Data Glacier

Your Deep Learning Partner

Resume Parsing and Classification Using Named Entity Recognition (NER)

Project Presentation

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: 02.07.2024

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:

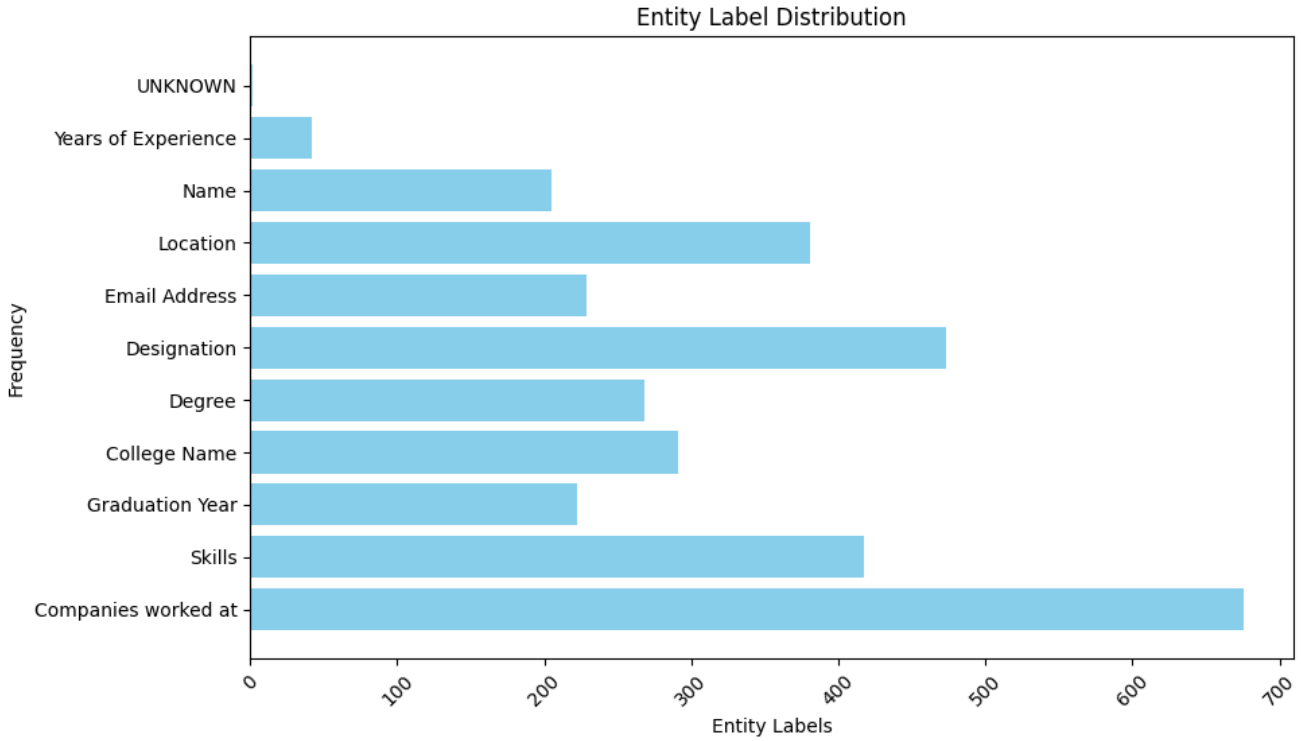
label: the type of entity

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



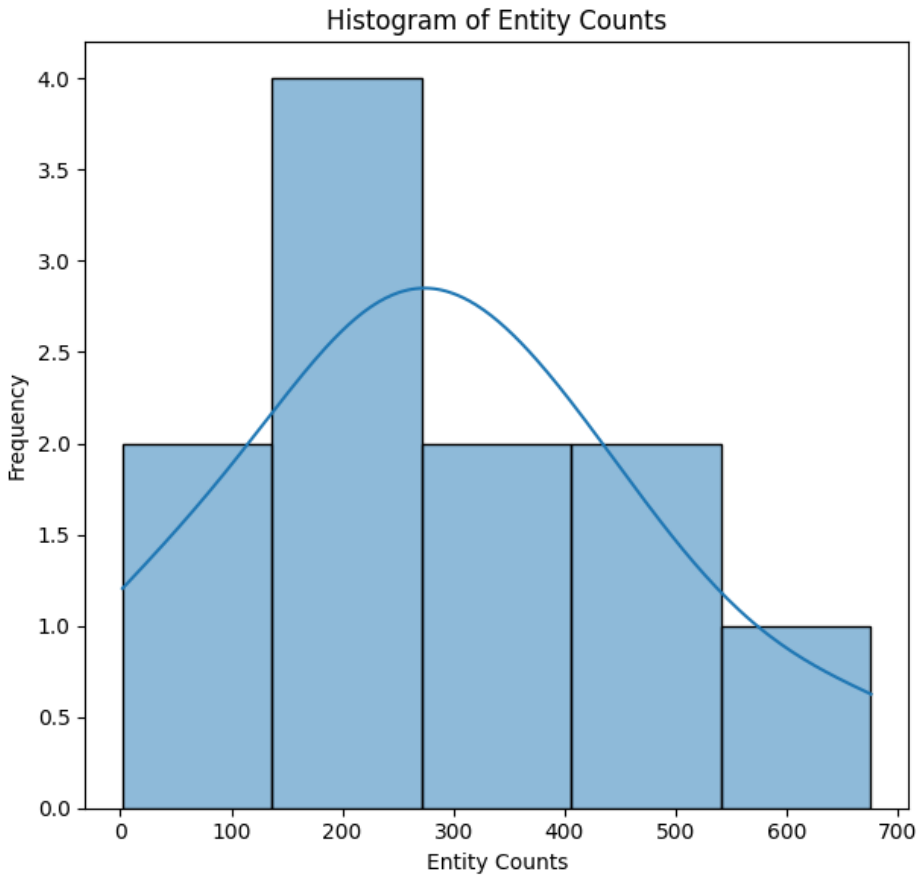

```
'Companies worked at': 676,
'Skills': 417,
'Graduation Year': 222,
'College Name': 291,
'Degree': 268,
'Designation': 473,
'Email Address': 229,
'Location': 381,
'Name': 205,
'Years of Experience': 42,
'UNKNOWN': 2
```

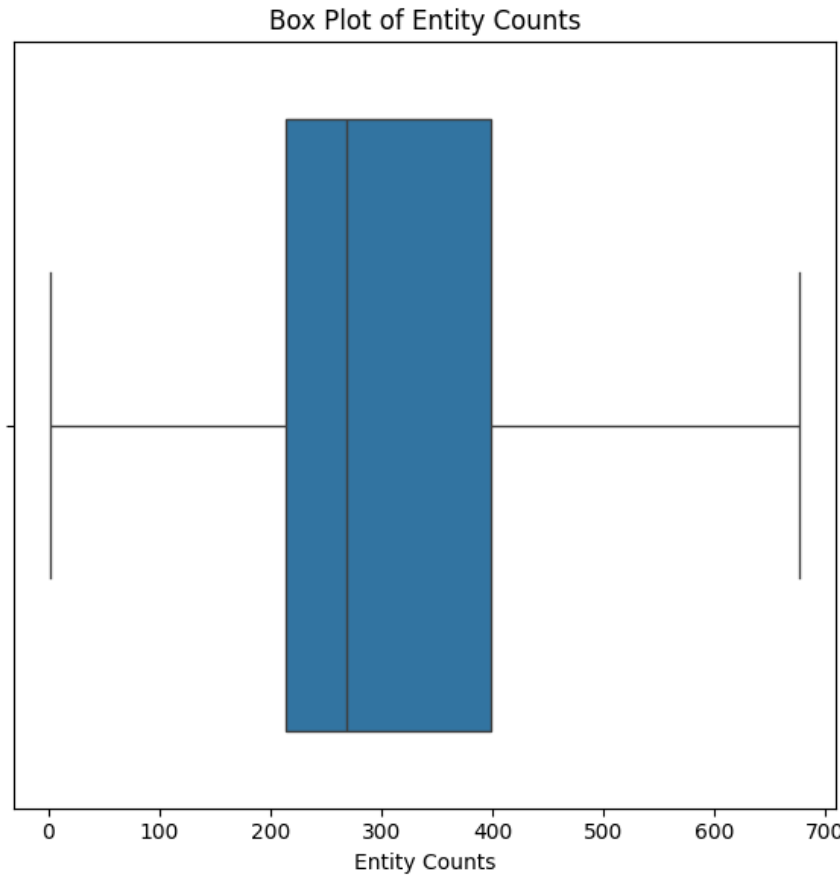
Distribution Analysis of Textual Features: Insights on Skewness and Kurtosis

•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.

