Training AI on Online Posts to Detect Stolen Car Parts:

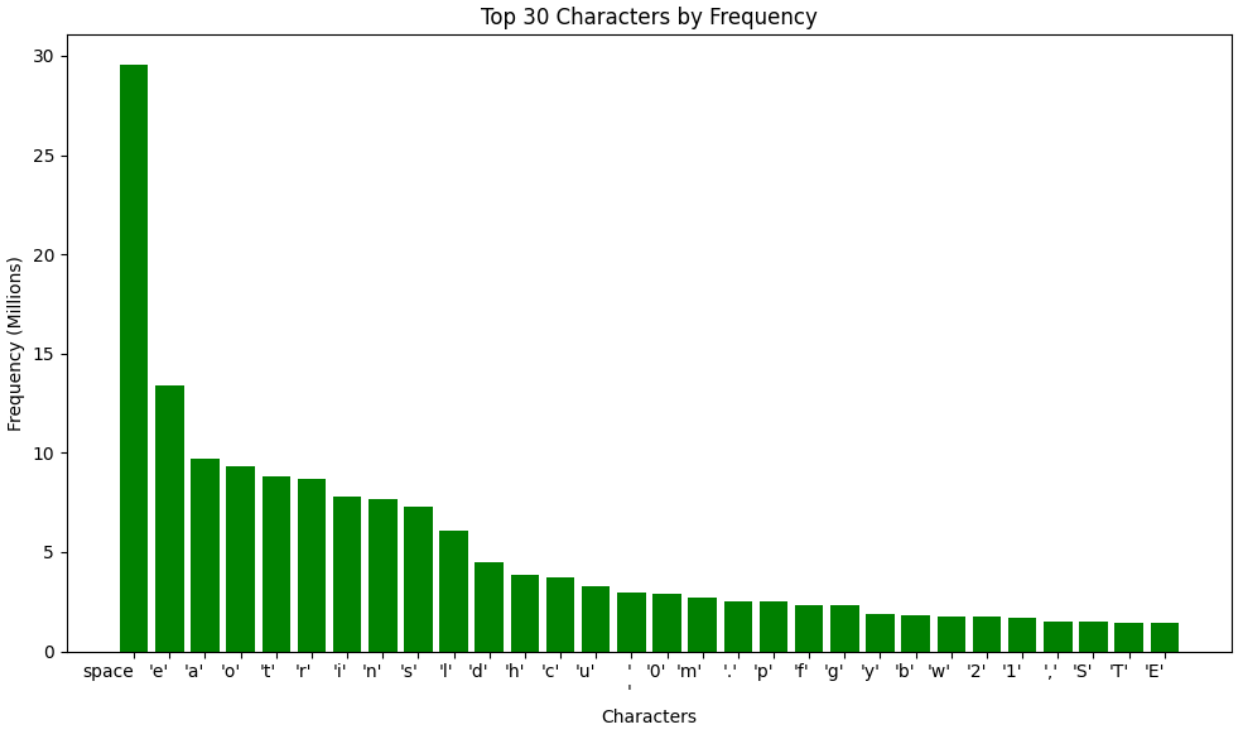
Research Question and Data Visualizations

Principal Investigator: Dr. Pablo Rivas

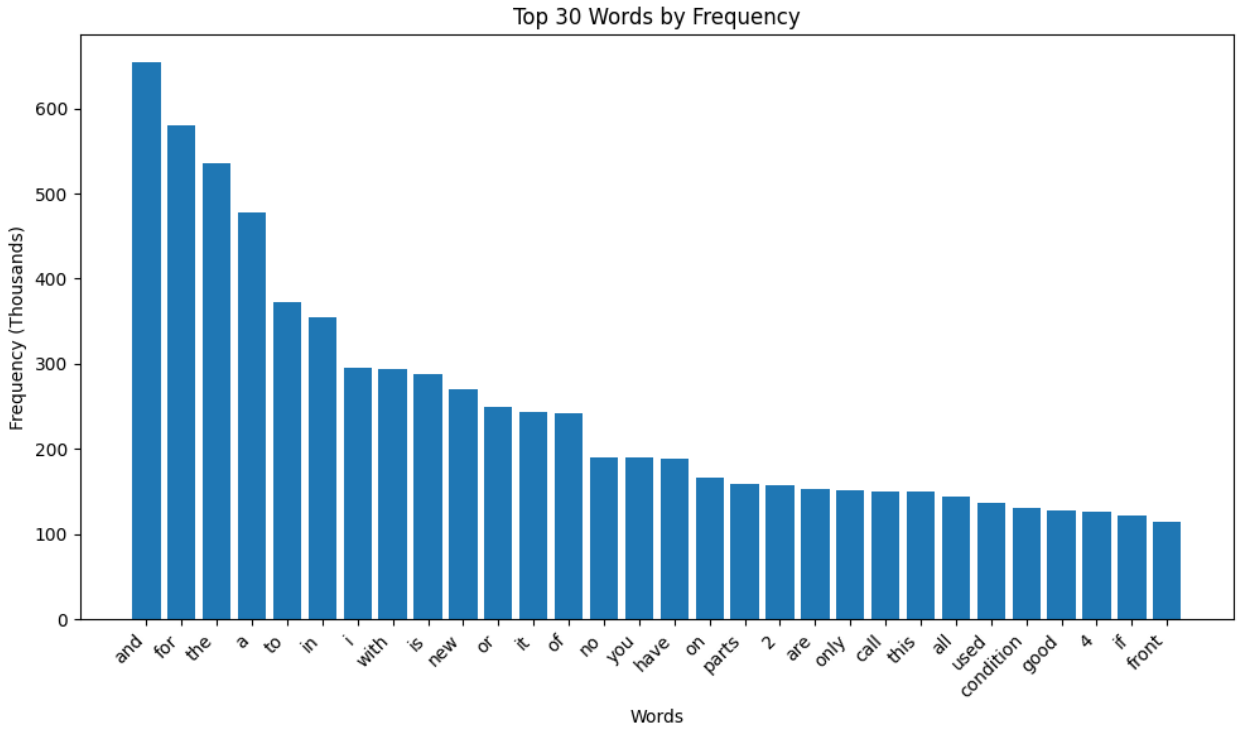
Harmond Drenth, Kurt Wokoek, Warren Burrus, Courtney Hodge, Denny Lee, and Tacoma Velez

