INTERNSHIP PROJECT - SLINTEL

-By Mohit Bagaria (17EX20017)

SLINTEL

Table Of Contents

ABOUT - SLINTEL

- A Sales Intelligence Platform, which helps to Identify Potential buyers in their Market & uncover top 5% prospects in their segment, using recent data
- Use Slintel to identify the best selling opportunities that you can reach out to today
- Get verified emails and direct dials for active, high intent buyers in your target markets
- Understand buyer behaviour and pain points using buyer journeys and keyword insights
- Streamline your pitch with technology adoption data
- All in one place- company, contact, technology and intent data

Understanding the problem statement

01 Technologies Extraction From Linkedin Summaries

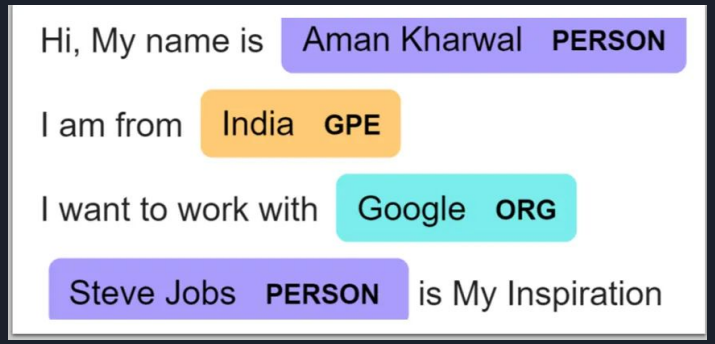
02 Data Available-

- Technologies Names
- Scraped Linkedin Profiles

Project objective

To extract the Organisation names and the Technological keywords mentioned in their LinkedIn About or Summary section

This Problem Statement is Termed as NAMED ENTITY RECOGNITION (NER)



SLINTEL

Use case

01 Slintel's product has a Dashboard which provides direct leads of the People from the targeted companies

02 At this Moment, the people sometimes were not relevant

03 Task is to find people Using specific Technologies from a Specific Organization (Company or our target prospect)

Requirement

Tagged Dataset for Obtaining Accuracy Matrix

Solution: Manually Tagged Data with Context | String Match Run

Drawback:

1. Human Intervention
2. Takes high amount of time

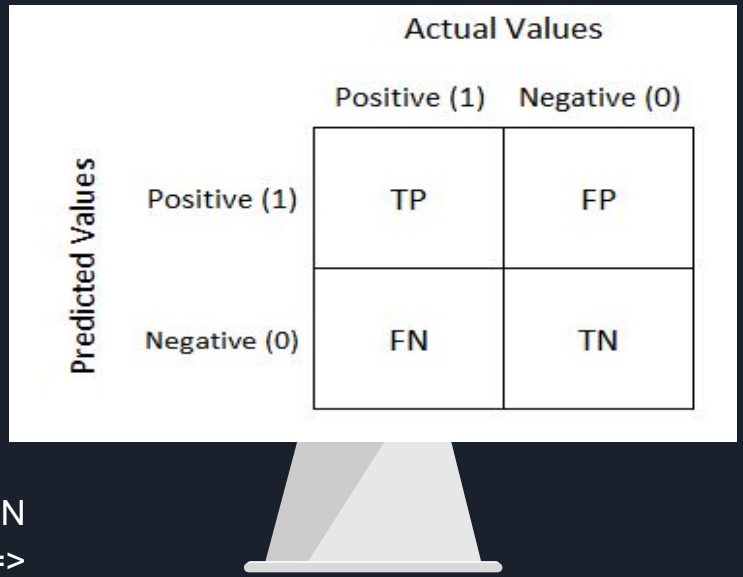
Introducing: Confusion Matrix

Accuracy Measure:

Decided to Use Confusion Matrix

- True Positive (TP)
- False Positive (FP)
- False Negative (FN)
- True Negative (TN)
- Precision (true positives/predicted positives)=TP/TP+FP
- Recall (true positives/all actual positives)=TP/TP+FN
- F Score : Harmonic Mean of Precision and Recall => 2*(precision*recall)/(precision+recall)

Range : (0,1) Higher the better



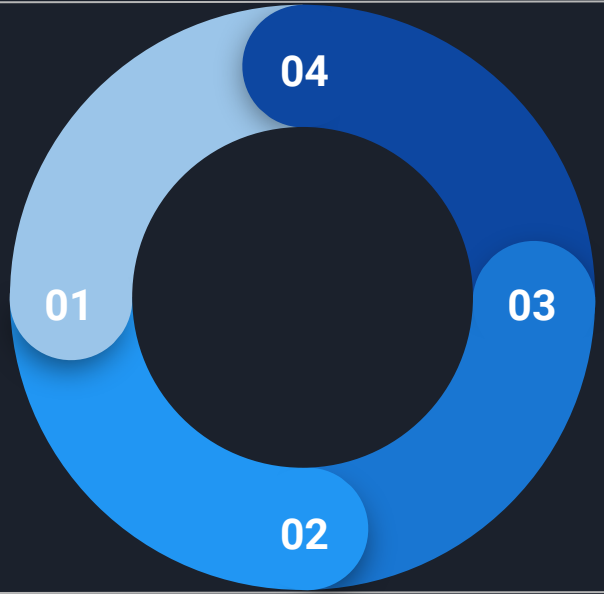
Cycle diagram

Extract Summary

LinkedIn About or Summary sections

Convert Data

Cleaning of Text and converting it to required format for model training



Model Training

Using SOTA BERT model for NER purposes specifically for our use case

Get feedback

Cross checked with Business Analyst team, which confirmed AI performs better than conventional string match

