an open window.

It's interesting, isn't it?

Even though the page is dedicated to men's fashion, posts related to attractive women seem to garner the most attention, views and likes from the male audience.

2.2. Sentimental Analysis Based on location and time

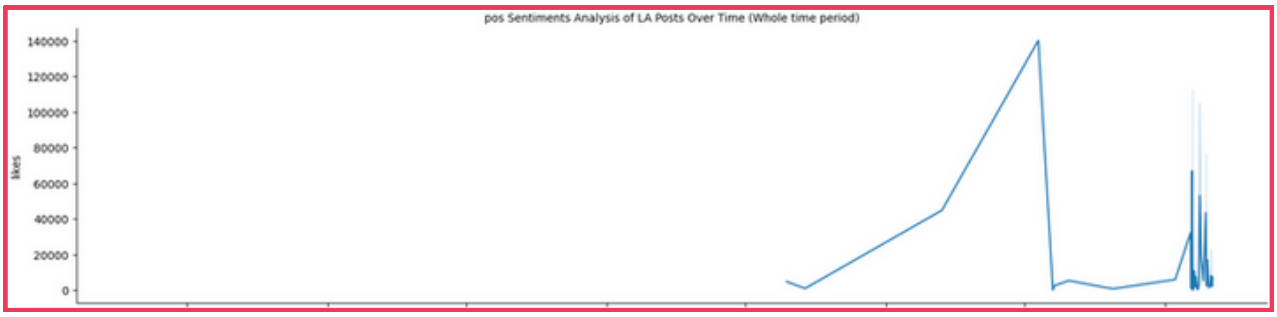
In this section, I analyzed the expressed emotions within a specific **geographical area**. I examined trends and changes in emotions over time within this area.
The initial step is to identify the location with the highest number of posts.

```
locational_data.groupby(by="location", as_index=False).agg({"shortcode":"count"}).sort_values(by="shortcode", ascending=False)
```

	location	shortcode
688	los-angeles-california	83
933	paris-france	78
762	milan-italy	67
848	new-york-new-york	31

We nailed it.

The location with the highest number of posts is LA, California, with a total of 83 posts. I segregated these posts into a new data frame. Next, I created a line plot to visualize the changes in each sentiment over time. This diagram also highlights the significant incompleteness and lack of integrity within the dataset.



