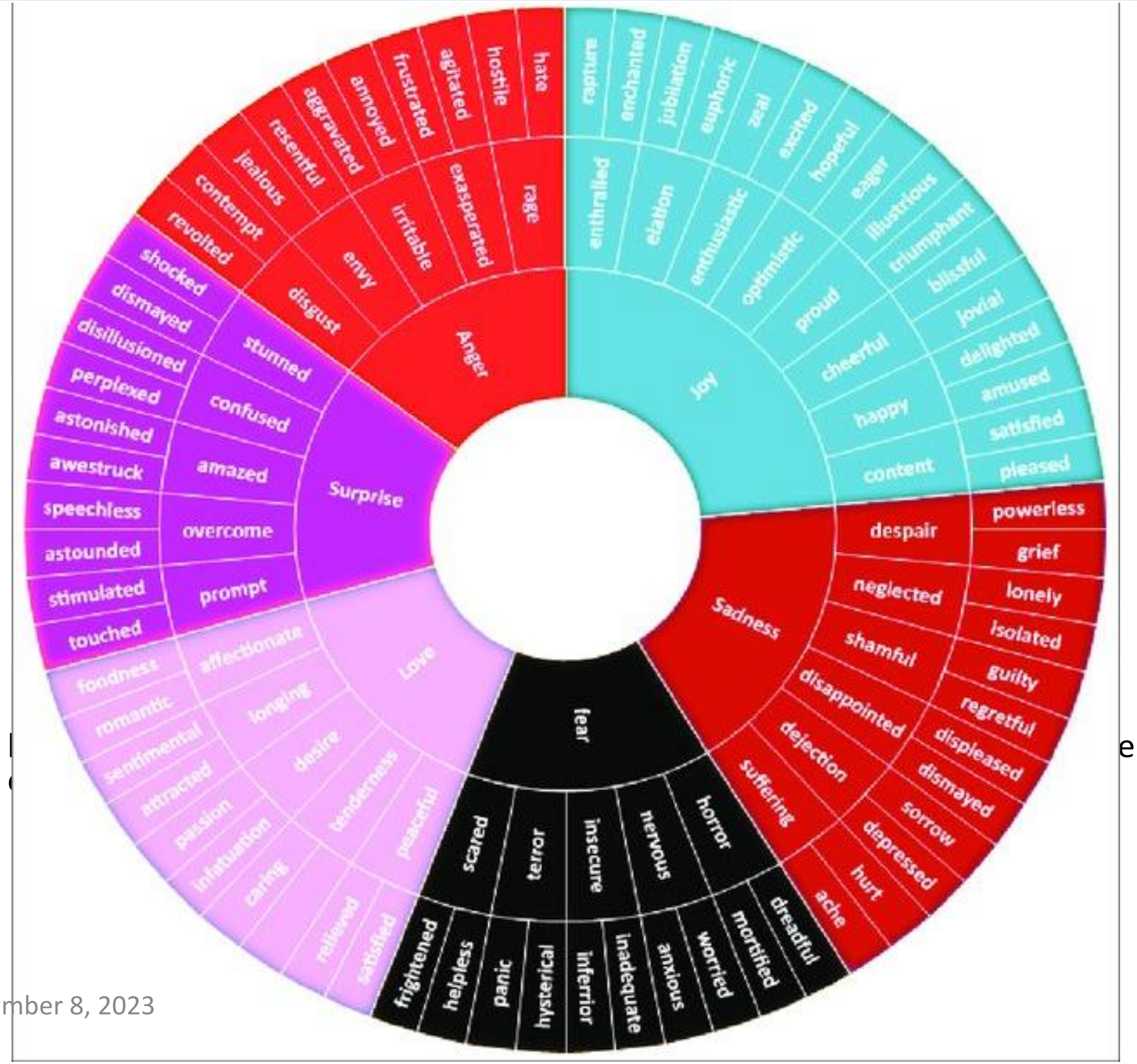
Introduction to Data Science with Python

