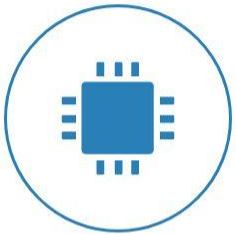
Natural Language Processing - 101

Concepts, Examples & Case-Studies

Session Agenda

Introduction

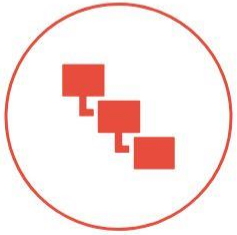
What is Natural Language



What is NLP?

Why NLP?

NLP Applications



Standard NLP Workflow

Hands-on Tutorials

Industry Case-studies


Introduction

About Me

Dipanjan Sarkar

Data Science Lead, Author, Google Developer Expert - ML

APPLIED MATERIALS®

Springboard

Experts

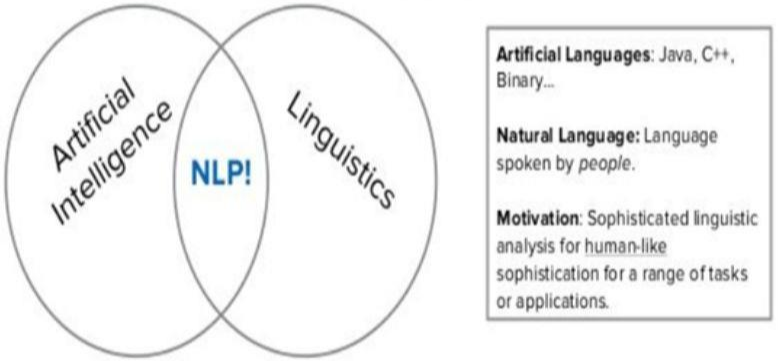
Machine Learning


What is Natural Language?

- A mechanism by which humans communicate with each other (and now with machines!)
- Very different from a constructed or a programming language
- Highly unstructured in nature - text & speech
- Difficult to parse and comprehend by machines!
- Text data is available in various forms
  - Emails
  - Messages
  - Documents and so on...


What is Natural Language Processing (NLP) ?

What is NLP?



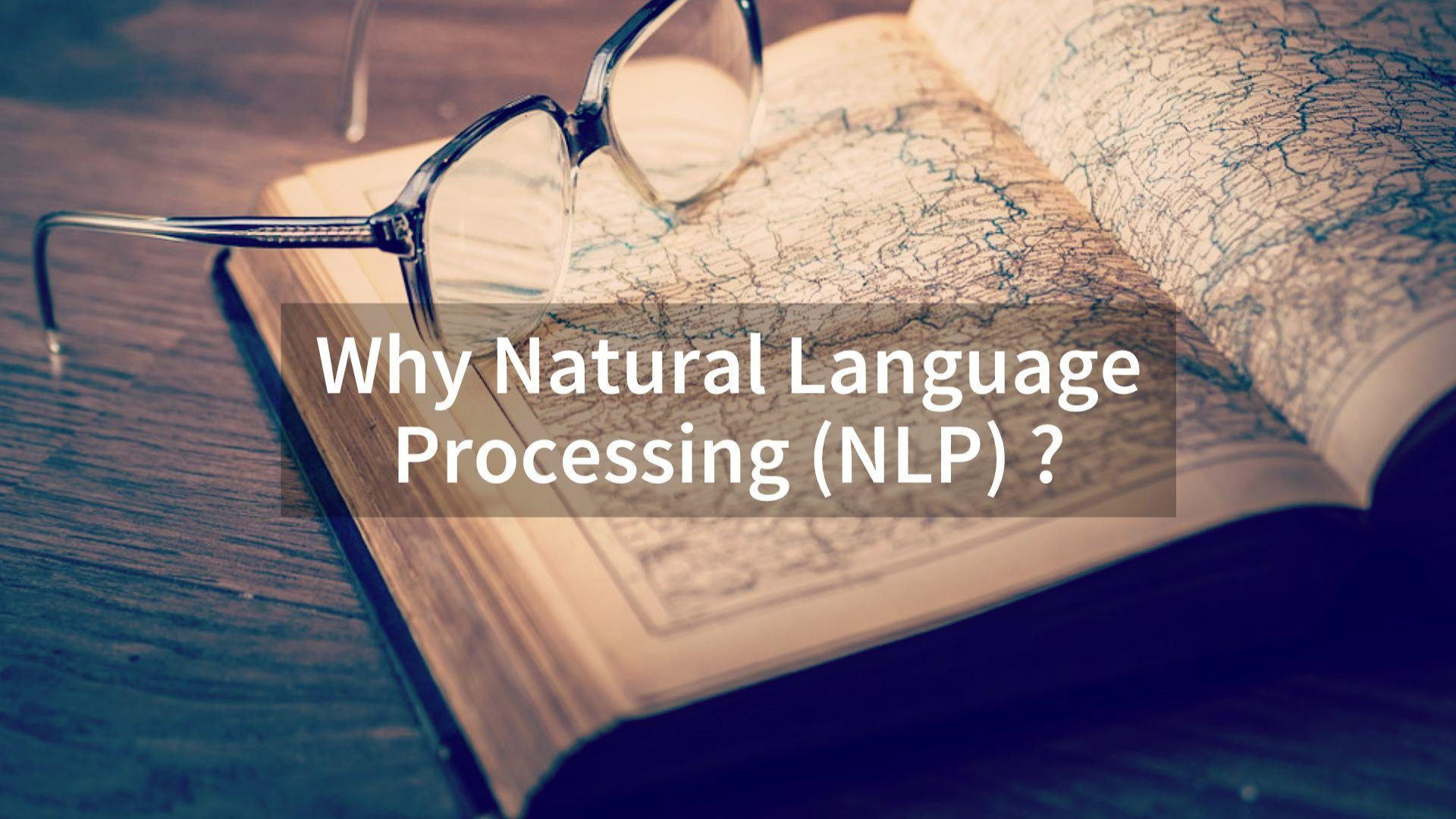
Goal: have computers *understand* natural language in order to perform useful tasks