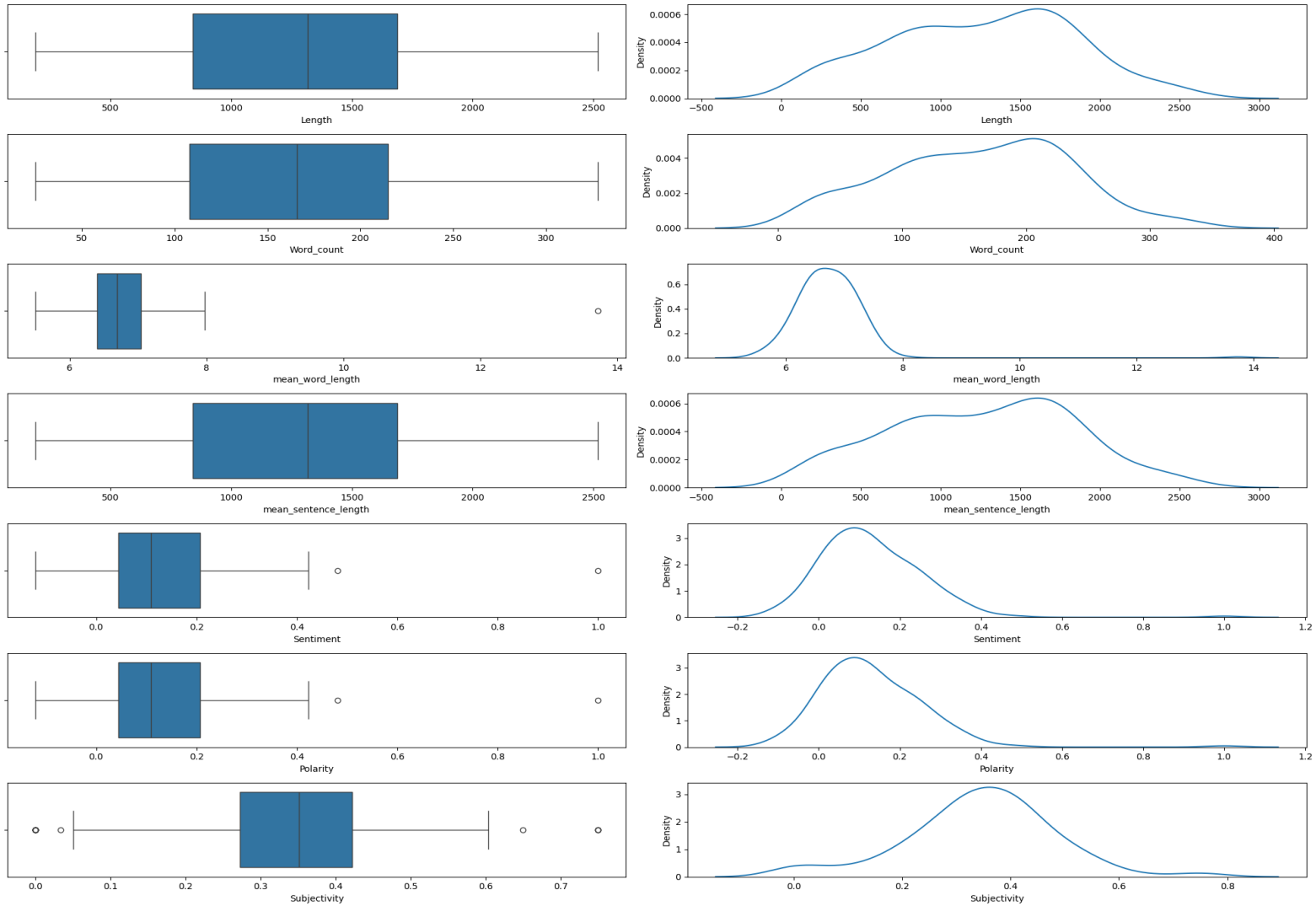
- **Positive Skew**: The tail on the right side of the distribution is longer or fatter than the left side.
- **Negative Skew**: The tail on the left side is longer or fatter than the right side.
- **Zero Skew**: Symmetrical distribution (normal distribution).

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.

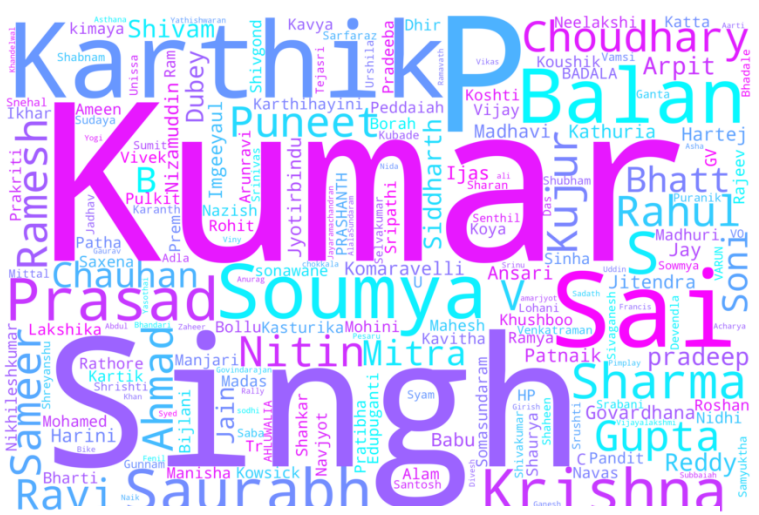
Positive Skewness for Mean Word Length suggests that a few resumes use significantly longer words, which is less common.

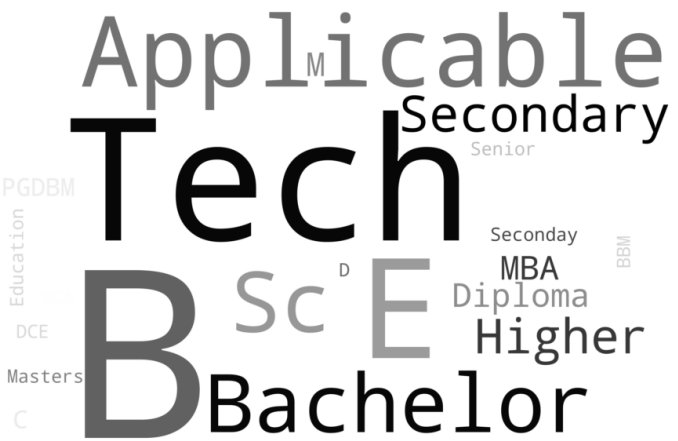
Low Kurtosis for Length, Word Count, and Mean Sentence Length implies that these distributions are relatively flat with fewer outliers, indicating a more consistent dataset.

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.