Week 12: Statistical Natural Language Processing for Sentiment Analysis

Outlines

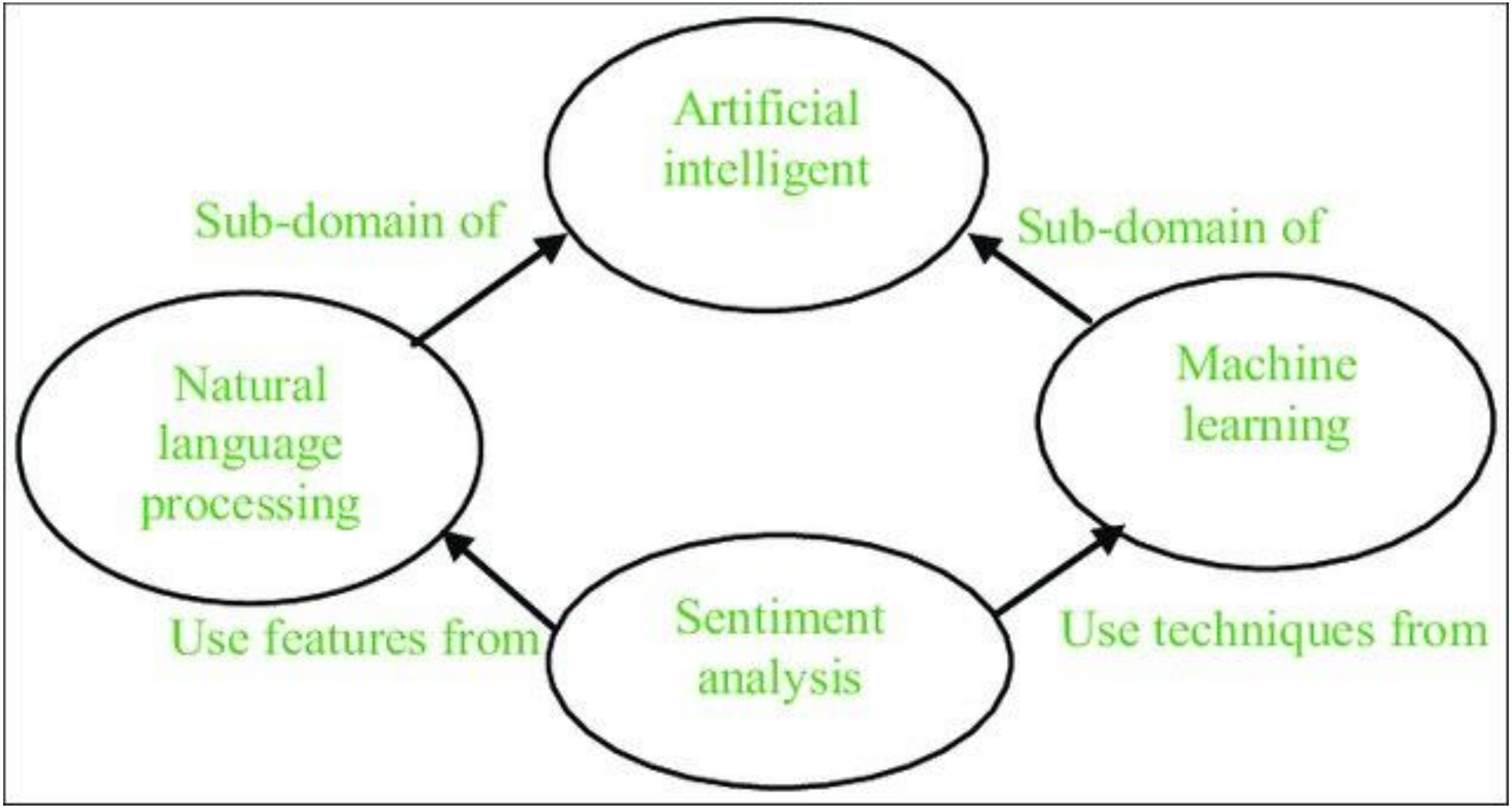
- **Statistical Natural Language Processing**
- **Sentiment Analysis**
- Data Cleaning,
- Text Representation,
- Bi-Grams and n-Grams.

Shades of Sentiments



- I ent_analysis_in_psych
  (

AI/ML/NLP for Sentiment Analysis



Sentiment Analysis

- The term sentiment analysis (or opinion mining) refers to the analysis from data of the attitude of the subject with respect to a particular topic.
- The goal of sentiment analysis is to understand and classify the subjective information in the text as positive, negative, or neutral.
- Sentiment analysis is a natural language processing (NLP) technique used to determine the sentiment or emotional tone expressed in a piece of text.
- This analysis can be applied to various types of textual data, such as customer reviews, social media posts, news articles, and more.

Sentiment Analysis[Cont’d]

- It means to identify the view or emotion behind a situation.
- It basically means to analyze and find the emotion or intent behind a piece of text or speech or any mode of communication.
- We, humans, communicate with each other in a variety of languages, and any language is just a mediator or a way in which we try to express ourselves. And, whatever we say has a sentiment associated with it. It might be positive or negative or it might be neutral as well.

Example1

- Suppose, there is a fast-food chain company and they sell a variety of different food items like burgers, pizza, sandwiches, milkshakes, etc. They have created a website to sell their food and now the customers can order any food item from their website and they can provide reviews as well, like whether they liked the food or hated it.
  - User Review 1: I love this cheese sandwich, it’s so delicious.
  - User Review 2: This chicken burger has a very bad taste.
  - User Review 3: I ordered this pizza today.
- So, as we can see that out of these above 3 reviews.

Example1 [Cont’d]

- The first review is definitely a **positive** one and it signifies that the customer was really happy with the sandwich.
- The second review is **negative**, and hence the company needs to look into their burger department.
- And, the third one doesn’t signify whether that customer is happy or not, and hence we can consider this as a **neutral** statement.
- By looking at the above reviews, the company can now conclude, that it needs to focus more on the production and promotion of their sandwiches as well as improve the quality of their burgers if they want to increase their overall sales.
- **But, now a problem arises, that there will be hundreds and thousands of user reviews for their products and after a point of time it will become nearly impossible to scan through each user review and come to a conclusion.**