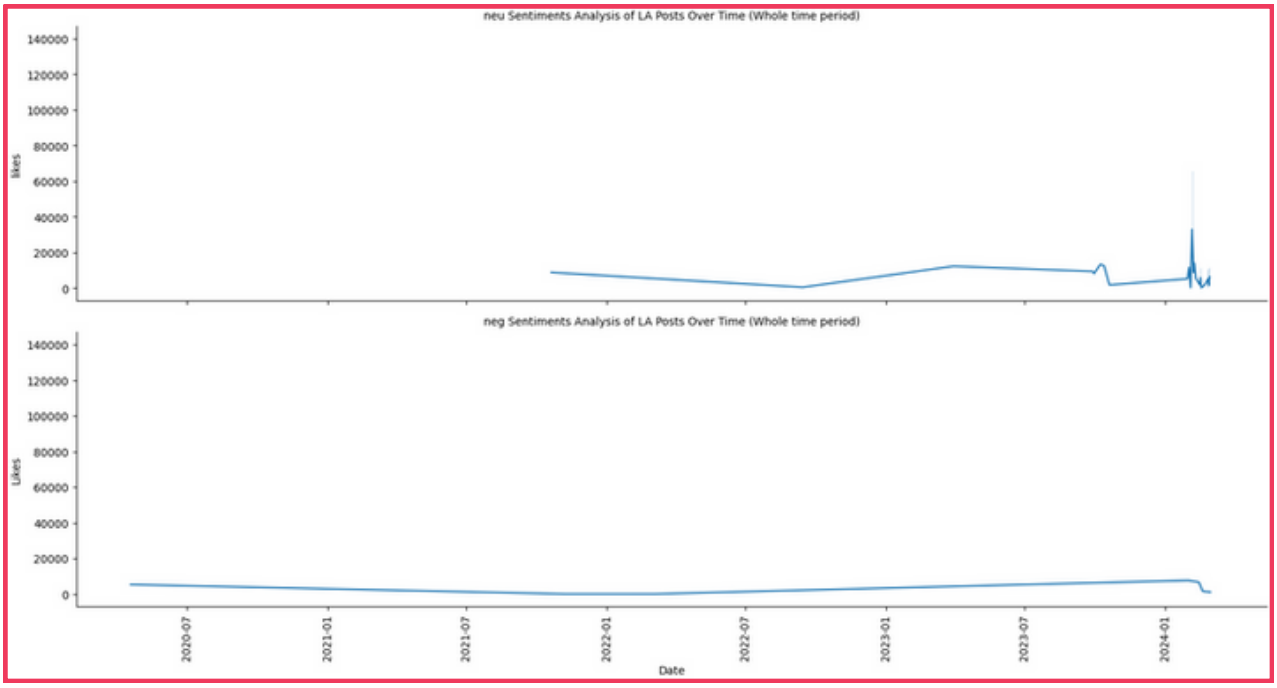


Diagram Two

There aren't sufficient data points available in the time interval from July 2020 to the first of January 2024. It would be more appropriate to draw a chart for the time interval from July 2023 to the present.