- Natural language processing (NLP) is an intersection between the fields of computer science, linguistics and artificial intelligence
- NLP is concerned with the interaction between computers and human (natural) languages
- Key focus is to train computers to process, analyze and model large amounts of natural language data

NLP & Text Analytics & Mining

Text analytics automates what researchers, writers, scholars, and all the rest of us have been doing for years. Text analytics --

- ***Applies linguistic and/or statistical techniques to extract concepts and patterns*** *that can be applied to categorize and classify documents, audio, video, images.*
- ***Transforms "unstructured" information into data*** *for application of traditional analysis techniques.*
- ***Unlocks meaning and relationships*** *in large volumes of information that were previously unprocessable by computer.*


Why Natural Language Processing (NLP) ?

Motivation for NLP

"The bulk of information value is perceived as coming from data in relational tables. The reason is that data that is structured is easy to mine and analyze."

-- Prabhakar Raghavan, Yahoo Research

Yet it's a truism that 80% of enterprise-relevant information originates in "unstructured" form.

Motivation for NLP

80% of the world's data is unstructured or semi-structured text

Social media is rife with information about products and services

- Discussions, blogs, tweets…

Applications often lock up useful information in blobs, description fields and semi-structured records that are difficult or impossible to open up for analysis

- Call center records, log files…

How do you get a metrics based understanding of facts from unstructured text?

Motivation for NLP

Text is everywhere!

Images get all the hype

When is the last time you had to detect stop signs at work?

OR

When was the last time you had to find out the total vehicles on the road?

Most useful data is text

Public (Tweets, Reddit, Blogs, News)

Proprietary (emails, slack, databases, logs)

Hybrid (reviews and comments, surveys)