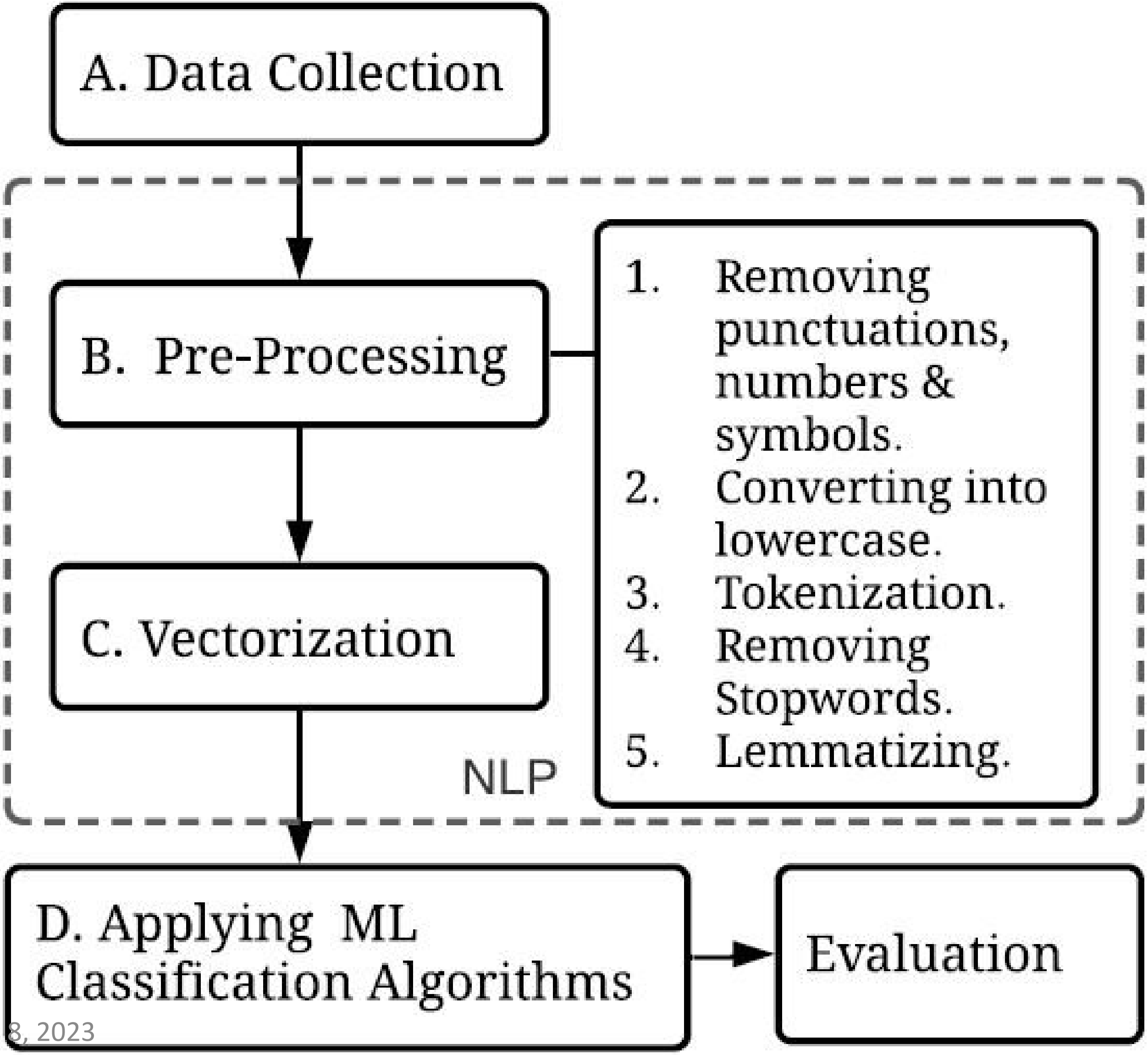
Example2

- There was a time when the social media services like Facebook used to just have two emotions associated with each post, i.e You can like a post or you can leave the post without any reaction and that basically signifies that you didn’t like it.
- But, over time these reactions to post have changed and grew into more granular sentiments which we see as of now,
  such as “like”, “love”, “sad”, “angry” etc.

Example2 [Cont’d]

- And, because of this upgrade, when any company promotes their products on Facebook, they receive more specific reviews which will help them to enhance the customer experience.
- And because of that, they now have more granular control on how to handle their consumers, i.e. they can target the customers who are just “sad” in a different way as compared to customers who are “angry”, and come up with a business plan accordingly because nowadays, just doing the **bare minimum is not enough**.