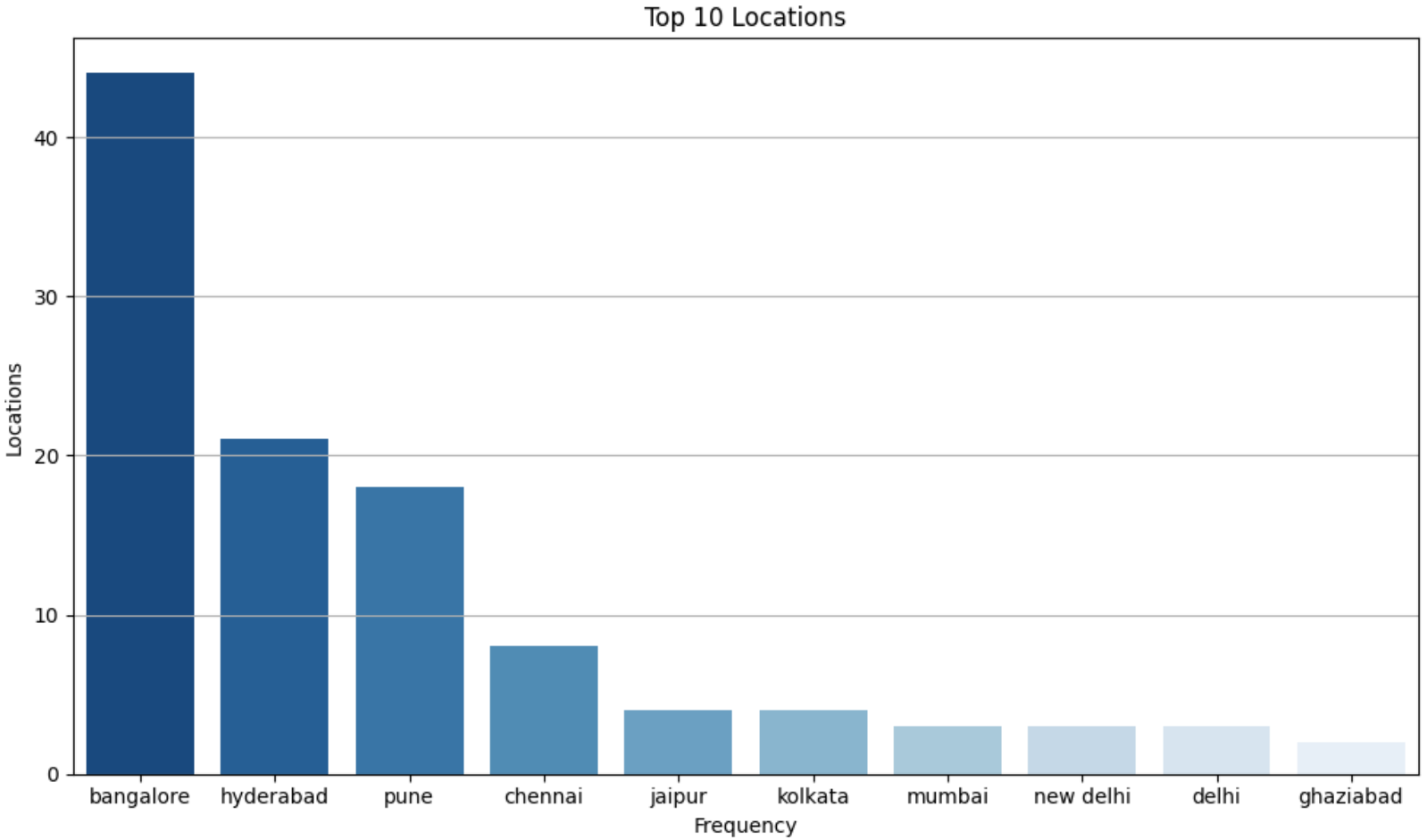
Insights Generation

WordCloud for entities



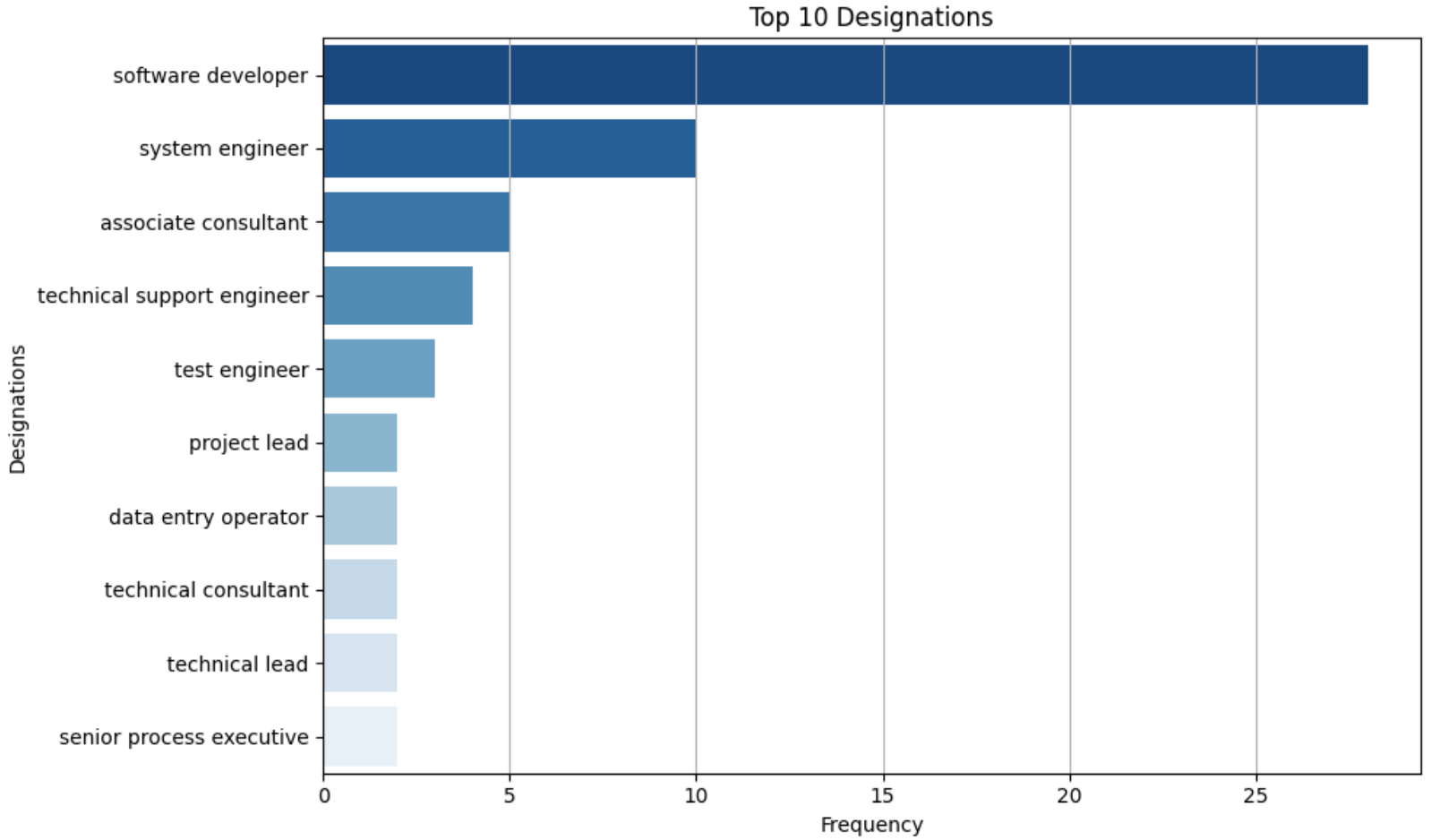


Entity Distribution: Location



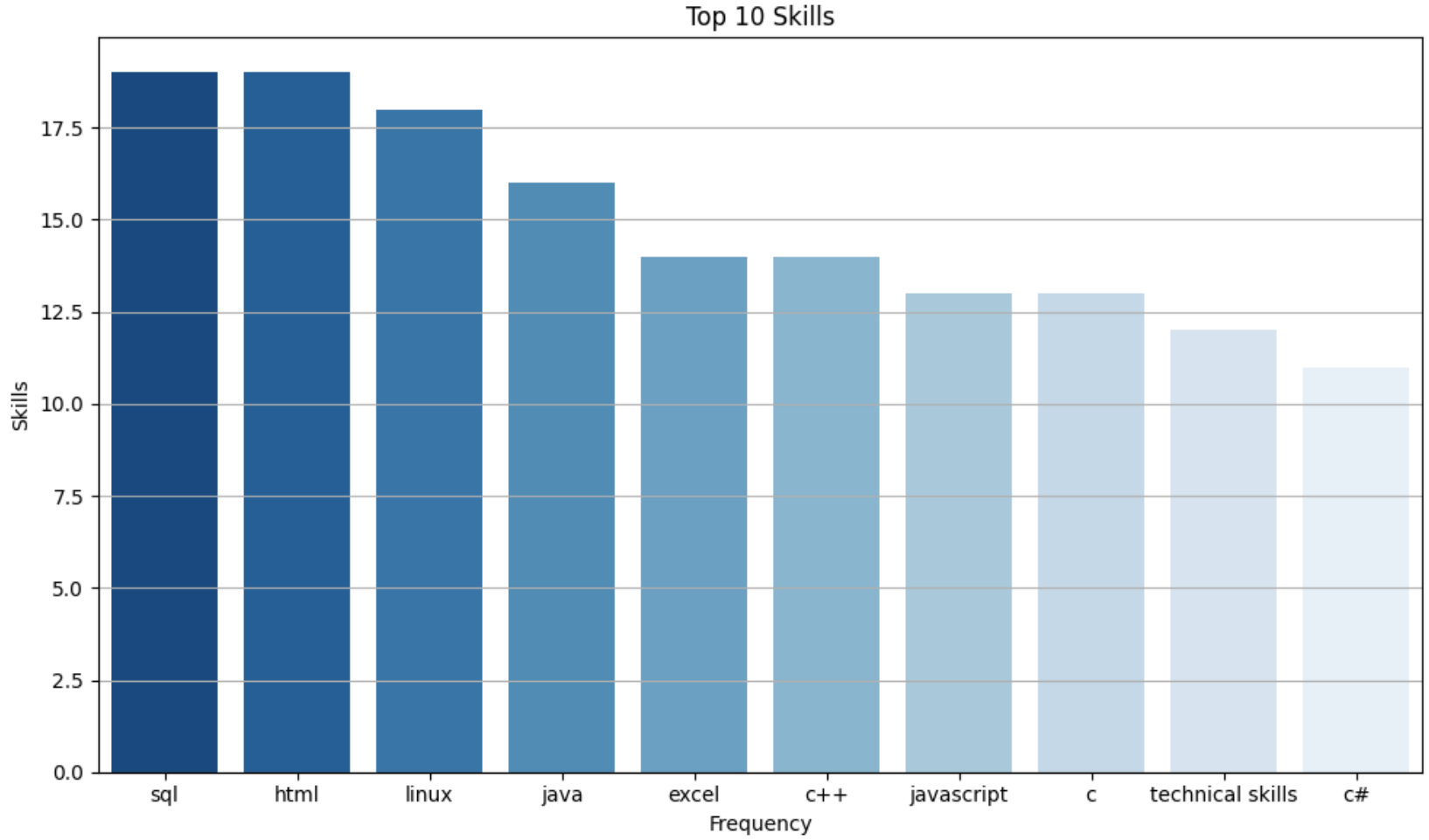
A significant portion (over 40+) applicants are located in Bangalore. This might be due to the city's reputation as a tech hub. Hyderabad and Pune also have a good pool of talent we can tap into.

Entity Distribution: Designation



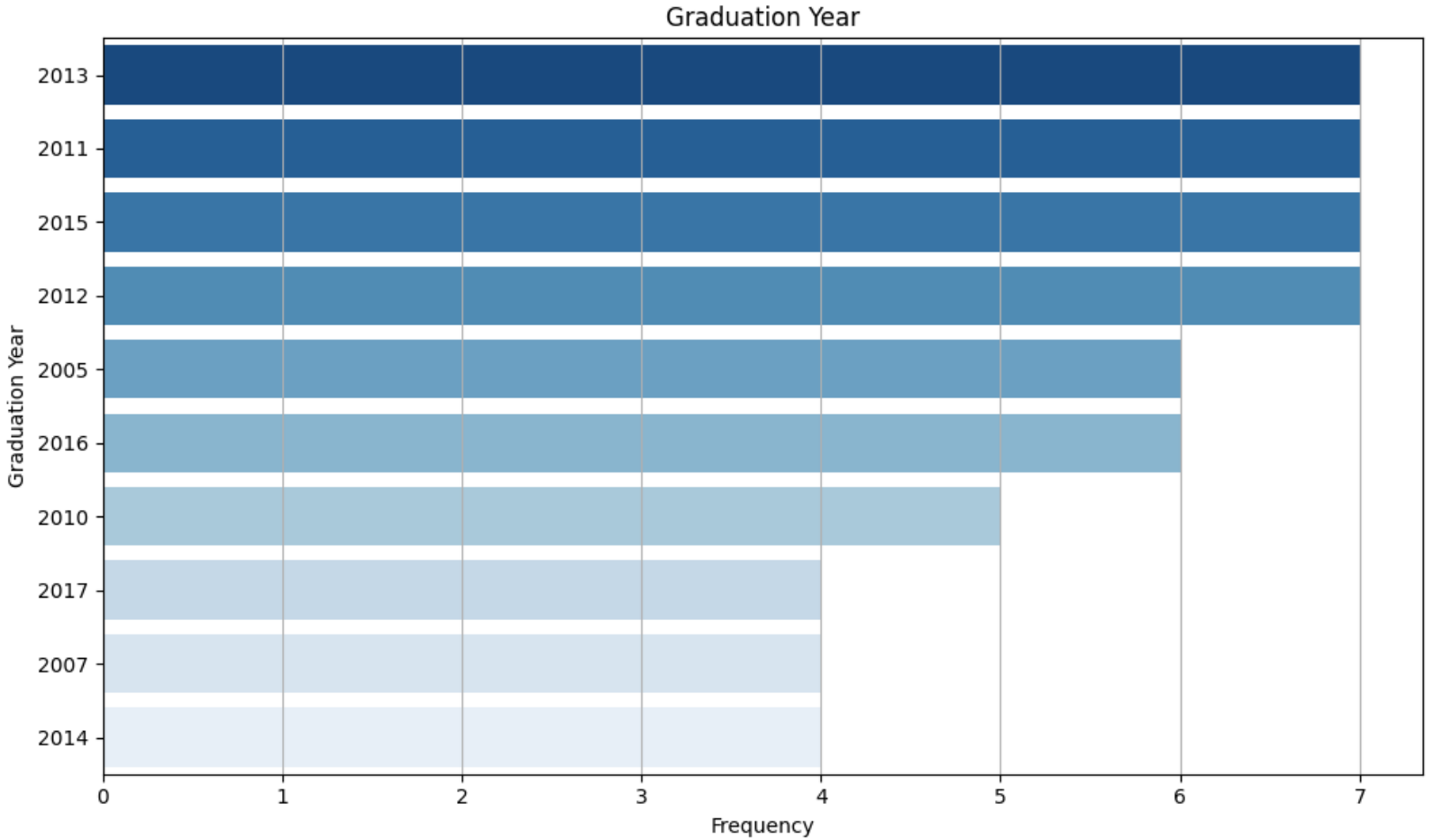
Here, we see Software Developer and System Engineer as the most common designations. This aligns well with the tech industry focus of the applicant pool we saw in the previous slide.