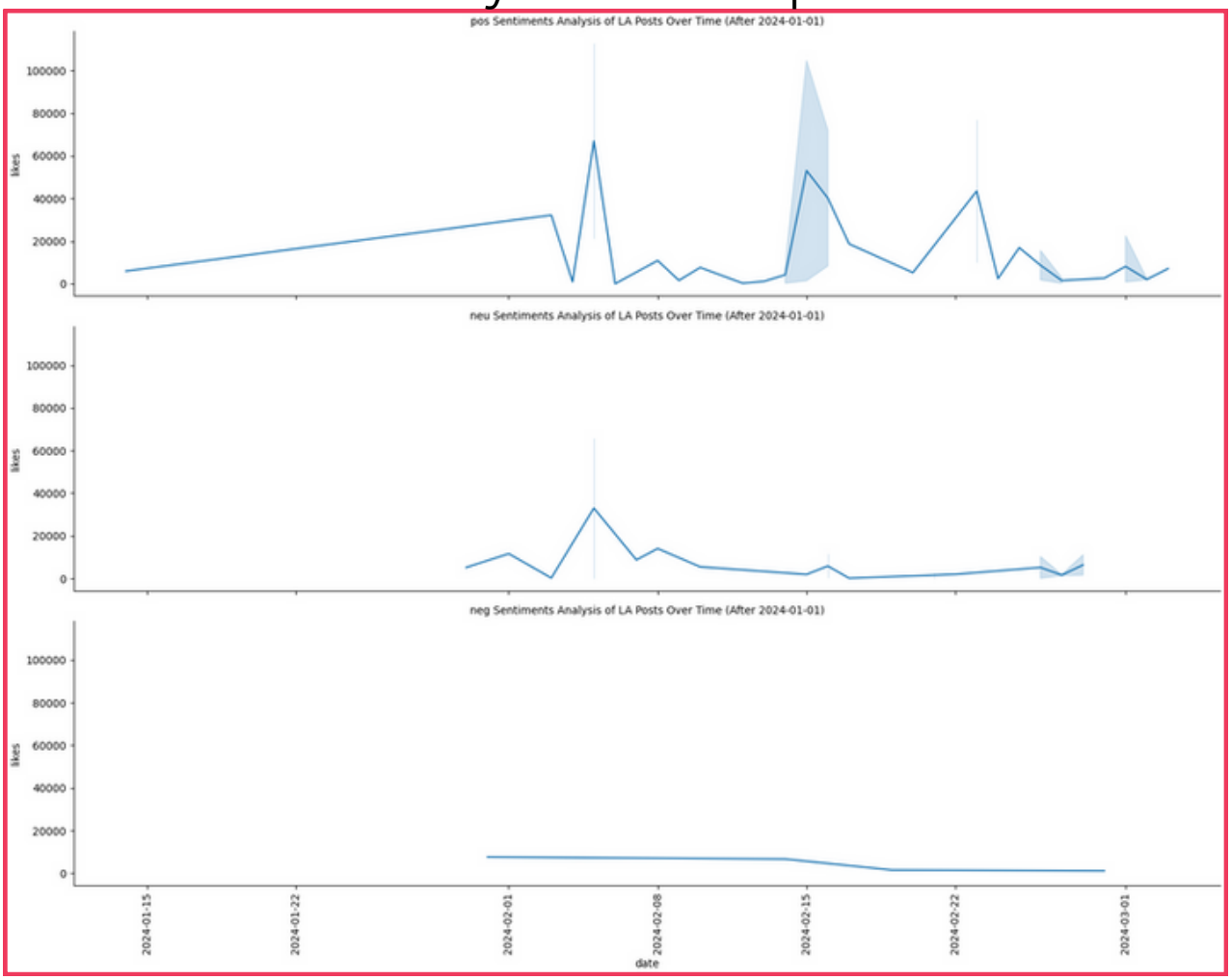


Diagram Three

Even within a narrow time interval, our dataset fails to provide a valid analysis of sentiments over a certain location. This limitation stems from a lack of sufficient data points.

The absence of data points is so pronounced that even when selecting posts with high engagement rates, there are almost none available for analysis.

However, if the dataset were to include an ample number of data points, it would enable us to fit a regression line over the time series of Diagram Two and Diagram Three.

3. Sentimental Analysis without considering location

After examining the behavior of users within a specific location, it's time to analyze sentiments across the globe and conduct a regression analysis to assess changes in expressed emotions worldwide over time.