Baylor University

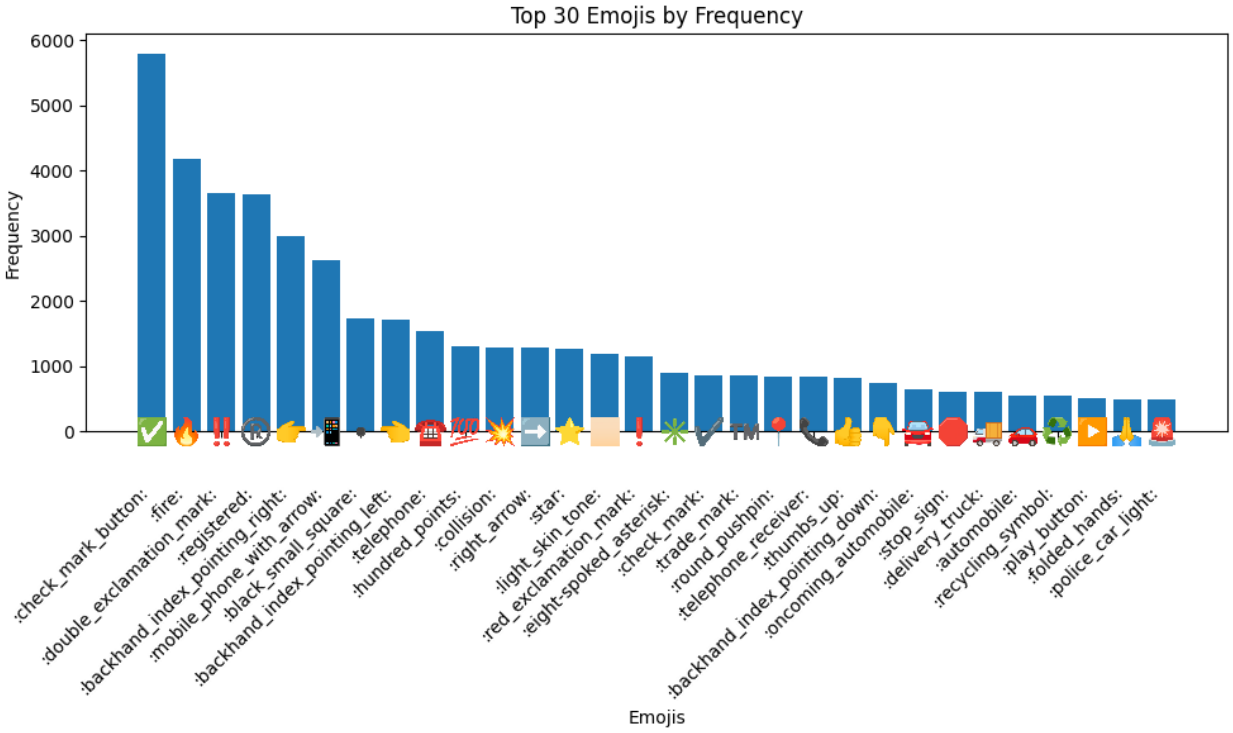
What are the most frequent characters used in the ad posts?



What are the most frequent words used in the ad posts?

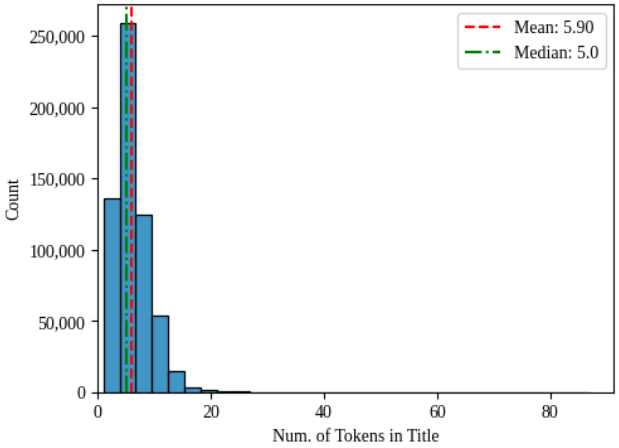


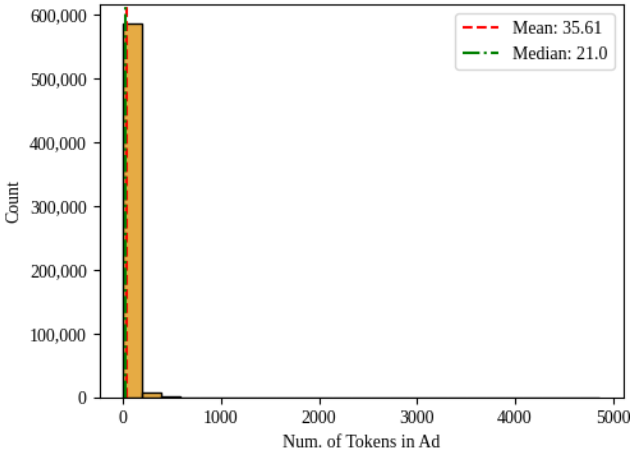
What are the most common emojis in posts?

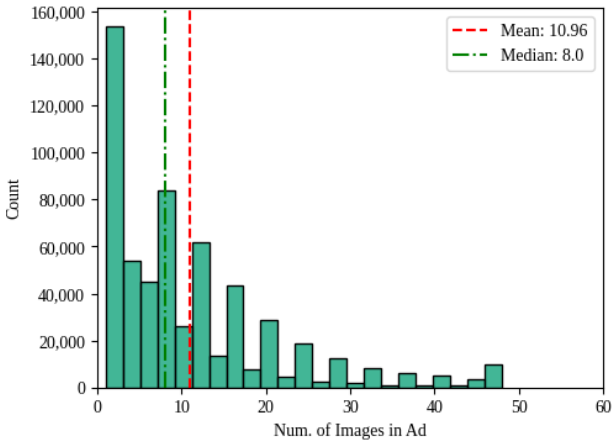


What is the distribution of post lengths?