Entity Distribution: Skills



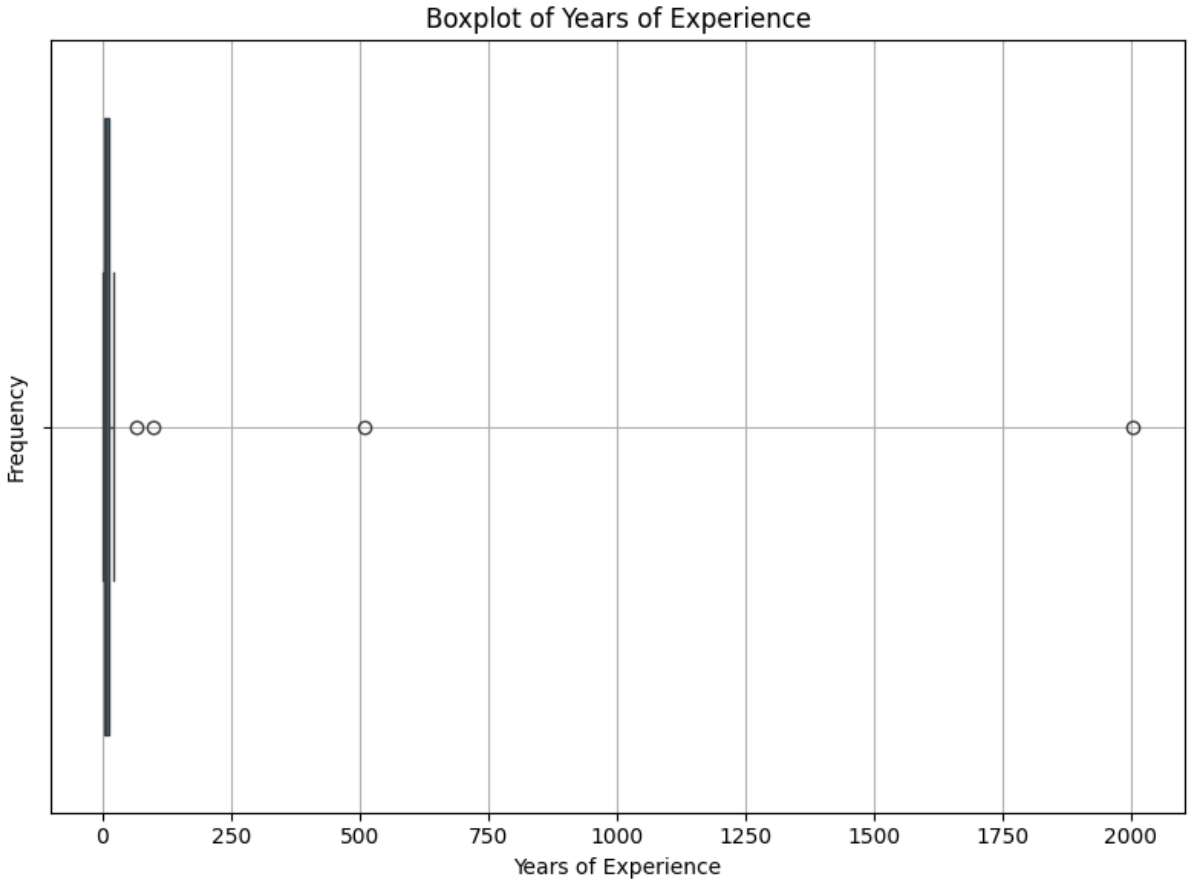
This slide highlights the most sought-after skills among the applicants. These skills are all in high demand within the tech industry. SQL, HTML, Linux, and Java are the top skills listed by applicants, with over 17 individuals mentioning each

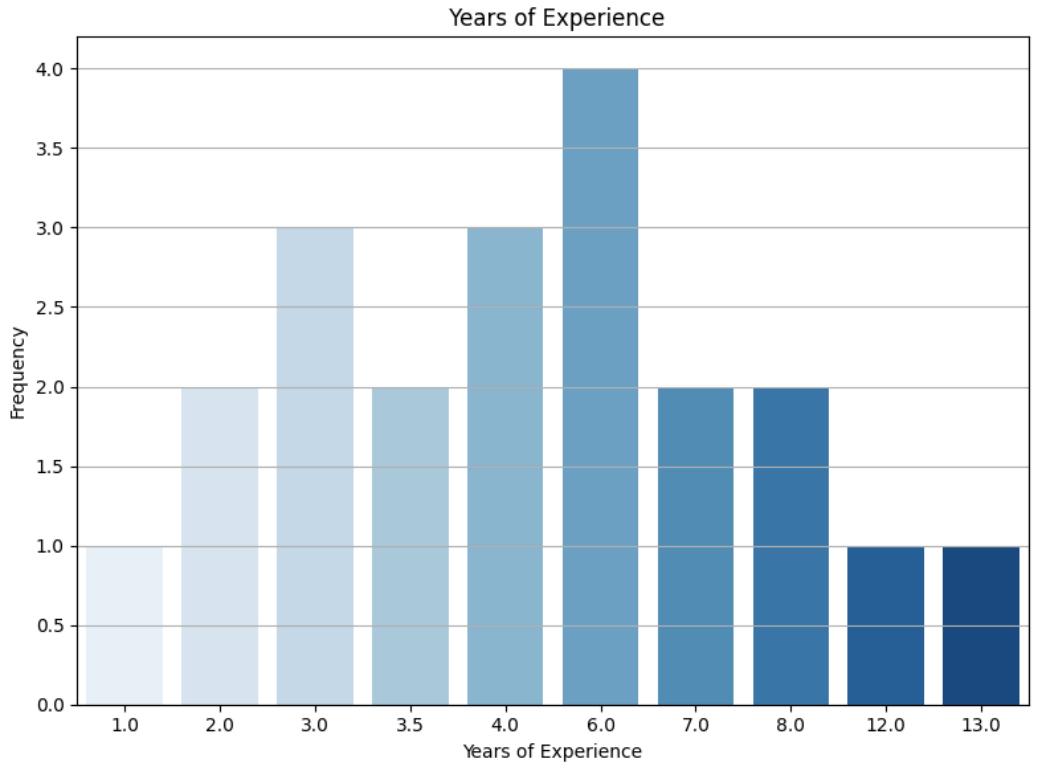
Entity Distribution: Graduation years



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.

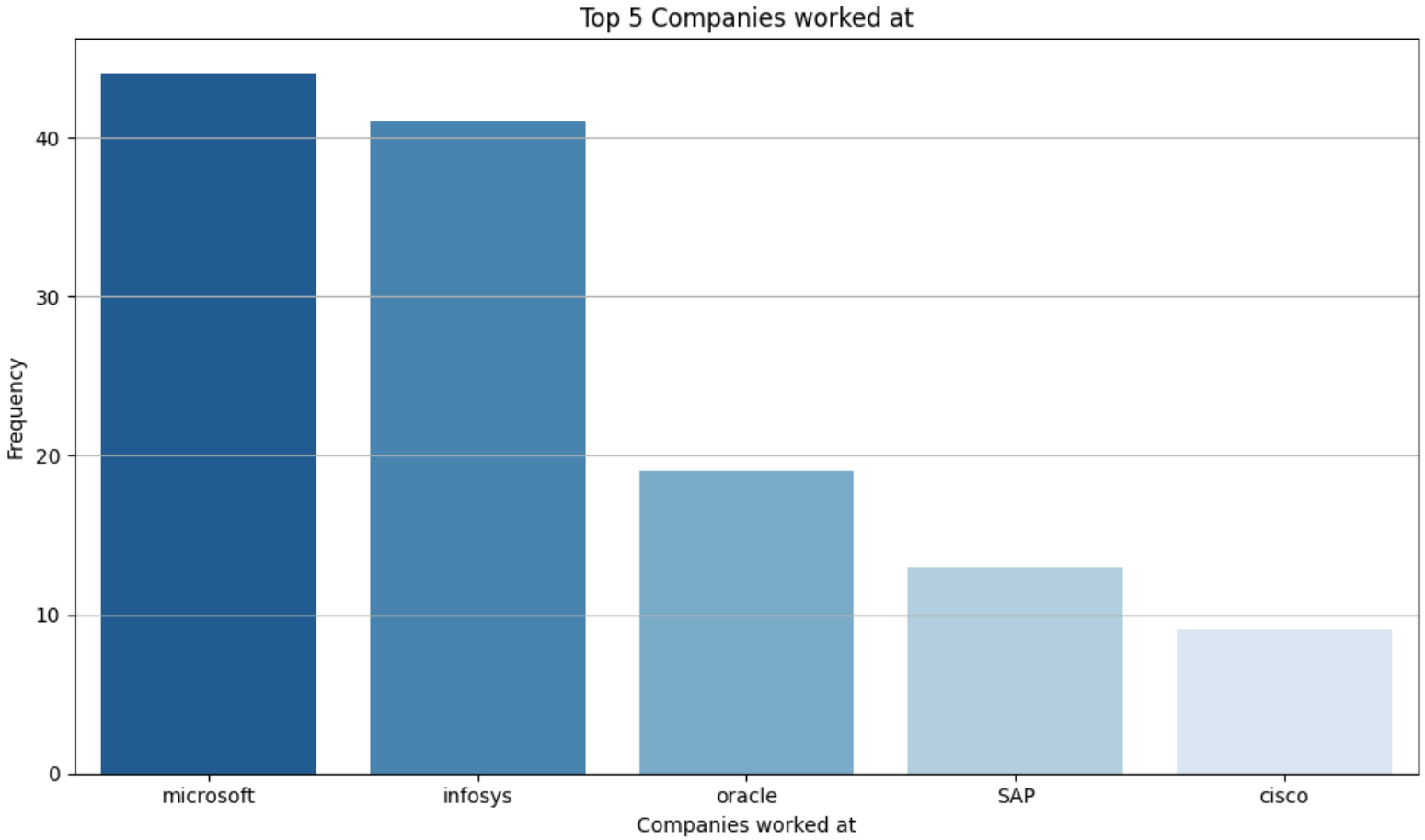
Entity Distribution: Years of Experience





