Data Glacier

Your Deep Learning Partner

Resume Parsing and Classification Using Named Entity Recognition (NER)

Exploratory Data Analysis

Internship: Data Science Intern
Specialization: NLP
Name: Monisha Shree Senthil Nathan
Unviversity: IU International University of Applied Sciences, Germany
Batch: LISUM32, Data Glacier
Email: monishashree.career@gmail.com
Date: 01.07.2024

Problem Statement

HR departments face the challenge of manually processing a large number of resumes, which is both time-consuming and labour-intensive. Each resume contains various sections such as personal details, education, work experience, and skills. By using Named Entity Recognition (NER) models in Natural Language Processing (NLP), we can automate the extraction and classification of these entities, streamlining the resume screening process and making it more efficient and accurate.

Introduction

This presentation highlights the findings from entity analysis and natural language processing (NLP) on resumes using Named Entity Recognition (NER). The key areas covered include top companies worked at, years of experience, graduation years, top skills, top locations, n-gram analysis, and sentiment analysis.

Data Collection and Understanding

Type of Data:

The data provided is a JSON structure containing resume information.
content: A string with the full text of the resume.
annotation: A list of dictionaries containing labels, text spans, and other metadata about the resume content.

Each labelled entity contains the following fields:

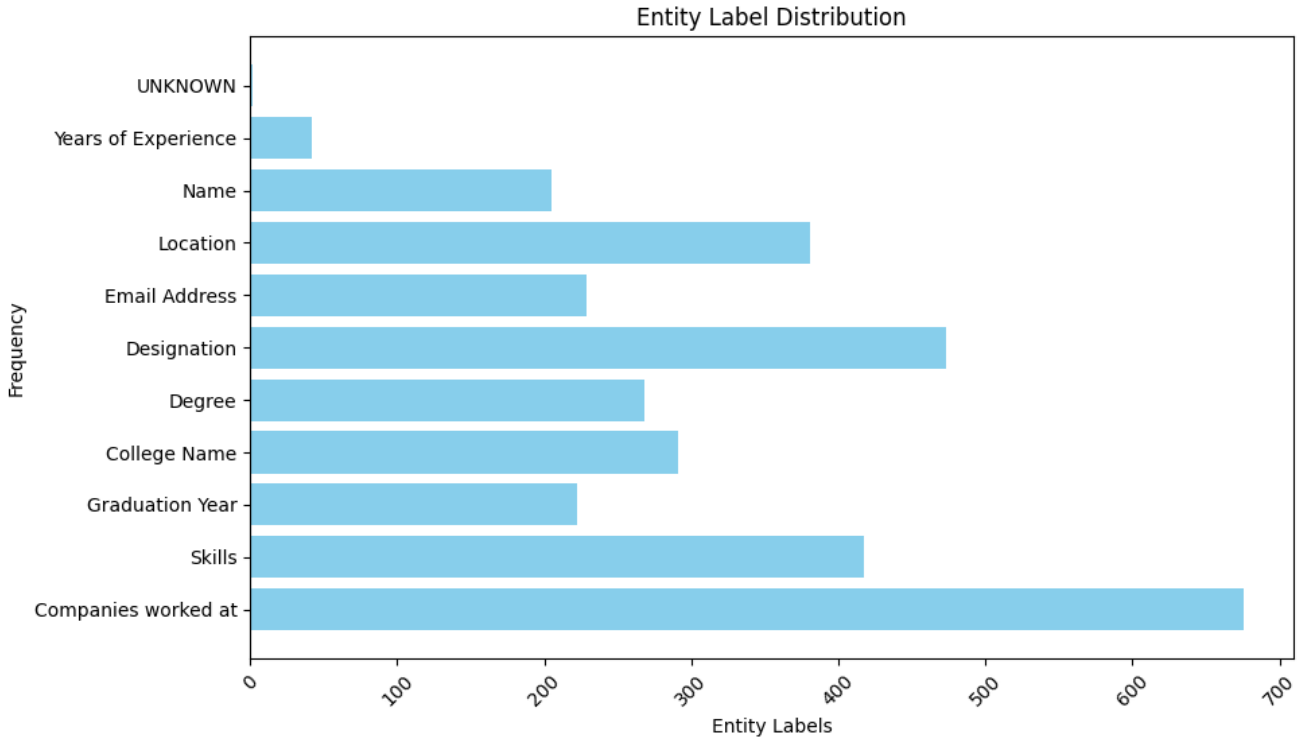
label: the type of entity

points: a list of character offsets indicating
the start and end positions of the entity in the resume text. It also includes the corresponding entity text

Name: The name of the person.
Email Address: Email address of the person.
Skills: Technical skills of the person.
College Name: The name of the college or university the person attended.
Degree: The qualification obtained by the person.
Designation: Job title or designation of the person.
Companies worked at: Companies where the person has worked.
Location: The location of the person.
Graduation Year: Graduation year.
Year of experience : The total years of experience.

EXPLORATORY DATA ANALYSIS

Entity Label distribution



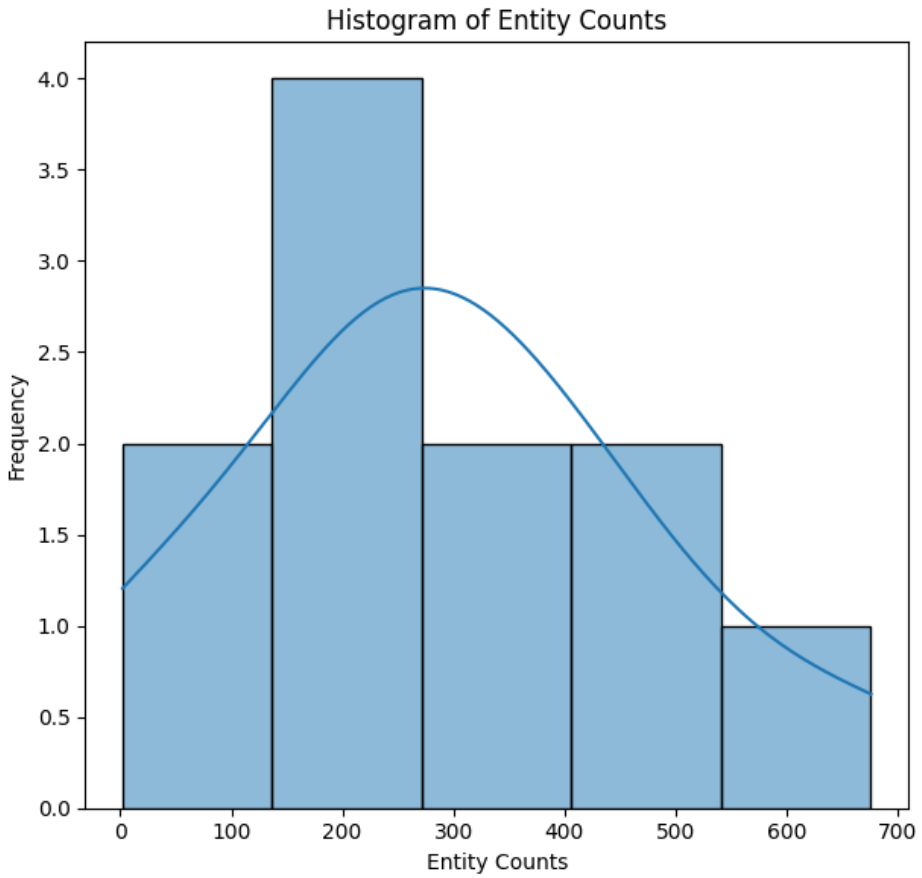

```
'Companies worked at': 676,
'Skills': 417,
'Graduation Year': 222,
'College Name': 291,
'Degree': 268,
'Designation': 473,
'Email Address': 229,
'Location': 381,
'Name': 205,
'Years of Experience': 42,
'UNKNOWN': 2
```

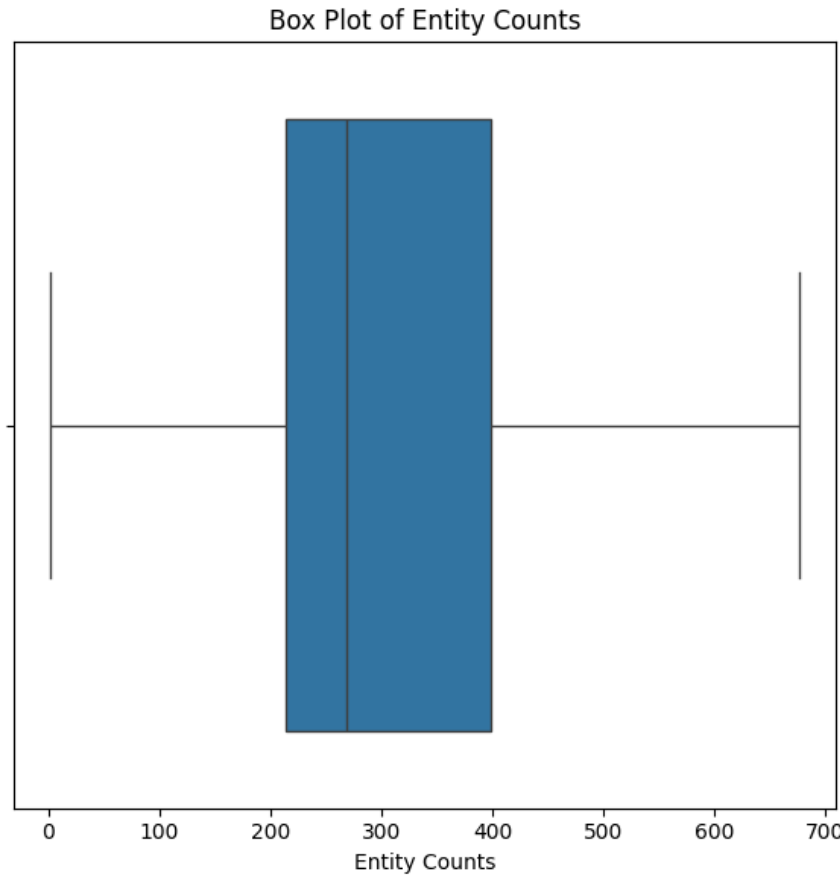
Distribution Analysis of Textual Features: Insights on Skewness and Kurtosis

•The majority of entity counts are concentrated between 200 and 300, indicating a common range for most entities.

