# Developer Sentiment Analysis (EDA)

## Project Overview

This project takes a closer look at over **4.6 million Stack Overflow posts** to figure out what really makes developers engage with a topic. The idea was simple — find out which kinds of discussions get people talking, and how the tone of a post (whether emotional or factual) changes how others react to it.

While working through the data, I focused on keeping it clean, scalable, and readable. The dataset included **35 tags and 40 topic areas**, so I spent time making sure everything was consistent before diving into analysis.

One clear pattern stood out — posts that felt **personal or emotional** (like “Teenage” or “Nature”) drew far more interest than neutral ones like “Work.” It shows that even in a technical space, people connect most when a post feels human.

The findings from this analysis can help **community teams, product owners, and analysts** understand what kind of content keeps people engaged and how emotional tone can shape user sentiment.

## Key Finding

The data made one thing very clear — **it’s not how often a topic appears that drives attention, but what it’s about and how it feels.**

Posts that show a bit of emotion or personality — like *“Teenage”* or *“Nature”* — perform much better than neutral or work-related ones. Even though “Work” was one of the most discussed tags, it didn’t connect with readers as much.

In short, **people respond to people** — not just information. When a topic sounds relatable, honest, or personal, engagement goes up naturally.

## Tools Used

I used Python and its core libraries — Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn — to clean, explore, and visualize the data. Since the dataset contained millions of rows, I focused on keeping the process fast, memory-efficient, and easy to scale.

**Python 3**  
- ***Pandas*** – to clean and work with data  
- ***NumPy*** – for number calculations  
- ***Matplotlib*** and ***Seaborn*** – to make charts  
- ***Scikit***-***learn*** (CountVectorizer) – to count and study words  
- ***Re*** (Regular Expressions) – to clean text, remove links and symbols

## Dataset

**File Name**: new\_tag\_data.csv (from Kaggle)  
**Total Posts**: 4,594,008  
**After Cleaning**: 4,376,084 unique posts (removed 217,924 duplicates)

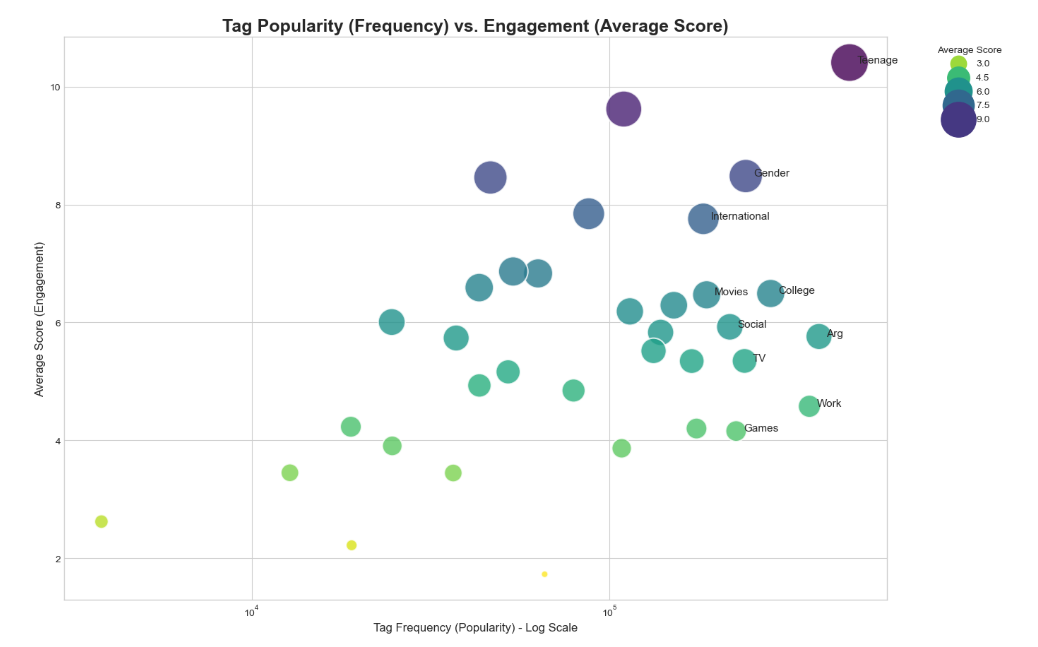
## Columns Studied

**score** – How much people liked the post  
**body** – The full text of the post  
**Topic** – A number showing the general topic (40 in total)  
**Tag** – Text label for the topic (35 in total)

## Main Insights

Before jumping into charts and numbers, here’s what I noticed while exploring patterns in the data. I compared how often topics appeared with how well they scored, and grouped them into categories based on their behavior. This made it easier to see which tags consistently attract attention — and which ones quietly fade into the background.