* Tokens in this instance refer to words

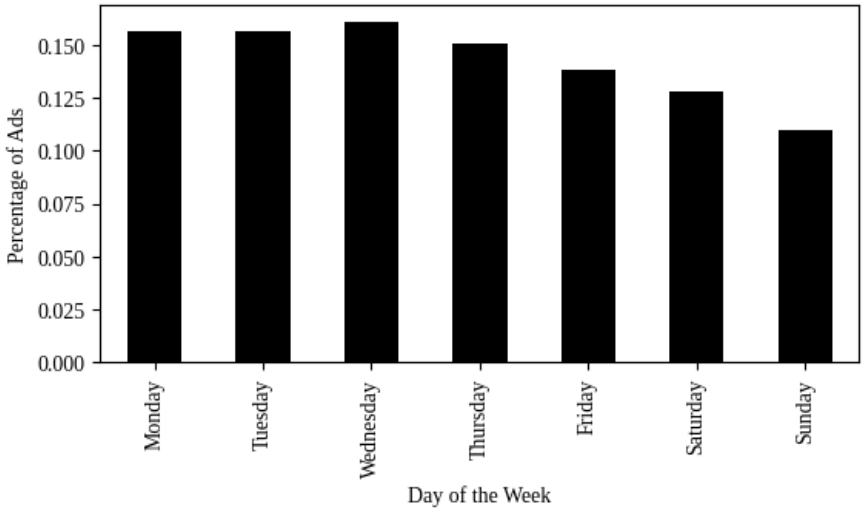


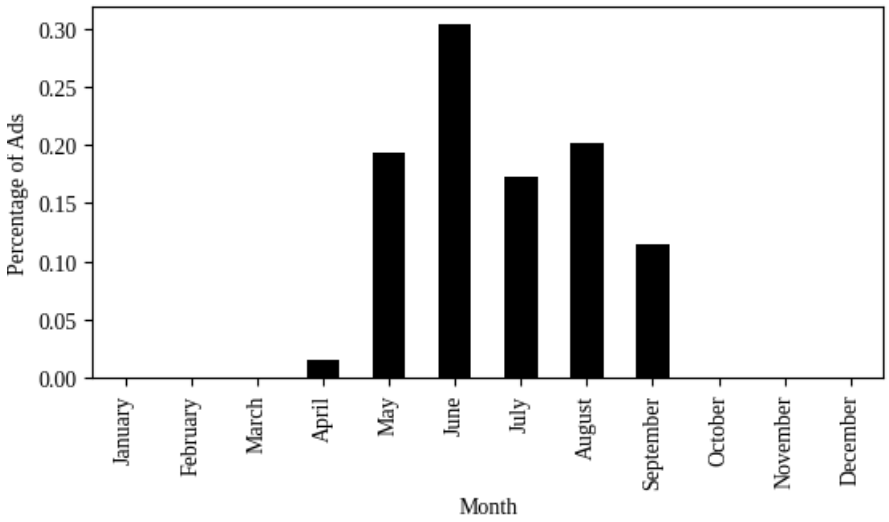




Are there any patterns in the day of the week, or month of the year when the ads are posted?

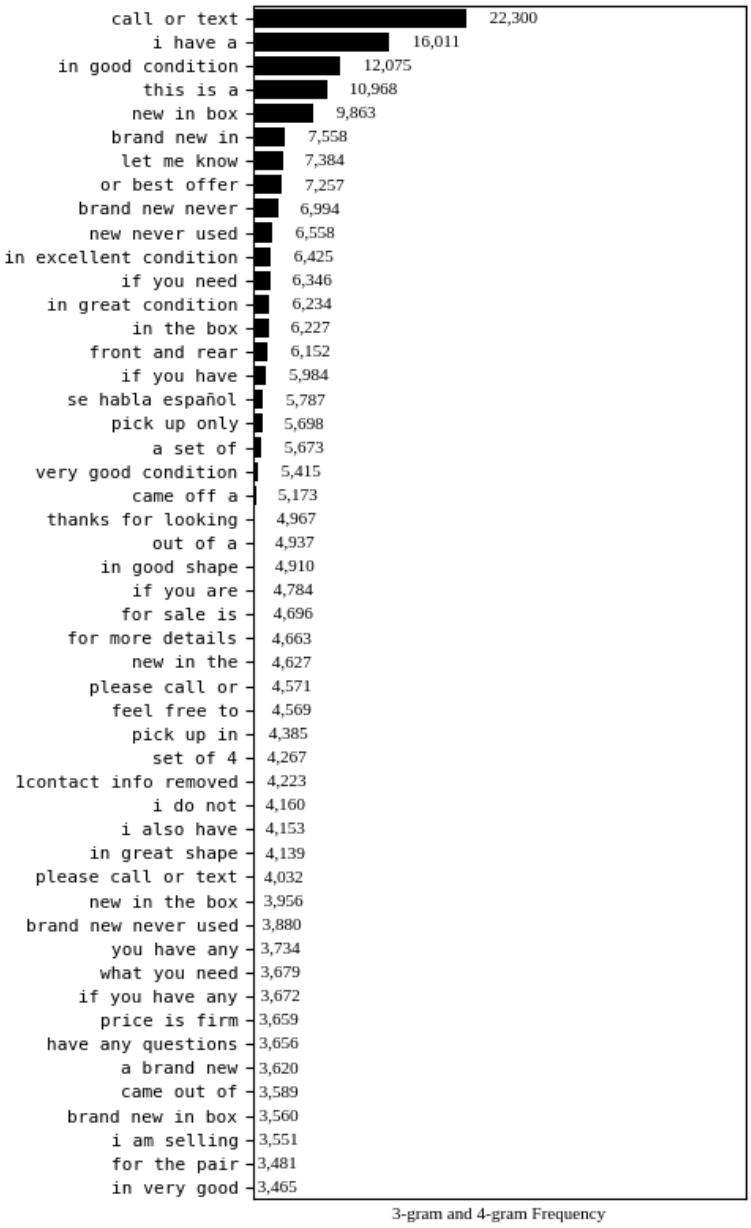
* Sample data used for visualizations does not contain posts from January, February, October, November, December





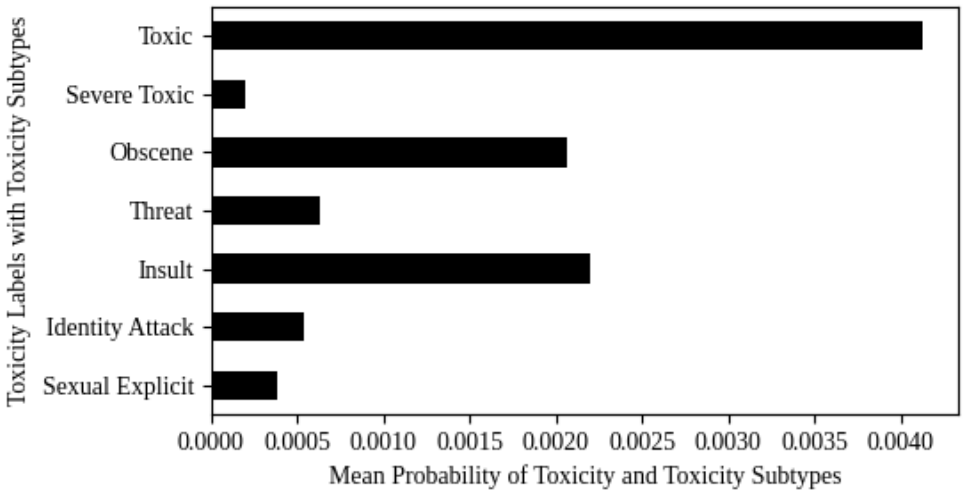
What are the most common phrases used in the titles and posts of the ads?