Workflow of Sentiment Analysis using NLP

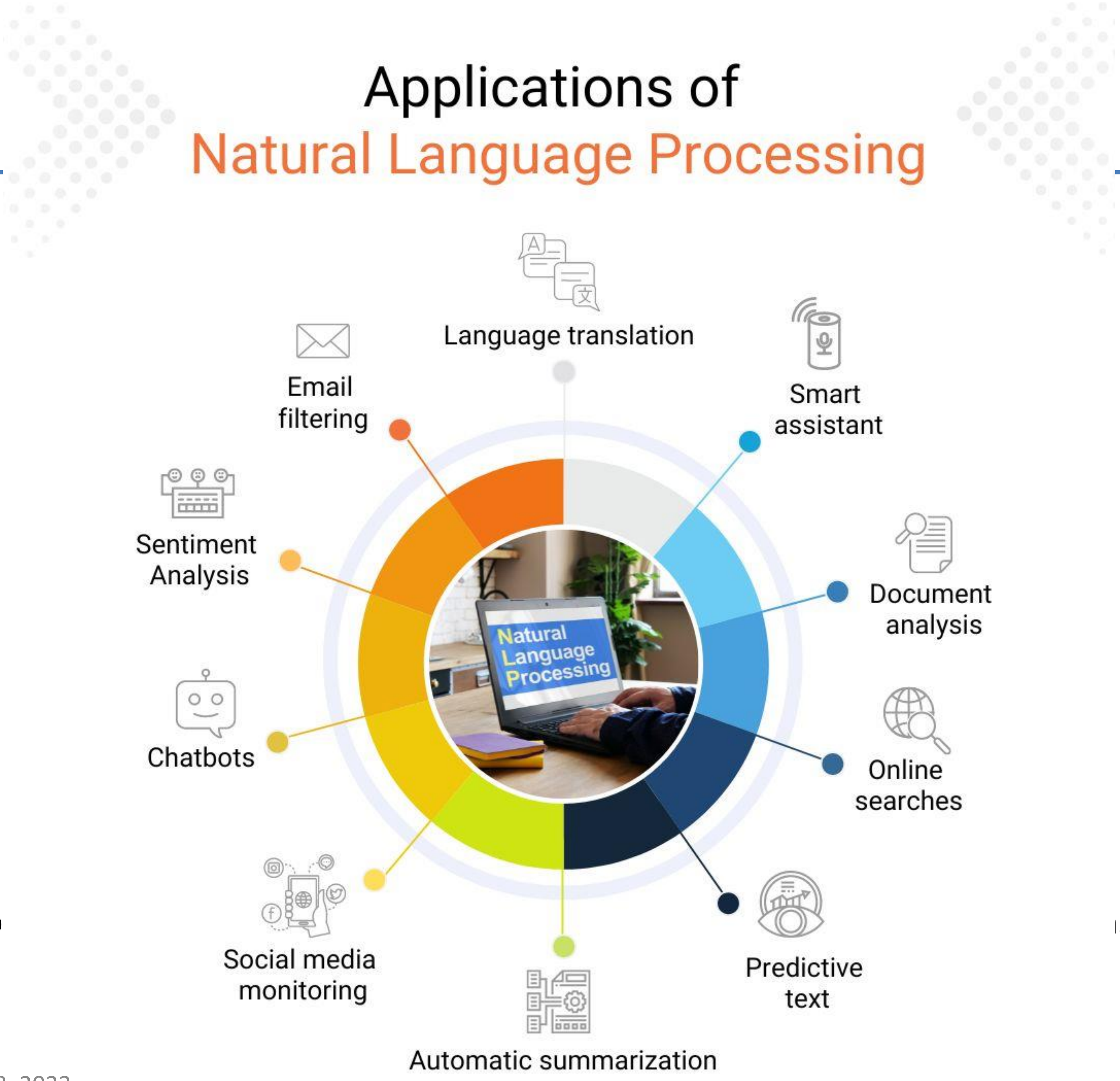


g-NLP-and-

H.

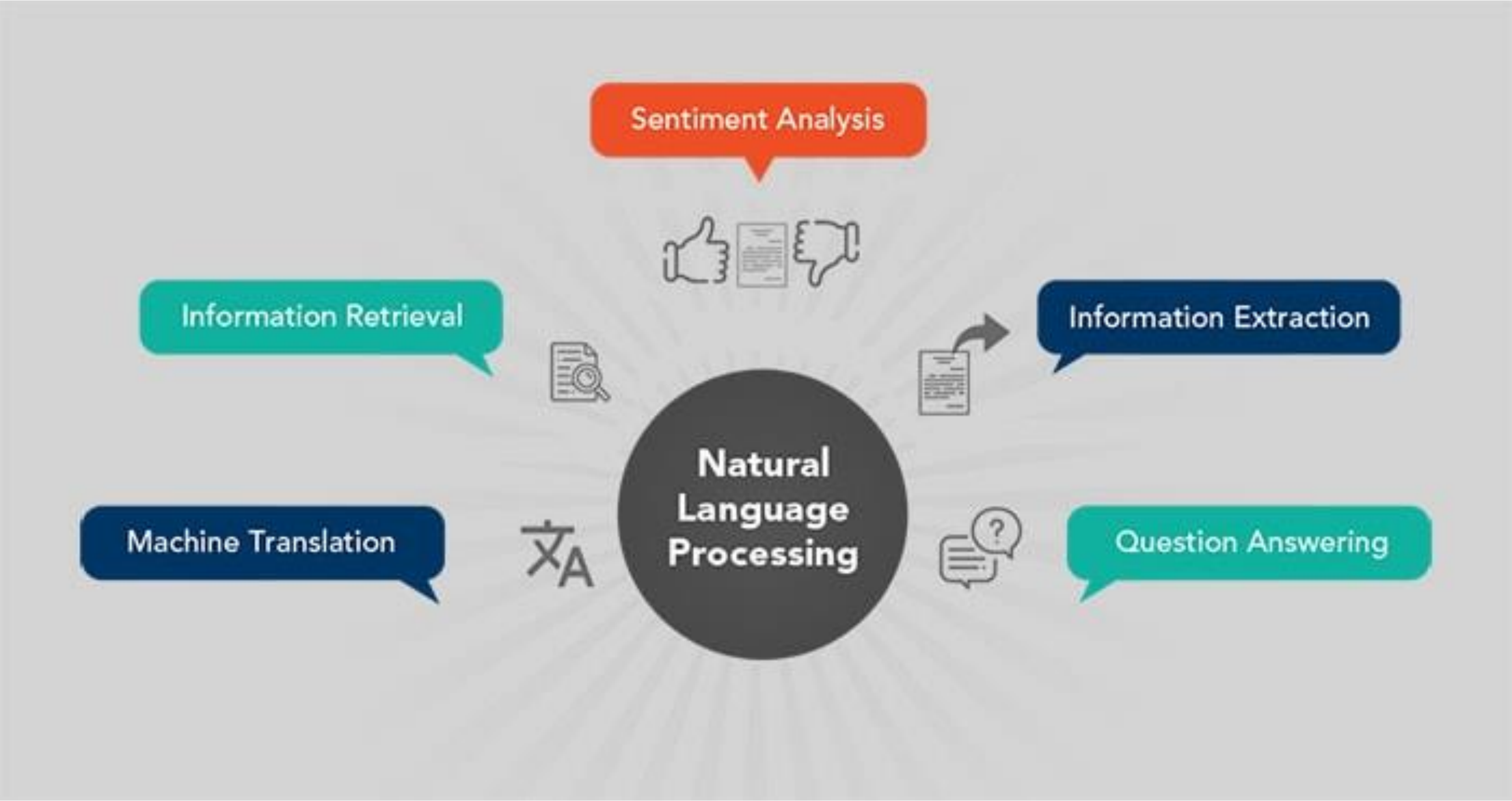
Natural Language Processing

- Natural language processing is an interdisciplinary subfield of computer science and linguistics. More specifically, NLP refers to the branch of artificial intelligence or AI.
- It is primarily concerned with giving computers the ability to understand, support and manipulate human language.
- It is the application of computational techniques to the analysis and synthesis of natural language and speech.
- NLP **combines computational linguistics—rule-based modeling of human language—with** statistical, machine learning, and deep learning models.
- Sentiment Analysis is a sub-field of NLP and with the help of machine learning techniques, it tries to identify and extract the insights.