Practical

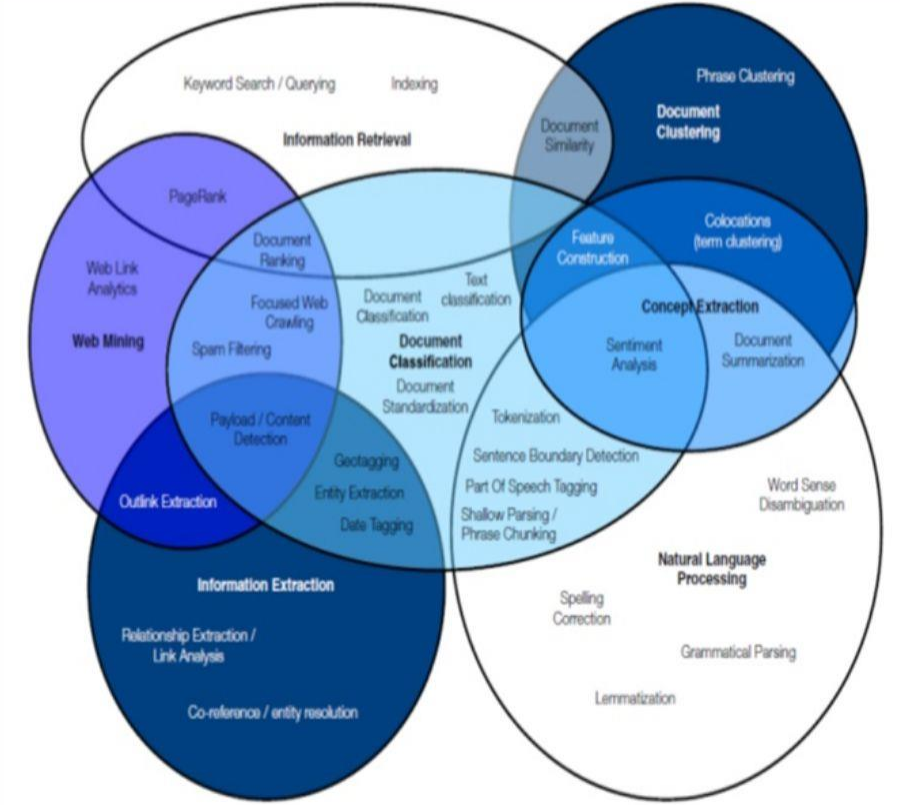
Most NLP models are realistic and affordable

Most NLP models are deployable


NLP Applications

Applications of NLP

Applications of NLP



Common NLP Use-Cases

- **Text Classification**

  Support Ticket Classification, News Article Categorization

- **Text Clustering & Similarity**

  Recommender systems, Duplicate Detection with Fuzzy Matching

- **Search and Information Retrieval**

  Search Engines, Document Ranking

- **Parsing and Named Entity Recognition**

  Entities from health records, legal documents

- **Text Summarization**

  Topic models, summarizing entire documents

- **Machine Translation**

  Speech to Text, Language Translation

- **Conversational Interfaces**

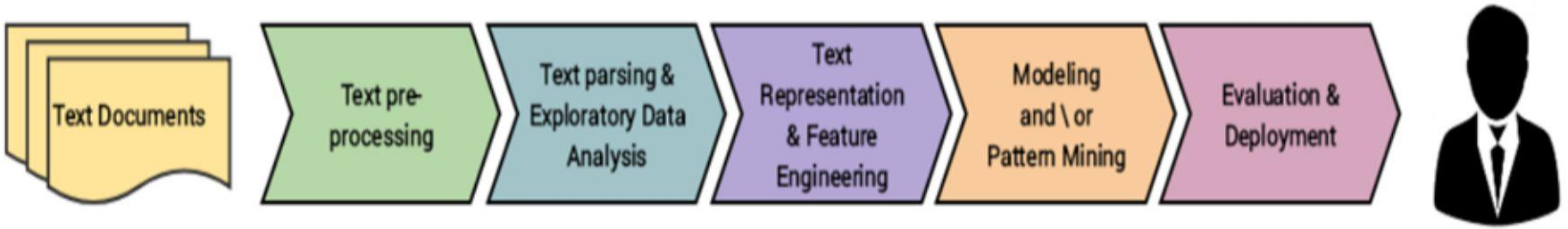
  Chatbots, Personal Assistants, Q&A Systems