•The right skewness in the histogram suggests that while most entities have moderate counts, a few have significantly higher counts.

•The absence of outliers in the box plot further supports a relatively stable distribution of entity counts.





Skewness and Kurtosis

Skewness

Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable. It helps to understand the direction and degree of deviation from the normal distribution.

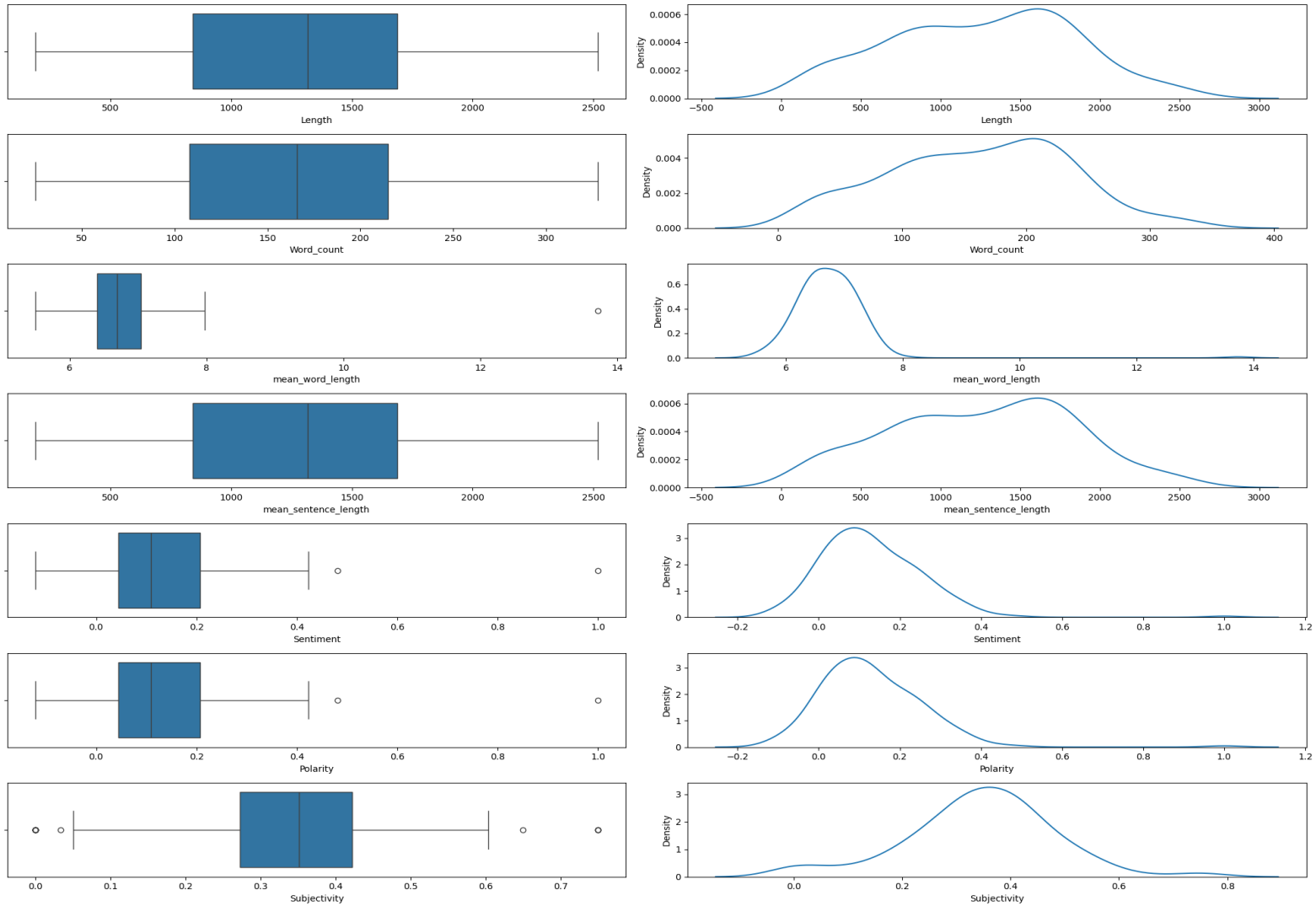
- **Positive Skew**: The tail on the right side of the distribution is longer or fatter than the left side.
- **Negative Skew**: The tail on the left side is longer or fatter than the right side.
- **Zero Skew**: Symmetrical distribution (normal distribution).

Kurtosis

Kurtosis is a measure of the "tailedness" of the probability distribution of a real-valued random variable. It indicates the presence of outliers and the sharpness of the peak of the distribution.

- **High Kurtosis**: Distribution has heavy tails or outliers.
- **Low Kurtosis**: Distribution has light tails or fewer outliers.
- **Normal Kurtosis**: Kurtosis close to 3 (mesokurtic distribution).

Distribution Analysis of Textual Features: Insights on Skewness and Kurtosis



Length:
Skew: -0.0657
Negative skew, indicating a slight longer tail on the left side; a few resumes are slightly shorter than most
Kurtosis: -0.7457 Low kurtosis, indicating lighter tails; fewer outliers and a flatter distribution.

Word Count:

Skew: -0.0690
Negative skew, indicating a slight longer tail on the left side; a few resumes have slightly fewer words than most.
Kurtosis: -0.6160
Low kurtosis, indicating lighter tails; fewer outliers and a flatter distribution

Mean Sentence Length:

Skew: -0.0657
negative skew, indicating a slight longer tail on the left side; a few resumes have slightly shorter sentences.
Kurtosis: -0.7457
kurtosis, indicating lighter tails; fewer outliers and a flatter distribution

Mean Word Length:

Skew: 5.6448
High positive skew, indicating a significant longer tail on the right side; a few resumes use much longer words.
Kurtosis: 58.2875
Very high kurtosis, indicating extreme outliers; some resumes use significantly longer words

Distribution Analysis of Textual Features: Insights on Skewness and Kurtosis

Overall Interpretation

Negative Skewness for Length, Word Count, and Mean Sentence Length indicates that the distribution of these variables has a longer tail on the left side, meaning there are a few resumes that are significantly shorter, have fewer words, or shorter sentences compared to the average.

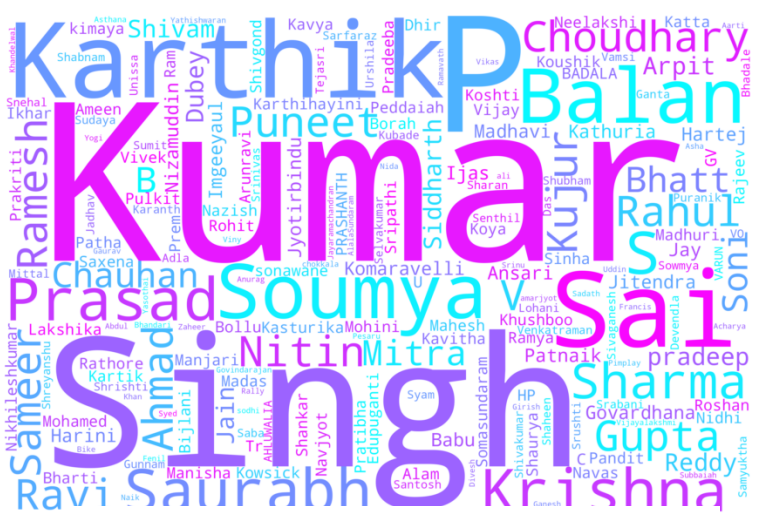
Positive Skewness for Mean Word Length suggests that a few resumes use significantly longer words, which is less common.

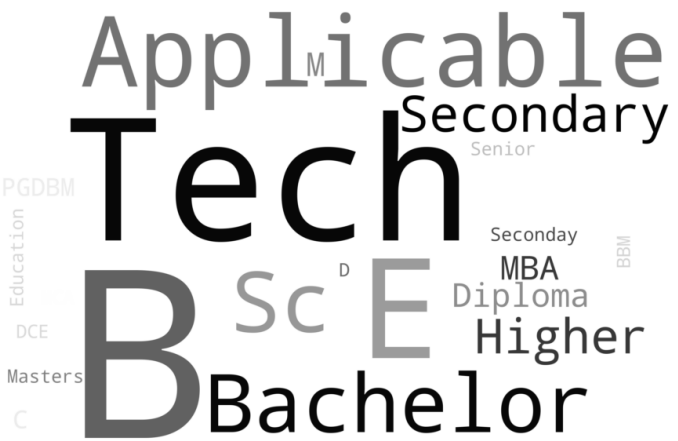
Low Kurtosis for Length, Word Count, and Mean Sentence Length implies that these distributions are relatively flat with fewer outliers, indicating a more consistent dataset.