Applications of
Natural Language Processing
Language translation
Smart assistant
Document analysis
Online searches
Predictive text
Automatic summarization
Social media monitoring
Chatbots
Sentiment Analysis
Email filtering
Natural Language Processing

- http s/

Types of NLP


Sentiment Analysis
Information Retrieval
Information Extraction
Natural Language Processing
Machine Translation
Question Answering

Data Cleaning

- In order to perform sentiment analysis, first we need to deal with some processing steps on the data.
- data cleaning is to remove those characters considered as noise in the data mining process, or removing irrelevant text items those are not associated with sentiment information.
  - For instance, comma or colon characters.

Data Preprocessing

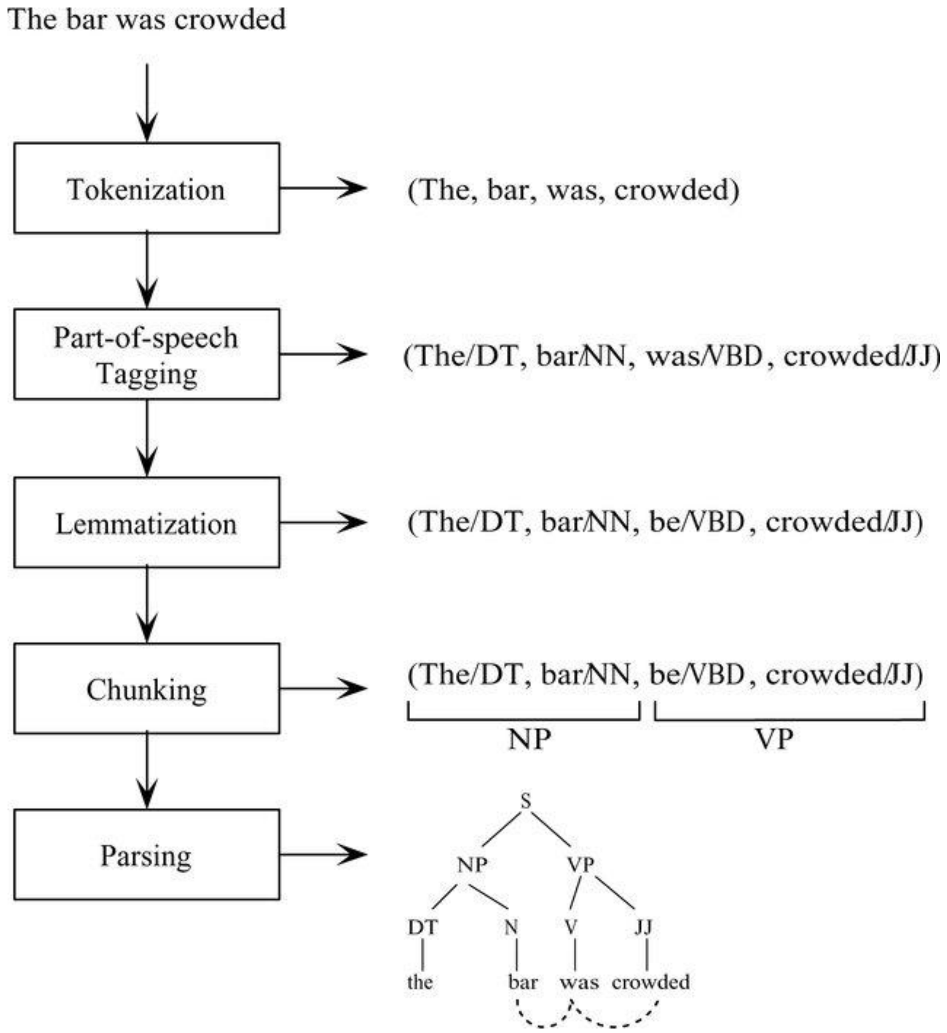
- **NLP (Natural Language Processing)**
- **Tokenization**
  - Tokenization is a process in natural language processing (NLP) that involves breaking down a text into smaller units called tokens. Tokens are the basic building blocks of a language and can be as short as one character (e.g., punctuation or individual letters) or as long as a word or even a group of words. The main purpose of tokenization is to simplify the text and make it easier to process or analyze.
- **Lemmatization**
  - a text pre-processing technique used in natural language processing (NLP) models to break a word down to its root meaning to identify similarities.
- **Stemming**
  - the process of reducing a word to its word stem that affixes to suffixes and prefixes or the roots.
- **Stop Words**
  - Stop words are common words in any language that occur with a high frequency but carry much less substantive information about the meaning of a phrase. Examples of some common stop words include: a, the, and , or , of , on , this , we , were, is, not

- **POS (Part-of-Speech) Tagging**
  - Part-of-Speech (POS) tagging is a preprocessing step in natural language processing (NLP) that involves assigning a grammatical category or part-of-speech label (such as noun, verb, adjective, etc.) to each word in a sentence.
- **NER (Named Entity Recognition)**
  - Named-entity recognition is a subtask of information extraction that seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories such as person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc.
- **Corpus**
  - a collection of machine-readable authentic texts (including transcripts of spoken data) that is sampled to be representative of a particular natural language or language variety
- **Vectorization**
  - the process of converting textual data, such as sentences or documents, into numerical vectors that can be used for data analysis, machine learning, and other computational tasks.
- **Bag-of-Words (BoW)**
  - a statistical language model used to analyze text and documents based on word count. The model does not account for word order within a document. BoW can be implemented as a Python dictionary with each key set to a word and each value set to the number of times that word appears in a text.