- **Sentiment Analysis**

  Survey result analysis, NPI analysis

Standard NLP Workflow

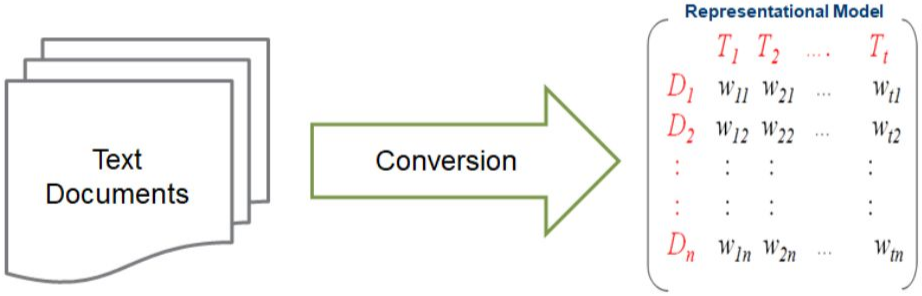
NLP Workflow



Text Wrangling \ Pre-processing

- Removing HTML tags
- Remove Extra Whitespace and Newlines
- Remove special characters and symbols (optionally numbers)
- Convert accented characters to ASCII
- Stemming OR Lemmatization
- Removing Stop Words
- Tokenization if needed
- Spell Check & Grammar Check

Text Representation Models



- ML\DL models at heart are mathematical functions and cannot understand unstructured text
- Hence we need to convert text into some numeric representations which can be understood by machines
- Commonly known as Vector Space Models where text is converted to numeric vectors
  - Bag of Words, TF-IDF
  - Topic Models
  - Similarity
  - Word Embeddings - Word2Vec, GloVe, FastText etc.

Traditional Text Representation Models

1 Bag of Words

- Each document is represented by a vector (bag) of words
- Depicts the number of times each word occurs in that document

2 Bag of N-grams

- Same as the Bag of Words model
- Instead of words, we also have n-grams and counts for them in the vector

3 TF-IDF

- Similar to the basic Bag of Words (TF) model
- Normalizes counts using the inverse document frequency (IDF) to downplay effect of frequently occuring words

4 Document Similarity

- Derived attribute \ feature from bag of words based features
- Assign scores to each document w.r.t how similar it is to other documents (based on their BOW vectors)

Word Embedding Models

1 Word2Vec

- Generates high quality dense vector representations of words
- Looks at each word and their surrounding words to generate representations

2 GloVe

- Unsupervised learning model which can be used to obtain dense word vectors
- Considering the Word-Context (WC) matrix, Word-Feature (WF) matrix and Feature-Context (FC) matrix, we try to factorize WC = WF x FC

3 FastText

- Considers each word as a Bag of Character n-grams to generate vector representations
- Taking the word **where** and **n=3** (tri-grams) as an example, it will be represented by the character n-grams: **<wh**, **whe**, **her**, **ere**, **re>** and the special sequence **< where >**

Downstream ML\DL Tasks after Feature Engineering

- Classification
- Recommender Systems
- Clustering
- Topic Modeling
- Sentiment Analysis
- Semantic Analysis

Real-World Challenges in the Industry

Framing a task as

- Classification (sentiment analysis, spam detection, code classification)
- Named Entity Recognition, Information extraction, Semantic Analysis
- Embeddings (recommendations, search)
- Clustering, Summarization, Topic Models

Debugging

- Most tutorials stop at p% accuracy.
- How do you know it is ready for deployment
- How do you know it won't be really bad unless you try it out


Hands-on
Tutorials

Hands-on Tutorials

1. ## Movie Recommender System

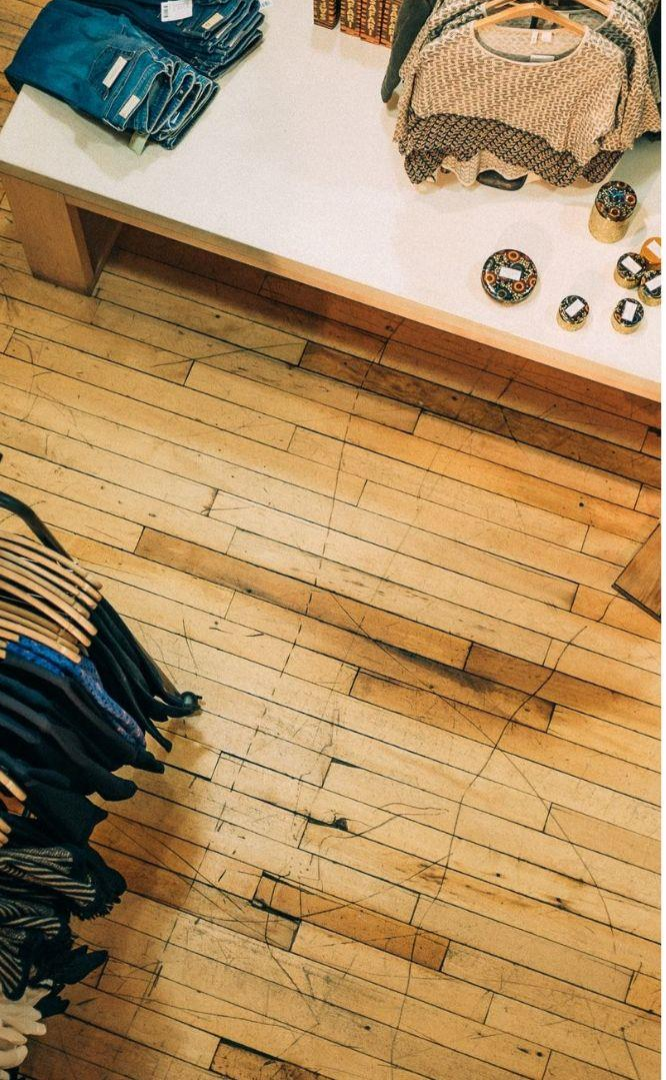
   Given a whole database of movies and their descriptions, can we use NLP to recommend similar movies to a user based on a movie of their choice?