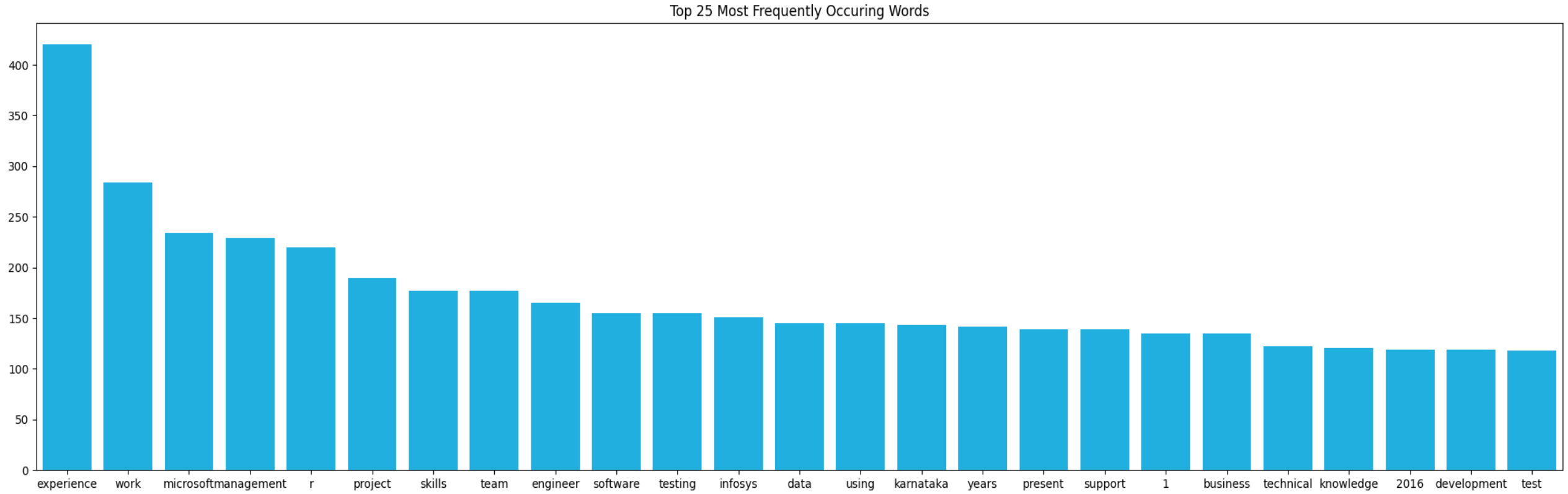
Years of experience had outliers, because it had the year ‘511’ and ‘2004’. fter removing them, this slide presents a clearer picture of the distribution of experience levels within the applicant pool.

Entity Distribution: Companies worked at



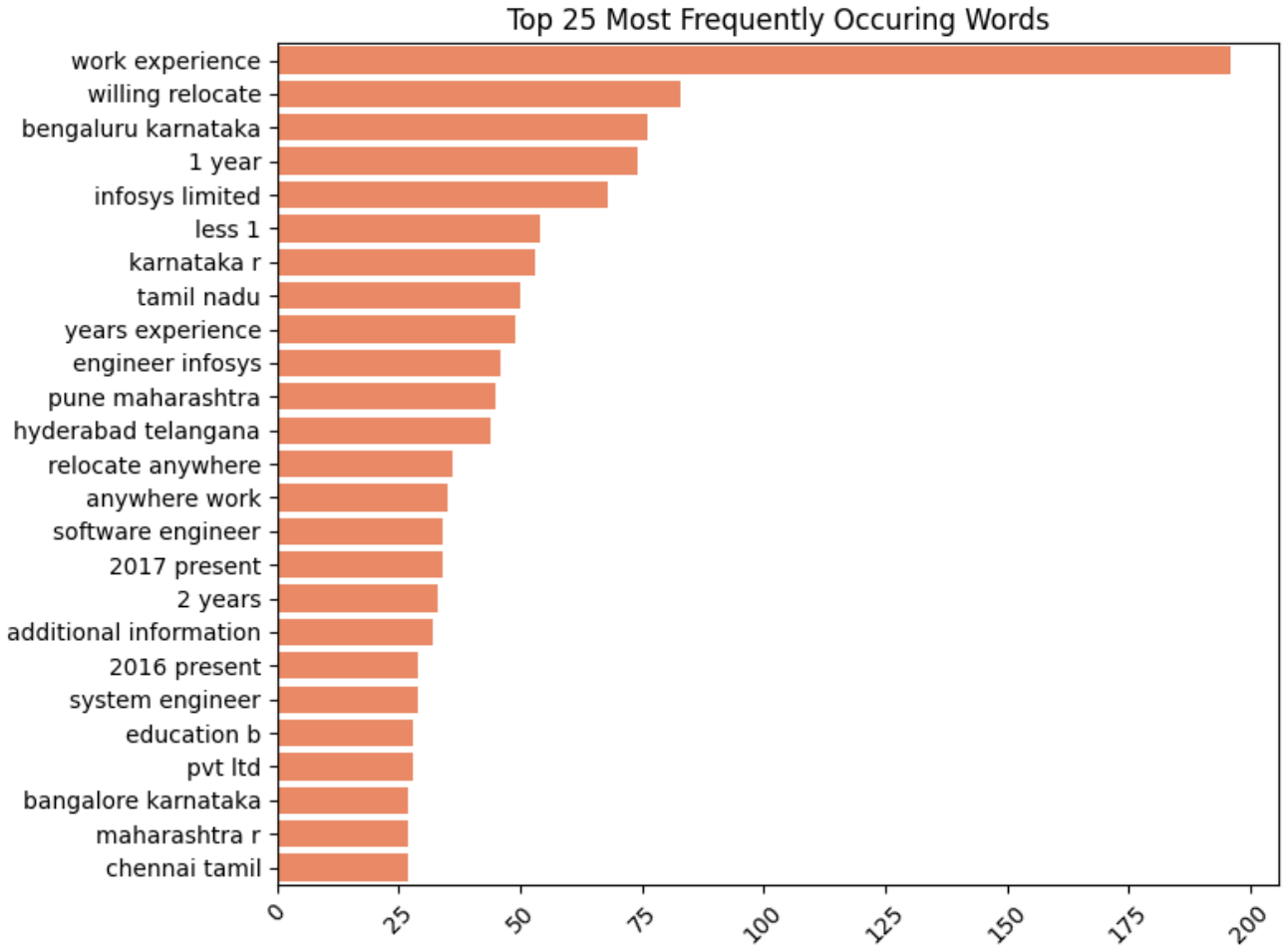
Analysis of the data reveals that a significant number of applicants (over 40) have prior experience at Microsoft and Infosys. Oracle is another frequently mentioned company.

N gram: Unigram



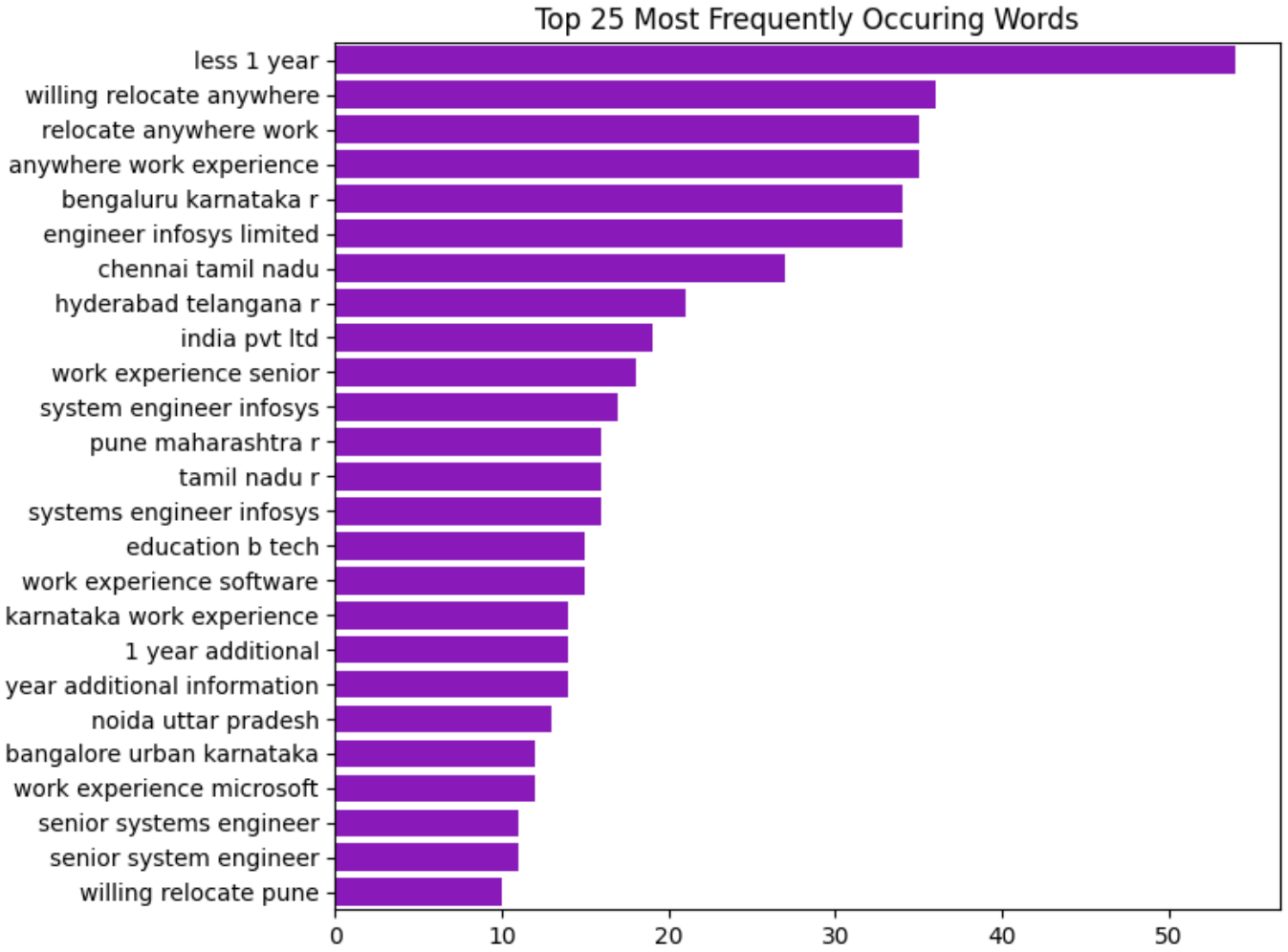
Unigram: Top words are experience, work, Microsoft and management