- **TF-IDF (Term Frequency-Inverse Document Frequency)**
  - Term frequency is how common a word is, inverse document frequency (IDF) is how unique or rare a word is. Example, Consider a document containing 100 words wherein the word apple appears 5 times. The term frequency (i.e., TF) for apple is then (5 / 100) = 0.05.
- **N-gram**
  - N-grams are continuous sequences of words or symbols, or tokens in a document. An N-gram **means a sequence of N words**. So for example, “Medium blog” is a 2-gram (a bigram), “A Medium blog post” is a 4-gram.
- **Syntax Tree**
  - A Syntax tree or a parse tree is a tree representation of different syntactic categories of a sentence. It helps us to understand the syntactical structure of a sentence.
- **Perplexity**
  - In the context of Natural Language Processing, perplexity is **one way to evaluate language models**. perplexity is a measurement of how well a probability distribution or probability model predicts a sample. It may be used to compare probability models. A low perplexity indicates the probability distribution is good at predicting the sample.

Example: Data Preprocessing

- .



The bar was crowded

Tokenization → (The, bar, was, crowded)

Part-of-speech Tagging → (The/DT, bar/NN, was/VBD, crowded/JJ)

Lemmatization → (The/DT, bar/NN, be/VBD, crowded/JJ)

Chunking → (The/DT, bar/NN, be/VBD, crowded/JJ)

NP

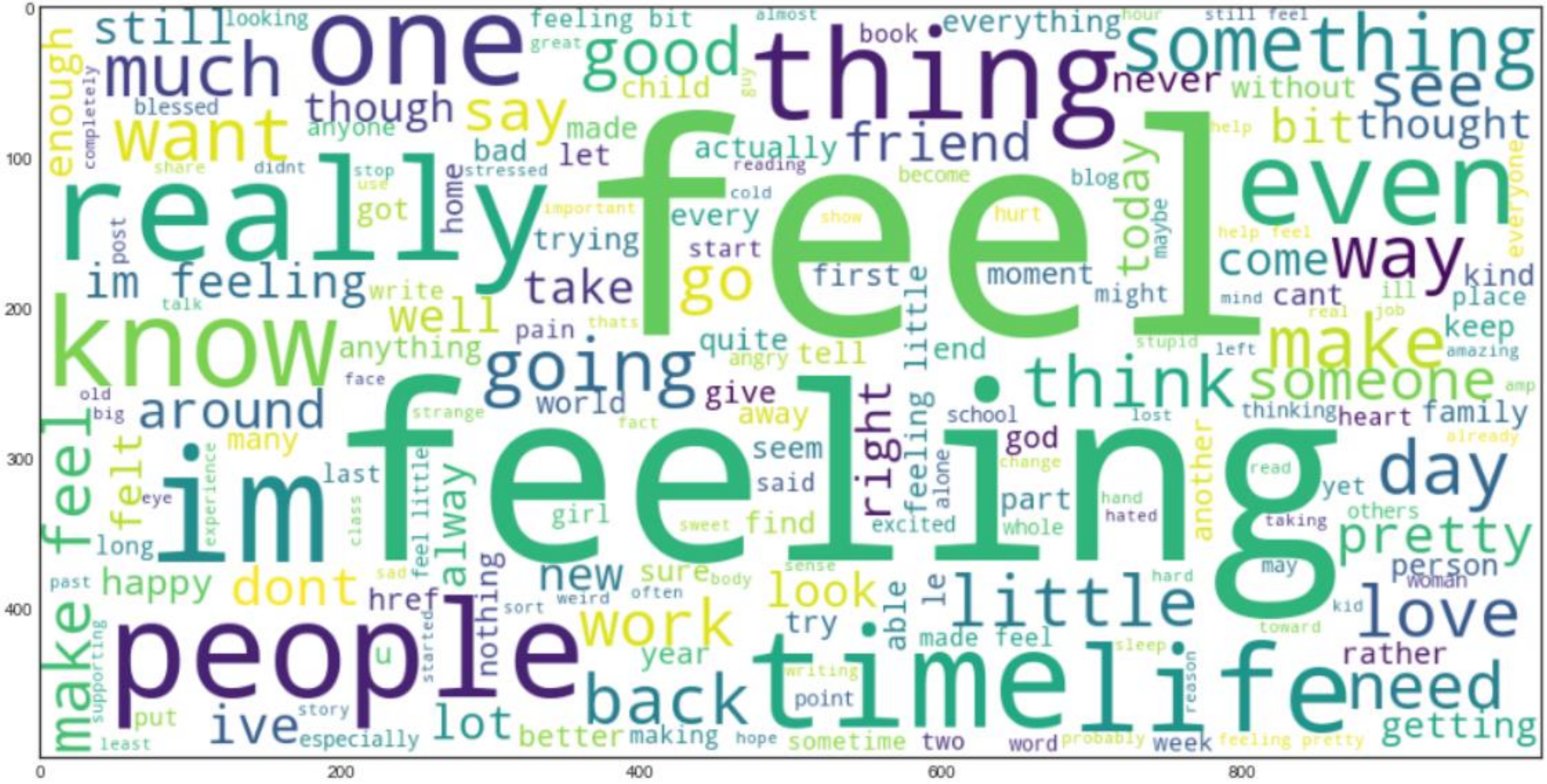
VP

Parsing →

Word Cloud

- It is a data visualization technique used to depict text in such a way that, the more frequent words appear enlarged as compared to less frequent words. This gives us a little insight into, how the data looks after being processed through all the steps until now.