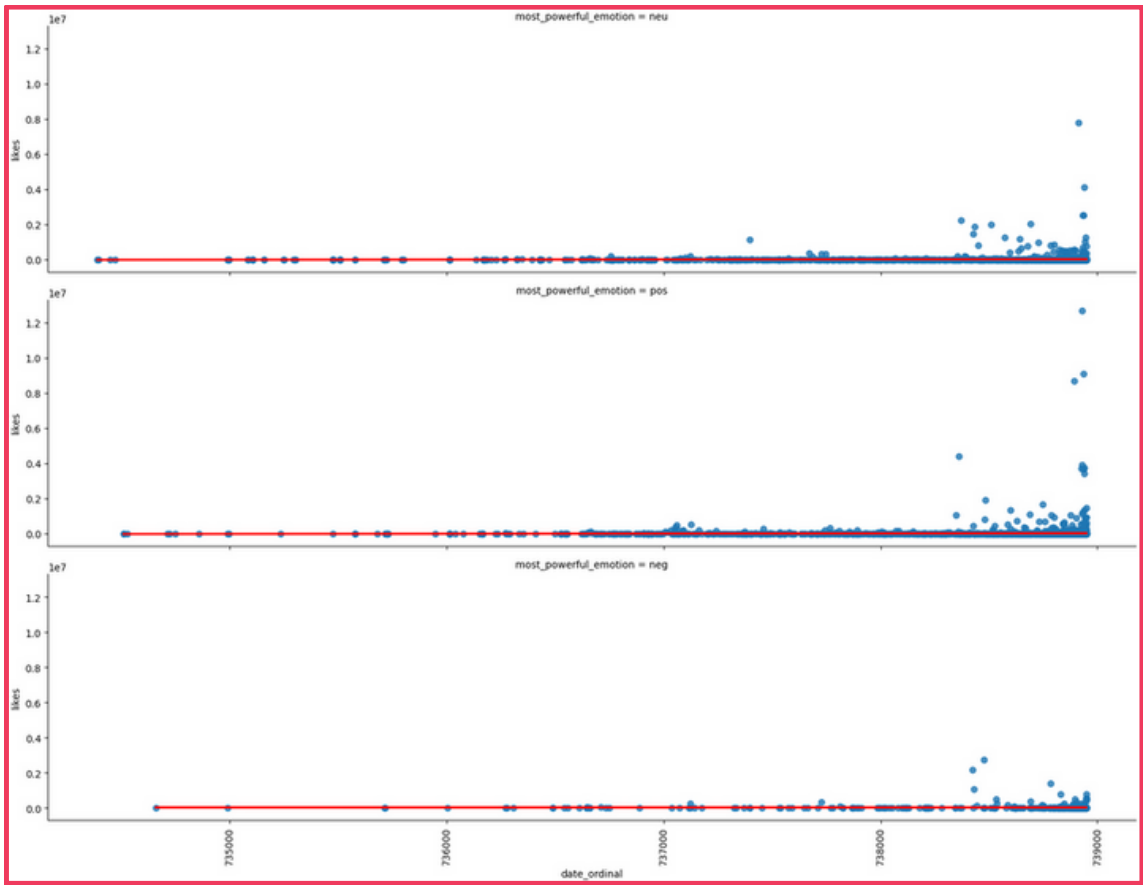


Diagram Four

Apparently, the condition of the dataset is suitable for this analysis, with the minimum required data points available to observe a trend. In this analysis, we can filter posts based on both time and Engagement Rate.

For the time interval, I selected posts posted after **2024-01-01** to ensure an adequate number of data points for a valid regression analysis. Additionally, I filtered posts with an engagement rate above 3.5. **Why 3.5?** Because for sentimental analysis aiming to identify the most dominant emotion across the population, it's essential to focus on posts with an engagement score exceeding 3.5.

By selecting data points meeting these criteria, we can generate a linear regression for each emotion:

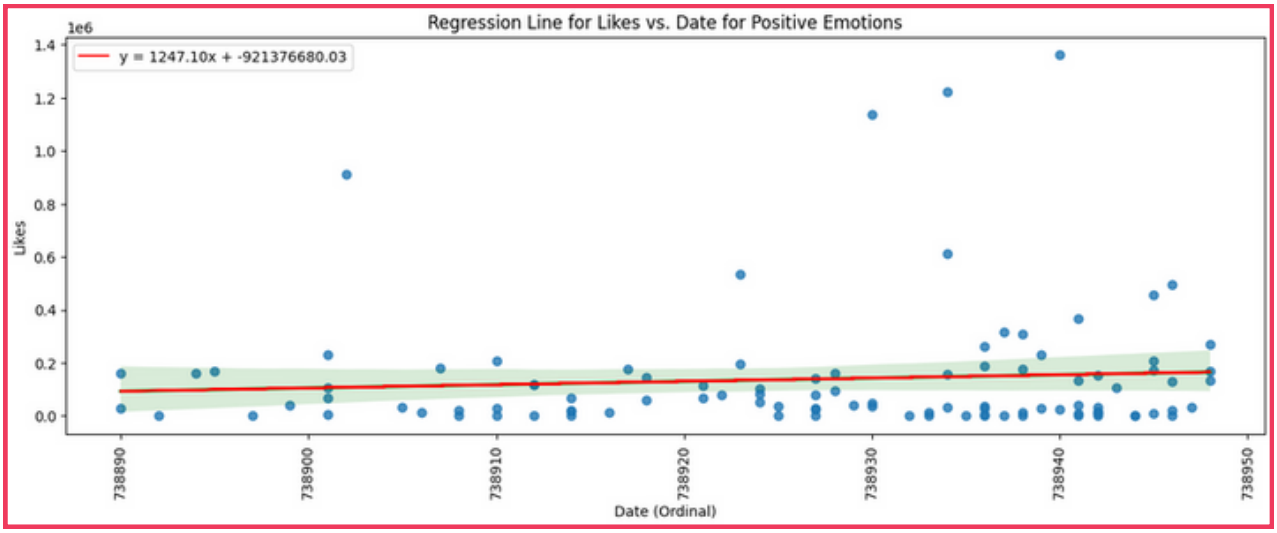


Diagram Five

The slope of the trend for **positive emotions** is **1247.1**. This indicates a notable increase in positive emotions among Instagram users.

