Linguistic Analysis of SOS Messages using spaCy

Introduction
This analysis examines a corpus of SOS messages using the spaCy library to perform Part-of-Speech (POS) tagging and Named Entity Recognition (NER). The goal is to understand the linguistic structure of emergency messages and identify patterns in how information is communicated.

Data Preprocessing
The dataset was preprocessed to improve the quality of linguistic analysis. The preprocessing steps included:

- Converting all text to lowercase for consistency
- Removing stopwords to retain meaningful content
- Eliminating punctuation and extra spaces
- Unmasked masked names (B***y) for NER analysis

These steps ensured that the analysis focused on semantically relevant tokens rather than noise.

Part-of-Speech (POS) Analysis

POS tagging was applied to the cleaned corpus to identify the grammatical structure of the SOS messages. The frequency distribution of POS tags revealed the following patterns:

- Nouns (NOUN) were the most frequent tag, indicating that messages heavily focus on entities such as people, places, and resources.
- Verbs (VERB) were also common, reflecting action-oriented language (e.g., requests for help, movement, or needs).
- Proper nouns (PROPN) appeared frequently, suggesting references to specific locations or names.
- Adjectives (ADJ) and adverbs (ADV) were less frequent, indicating that messages tend to be concise and informational rather than descriptive.

Interpretation
The dominance of nouns and verbs suggests that SOS messages are goal-directed and information-dense, prioritizing clarity and urgency over descriptive language. This aligns with the functional purpose of SOS communication, where conveying key facts quickly is critical.

Named Entity Recognition (NER) Analysis
NER was used to extract and categorize entities within the messages. The most common entity types included:

- CARDINAL: Numbers not included in dates
- PERSON: Names of individuals
- DATE/TIME: Temporal references
- ORG: Organisations
- GPE (Geo-Political Entities): Locations such as cities or countries

Interpretation

The high frequency of CARDINAL entities suggests that numerical information plays a central role in SOS messages. This may include counts of people affected, resource quantities, or other critical numeric details relevant to emergency situations.

The prominence of PERSON entities indicates that many messages include identifiable individuals, either victims, reporters, or contacts. This reflects the human-centered nature of SOS communication.

The presence of DATE entities highlights the importance of temporal information, which can help establish urgency or provide context for when events occurred.

The amount of ORG entities represents the hospitals and government agencies which helps show where resources are needed.

Although GPE entities appear less frequently than expected, their presence still underscores the role of location in emergency communication. The relatively low count may be due to informal language or incomplete location references that are not always recognized by the NER model.

The smaller counts for other entity types suggest that SOS messages are highly focused and avoid unnecessary detail, reinforcing the idea that they prioritize essential, actionable information.

Overall Insights

The linguistic patterns observed in the SOS messages reveal several key characteristics:

Conciseness and Efficiency: Messages are structurally simple, with a strong emphasis on nouns and verbs. This reflects the need for rapid communication in emergency situations.

Information Prioritization: The frequent use of location entities demonstrates that where an event is happening is often the most critical piece of information.

Action-Oriented Language: The presence of verbs indicates that messages often include requests, needs, or ongoing situations, emphasizing immediacy.

Low Descriptive Complexity: The relatively low frequency of adjectives and adverbs suggests that messages avoid unnecessary detail, focusing instead on essential information.

Conclusion

The analysis shows that SOS messages are linguistically optimized for urgency and clarity. POS tagging reveals a structure dominated by nouns and verbs, while NER highlights the importance of location and entity information. Together, these findings demonstrate that emergency communication prioritizes efficiency, precision, and actionable content.

This type of linguistic analysis can be useful for designing automated systems that detect and prioritize emergency messages, as well as improving information extraction in crisis response scenarios.