Example: Word Cloud



Bag of Words

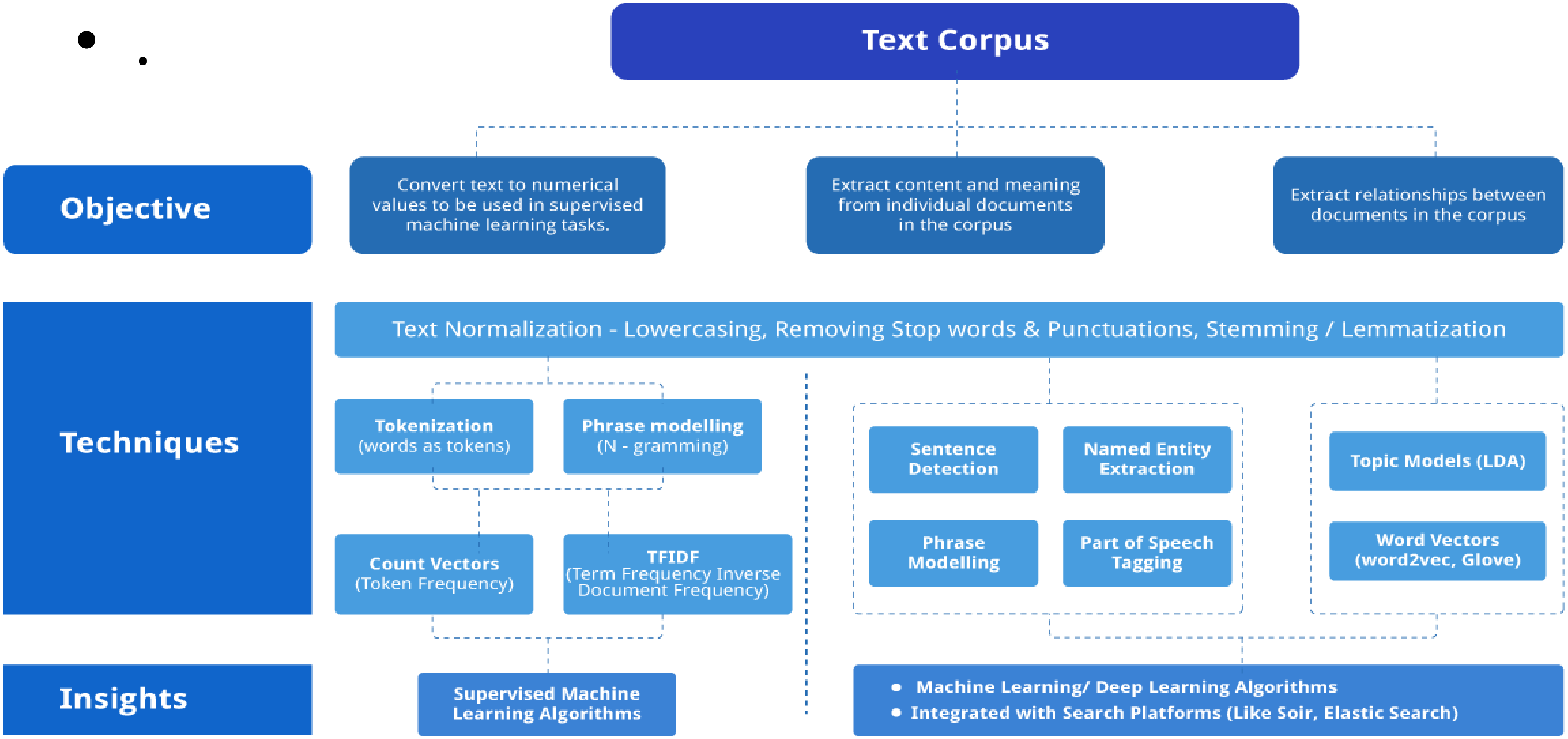
- Bag of Words Model(BOW), which is used to represent the text in the form of a bag of words,i.e. the grammar and the order of words in a sentence are not given any importance, instead, multiplicity,i.e. (the number of times a word occurs in a document) is the main point of concern.
- Basically, it describes the total occurrence of words within a document.

CountVectorizer

- **Scikit-Learn** provides a neat way of performing the bag of words technique using **CountVectorizer** that will convert the text data into vectors, by fitting and transforming the corpus that we have created.

Text Corpus

- .



Text Corpus

Objective

Convert text to numerical values to be used in supervised machine learning tasks.

Extract content and meaning from individual documents in the corpus

Extract relationships between documents in the corpus

Techniques

Text Normalization - Lowercasing, Removing Stop words & Punctuations, Stemming / Lemmatization

Tokenization (words as tokens)

Phrase modelling (N - gramming)

Count Vectors (Token Frequency)

TFIDF (Term Frequency Inverse Document Frequency)

Sentence Detection

Named Entity Extraction

Phrase Modelling

Part of Speech Tagging

Topic Models (LDA)

Word Vectors (word2vec, Glove)