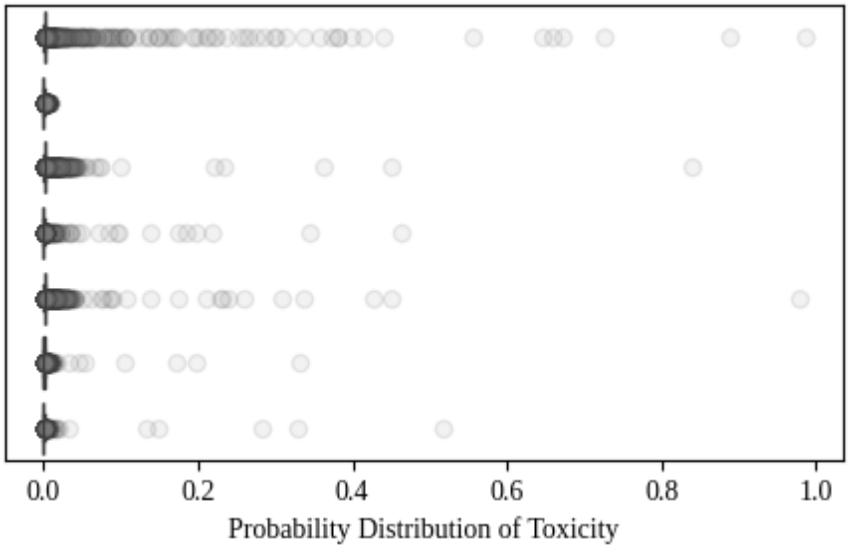
* 3 and 4-gram frequencies represent 3 and 4 word phrases in this instance of titles and posts



Can we quantify the level of toxicity in text?

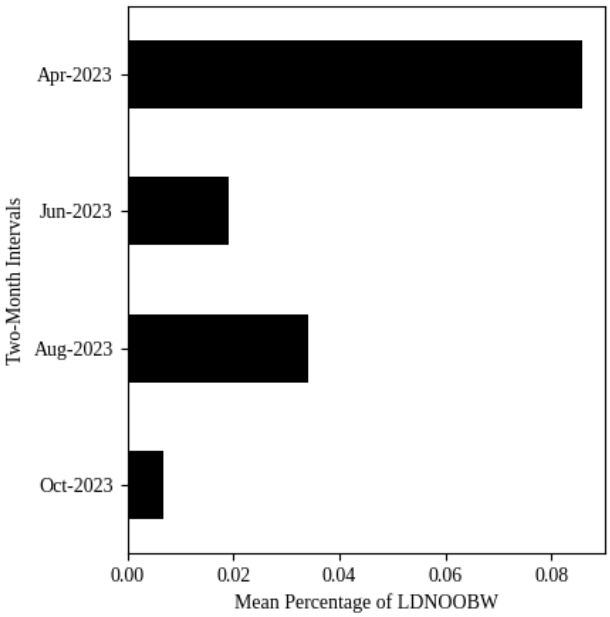
* Toxic represents anything in the text that is rude or disrespectful. Severe Toxic, Obscene, Threat, Insult, Identity Attack, and Sexual Explicit are all sub-categories of Toxic for more specific instances of disrespectful wording

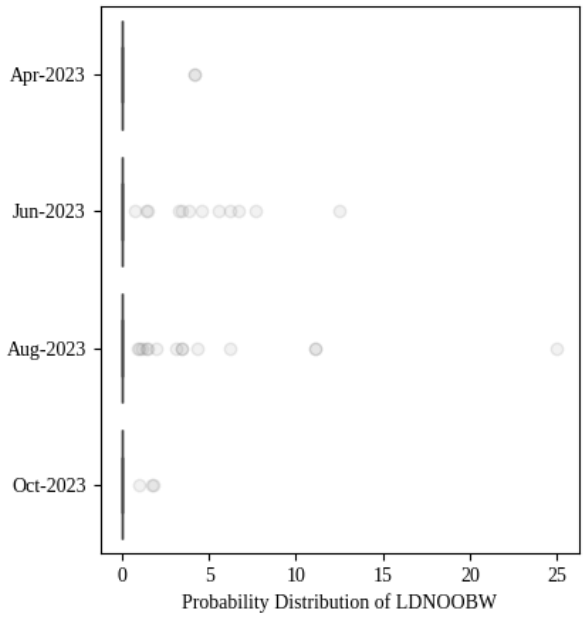




Is there a way to quantify how much profanity is there in text?

* LDNOOBW stands for "List of Dirty, Naughty, Obscene, and Otherwise Bad Words", and a list of the words can be found at: https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words

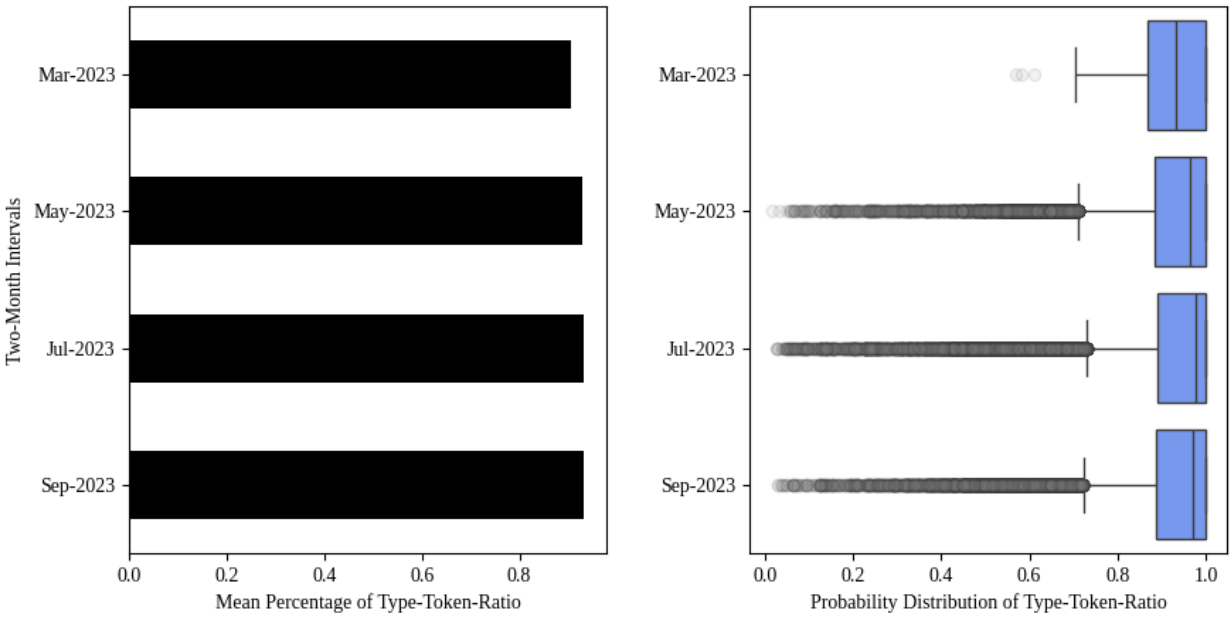




Can we measure the text diversity with the type-token-ratio (TTR)?