N gram: Bigram



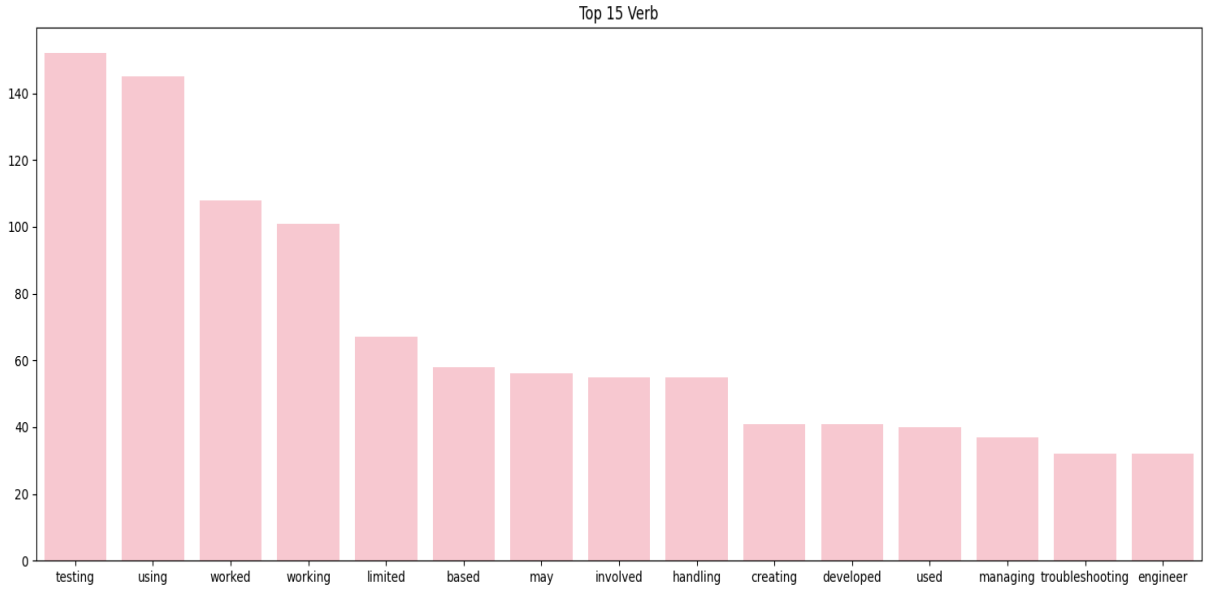
Top Bigrams are ‘Work Expereience’, ‘willing relocate’,’Bengaluru karnataka’

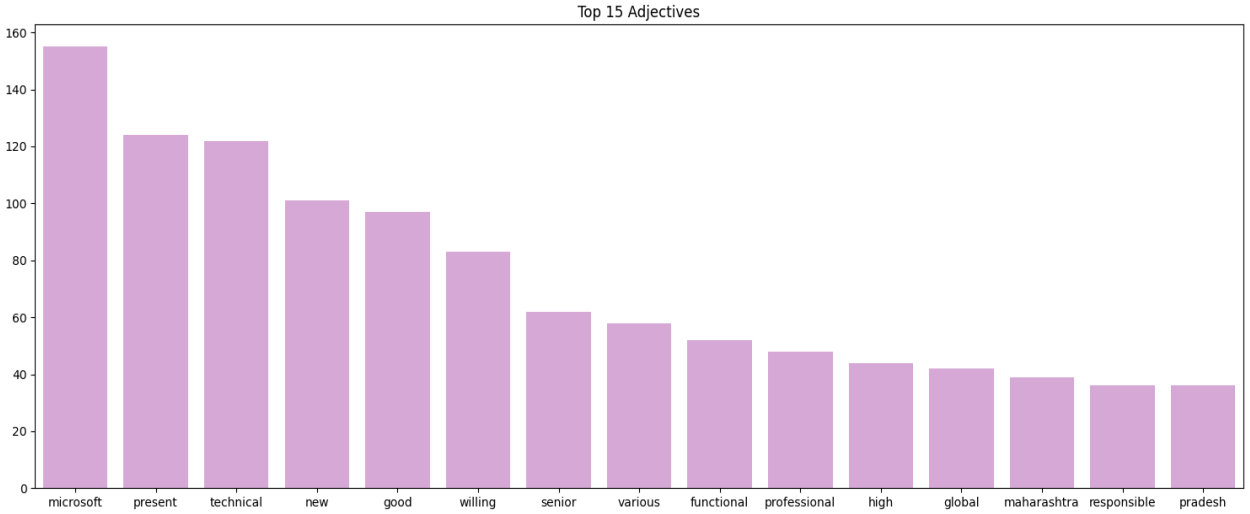
N gram: Trigram



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.

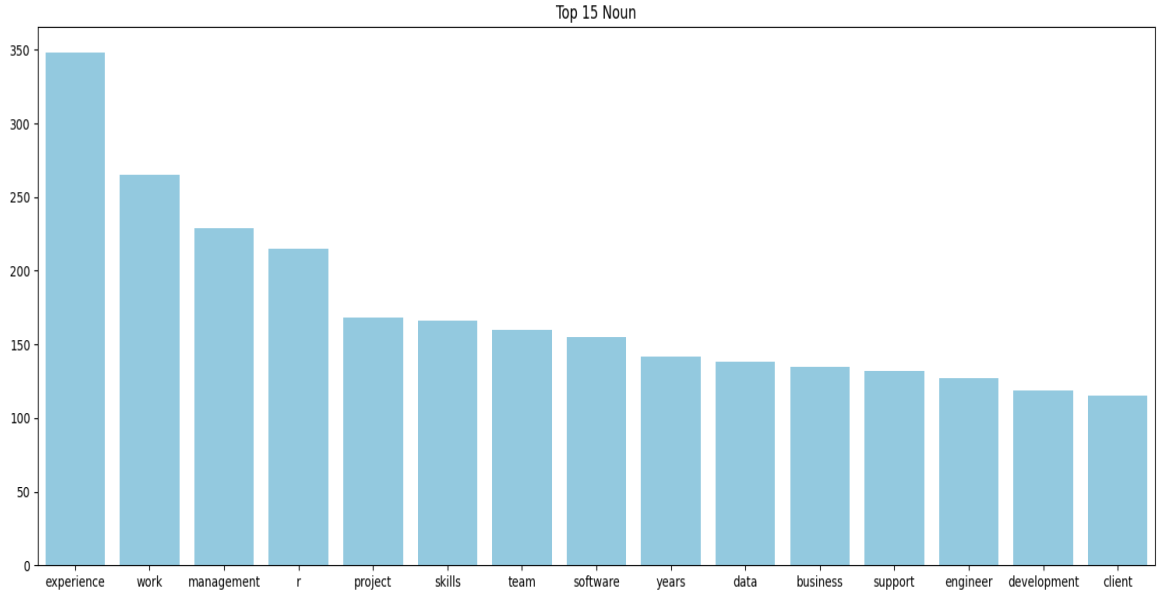
Parts of Speech





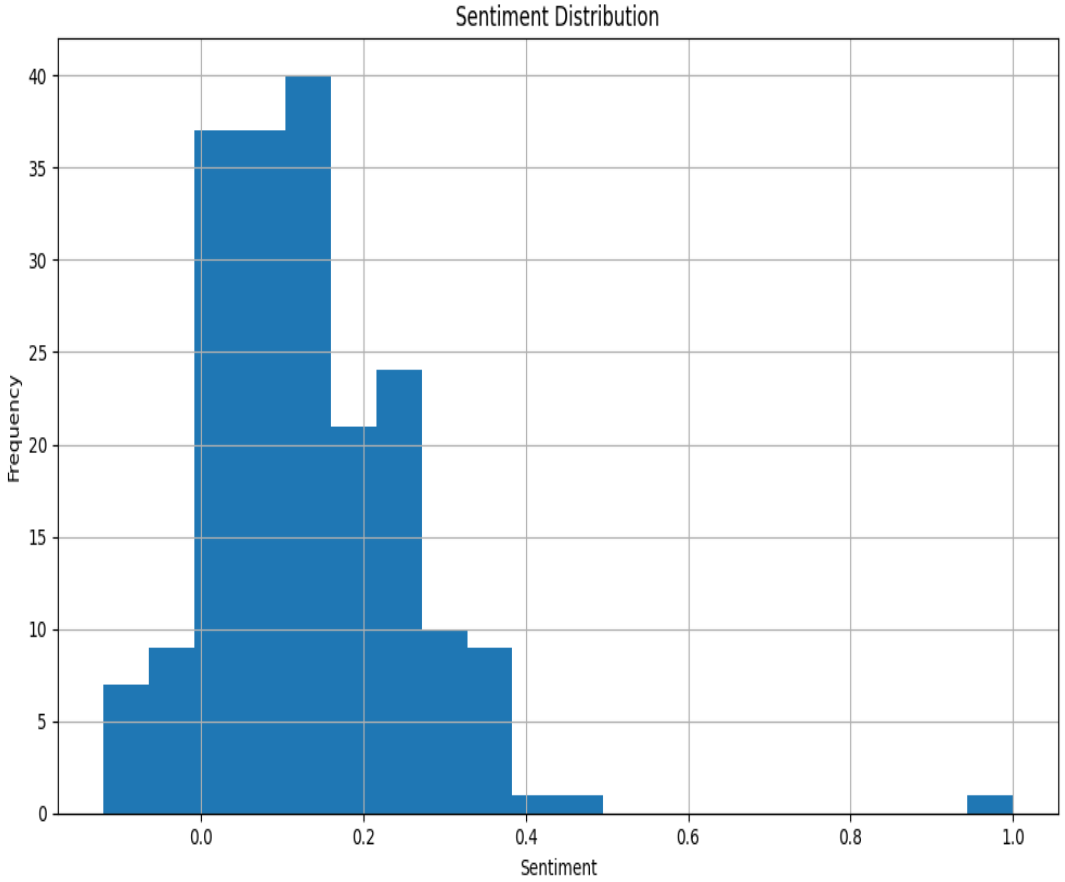
Parts of Speech refer to the categories of words based on their function in a sentence. The main POS categories are:

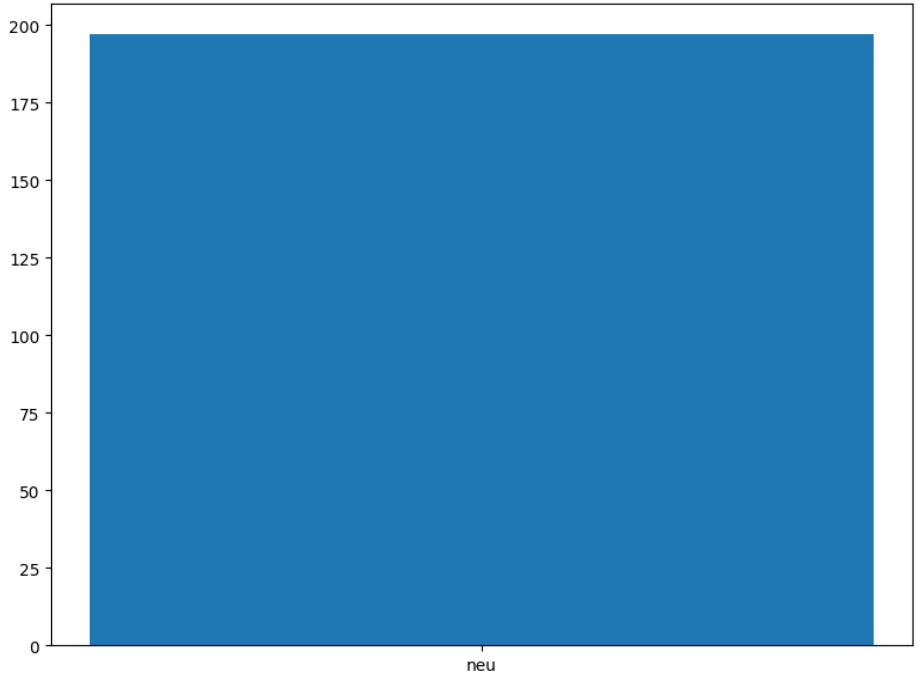
Nouns: Words that name people, places, things, or ideas
Verbs: Words that express actions or states of being
Adjectives: Words that describe or modify
Adverbs: Words that modify verbs, adjectives, or other adverbs



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.

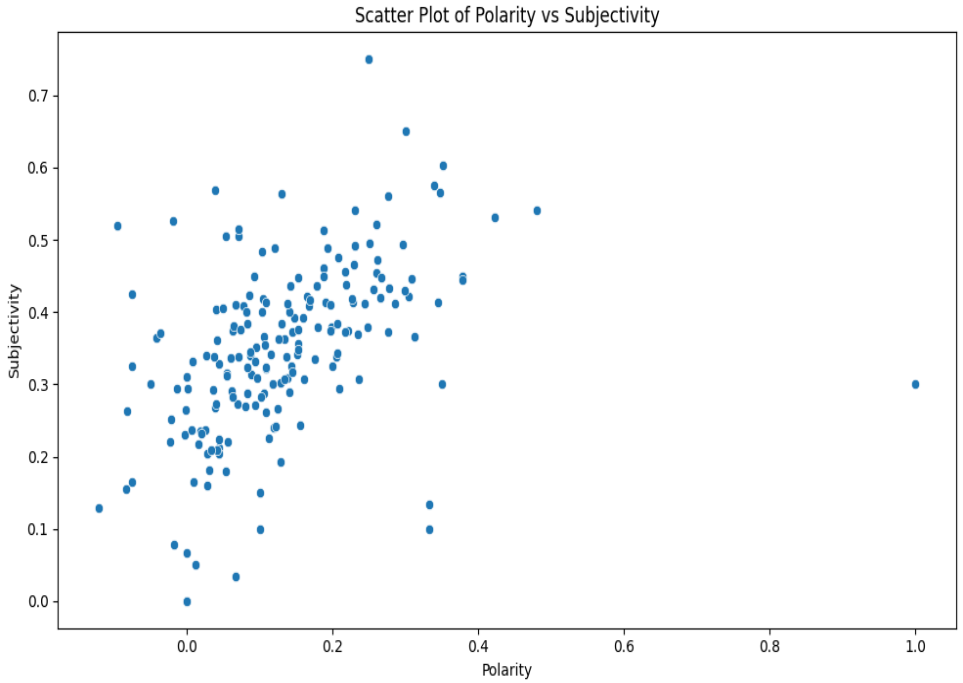
Sentiment Analysis

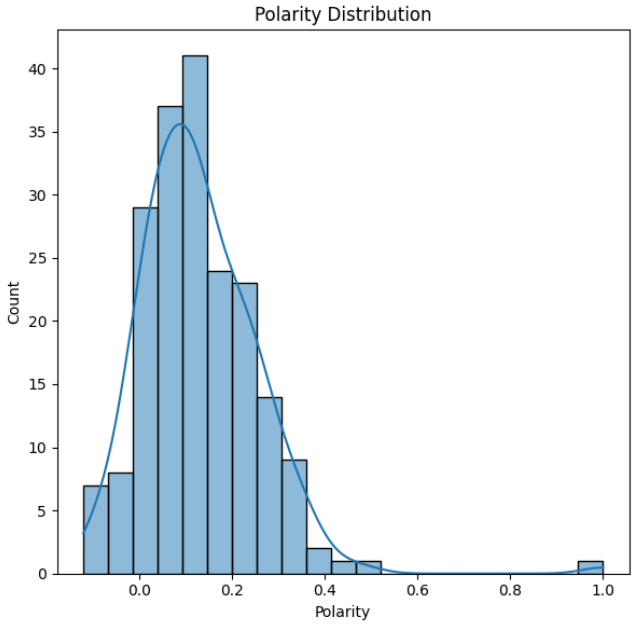


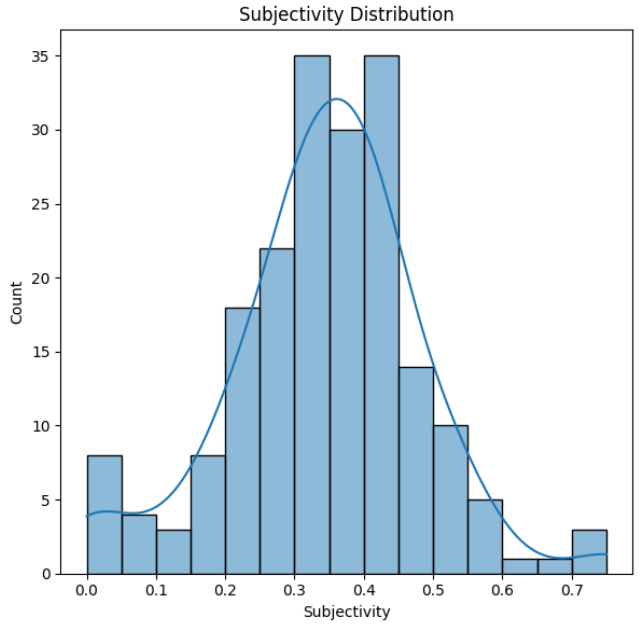


The data appears to have a relatively neutral sentiment distribution, with a peak around a score of 0.5 on the histogram.
However, there are data points with scores on either side of the neutral range, indicating the presence of both positive and negative sentiment.

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.

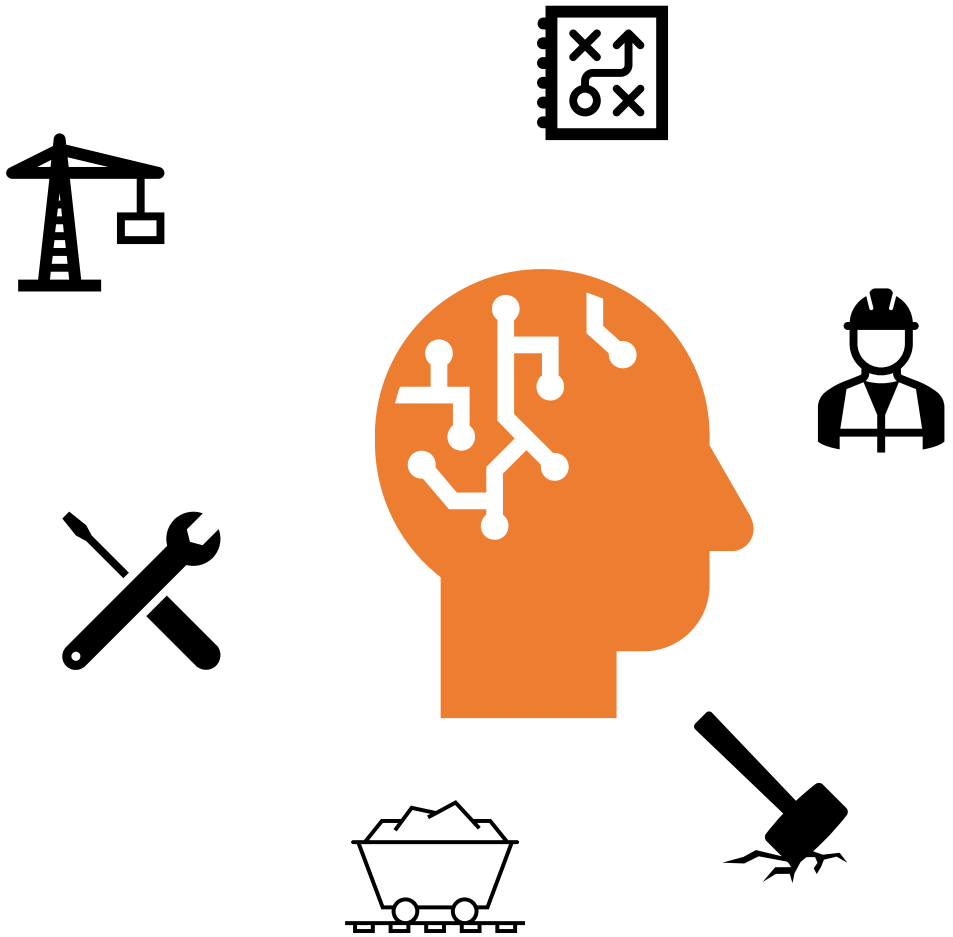
BERT (Bidirectional Encoder Representations from Transformers)

- 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.