High Kurtosis for Mean Word Length highlights the presence of extreme outliers, suggesting that some resumes have unusually long words, which could be due to technical jargon or specific industry terms.

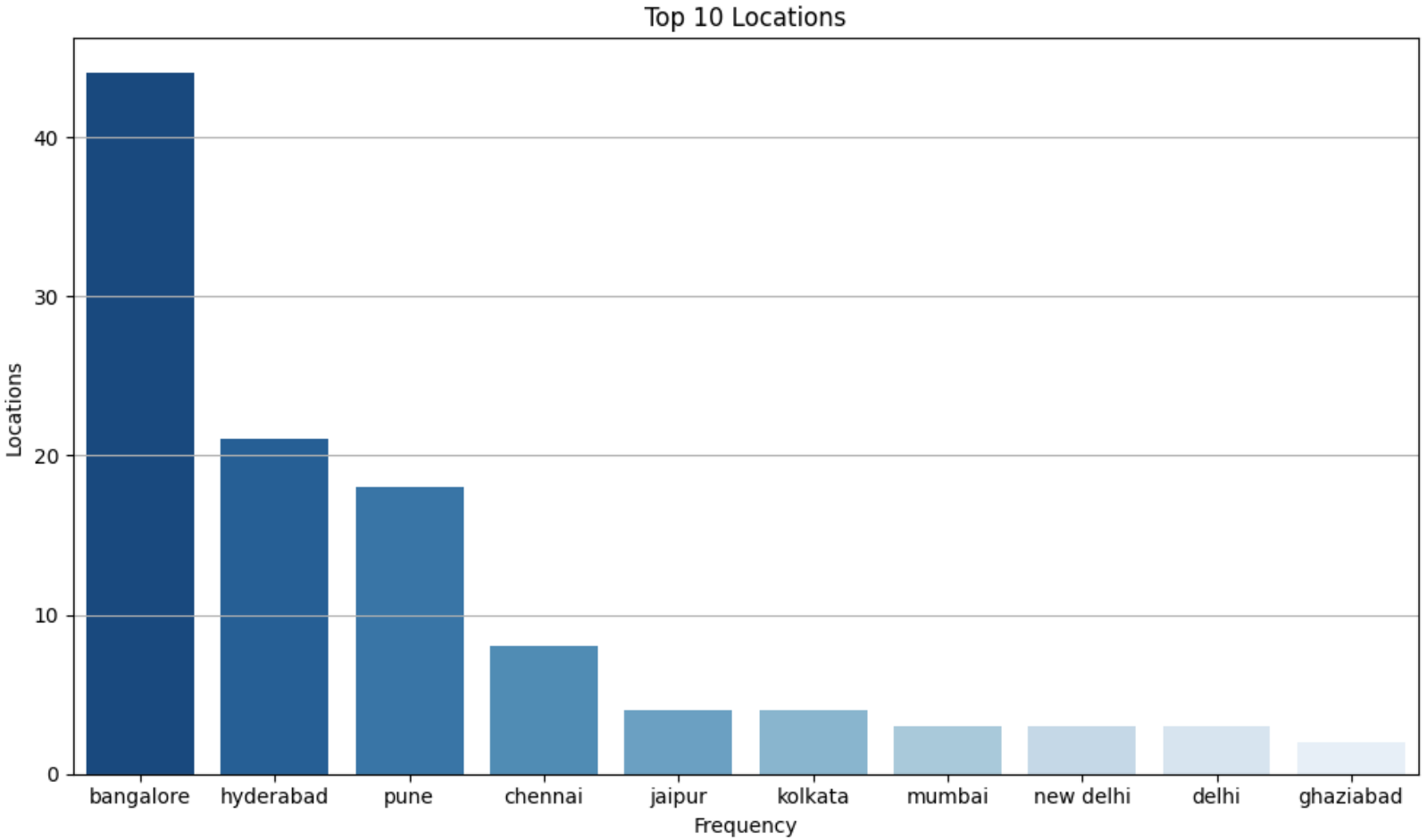
Insights Generation

WordCloud for entities



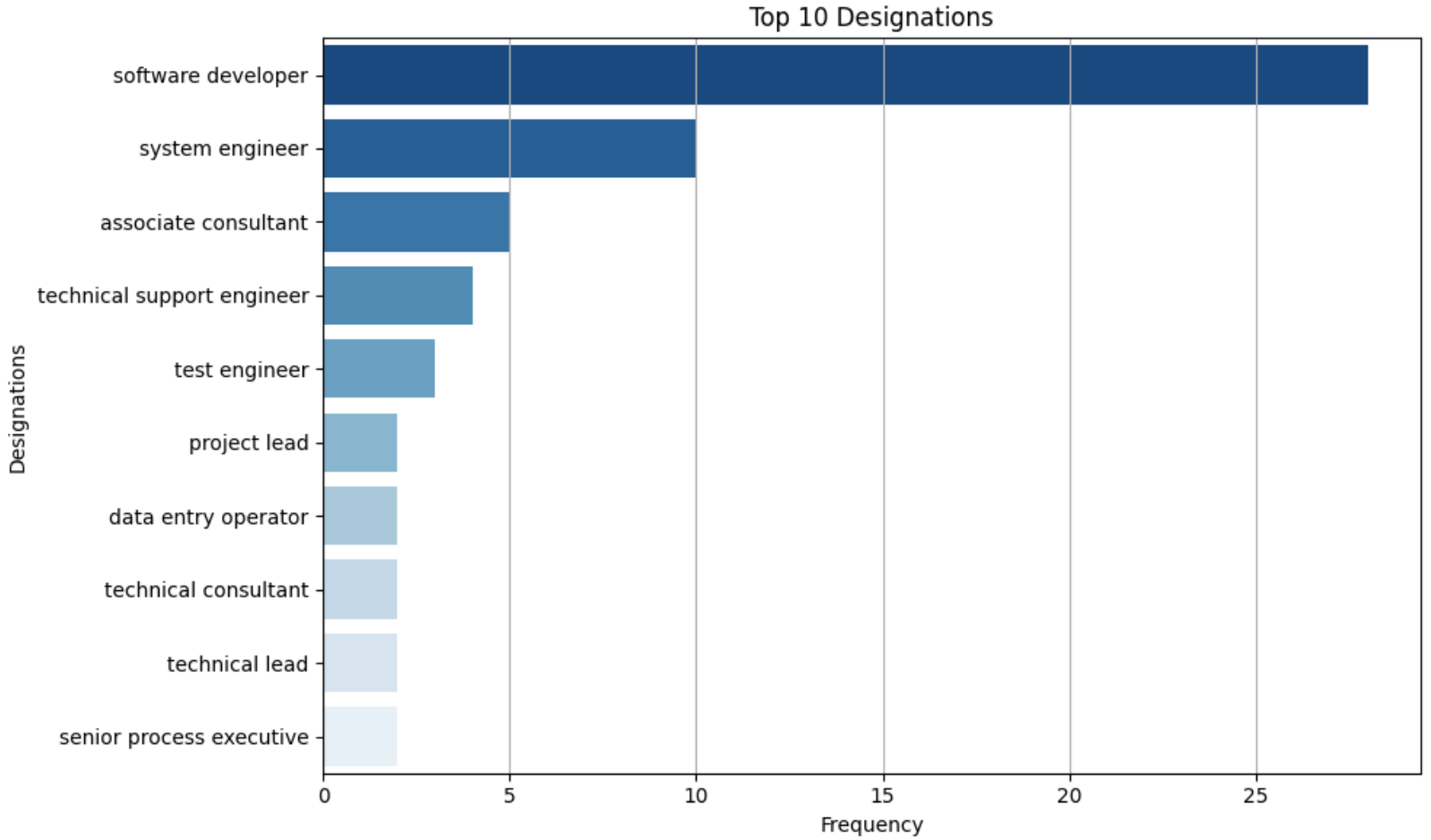


Entity Distribution: Location



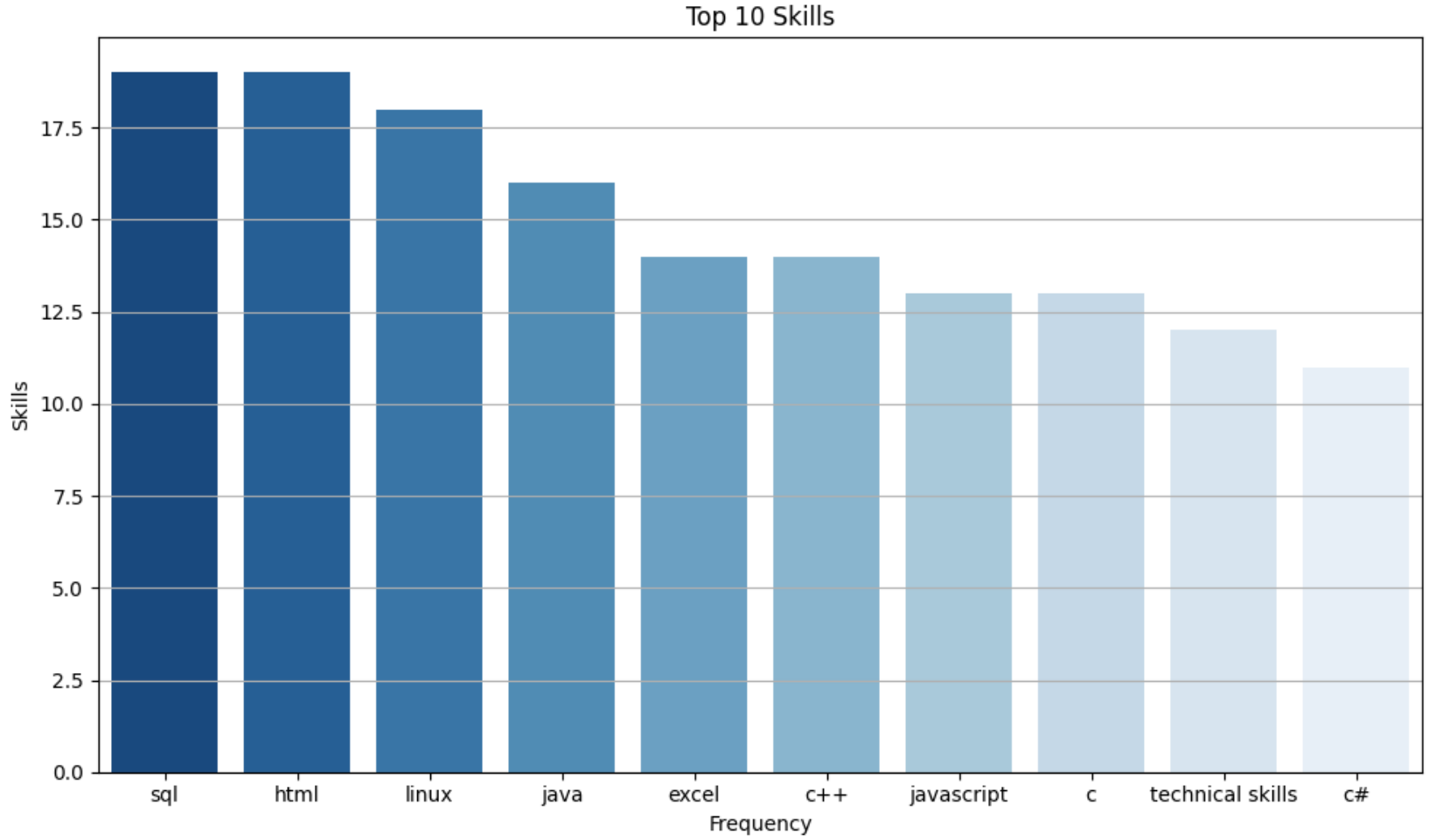
A significant portion (over 40+) applicants are located in Bangalore. This might be due to the city's reputation as a tech hub. Hyderabad and Pune also have a good pool of talent we can tap into.

