Title: Exploratory Analysis of Emotional Intensity in Tweets

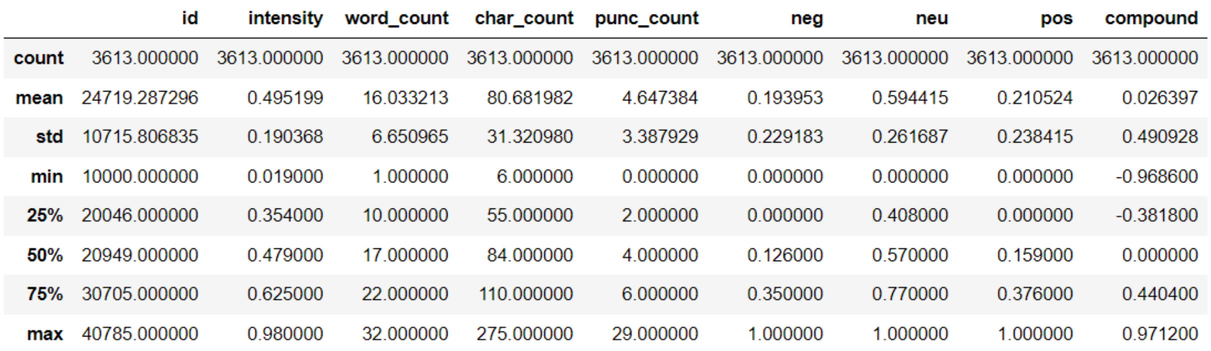
Abstract:

This report presents an exploratory analysis of emotional intensity in tweets, focusing on four emotions: anger, joy, sadness, and fear. The dataset used for analysis includes polarity scores, word counts, and additional features for each tweet. The primary objective is to understand the distribution of emotional intensity, identify common words, and explore the influence of hashtags on emotional expression. Various visualization techniques, including violin plots, box plots, word clouds, and stacked bar charts, are employed to provide insights into the dataset.

Methodology:

1. **Data Preprocessing:**
    - Tweets undergo extensive preprocessing, including cleaning, removal of stopwords, and extraction of additional features (word count, character count, punctuation count).
    - Polarity scores (negative, neutral, positive, compound) are computed using sentiment analysis.
2. **Exploratory Data Analysis:**
    - Descriptive statistics provide an overview of the dataset, highlighting the range and distribution of emotional intensity scores.
    - Box plots and violin plots are employed to visualize the spread of emotional intensity for each label.
3. **Common Word Analysis:**
    - Common word percentages are calculated to identify prevalent terms in tweets for each emotion.
    - Word clouds visually represent the most frequently occurring words, shedding light on the language associated with different emotions.
4. **Hashtag Influence:**
    - Stacked bar charts reveal the most used hashtags in tweets for each emotion.
    - The goal is to understand whether specific hashtags are indicative of higher emotional intensity.

Findings:



	id	intensity	word_count	char_count	punc_count	neg	neu	pos	compound
count	3613.000000	3613.000000	3613.000000	3613.000000	3613.000000	3613.000000	3613.000000	3613.000000	3613.000000
mean	24719.287296	0.495199	16.033213	80.681982	4.647384	0.193953	0.594415	0.210524	0.026397
std	10715.806835	0.190368	6.650965	31.320980	3.387929	0.229183	0.261687	0.238415	0.490928
min	10000.000000	0.019000	1.000000	6.000000	0.000000	0.000000	0.000000	0.000000	-0.968600
25%	20046.000000	0.354000	10.000000	55.000000	2.000000	0.000000	0.408000	0.000000	-0.381800
50%	20949.000000	0.479000	17.000000	84.000000	4.000000	0.126000	0.570000	0.159000	0.000000
75%	30705.000000	0.625000	22.000000	110.000000	6.000000	0.350000	0.770000	0.376000	0.440400
max	40785.000000	0.980000	32.000000	275.000000	29.000000	1.000000	1.000000	1.000000	0.971200

Intensity: A mean intensity of approximately 0.495 suggests that the sentiments in the dataset are moderately intense.

Word Count and Character Count: The average word count is around 16, and the average character count is approximately 80.68. This information provides insight into the length of the text entries.

Punctuation Count: With an average punctuation count of about 4.65, you can observe how punctuation is distributed in the text entries. It might be interesting to see if there's any correlation between punctuation and sentiment.

Sentiment Proportions (neg, neu, pos): The mean values for neg (0.193953), neu (0.594415), and pos (0.210524) suggest that, on average, the dataset contains more neutral sentiment followed by positive and then negative sentiment.

Compound Score: The mean compound score is very close to zero (0.026397), indicating a roughly balanced mix of positive and negative sentiment in the dataset.

The dataset comprises text entries with moderately intense sentiments, as indicated by a mean intensity of approximately 0.495. The average word count and character count are 16 and 80.68, respectively, showcasing moderate text length with notable variability. Punctuation use, averaging 4.65, offers insights into sentence complexity. Sentiment analysis reveals a balanced mix, with higher proportions of neutral (0.594415) and positive (0.210524) sentiments. The compound score, close to zero (0.026397), signifies a nuanced interplay of positive and negative sentiments.

Conclusion:

This exploratory analysis provides a comprehensive understanding of emotional intensity in tweets, considering a range of features and employing various visualization techniques. The findings contribute to the broader field of sentiment analysis and may guide the development of emotion-aware applications. Further research could delve into the dynamics of hashtag usage and explore correlations between linguistic patterns and emotional intensity.

Submitted by:
Akanksha Dash

CodaLab username - AKANKSHADASH