2. ## Predict Product Ratings from their Reviews

   Given an e-commerce store with various products which are reviewed and rated by consumers can we predict the quantitative rating from the review text?

Movie Recommender System

- Get Movie Dataset
- Clean Movie Descriptions
- Build TF-IDF Features per Movie Description
- Compute Document Similarity (pairwise)
- Recommend Similar Movies based on Movie Description

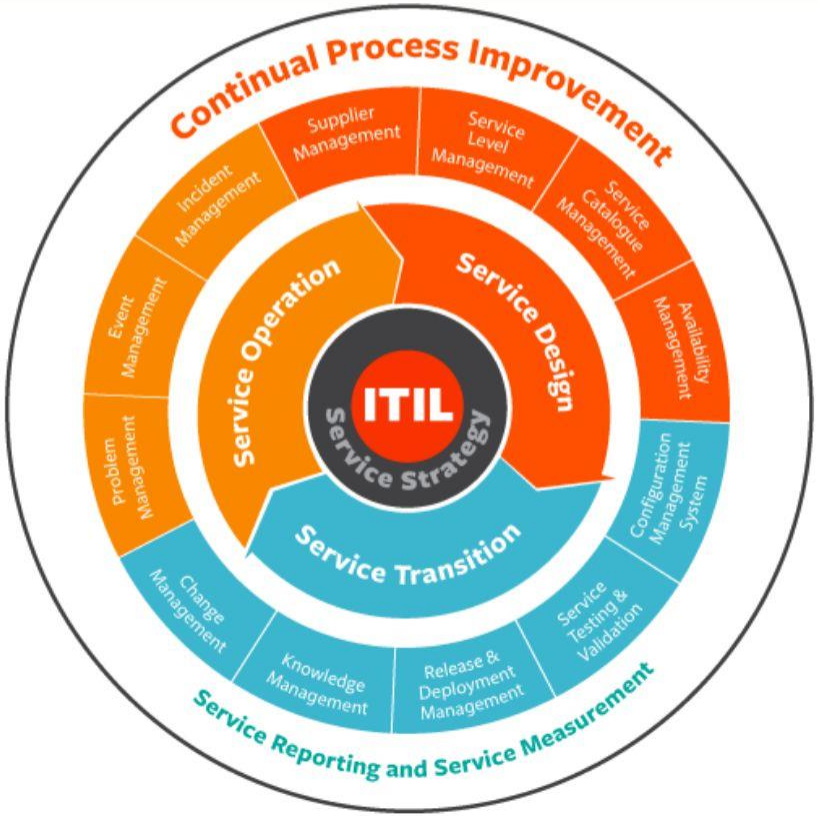


Product Rating Prediction System

- Get e-commerce review-ratings dataset
- Basic Data pre-processing
- Feature Engineering for Reviews
  - NLP count based features
  - Text Sentiment features
  - Bag of words based features
- Train Models on combination of above features
- Train, Evaluate and Re-iterate


Industry Case-Studies

Understanding IT Service Management



- Incident Management is one of the key processes in ITSM
- Goal of the incident management process is to restore a normal service operation as quickly as possible
- Focus on minimizing the impact on business operations
- Ensuring that the best possible levels of service quality and availability are maintained