Entity Distribution: Designation



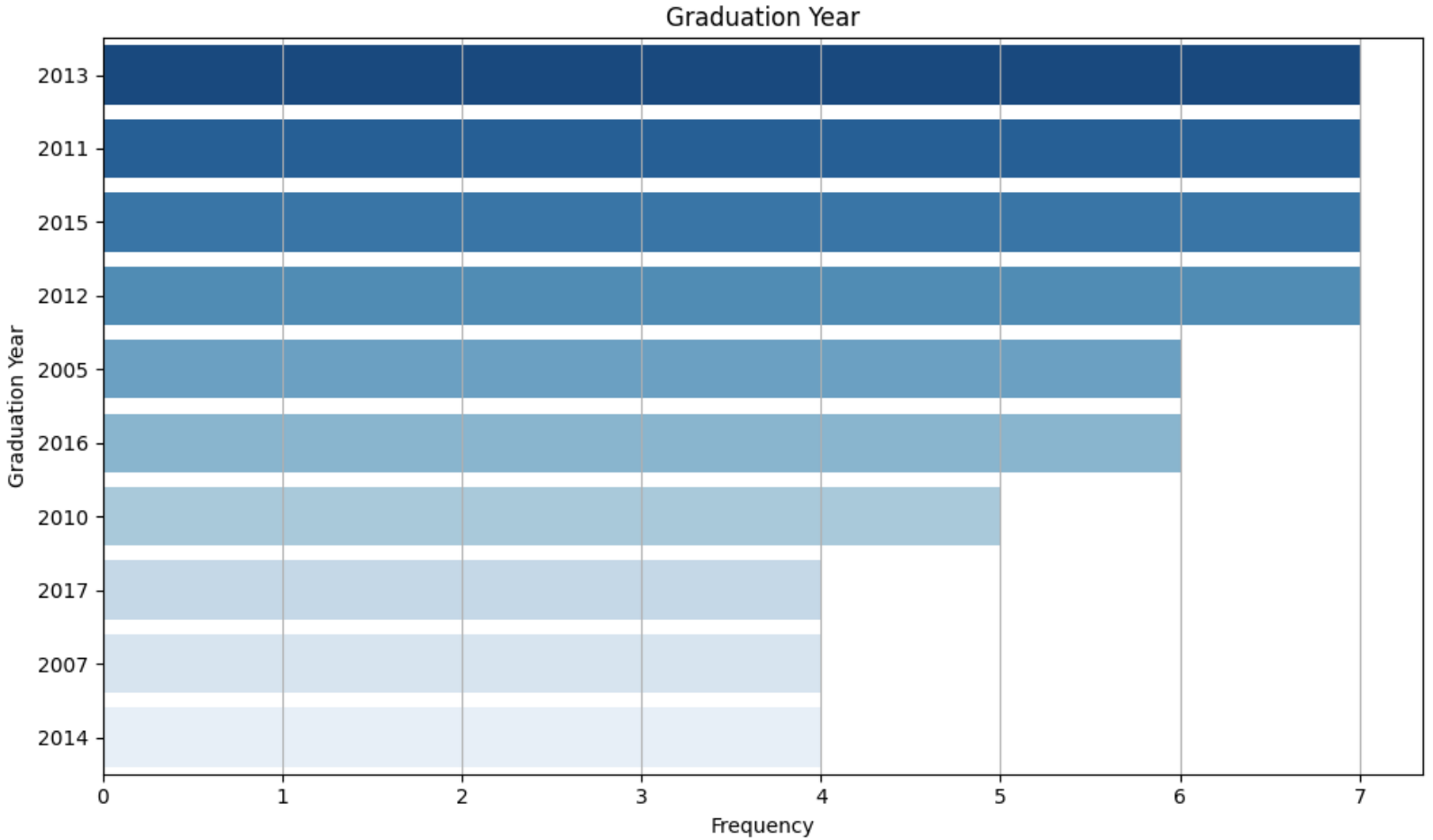
Here, we see Software Developer and System Engineer as the most common designations. This aligns well with the tech industry focus of the applicant pool we saw in the previous slide.

Entity Distribution: Skills



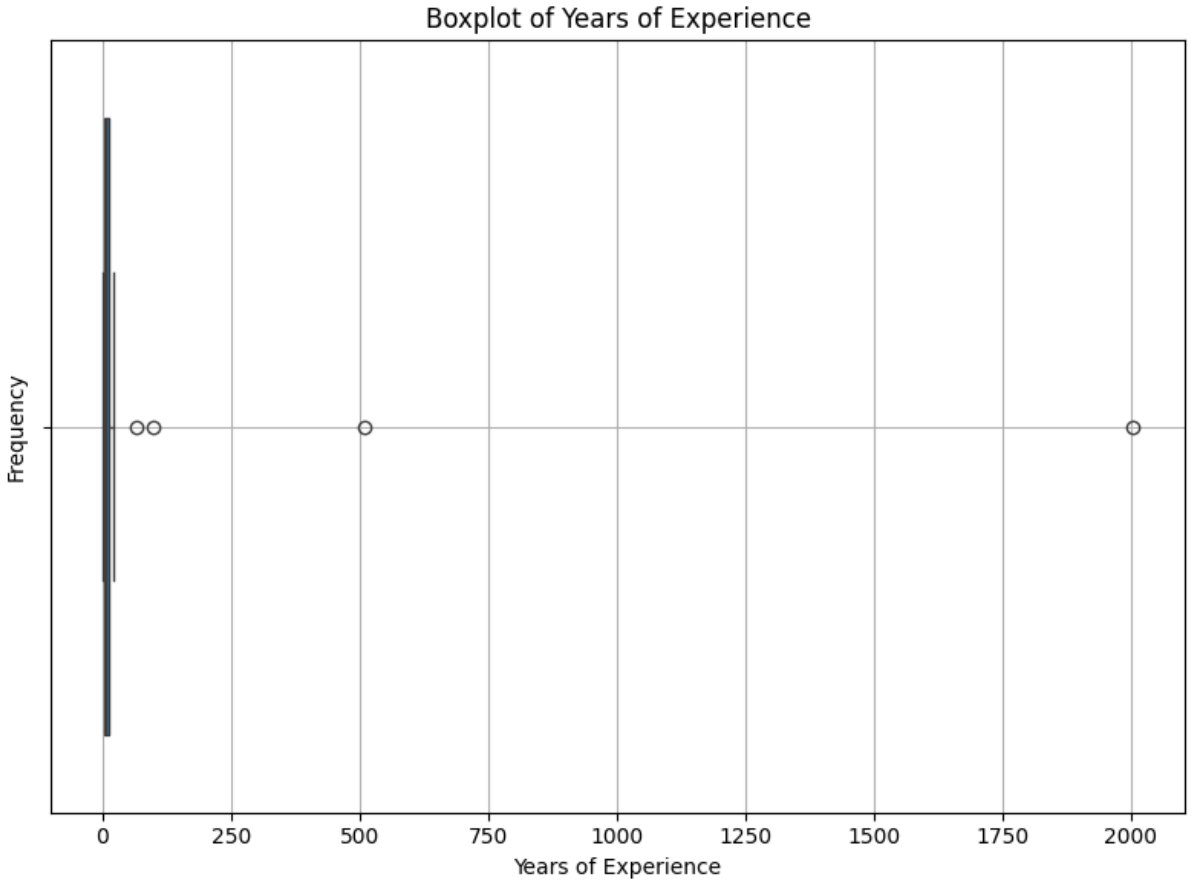
This slide highlights the most sought-after skills among the applicants. These skills are all in high demand within the tech industry. SQL, HTML, Linux, and Java are the top skills listed by applicants, with over 17 individuals mentioning each

