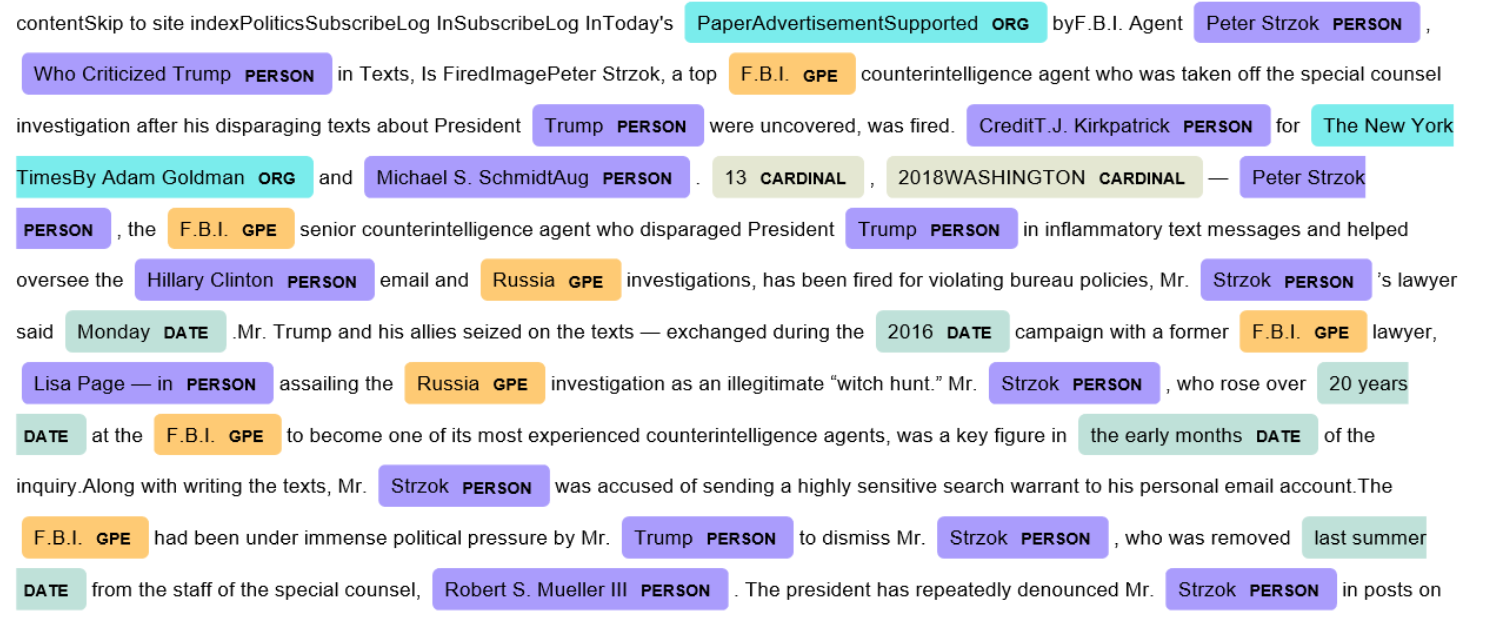
## horizontal line



Technology Extraction - NER

May - Aug 2021

**─**

**SUMMER INTERNSHIP REPORT- 2021**

Mohit Bagaria

17EX20017

Geology & Geophysics

Data Science Intern

Slintel

# Overview of what SLINTEL does?

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

# Problem Statement

1. Technologies Extraction From Linkedin Summaries
2. 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

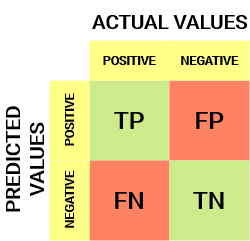
This Problem Statement is Termed as NAMED ENTITY RECOGNITION (NER)

# Requirements:

## Tagging Data:

Each word has to be tagged using String Match and then with the help of Human Intervention he/she has to check contextually whether the word is in real referring to an Organization or Technology

## Confusion Matrix - used for comparison among different approaches

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making.

**True Positive (TP) :**

The predicted value matches the actual value

The actual value was positive and the model predicted a positive value

**True Negative (TN) :**

The predicted value matches the actual value

The actual value was negative and the model predicted a negative value

**False Positive (FP):** – Type 1 error

The predicted value was falsely predicted. The actual value was negative but the model predicted a positive value

**False Negative (FN):** – Type 2 error

The predicted value was falsely predicted. The actual value was positive but the model predicted a negative value

**Precision:**

Precision tells us how many of the correctly predicted cases actually turned out to be positive

(true positives/predicted positives)=TP/(TP+FP)

**Recall:**

Recall tells us how many of the actual positive cases we were able to predict correctly with our model.

(true positives/all actual positives)=TP/(TP+FN)

**F Score :**

F1-score is a harmonic mean of Precision and Recall, and so it gives a combined idea about these two metrics. It is maximum when Precision is equal to Recall. The interpretability of the F1-score is poor. This means that we don’t know what our classifier is maximizing – precision or recall? So, we use it in combination with other evaluation metrics which gives us a complete picture of the result.

Harmonic Mean of Precision and Recall => 2\*(precision\*recall)/(precision+recall)

Range of Precision, Recall and F1 score is between (0,1)

# MODEL CYCLE-

## 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

Then remove all the words which contextually meant something else,i.e., not referring to any Organization or Technology

BIO Format example

## 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.)

## Step 3: Fine Tuning BERT

BERT (Bidirectional Encoder Representations from Transformers) is a recent paper published by researchers at Google AI Language. It has caused a stir in the Machine Learning community by presenting state-of-the-art results in a wide variety of NLP tasks, including Question Answering (SQuAD v1.1), Natural Language Inference (MNLI), etc.

BERT makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text.

Loaded Google’s pre-trained model with the help of Huggingface library and then Fine-Tuned Bidirectional Encoder Representations from Transformers (bert-ner-uncased) model on our data to generate future predictions from text.

## Step 4: Accuracy Analysis

PRECISION -0.957

RECALL - 0.725

F1\_SCORE- 0.825

## Thank You!

-By Mohit Bagaria

17EX20017

GEOLOGY & GEOPHYSICS