Incident Management Use-Cases for AI

- Automated Incident Priority Categorization
- Incident Root Cause Prediction
- Automated Incident Resolution
- Key Themes of Daily Incidents

Predicting Incident Root Cause and Priority

- Enterprise Services have their own operations team with support agent(s) who manually classify, tag and label the root cause and priority of each ticket daily\weekly

- Leads to a lot of manual effort, wasted efficiency, distraction from their true work of resolving problems

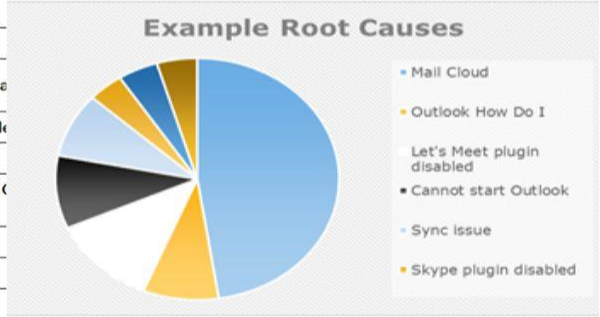
- Based on historical incident data per service, can we model on this data to predict the root cause\priority for future tickets?

Opportunity - ML\DL models to predict root cause

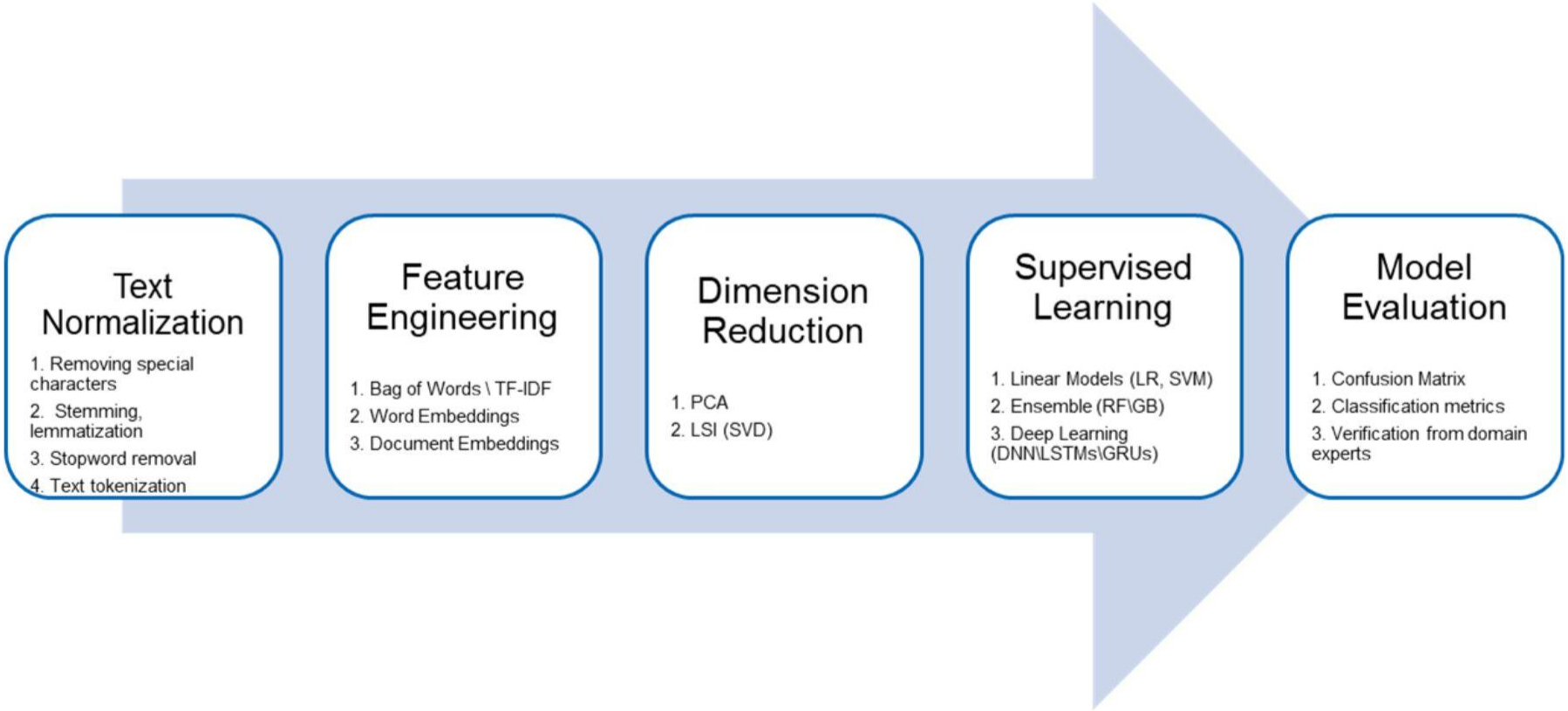
- Automated capability to process and classify tickets by leveraging machine learning or deep learning to train a model for this task
- Leverage historical ticket data with categorized root causes
- Parse and process ticket descriptions, other structured fields like location, product and so on
- Generate text features and emebddings with NLP
- Classification Models - ML\DL