MODEL BUILDING AND TESTING

Sample input text to models

“””Pratibha P SOFTWARE ENGINEER CONTACT cmcturland@indeed.com (123) 456-7890 Chennai, India EDUCATION B.S.Computer Science University of Pittsburgh September 2008 - April 2012 Kochin, Kerela SKILLS Python (Django) Javascript (NodeJS ReactJS, jQuery) SQL (MySQL, PostgreSQL, NoSQL) HTML5/CSS AWS Unix, Git WORK EXPERIENCE Software Engineer January 2015 - current / Delhi, India Worked with product managers to re-architect a multi-page web app into a single page web-app, boosting yearly revenue by $1.4M Constructed the logic for a streamlined ad-serving platform that scaled to our 35M users, which improved the page speed by 15% after implementation Tested software for bugs and operating speed, fixing bugs and documenting processes to increase efficiency by 18% Iterated platform for college admissions, collaborating with a group of 4 engineers to create features across the software MarketSmart Software Engineer April 2012 - January 2015 / Delhi, India Built RESTful APIs that served data to the JavaScript front-end based on dynamically chosen user inputs that handled over 500,000 concurrent users Built internal tool using NodeJS and Pupeteer.js to automate QA and monitoring of donor-facing web app, which improved CTR by 3% Reviewed code and conducted testing for 3 additional features on donor-facing web app that increased contributions by 12% Software Engineer Intern Marketing ScienceCompany April 2011 - March 2012 / Pittsburgh, PA Partnered with a developer to implement RESTful APIs in Django, enabling analytics team to increase reporting speed by 24% Using Selenium I built out a unit testing infrastructure for a client application that reduced the number of bugs reported by the client by 11% month over month PROJECTS Poker Simulation Built a full-stack web app to allow users to simulate and visualize outcomes of poker hands against opponents of different play styles using open source cards.js on the front-end Utilized sci-kit learn in Python to simulate possible outcomes under different scenarios that the user chose """

Spacy model predicted Output

Name
Pratibha P : 0 10

Designation
SOFTWARE ENGINEER : 12 29

College Name
University of Pittsburgh : 115 137

Skills
Python (Django) Javascript (NodeJS ReactJS, jQuery) SQL (MySQL, PostgreSQL, NoSQL) HTML5/CSS AWS Unix, Git : 200 312

Designation
Software Engineer : 324 341

Explanation:

•**Name**: Detected entity "Pratibha P" starting at position 0 and ending at position 10 in the text.
•**Designation**: Detected entity "SOFTWARE ENGINEER" starting at position 12 and ending at position 29.
•**College Name**: Detected entity "University of Pittsburgh" starting at position 115 and ending at position 137.
•**Skills**: Detected a list of skills starting at position 200 and ending at position 312.
•**Designation**: Detected entity "Software Engineer" starting at position 324 and ending at position 341.

The SpaCy model successfully identifies and extracts key entities such as personal name, job designation, educational institution, skills, and work experience designation from the resume text. This output demonstrates the model's capability to accurately recognize and categorize relevant information, facilitating automated extraction and classification of resume content.
This demonstrates how the SpaCy model can effectively automate the process of extracting structured information from resumes, aiding in tasks such as resume screening and candidate evaluation.

Spacy Model Performance

Entity	Precision	Recall	F1-score	Support
College Name	0.88	0.96	0.92	106
Companies worked at	0.58	0.89	0.71	110
Degree	0.58	1.00	0.74	57
Designation	0.46	0.80	0.58	140
Email Address	1.00	0.87	0.93	23
Graduation Year	0.56	0.50	0.53	20
Location	0.93	0.68	0.78	37
Name	1.00	1.00	1.00	40
Skills	0.94	0.56	0.70	1260
UNKNOWN	-	-	-	-
Years of Experience	1.00	0.55	0.71	11
O (non-entity)	0.94	0.97	0.96	10312
Micro Avg	0.92	0.92	0.92	12116
Macro Avg	0.74	0.73	0.71	12116
Weighted Avg	0.93	0.92	0.92	12116

The SpaCy model demonstrated competitive performance across most entity types.It excelled particularly in Name (precision=1.00, recall=0.83, f1-score=0.91), Email Address (precision=0.89, recall=0.89, f1-score=0.89), and Skills (precision=0.90, recall=0.50, f1-score=0.64). Notably, it achieved a high F1-score of 0.96 for the O (non-entity) category, indicating robust generalization.

BERT model predicted Output

Predicted Tags:

pr Name
##ati Name
p Designation
in College Name

•The predicted tags include partial entities such as "Prati" for Name, "p" for Designation, and "in" for College Name.

•Further refinement and training may be necessary to improve accuracy and completeness of entity recognition.

BERT Model performance

Entity	Precision	Recall	F1-score	Support
College Name	0.03	0.02	0.03	45
Companies worked at	0.03	0.02	0.02	53
Degree	0.07	0.02	0.03	52
Designation	0.54	0.38	0.45	84
Email Address	0.12	0.09	0.11	11
Empty	0.95	0.98	0.97	9652
Graduation Year	0.00	0.00	0.00	5
Location	0.25	0.11	0.15	9
Name	0.38	0.89	0.53	45
Skills	0.40	0.04	0.07	273
Years of Experience	0.00	0.00	0.00	11
Micro Avg	0.93	0.93	0.93	10240
Macro Avg	0.25	0.23	0.21	10240
Weighted Avg	0.92	0.93	0.92	10240

BERT model, while performing well on UNKNOWN and O categories with high precision and recall, struggled with lower support entity types.It achieved higher performance in Name (precision=0.38, recall=0.89, f1-score=0.53) and Designation (precision=0.54, recall=0.38, f1-score=0.45) compared to entities like College Name and Companies worked at.

CRF model predicted Output

Pratibha	B-Name
P	I-Name
SOFTWARE	B-Designation
ENGINEER	I-Designation
B.S.Computer	B-Degree
Science	B-College Name
University	I-College Name
Software	B-Designation
Engineer	I-Designation

Explanation:

•**B-Name:** "Pratibha P" is labeled as a name, with "Pratibha" marked as the beginning (B-Name) and "P" as inside the name (**I-Name**).
•**B-Designation and I-Designation:** "SOFTWARE ENGINEER" is tagged as a designation, where "SOFTWARE" is the beginning (B-Designation) and "ENGINEER" is inside the designation (I-Designation).
•**B-College Name and I-College Name:** "B.S. Computer Science" is identified as a degree and college name, with "B.S." marked as the beginning of the degree (B-Degree) and "Computer Science" as the college name (B-College Name). "University of Pittsburgh" is marked as inside the college name (I-College Name).

CRF Model Performance

Entity	Precision	Recall	F1-score	Support
B-College Name	0.90	0.45	0.60	20
I-College Name	0.86	0.59	0.70	61
B-Companies worked at	0.65	0.26	0.37	42
I-Companies worked at	0.43	0.12	0.19	25
B-Degree	1.00	0.62	0.77	16
I-Degree	0.95	0.95	0.95	37
B-Designation	0.83	0.60	0.70	25
I-Designation	0.96	0.57	0.72	40
B-Email Address	0.89	0.89	0.89	9
I-Email Address	1.00	0.78	0.88	9
B-Graduation Year	1.00	0.33	0.50	12
B-Location	0.33	0.36	0.35	11
I-Location	0.00	0.00	0.00	1
B-Name	1.00	0.83	0.91	12
I-Name	0.91	0.83	0.87	12
O	0.94	0.98	0.96	4568
B-Skills	0.90	0.50	0.64	18
I-Skills	0.77	0.67	0.72	491
B-Years of Experience	0.00	0.00	0.00	1
I-Years of Experience	0.00	0.00	0.00	1
Micro Avg	0.92	0.92	0.92	5411
Macro Avg	0.72	0.52	0.59	5411
Weighted Avg	0.92	0.92	0.91	5411

The CRF model achieved varying levels of precision, recall, and F1-score across different entity types.It showed strong performance for B-Degree (precision=1.00, recall=0.62, f1-score=0.77) and I-Degree (precision=0.95, recall=0.95, f1-score=0.95).
However, it struggled with entities like B-Companies worked at (precision=0.65, recall=0.26, f1-score=0.37) and I-Companies worked at (precision=0.43, recall=0.12, f1-score=0.19), where precision and recall were lower.

Conclusion

After evaluating the three models, the SpaCy model was chosen for its superior performance in precision, recall, and F1-score across most entity categories. While the CRF model performed well, particularly for certain entities, and the BERT model showed potential, especially for entities like 'Designation' and 'Name,' the SpaCy model consistently provided the most balanced and reliable results overall. This model will significantly enhance the efficiency and accuracy of the resume screening process, making it faster and less prone to human error.

Thank You