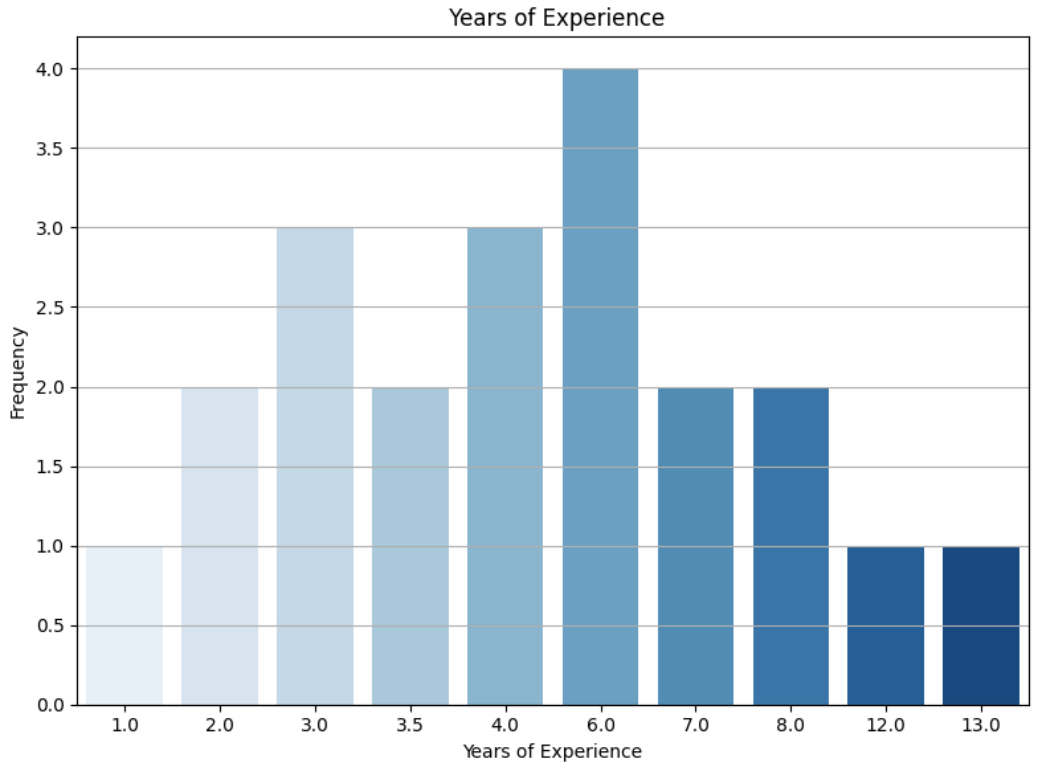
Entity Distribution: Graduation years



We can see a concentration of graduates from the years 2013, 2011, and 2012. This might indicate an influx of talent that graduated around that time.

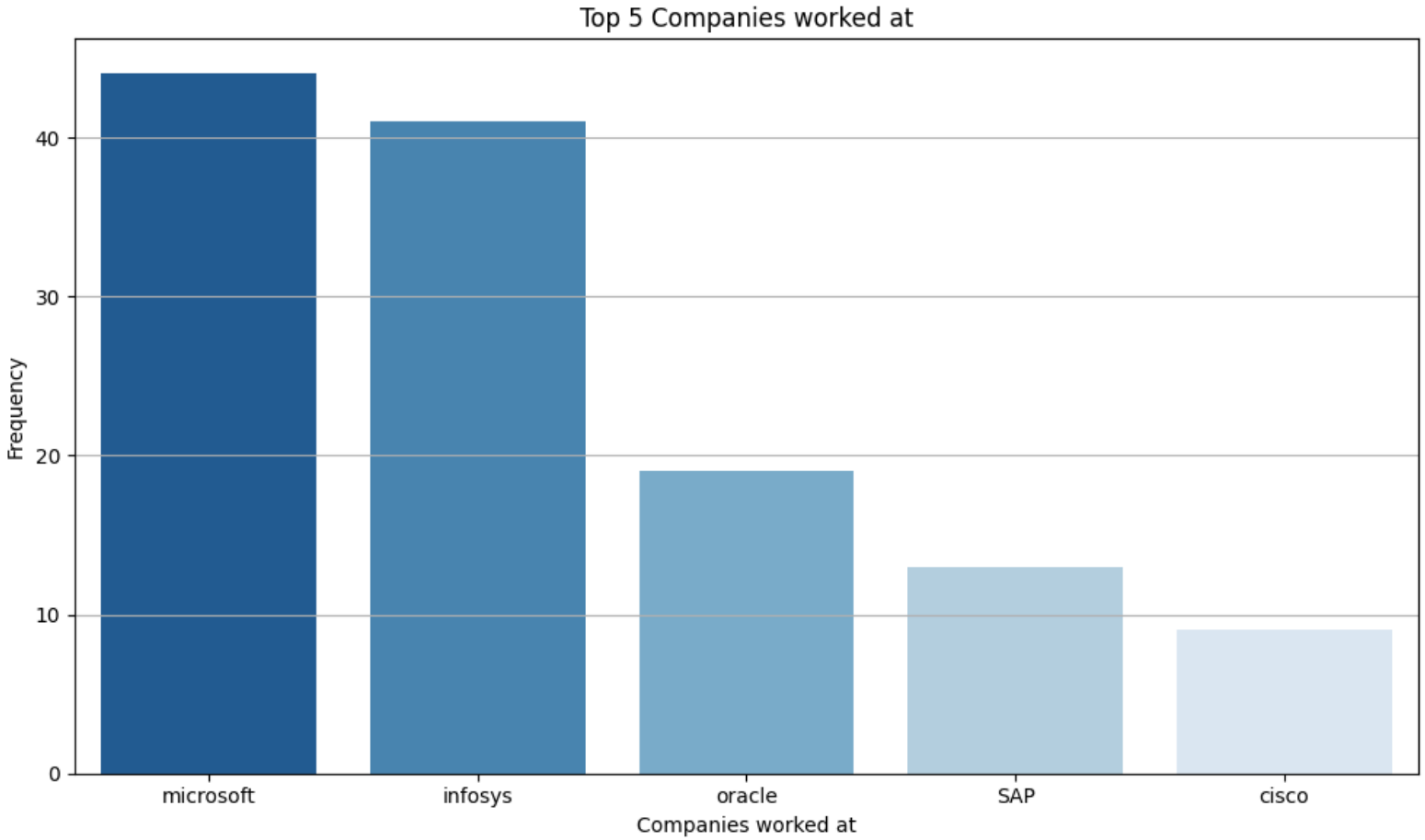
Entity Distribution: Years of Experience





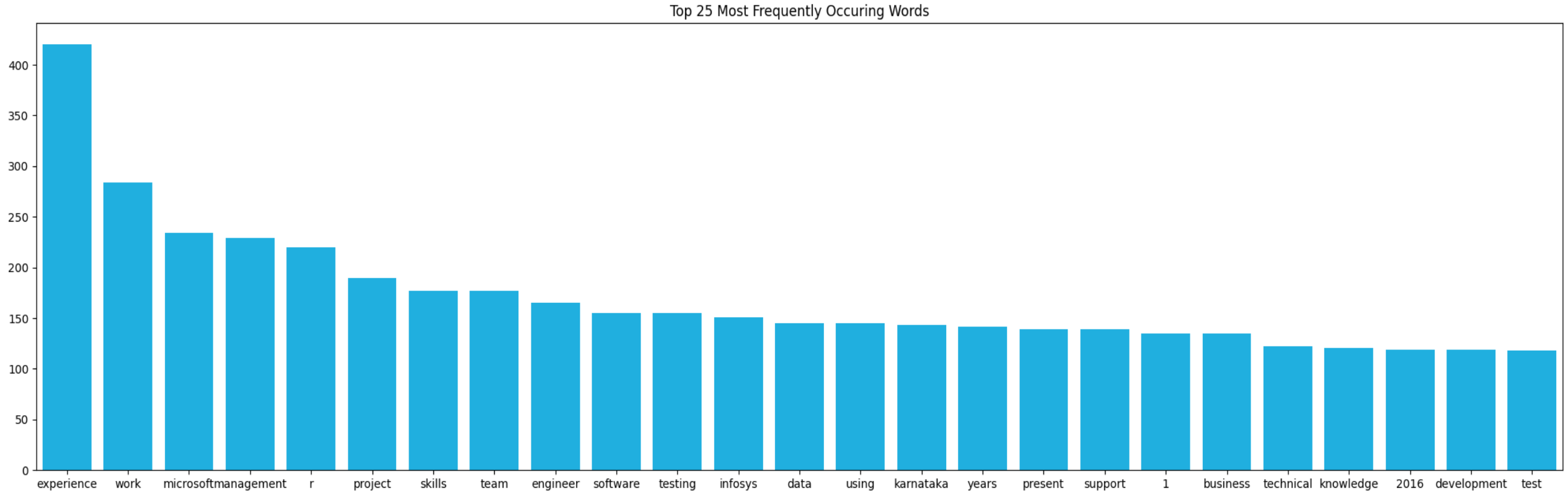
Years of experience had outliers, because it had the year ‘511’ and ‘2004’. fter removing them, this slide presents a clearer picture of the distribution of experience levels within the applicant pool.