* The Type-Token Ratio (Bender, 2013) is calculated by dividing the number of unique words (types) by the total number of words (tokens) in a given text. A higher TTR indicates a greater

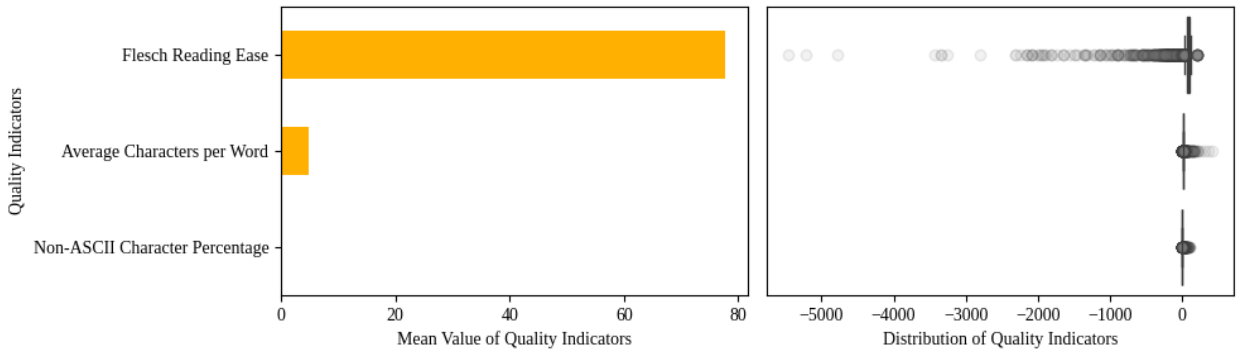
variety of words used in the text, while a lower TTR suggests a more limited vocabulary or a lot of repetition.



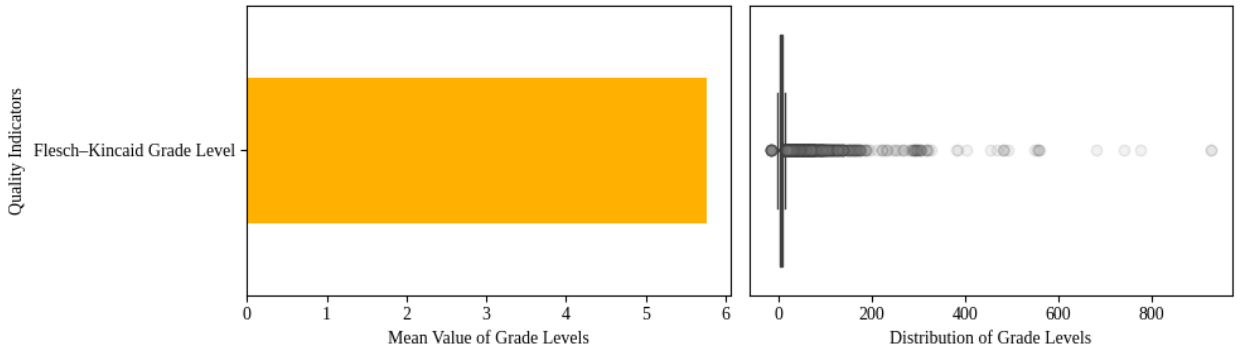
Can we get additional text statistics that relate to readability

* Flesch Reading Ease: Scored 0-100, with 100 being the easiest to read

Non-ASCII Character Percentage: Symbols and characters not included in the ASCII character set (Ex: ¢, é, ñ)

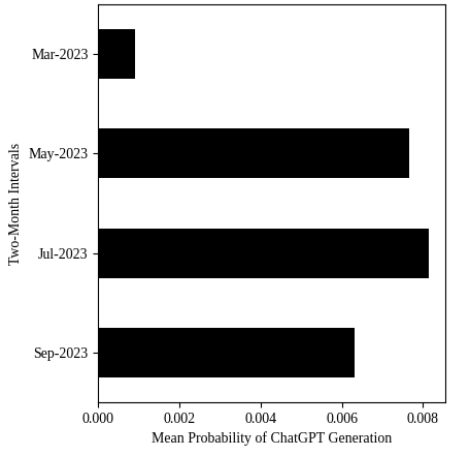


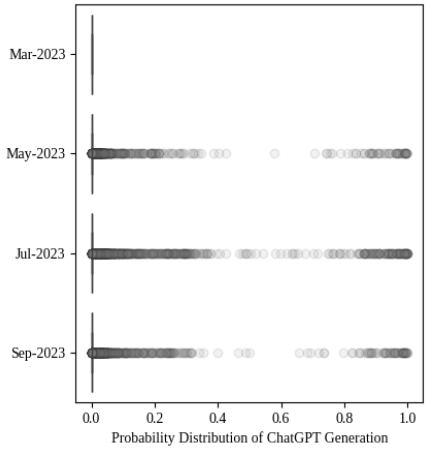
* Flesch-Kincaid Grade Level: Scored 0-18, where score corresponds to school grade level

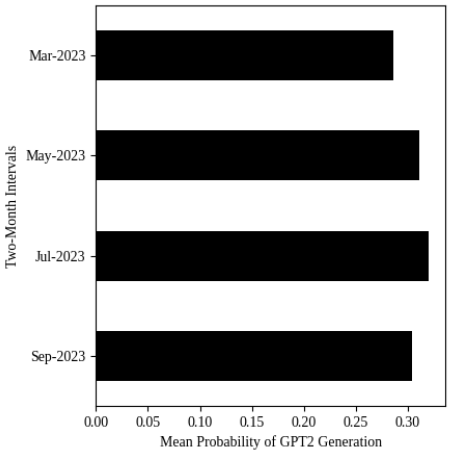


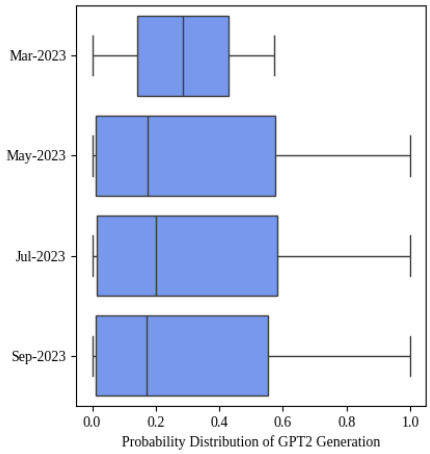
Is it possible to detect if text was generated by AI?