**Popularity vs. Engagement**  
A scatter chart compared how often tags appear vs. how engaging they are.

A) **Star Performers** (Popular + High Score): Teenage, Gender   
B) **General Chatter** (Popular + Low Score): Work, Arg, Games, Social, College   
C) **Hidden Gems** (Less Popular + High Score): Nature, Photography   
D) **Less Active** (Less Popular + Low Score): Hardware, Politics  
  
**Frequency vs. Engagement – The “Work” Tag**  
Posting a lot doesn’t mean people will like it. For example, “Work” is the 3rd most used tag but ranks 24th in score. “Teenage” is both the most common and most liked topic.  
  
Top 5 by Score:  
1. Teenage (10.38)  
2. Nature (9.60)  
3. Gender (8.40)  
4. Media (8.22)  
5. Treatment (7.90)  
Work Tag Score: 4.60 (Low)  
A graph of different colored rectangular shapes

AI-generated content may be incorrect.

A graph of a bar graph

AI-generated content may be incorrect.  
**Deep Dive on “Work” Tag**  
We checked 3.6 lakh posts with the “Work” tag. Common Words example just, like, people, time, know, work, job, company. These words are simple and practical, not emotional — that’s why the posts get lower scores.  
A graph of a graph

AI-generated content may be incorrect.

**Top 30 Words in All Posts**  
After removing links, symbols, and common filler words, the most used words were: just, like, dont, people, im, think, time, know, really, good. These words show that most posts are personal and written in a casual way.

A chart of a graph

AI-generated content may be incorrect.

## Business Uses

Here’s how these insights can be useful beyond just charts and analysis:

* **Growing communities:** Topics like *Nature* and *Photography* might not show up often, but when they do, people love them. That’s a strong signal for community managers to build more around them.
* **Improving engagement:** Posts with emotion or story-like tone connect better than purely factual ones. This helps content teams plan what kind of posts keep users active.
* **Content quality checks:** Emotional or sensitive tags like *Teenage* or *Gender* may attract a lot of responses, so they need careful moderation.
* **Better recommendations:** Instead of just showing similar tags, platforms can suggest posts that fall into the same engagement “cluster.”
* **Understanding language:** The way users talk about topics like “Work” or “Treatment” gives useful hints about what motivates or frustrates them.

## Steps Followed

***Step 1: Load and Clean Data***  
- Opened dataset with 4.6M posts  
- Removed 217K duplicate posts  
- Fixed data types and tag names  
- Reduced memory size for faster work  
  
***Step 2: Explore Data (EDA)***  
- Looked at how scores and tags are spread  
- Found which tags get most attention  
- Made scatterplots to compare popularity and engagement

***Step 3: Text and Relationship Study***  
- Used heatmaps to show link between Topics and Tags  
- Cleaned and analyzed words  
- Studied “Work” tag deeply to find reasons for low engagement

## How to Run the Project

***1. Clone the Repository***  
gitclone *https://github.com/analytics-ak/developer-sentiment.git*  
cd *developer-sentiment*  
  
***2. Install Tools***  
pip install *pandas* *numpy* *matplotlib* *seaborn* *scikit*-*learn*  
  
***3. Run Notebook***  
jupyter notebook "*developer-sentiment.ipynb*"

## Expected Outcome

This project looks at over **4.3 million Stack Overflow posts** to find what kind of topics get people most engaged.

You’ll see clear insights on which tags attract more attention — especially how **emotional topics** compare to **neutral ones** like “Work.”

Simple charts and summaries make it easy to spot which topics are popular, hidden gems, or not doing well.

The whole process also shows how Python can handle big text data smoothly and efficiently.

## Limitations

1. The data covers around **4.6 million posts** collected at one point in time, so newer trends may differ.
2. Only **English-language posts** were analysed.
3. Some cleaning steps might have removed useful posts or kept a few duplicates.
4. Emotion detection is based on common word use, so it may miss subtle tones or context.
5. The analysis finds **patterns and links**, not cause-and-effect relationships.

## Result Summary

In total, the analysis covered over 4.3 million unique posts and identified clear engagement patterns across 35 tags. Emotional or personal tags consistently scored 2–3× higher than neutral ones like “Work.”

## Final Summary

This project was more than just running numbers — it was about understanding how **real people interact** in a technical space. After cleaning and studying millions of posts, one pattern kept repeating: **emotion builds connection.**

Posts that sound real, relatable, or even a little personal tend to get noticed more. It’s a reminder that even on platforms built for technology, **human tone still wins.**

The work also shows how large datasets can reveal social behaviour at scale — something every community or product team can use to improve engagement and user experience. For me, it’s a clear example of how **data analysis can tell human stories** when you look beyond just numbers.

## Links

[**GitHub Repository:**](https://github.com/analytics-ak/developer-sentiment) https://github.com/analytics-ak/developer-sentiment  
[**Notebook:**](https://github.com/analytics-ak/developer-sentiment/blob/main/developer-sentiment.ipynb) developer-sentiment.ipynb  
[**Dataset: Kaggle Developer Sentiment Dataset**](https://www.kaggle.com/datasets/vaibhavsxn/reddit-comments-labeled-data)  
[**LinkedIn: Ashish Kumar Dongre**](https://www.linkedin.com/in/analytics-ashish/)

Created and maintained by Ashish Kumar Dongare — Senior Data Analyst