Entity Distribution: Companies worked at



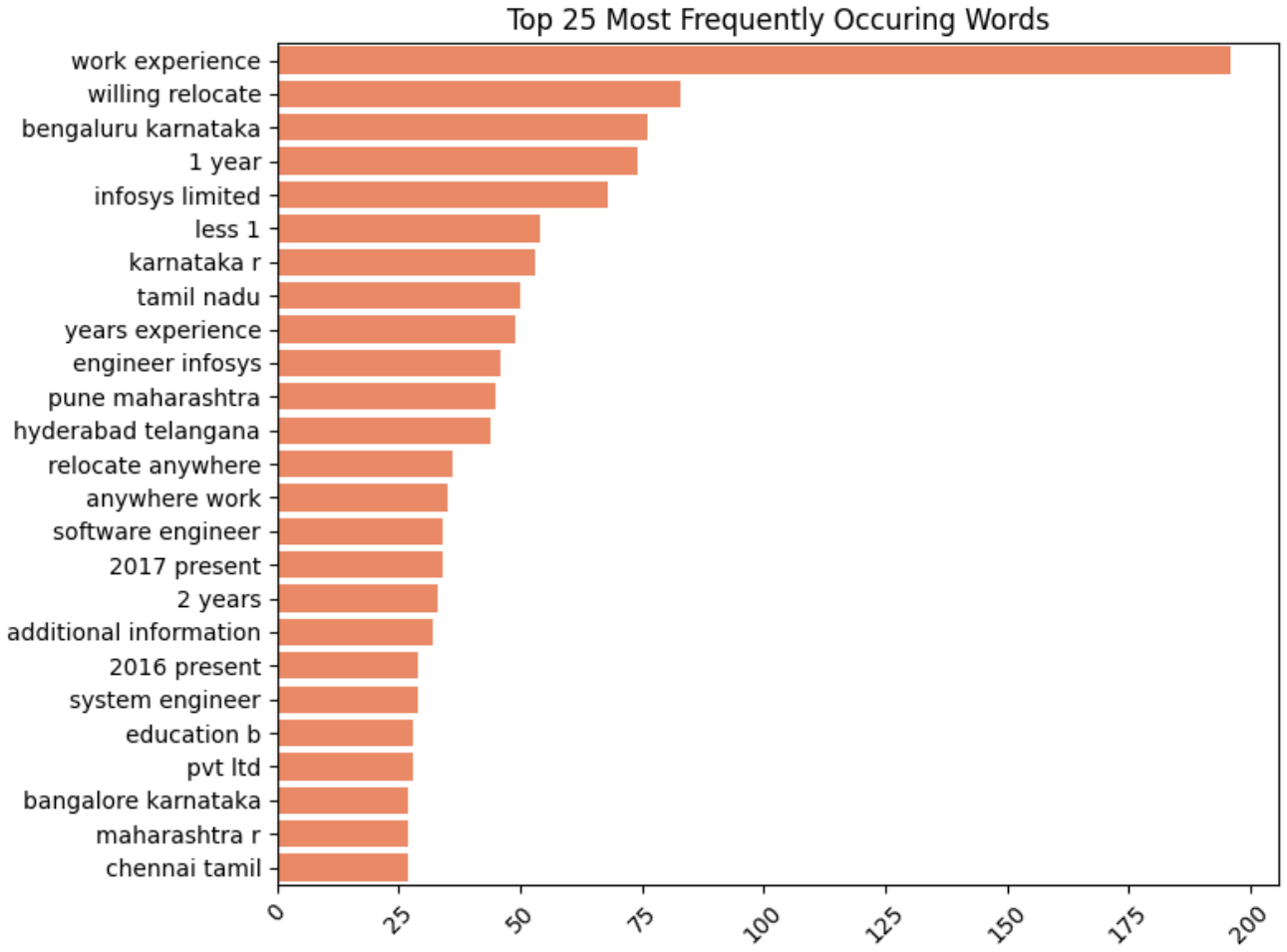
Analysis of the data reveals that a significant number of applicants (over 40) have prior experience at Microsoft and Infosys. Oracle is another frequently mentioned company.

N gram: Unigram



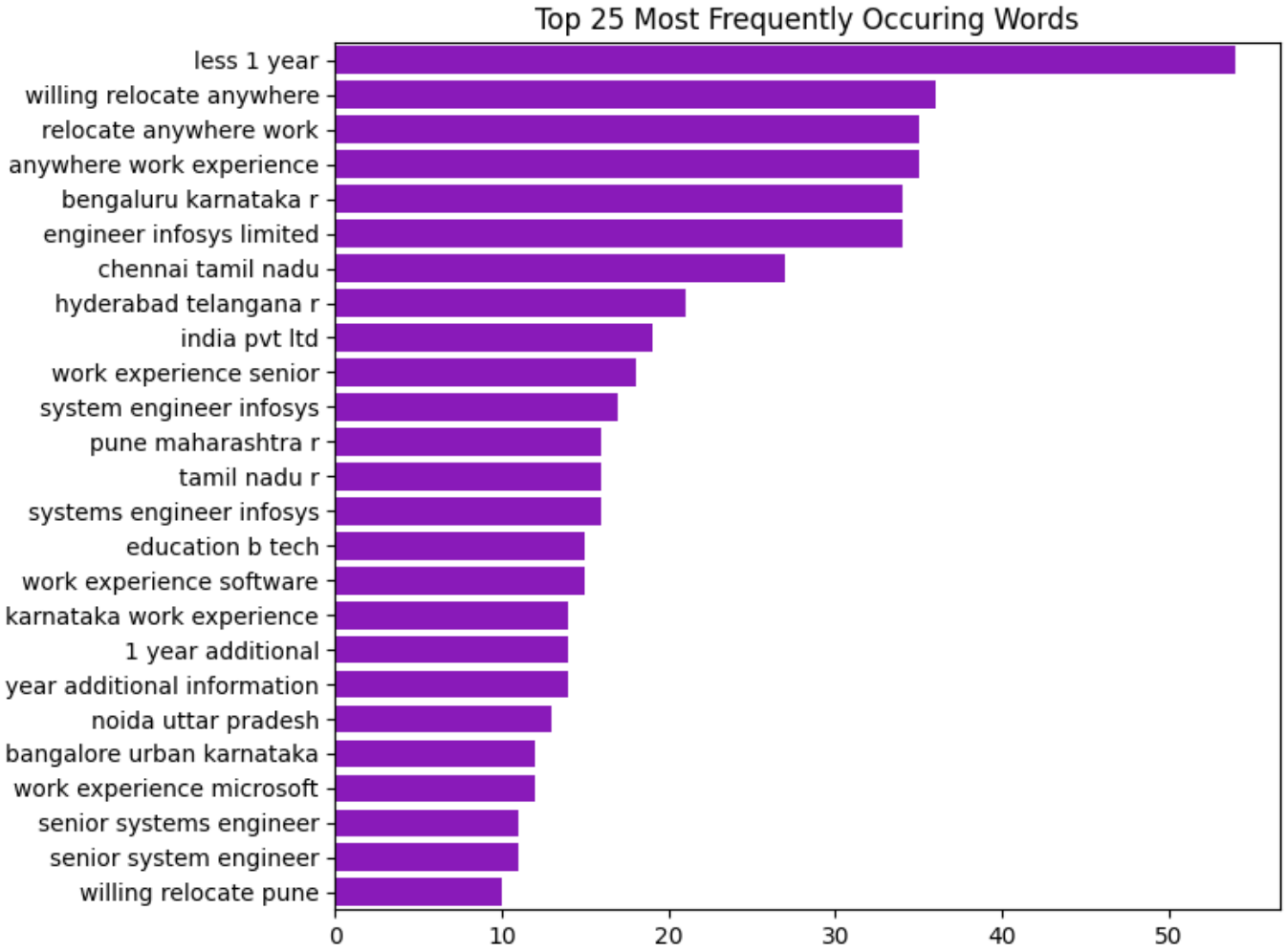
Unigram: Top words are experience, work, Microsoft and management