Let's look at **negative emotions**:

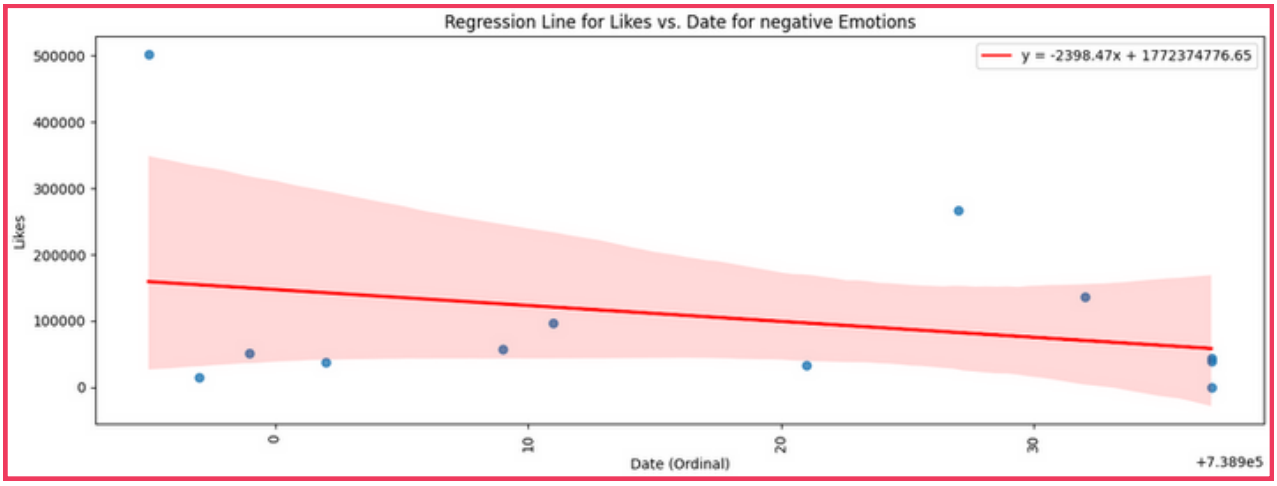


Diagram six

The slope of this line is **-2398.47**, indicating a significant decrease in negative emotion. This slope is much steeper than that of positive emotions, suggesting that the dominant emotion among Instagram users is **Positive**.

Let's look at **Neutral Emotions**:

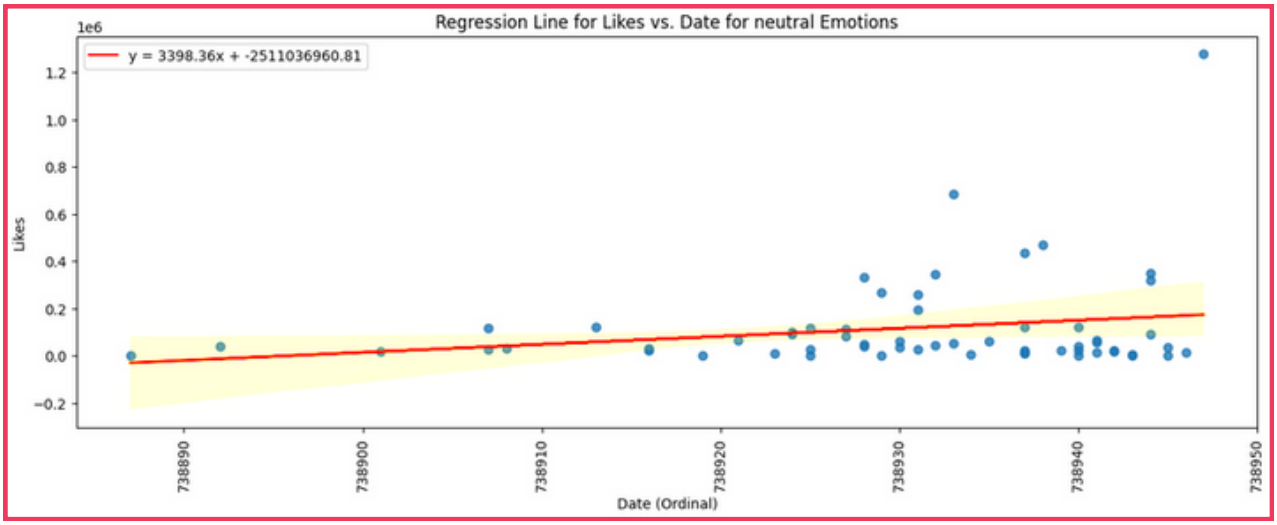


Diagram Seven

The slope of this trend line, **3398.36**, indicates an increase in **Neutral emotions**. This leads us to several conclusions:

1. Increasing number of Instagram users over time.
2. Increasing activities of Instagram users over time.
3. A combination of both conditions!

BUT WAIT A MINUTE!

These results are not based on a complete and integrated dataset. Therefore, if we continue collecting data, there is a reasonable chance of obtaining different slopes for each of these three lines.

4.HASHTAG ANALYSIS

Hashtags are used primarily on social media platforms to categorize content and make it easily discoverable by users interested in a particular topic or theme. When a hashtag is included in a post or comment, it becomes a clickable link that leads to a feed of other posts containing the same hashtag.

Analyzing related posts for a single hashtag or a combination of them allows us to predict trends, identify their causes, and analyze the dominant emotions associated with these hashtags.

In our dataset, I found that **#love** is the most frequently used hashtag, **appearing 79 times** in post captions. Following #love, **#ad** is the next most repeated hashtag, **appearing 56 times** throughout the dataset.

An interesting objective could be to analyze the emotions associated with each hashtag over time. For this purpose, linear regression proves to be a valuable tool.

4.1. #Love Sentiments over the time

Similar to the previous sections, we can draw three linear regression lines, with each regression representing a different emotion. By combining the slopes of these lines, we can discern the overall emotional trend for each hashtag.

4.1.1 Love - Positive

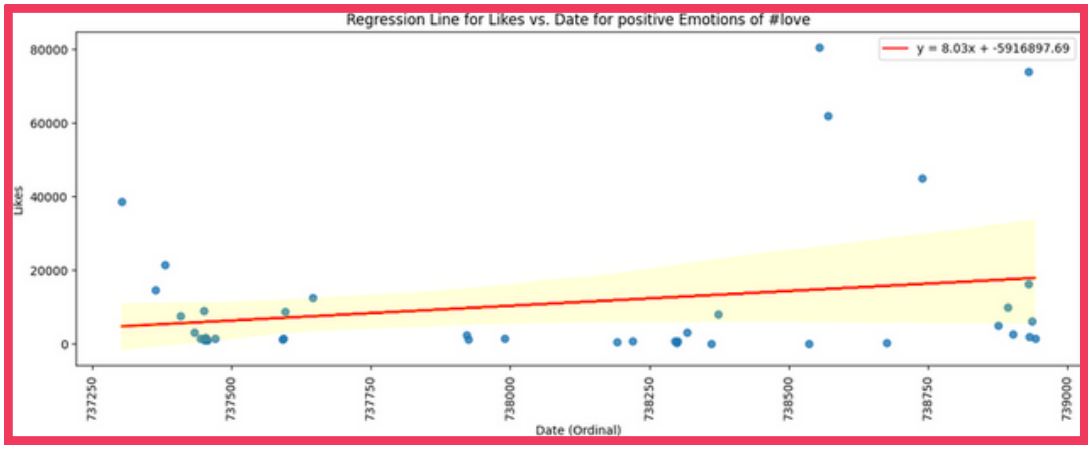


Diagram Eight

The number of likes on posts with the **#love** hashtag and positive emotions is increasing over time, but to draw conclusions, we must also consider the trends in the two other emotions.

