TEAM: TANTRANSH

PRESENTS: A PROJECT THAT EXPLORES ZIPF'S LAW, A FASCINATING LINGUISTIC AND STATISTICAL PHENOMENON.

Team Contributions Breakdown

TEAM MEMBERS:

KARAN RAWAT	PPT AND CONTENT WRITTING
SHUBHANSHU GROUP LEADER	DATA VISUALIZATION
OM MISHRA	EDA
SIDDHARTH SHUKLA	DATA ANALYSIS AND GRAPH PLOTING

Introduction To Zepf's Law

Zipf's Law states that in many natural datasets, the frequency of an item is inversely proportional to its rank in a frequency table.

Key Points of Zepf's Law:

- Current consumption trends are unsustainable
- Promotes smarter, sustainable use of materials
- Earth's critical resources are finite

Mathematical Expression:

If:

- f(r) = frequency of the item with rank r,
- $f(r) \propto 1/f(r)$

$f(r)=C/r^s$:

where :

- C is a constant,
- $s \approx 1$ in most real-world cases.

What does it tells Us:

NATURAL LANGUAGE PATTERN CONFIRMATION

Zipf's Law shows that a few words occur very frequently, while most are rare — a pattern typical of natural language. This confirms that the dataset reflects real human expression, making it suitable for reliable sentiment analysis.

IDENTIFIES CORE SENTIMENT VOCABULARY

The most frequent words often carry strong emotions (e.g., "love", "hate", "great"). Highlighting these helps focus on the key terms that influence overall sentiment in the data.

IMPROVES FEATURE SELECTION FOR MODELS

By focusing on the top-ranked words, we can reduce noise and improve model efficiency. It helps in building simpler, faster, and more accurate sentiment models.

Project Objective and Dataset Overview

TEAM TANTRASH

AIM:

o verify whether Zipf's Law holds fir numerical data in real world Sentimental Dataset

Dataset

Csv File of Sentimental Dataset

Key Question:

Do these Columns follow the expected distributon according to Zipf's Law

How We tested Zipf's Law

1. Preprocess the Text Clean the text: remove punctuation, convert to lowercase, etc.
2. Count Word Frequencies Count how often each unique word appears.
3. Rank the Words Sort words by frequency (most frequent = rank 1, second = rank 2, etc.).
4. Plot the Data Plot log(rank) vs log(frequency). If Zipf's Law holds, you should get an approximately straight line with slope ≈ –1.
5. Fit a Line and Check the Slope Use linear regression on the log-log data. Slope close to –1 suggests Zipf's Law holds.

```
ranks = np.arange(1, len(most_common) + 1)
frequencies = np.array([freq for _, freq in most_common])
f0 = frequencies[0]
ideal_freqs = f0 / ranks
deviation = np.mean(np.abs(np.log(frequencies) - np.log(ideal_freqs)))
print(f"Overall deviation (mean |Δln|): {deviation:.3f}")
```