Sample Incidents and Root Causes

	IncidentNumber	Reason	SupportOrg_BusinessService	ShortDescription
5	INC004928	Cannot start Outlook	Messaging	Outlook can not be started
6	INC004929	Recover Deleted Items	Messaging	Retrieve emails from old account. New employee who has 2 wwid previously and pointing to the same em
7	INC004929	Outlook client slow	Messaging	outlook 2013 start up slow
8	INC004929	Retention policy How Do I	Messaging	Hi- how do i make sure my outlook mails dont get deleted automatica
9	INC004929	Mail Cloud	Messaging	User mail cloud utility migrated all the files but it is not getting compl
10	INC004929	Password	Messaging	outlook was prompting for password
11	INC004929	OWA	Messaging	Getting error message while trying to set sender restriction for "SAP users" via OWA
12	INC004929	Mail Cloud	Messaging	Missing Mails from inbox , can't find them can you look into it
13	INC004929	Mail Cloud	Messaging	cannot move emails to folders
14	INC004930	Cannot start Outlook	Messaging	Outlook is not starting up, it is just showing loading page and not opening
15	INC004930	Mail Cloud	Messaging	Mail cloud - missing emails
16	INC004930	Outlook not connecting	Messaging	my Outlook cannot connect to server



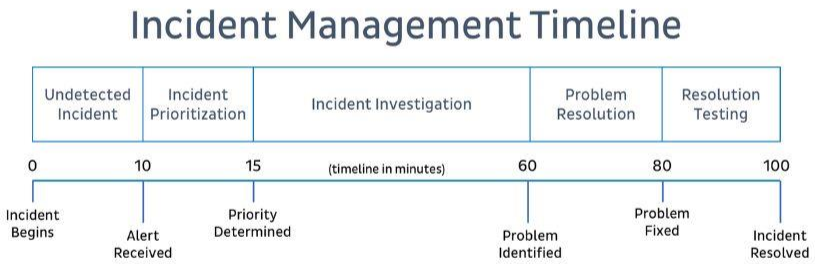
Classification Model Workflow



Sample Predictions

	ShortDescription	PredictedReason	ActualReason
0	outlook is not responing	Let's Meet plugin disabled	Let's Meet plugin disabled
1	Outlook doesn't pull mails from server though looks connected	Let's Meet plugin disabled	Let's Meet plugin disabled
2	Outlook can not be started	Cannot start Outlook	Cannot start Outlook
3	Retrieve emails from old account. New employee who has 2 wwid previously and pointing to the same em	Outlook How Do I	Outlook How Do I
4	Hi- how do i make sure my outlook mails dont get deleted automatically.	Others	Password
5	Outlook is not starting up, it is just showing loading page and not opening	Others	Others
6	Mail Cloud Query	Skype plugin disabled	Skype plugin disabled
7	I cant find a folder in my inbox	Mail Cloud	Mail Cloud
8	getting error massage for sender not receiving the mail - the error return back from 2 weeks ago	Mail Cloud	Others
9	Cannot open any power point attachment from outlook	Others	Outlook crashing
10	user wants to save emails in the mail cloud	Mail Cloud	Mail Cloud
11	let's meet: outlook plugin disappears from my meetings window? How do I bring it back?	Mail Cloud	Mail Cloud
12	Outlook 2013 / Unable to update the name on my outlook profile after switched from CW to BB	Let's Meet plugin disabled	Let's Meet plugin disabled
13	Mail cloud client utility - Understanding how mail cloud works, Need to move mails into mail cloud	Sync issue	Sync issue
14	Outlook 2013 / User does not receive incoming email / Creat new outlook profile	Cannot start Outlook	Cannot start Outlook