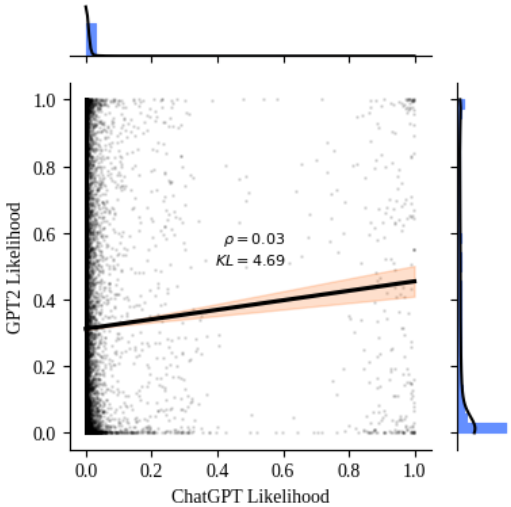
* Both ChatGPT and GPT-2 are large language models developed by OpenAI. ChatGPT was developed on GPT-3.5 and GPT-4, released in 2022. GPT-2 was never incorporated in ChatGPT and was released in 2019.





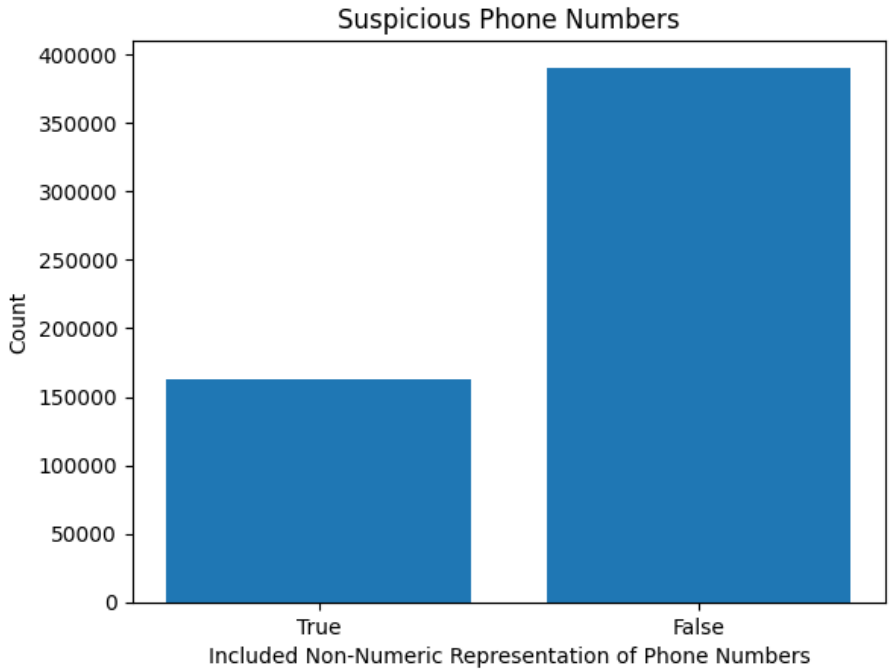




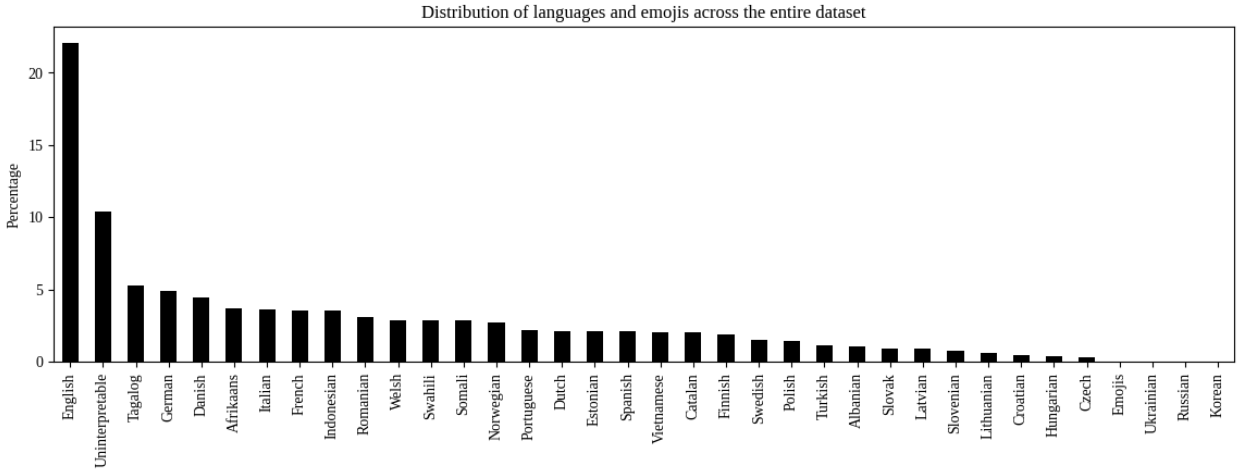


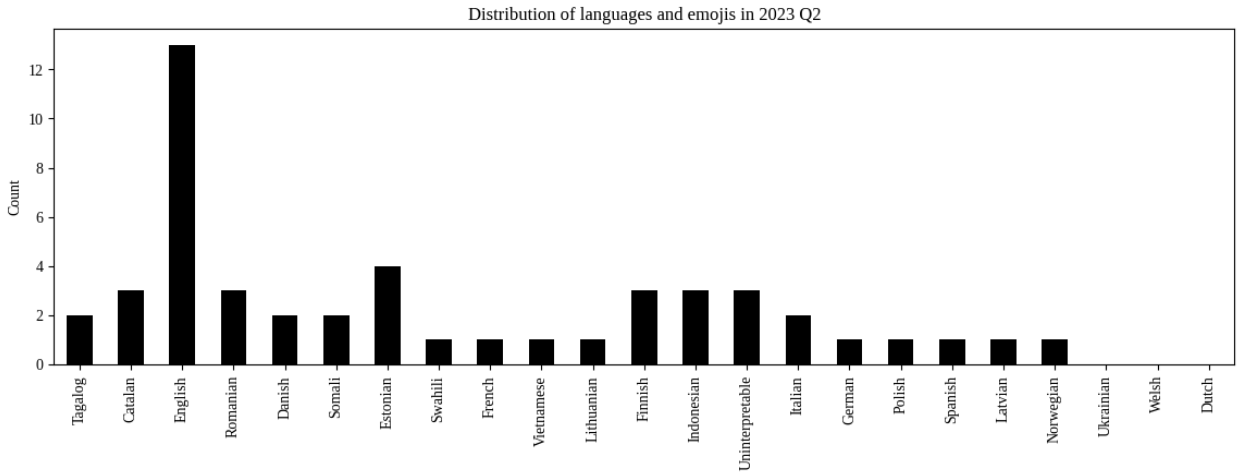
Can you identify any phone numbers or social media accounts being advertised? Are there any patterns in the contact information?