4.1.2 Love - Negative

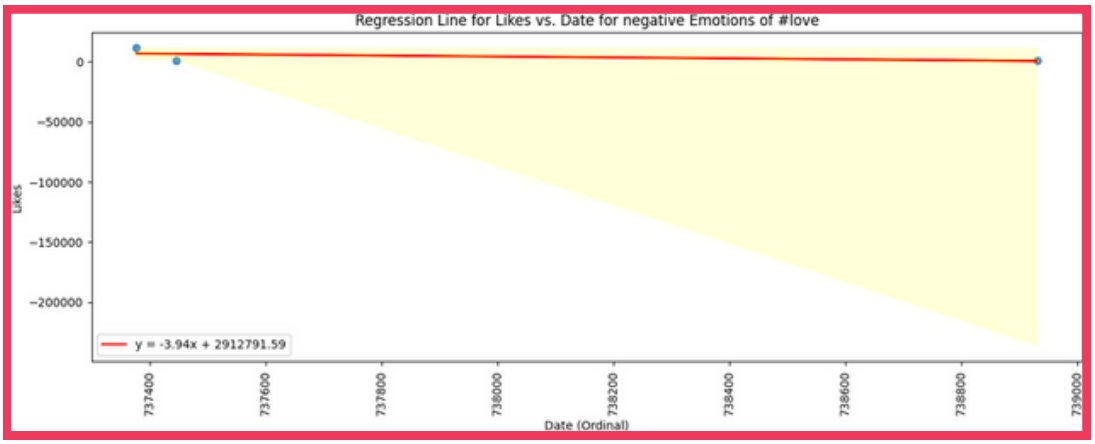


Diagram Nine

The slope indicates a negative trend, suggesting an increase in negative emotions over time in posts with the **#love** hashtag. However, it's crucial to note that this line may not accurately reflect the real-world situation, especially given that there are only three data points in this linear regression.

4.1.3 Love - Neutral

Diagram Ten

The decline in neutral emotions associated with the usage of the #love hashtag is evident. To draw any conclusions about this phenomenon, we must continue collecting data. Additionally, we need to examine other hashtags that are used alongside #love.

4.2. #ad Sentiments over the time

As mentioned previously, #ad has been used around 60 times. I have fitted three linear regression lines accordingly.