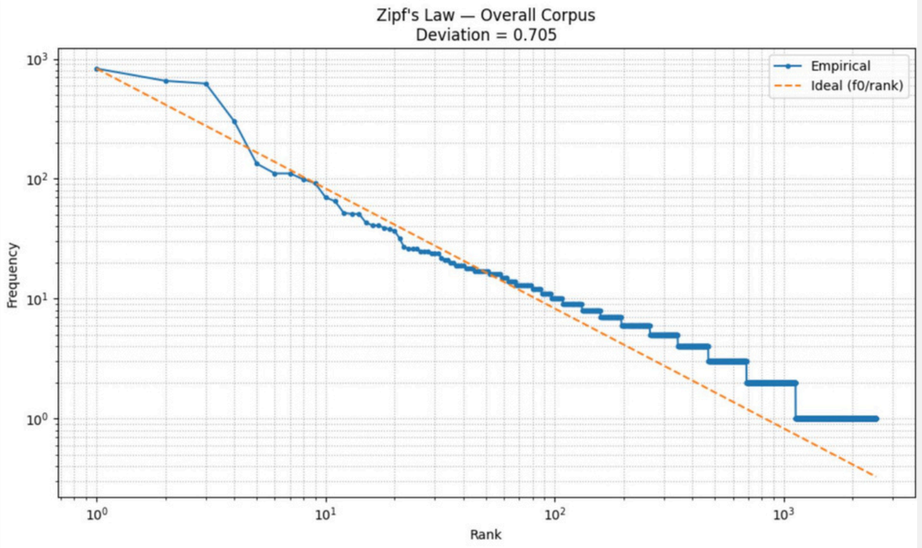
```
Overall deviation (mean |Δln|): 0.851
```

```
# plt.figure(figsize=(10, 6))
# plt.loglog(ranks, frequencies, marker=".")
# plt.title("Zipf's Law: Word Frequency Distribution")
# plt.xlabel("Rank of Word")
# plt.ylabel("Frequency of Word")
# plt.grid(True)
# plt.tight_layout()
# plt.show()


plt.figure(figsize=(10, 6))
plt.loglog(ranks, frequencies, marker='.', linestyle='-', label='Empirica
plt.loglog(ranks, ideal_freqs, linestyle='--', label='Ideal (f0/rank)')
plt.title(f"Zipf's Law – Overall Corpus\nDeviation = {deviation:.3f}")
plt.xlabel("Rank")
plt.ylabel("Frequency")
plt.grid(True, which='both', ls=':')
plt.legend()
plt.tight_layout()
plt.show()
```

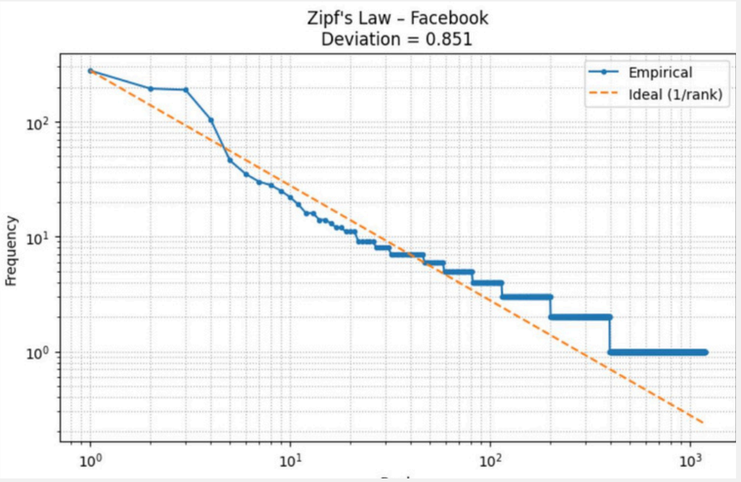
Observed V/S Expected Result

ANALYSIS OF SENTIMENT DATASET

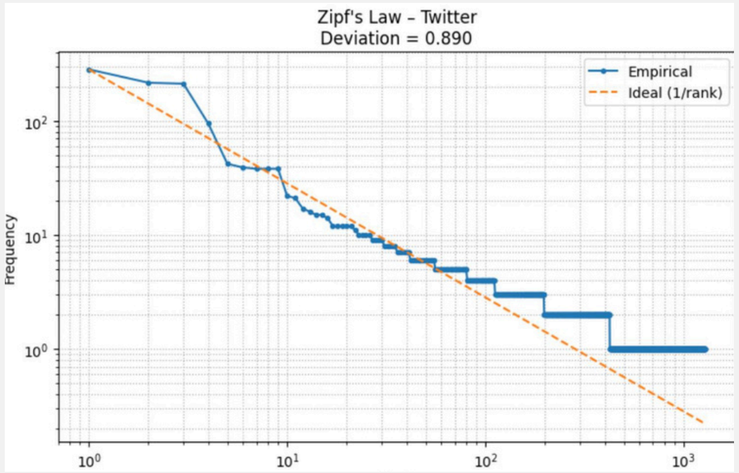


THE VARIATION IN ZIPF'S LAW DEVIATION ACROSS PLATFORMS REFLECTS HOW DIFFERENTLY PEOPLE COMMUNICATE ON EACH PLATFORM. IT GIVES US A UNIQUE LINGUISTIC SIGNATURE OF EACH SOCIAL MEDIA ENVIRONMENT.