Automating Incident Resolution

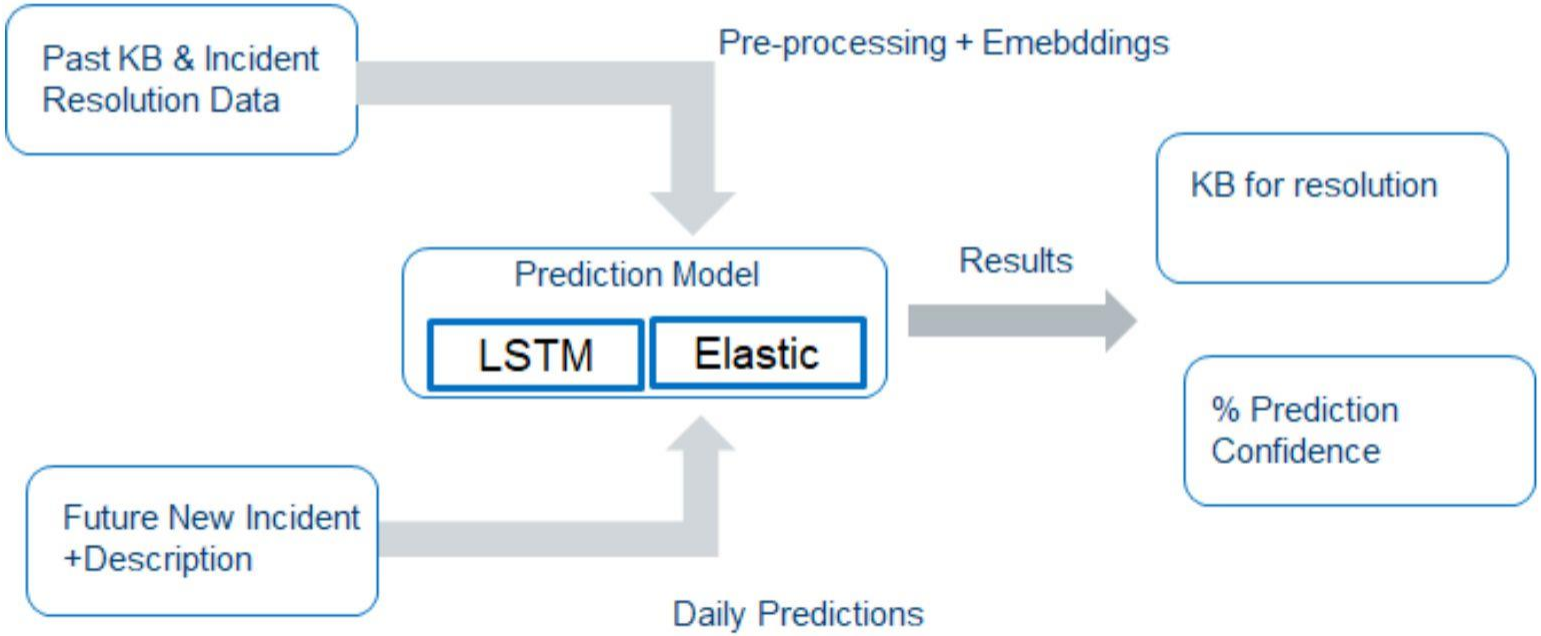


- Leverage reusable components from classification models to predict potential solution for a new incident
- Similar methodology as earlier except we map a problem to a known solution
- Automated scripts can self-heal system if model has enough confidence
- Model once deployed predicts potential knowledge article to solve a future incident

Automated Incident Resolution

- Leverage reusable components from classification models to predict potential solution for a new incident
- Can work in two ways - post-processing an already filed incident or providing a fix as the user enters a query so an incident doesn't need to be filed
- SC1: ML\DL Model is trained on historical incidents and their provided fix (a knowledge article)
- SC1: Model predicts potential knowledge article to solve a future incident
- SC2: Intelligent search provides potential knowledge articles from search bar as user types in the problem

Automated Incident Resolution - Scenario 1



Automated Incident Resolution - Scenario 2