Step 0:

Data Collection

Slintel's Data Engineers had scraped three kinds of data

1. LinkedIn Summaries (About section) of People

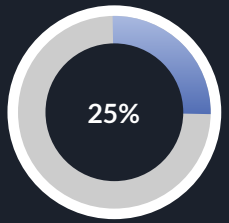
2. Technology Dictionary(, ~ 24000 rows; Technological words for which we are sure are Technical or Organizational words, Ex= GitHub, MongoDB since these words have no other meaning other than a Technical one)

3. English Dictionary, ~6000 (Contextual words which might not be a technology , ex-'**Python**' can referred as a snake and also as a Programming Language, WorkSpace)

Step 1: Cleaning & Tagging

I removed the unnecessary symbols like & - / , : ; which might not be relevant. Apart from this common words like is the can be removed since it is basically a Noise in our data .Converting all text to lowercase

Marking all the words using String Match which will be used to convert text into BIO Format

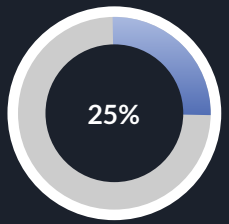


Data Processing

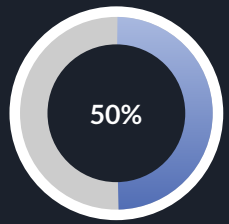
Step 2: Conversion to BIO format

All sentences are broken into words separated by space and tagged as B-org or I-org & O (The B- prefix before a tag indicates that the tag is the beginning of a chunk, & an I- prefix before a tag indicates that the tag is inside a chunk. An O tag indicates that a token belongs to no entity / chunk.)

5758	Sentence :19	national	O
5759	Sentence :19	and	O
5760	Sentence :19	international	O
5761	Sentence :19	market	O
5762	Sentence :19	playersmy	O
5763	Sentence :19	specialties	O
5764	Sentence :19	include	O
5765	Sentence :19	anaylytical	O
5766	Sentence :19	thinking	O
5767	Sentence :19	microsoft	B-ORG
5768	Sentence :19	power	I-ORG
5769	Sentence :19	point	I-ORG
5770	Sentence :19	microsoft	B-ORG
5771	Sentence :19	excel	I-ORG
5772	Sentence :19	microsoft	B-ORG
5773	Sentence :19	word	I-ORG
5774	Sentence :19	i	O
5775	Sentence :19	work	O
5776	Sentence :19	for	O
5777	Sentence :19	accenture	B-ORG
5778	Sentence :19	as	O
5779	Sentence :19	management	O
5780	Sentence :19	consulting	O
5781	Sentence :19	manager	O



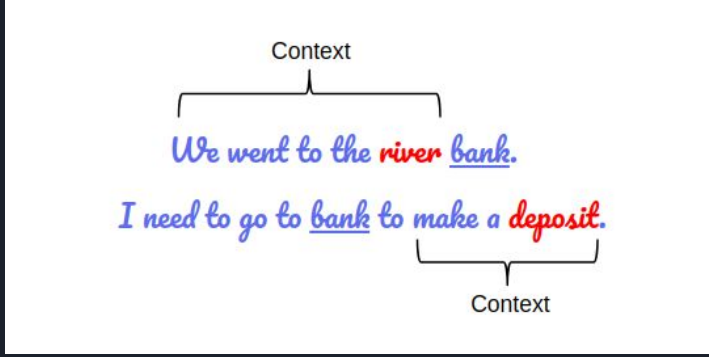
Data Processing



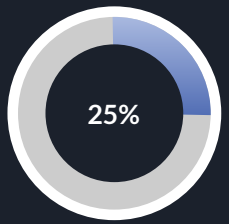
BIO Formatting

Step 3: Fine Tuning BERT

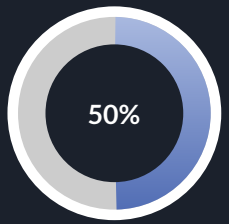
Fine-Tuned Bidirectional Encoder Representations from Transformers (BERT) model on our data to generate future predictions from text



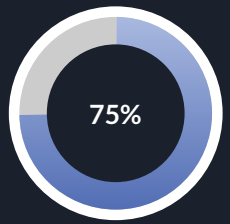
An example where BERT model differentiates same word based on the Context of the sentence (Here: BANK)



Data Processing



BIO Formatting

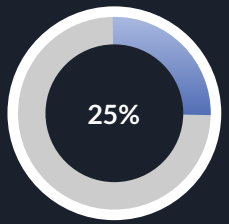


BERT Fine Tuning

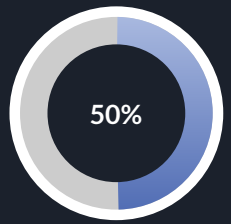
Step 4: Accuracy Analysis

Comparing accuracy for our use case on my trained model.

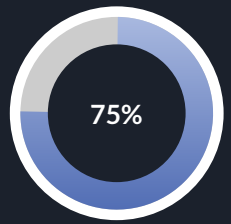
PRECISION -0.957
RECALL - 0.725
F1_SCORE- 0.825



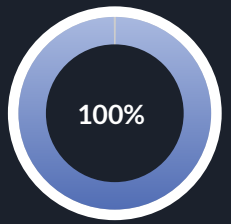
Data Processing



BIO Formatting



BERT Fine Tuning



Comparison

SLINTEL

Thank you!

-By Mohit Bagaria, 17EX200217

GEOLOGY & GEOPHYSICS

Data Science Intern at SLINTEL