N gram: Bigram



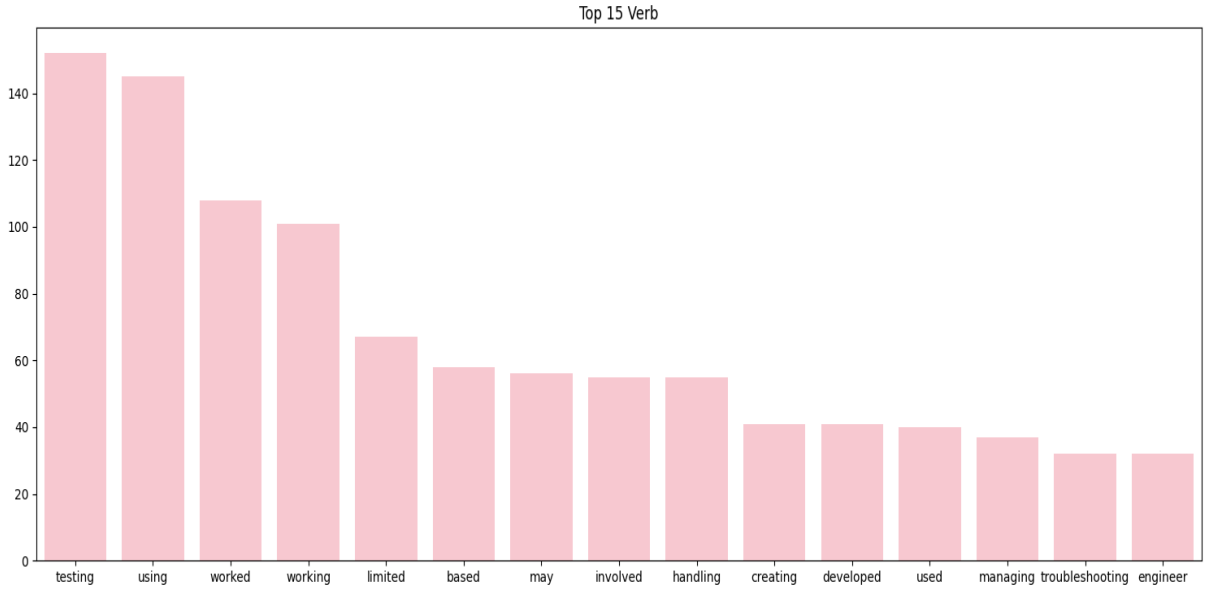
Top Bigrams are ‘Work Expereience’, ‘willing relocate’,’Bengaluru karnataka’

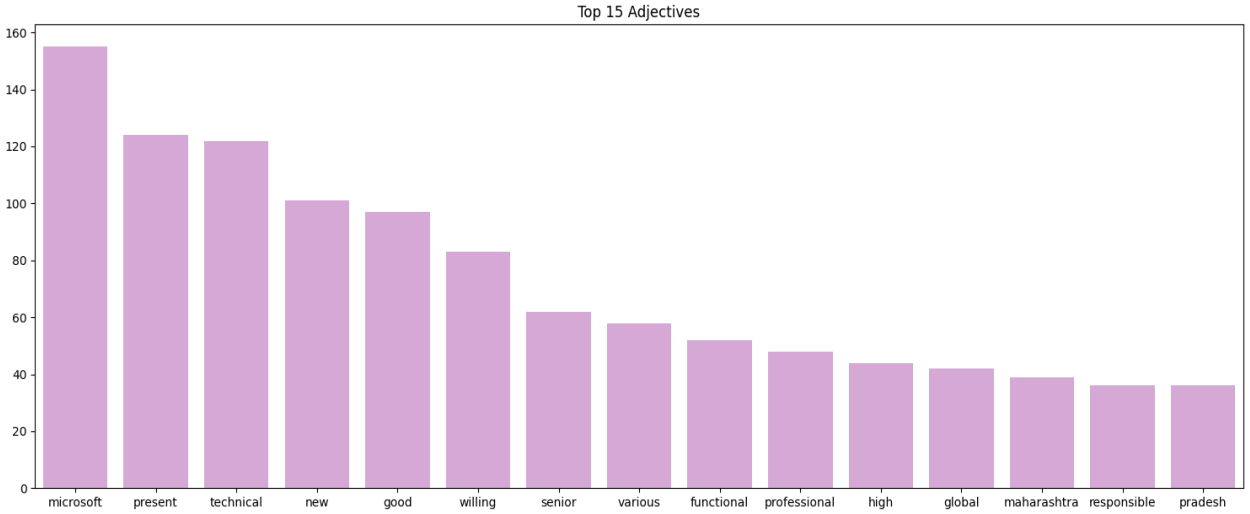
N gram: Trigram



Top trigrams are ‘less 1 year’, ‘willing relocate anywhere’ and ‘relocate anywhere work’. Keywords like "relocate" suggest candidates open to new opportunities. We can leverage this during outreach.

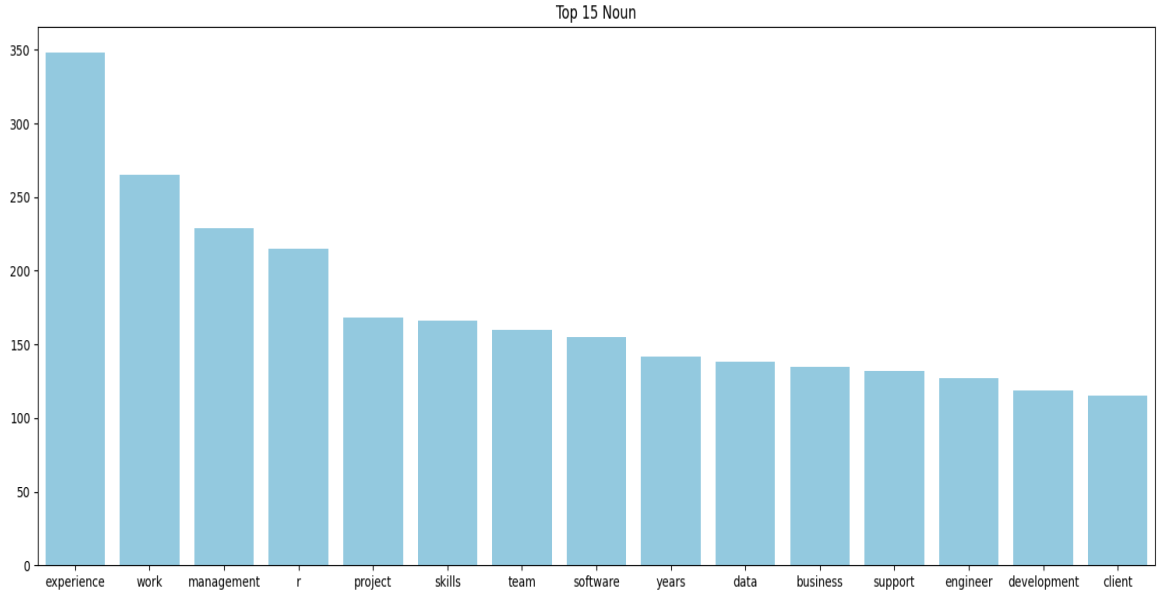
Parts of Speech





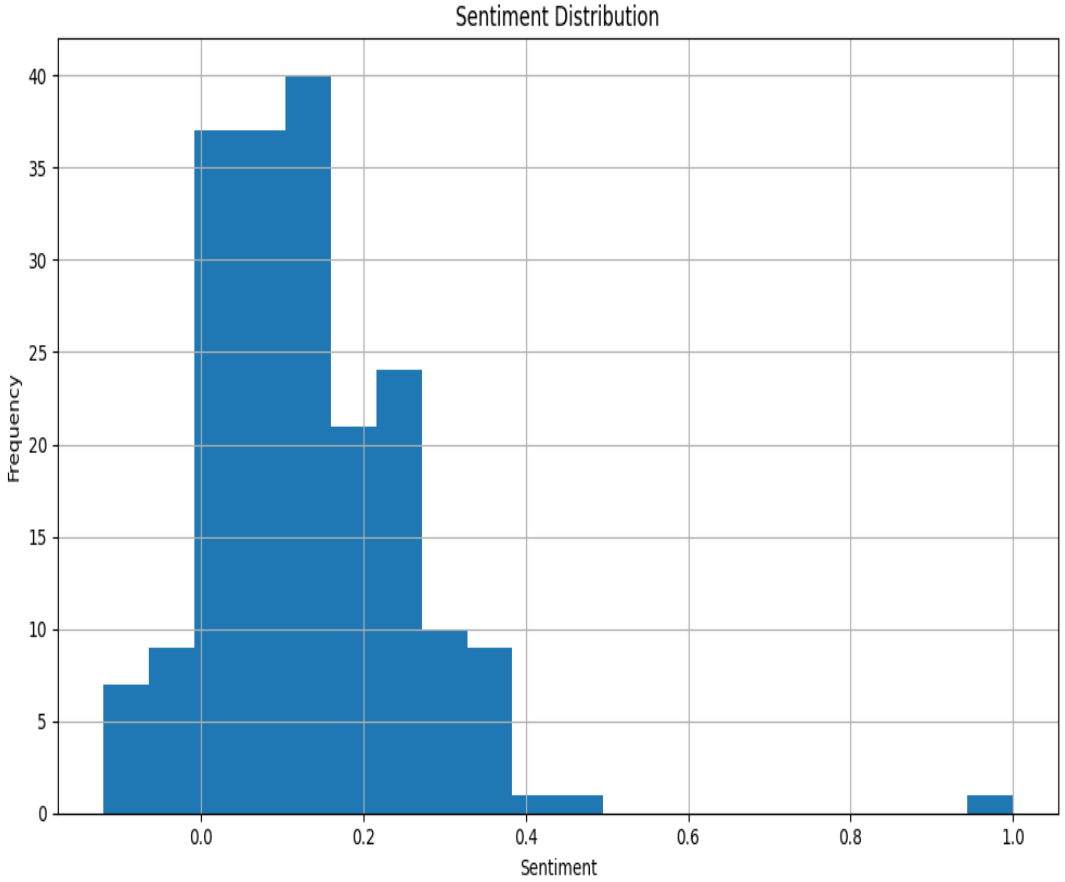
Parts of Speech refer to the categories of words based on their function in a sentence. The main POS categories are:

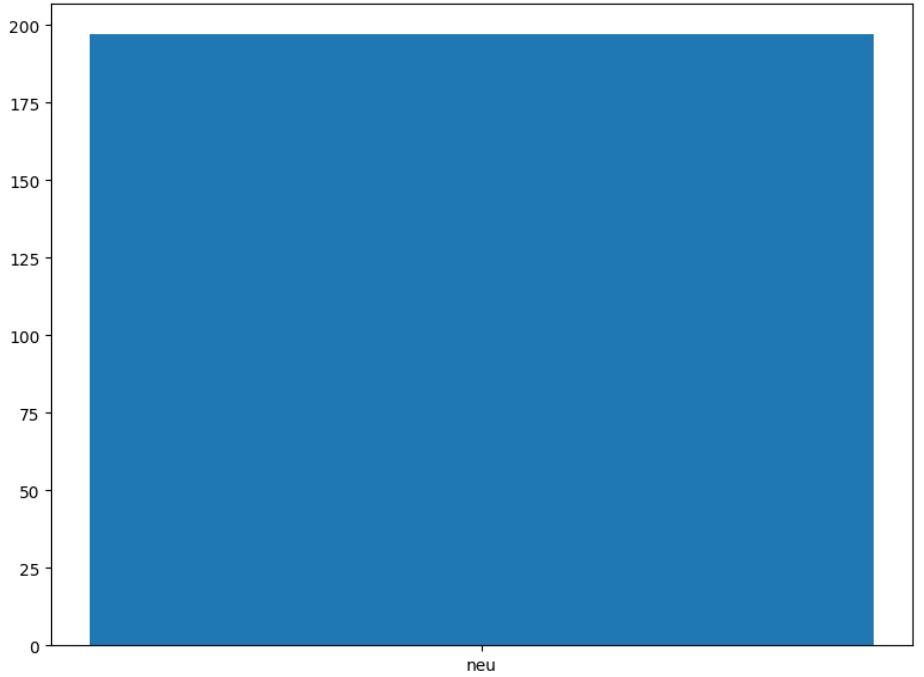
Nouns: Words that name people, places, things, or ideas
Verbs: Words that express actions or states of being
Adjectives: Words that describe or modify
Adverbs: Words that modify verbs, adjectives, or other adverbs



The identification of top verbs, adjectives, and nouns in resumes helps in understanding the common actions, qualities, and entities mentioned, which can aid in better resume parsing and keyword extraction.

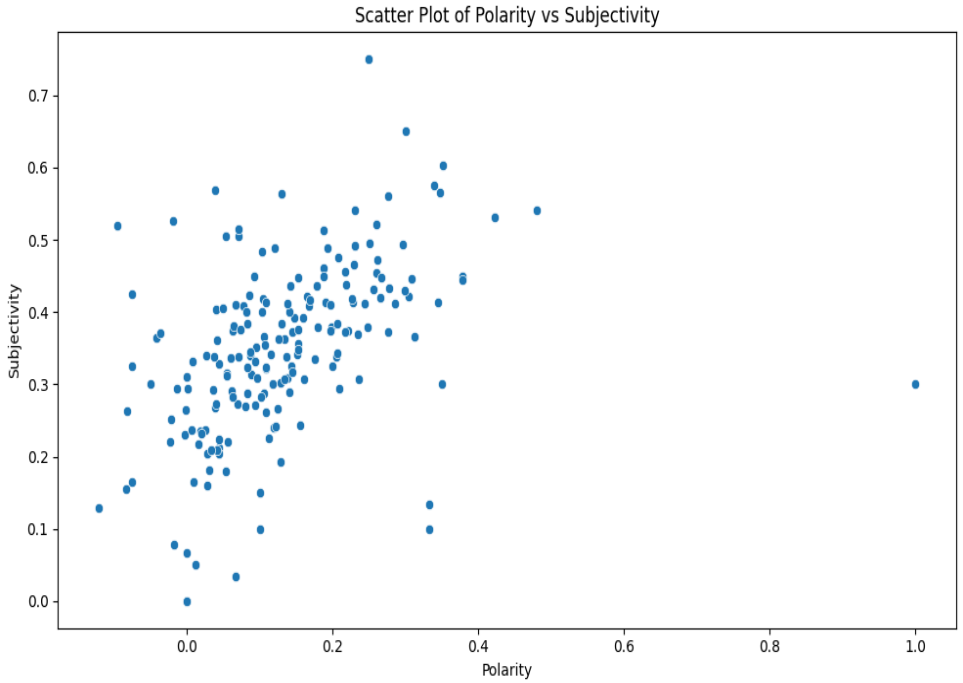
Sentiment Analysis

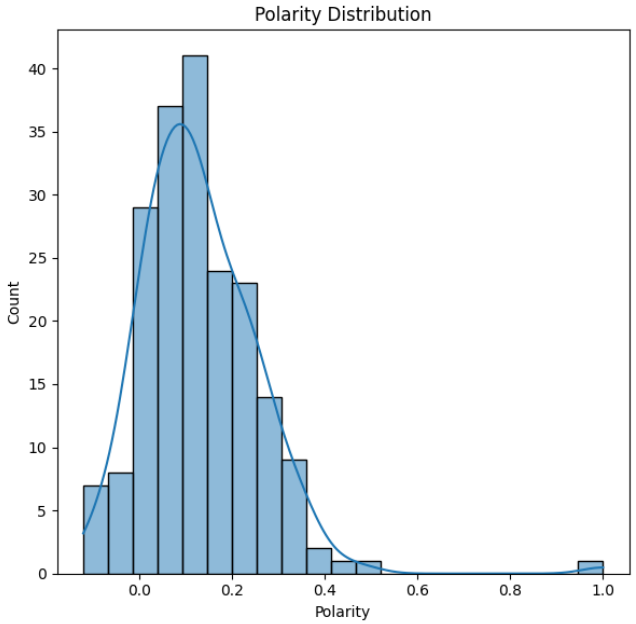


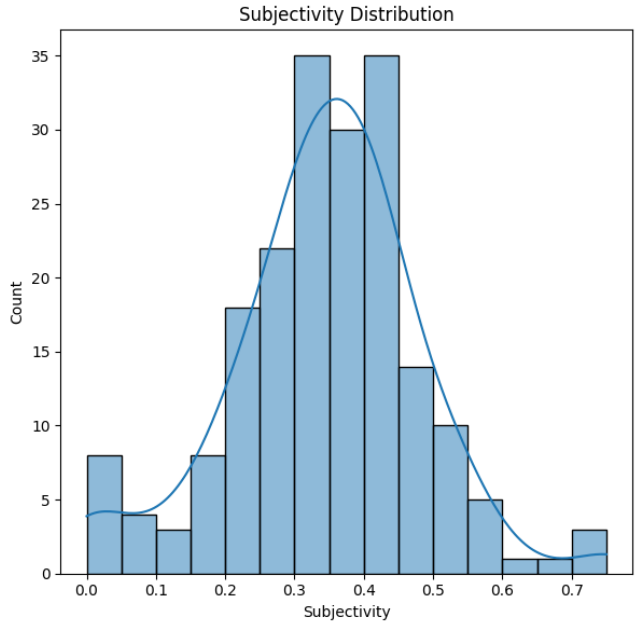


The data appears to have a relatively neutral sentiment distribution, with a peak around a score of 0.5 on the histogram.
However, there are data points with scores on either side of the neutral range, indicating the presence of both positive and negative sentiment.

Sentiment Analysis : Polarity and Sensitivity







Polarity Distribution:

- Polarity measures the sentiment expressed in the text, ranging from -1 (negative) to 1 (positive). However, in this plot, it seems to range from 0 to 1, indicating only positive sentiment has been considered.
- The histogram shows the frequency of various polarity scores.
- The majority of the data points are concentrated between 0 and 0.4, indicating that most of the text has low to moderate positive sentiment.
- The distribution is right-skewed, meaning there are fewer texts with high positive sentiment.

Subjectivity Distribution:

- Subjectivity measures how subjective or objective the text is, ranging from 0 (objective) to 1 (subjective).
- The histogram shows the frequency of various subjectivity scores.
- The data points are concentrated around the middle of the range (0.3 to 0.5), indicating that most of the text has a moderate level of subjectivity.
- The distribution appears to be approximately normal, centered around a subjectivity score of 0.4.

Recommended Models for Resume Parsing and Classification

To automate the extraction and classification of entities from resumes, the following models are recommended for their robustness, accuracy, and efficiency in Named Entity Recognition (NER) tasks:

Models:

SpaCy

- Strengths: Open-source library with pre-trained NER models for various languages, efficient for real-time applications.
- Weaknesses: Limited customization options for complex NER tasks compared to deep learning models.

BERT (Bidirectional Encoder Representations from Transformers)

- Strengths: Powerful deep learning model for various NLP tasks, including NER. Can be fine-tuned on domain-specific data for improved accuracy.
- Weaknesses: Requires significant computational resources for training and inference, can be a black box for interpretability.

Conditional Random Fields (CRF)

- Strengths: Probabilistic graphical model excelling at sequence labeling tasks like NER. Offers good accuracy with efficient training compared to deep learning models.
- Weaknesses: Requires feature engineering expertise for optimal performance, may not capture complex relationships between words as effectively as deep learning models.

The best model choice depends on the specific requirements and dataset characteristics. Experimentation is key to finding the optimal solution for the resume parsing and classification task.

Thank You