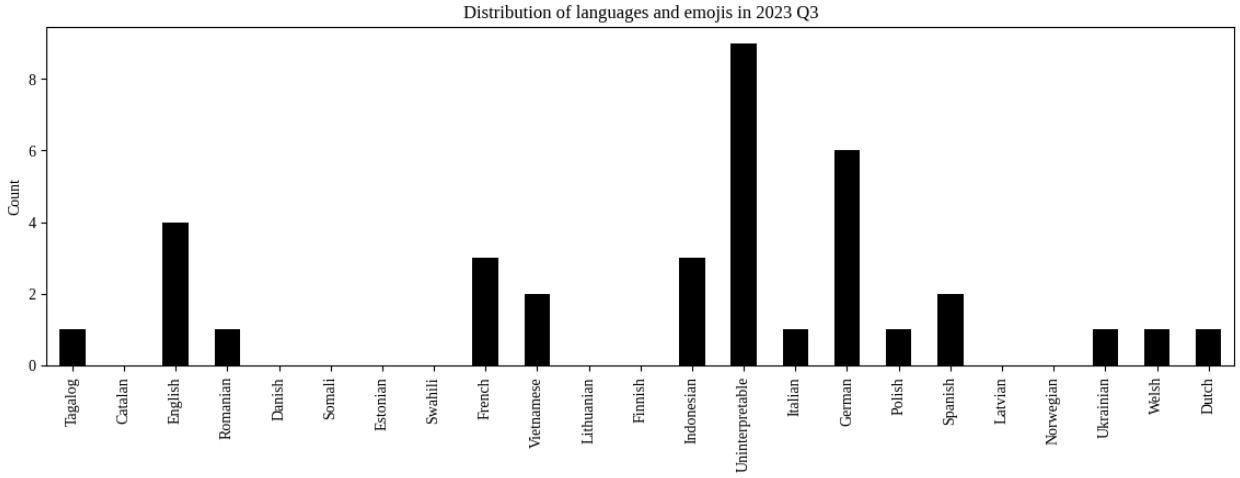
*Non-Numeric Representations of Phone Numbers are numbers that included textual representations (Ex. Two3six)



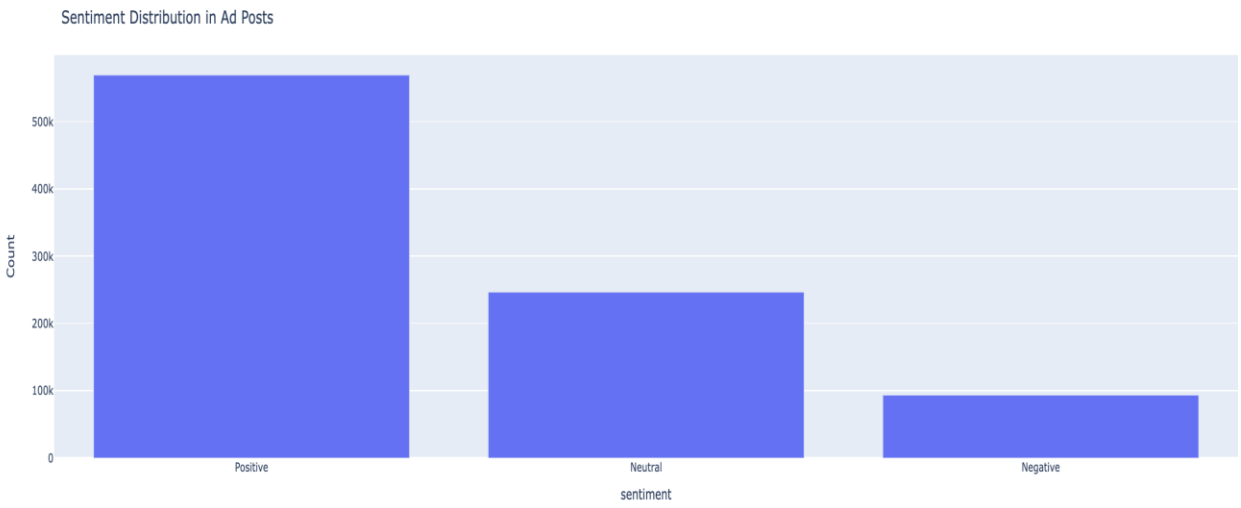
What is the distribution of ads with respect to languages and emojis?