4.2.1 Ad - Positive

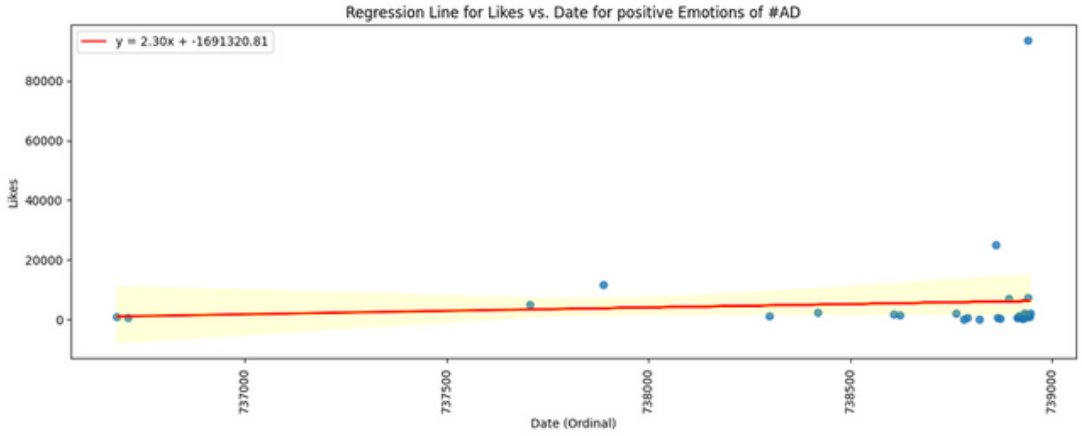


Diagram 11

4.2.2 Ad - Negative

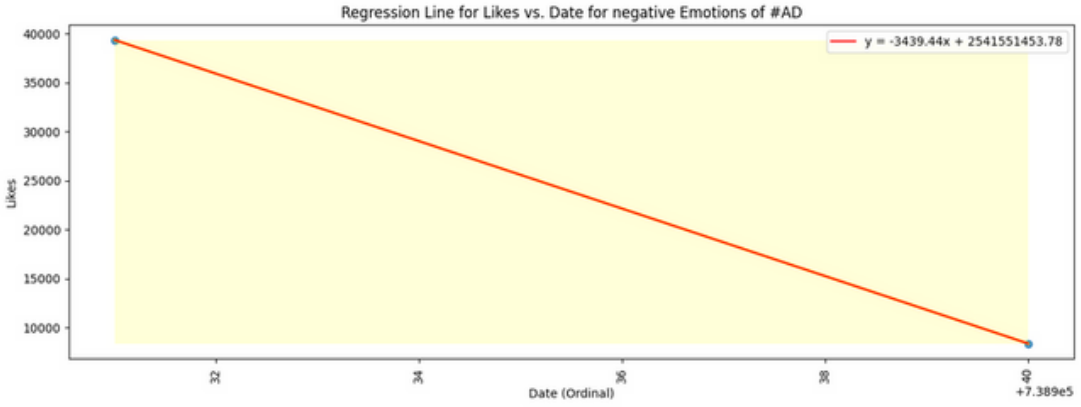


Diagram 12

4.2.3 Ad - Neutral

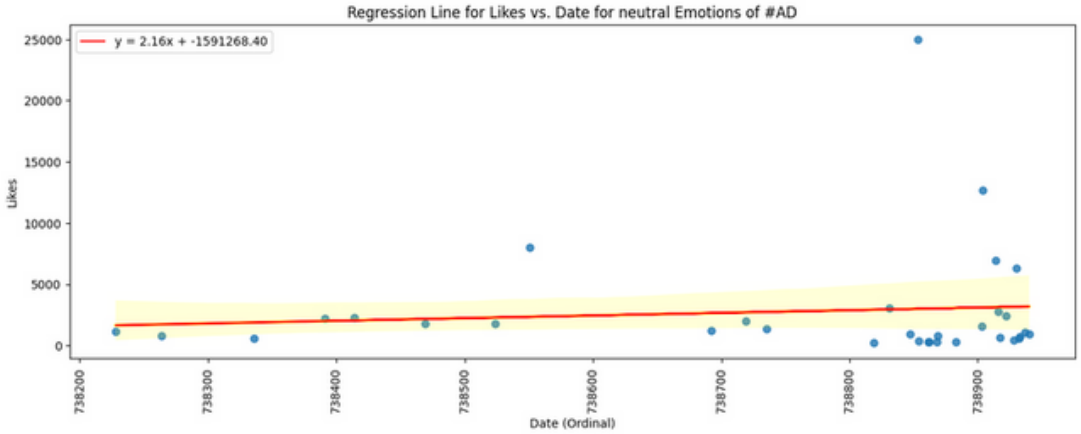


Diagram 13

5. conclusion

Although the dataset is currently not in optimal condition and lacks integrity and maturity, the overall concept is promising. Tracking posts from public accounts and analyzing their conditions based on the factors I've examined could provide valuable insights for making business decisions.