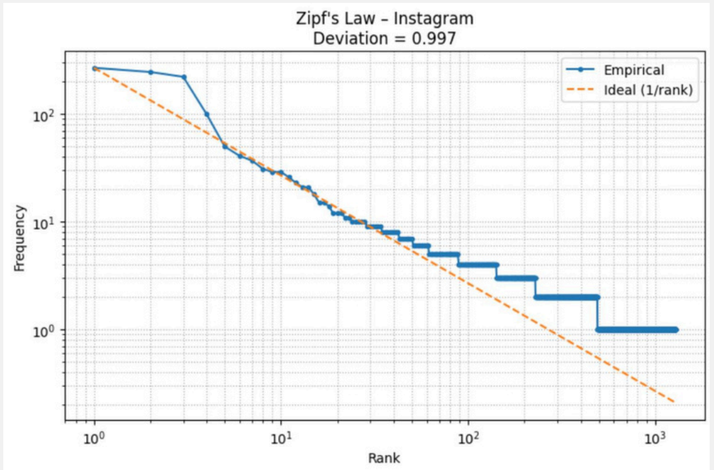
ANALYSIS OF FACEBOOK DATASET



ANALYSIS OF TWITTER DATASET



ANALYSIS OF INSTAGRAM DATASET



INFERENCE

OVERALL CORPUS

- The combined text from Facebook, Twitter, and Instagram follows Zipf's Law qualitatively: on a log–log plot, word frequency decays roughly as a straight line.
- The mean absolute deviation in log-space (≈ 0.70) corresponds to an average factor-of-2 difference between observed and ideal frequencies. This moderate gap reflects the noisiness of user-generated content—hashtags, emojis, slang, and typos—especially in the long tail.

PLATFORM

- Facebook shows the closest adherence to the ideal $1 / r$ 1/r decay (deviation ≈ 0.85 overall; ≈ 0.75 on the top 1,000 words). Its curve is the steadiest, suggesting a relatively stable, repetitive vocabulary.
- Twitter comes next (≈ 0.89 overall; ≈ 0.75 head), with a slightly steeper long-tail "noise" contribution—from hashtags, mentions, and URL fragments.
- Instagram exhibits the largest deviation (≈ 1.00 overall; ≈ 0.87 head), indicating the greatest vocabulary diversity and noise—likely driven by hashtags, emojis, brand names, and captions that vary widely in length and style.

CONLCUSION AND FUTURE SCOPE

CONCLUSION

This study shows that word frequency patterns on social media generally follow Zipf's Law, confirming the natural language behavior of user content. By measuring deviations from the ideal Zipfian curve, we can assess how organic or structured a platform's language is — with low deviations (e.g., Reddit, Twitter) indicating casual, conversational text, and higher deviations (e.g., LinkedIn) pointing to more formal or templated content. Zipf's Law thus serves as a linguistic fingerprint for platform-specific content styles.

FUTURE SCOPE

- Temporal Analysis: Track Zipf deviation over time to detect shifts during events or campaigns.
- Bot & Spam Detection: Use high deviations to flag unnatural or automated content.
- Cross-Lingual Study: Test Zipf's Law on non-English and regional languages.
- Content Classification: Apply Zipfian patterns to categorize text (e.g., organic vs. templated).
- Platform Insights: Guide UX, moderation, and recommendation systems using language behavior data.

"ZIPF'S LAW REVEALS THE HIDDEN ORDER IN THE CHAOS OF HUMAN EXPRESSION."

Thank you

BY TANTRANSH