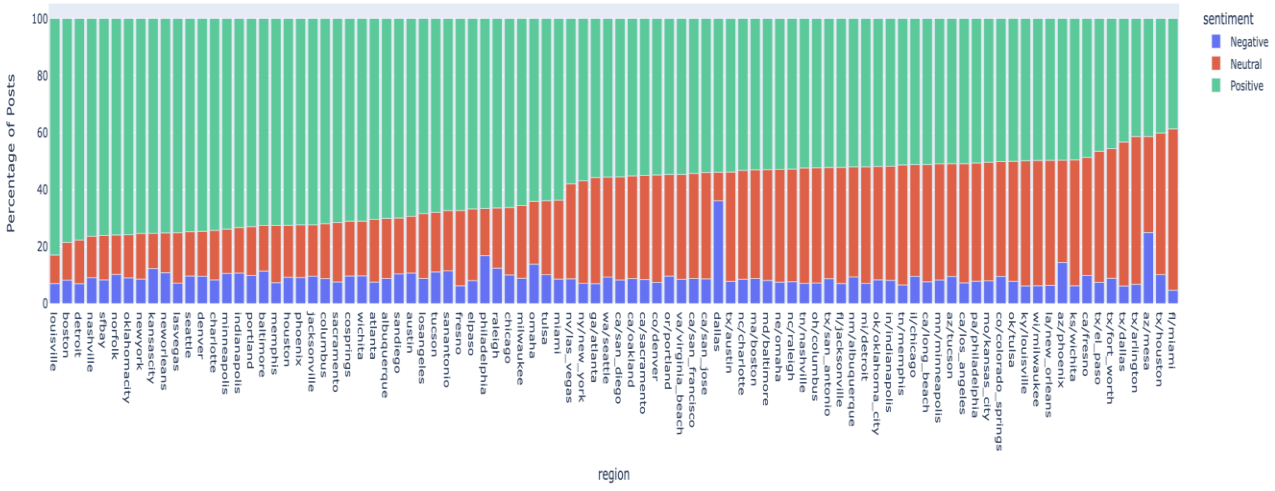




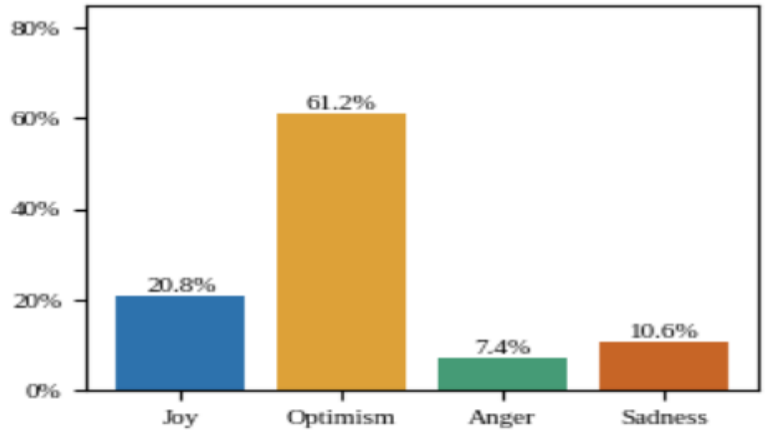
What are the sentiments and tones of the ad posts? Are they generally positive, negative, or neutral?



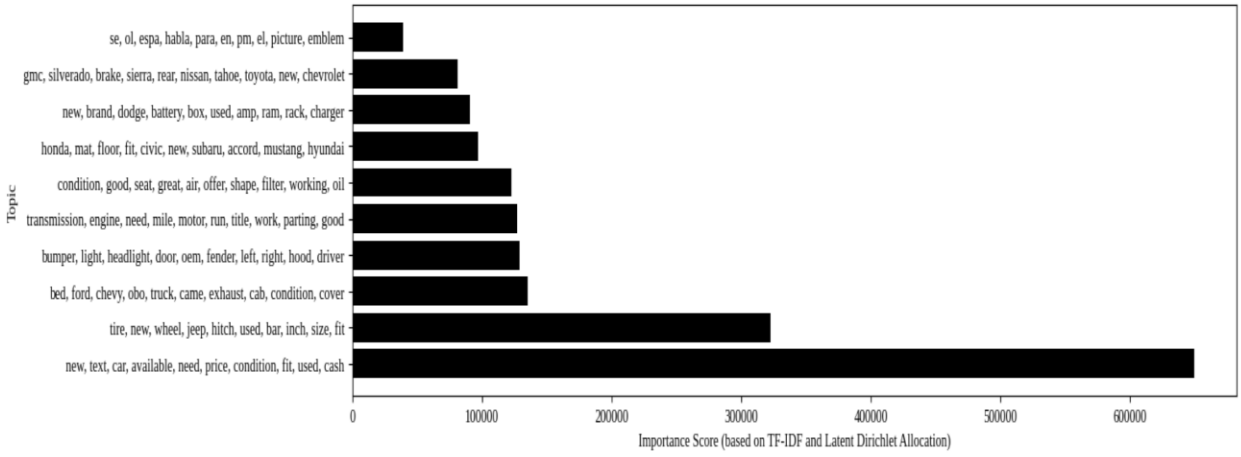
*Percentage of Sentiments by Region (Sorted by Positive Sentiment)



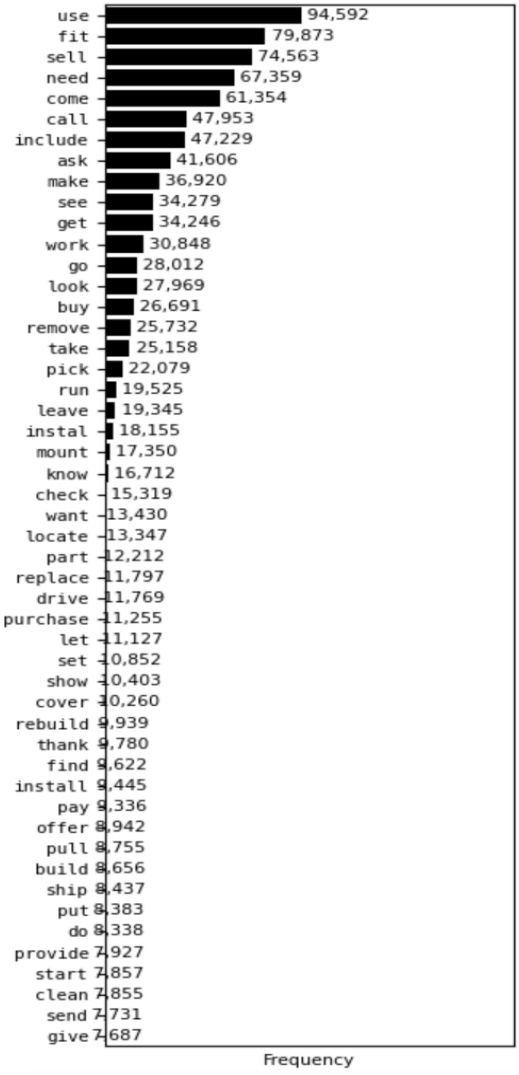
Can you classify the ads based on their emotion? Are there any patterns in the emotion of the ads?



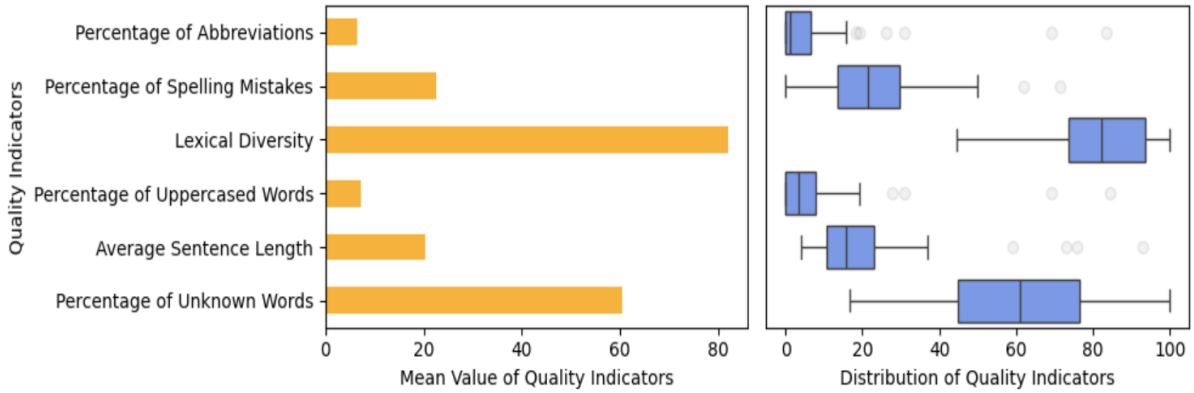
What are the most common topics or themes in the ad posts?



What are the most common calls-to-action (CTAs) used in the ad posts?



What are good text quality indicators in general terms?



Can we get additional text statistics that relate to readability?