Insights

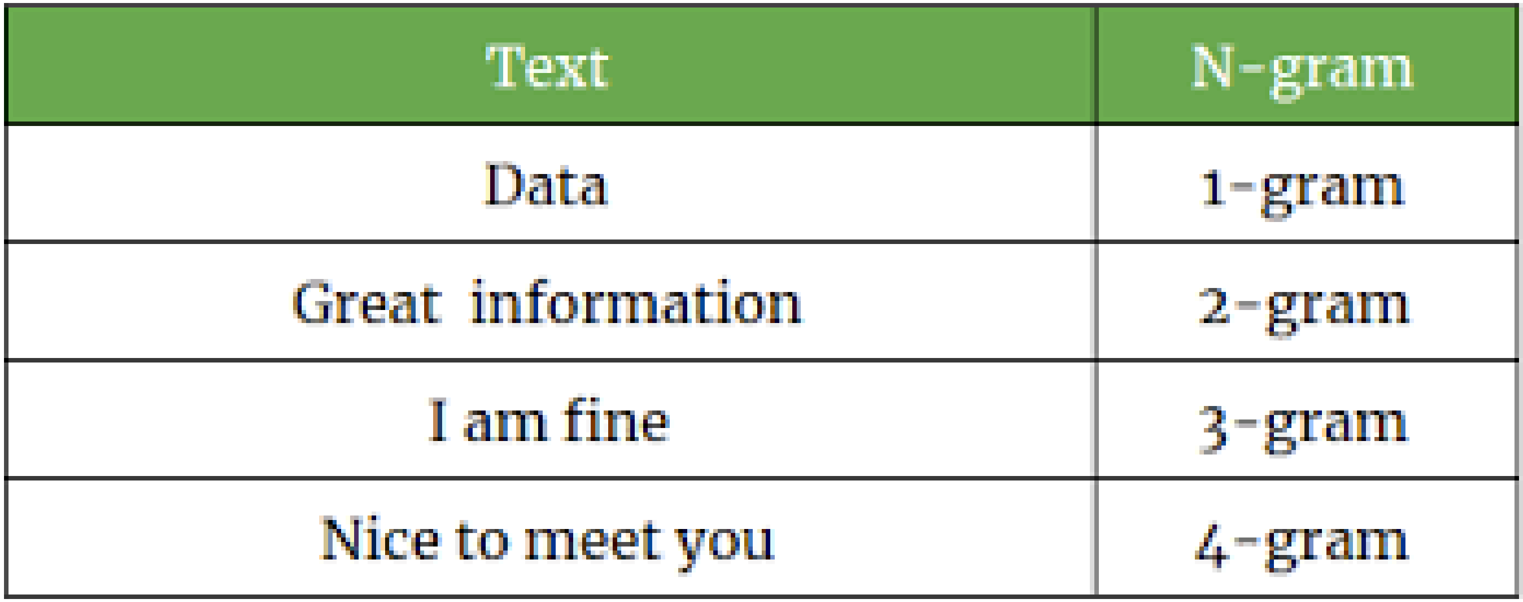
Supervised Machine Learning Algorithms

- Machine Learning/ Deep Learning Algorithms
- Integrated with Search Platforms (Like Soir, Elastic Search)

NLP infographic above: Intellectual property of Latentview.com

N-gram

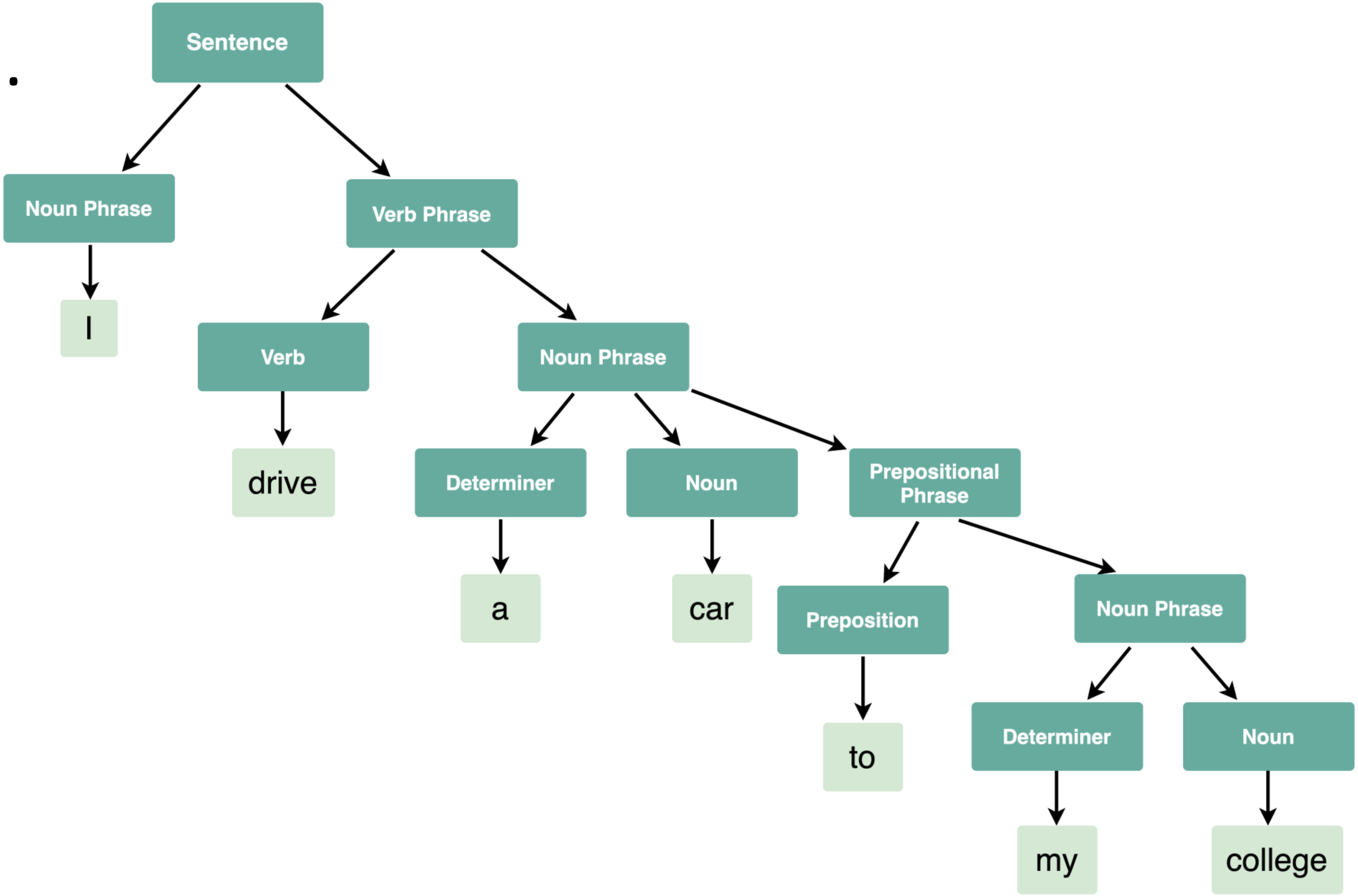
- .



Text	N-gram
Data	1-gram
Great information	2-gram
I am fine	3-gram
Nice to meet you	4-gram

Syntax tree

- .


Sentence
Noun Phrase
Verb Phrase
I
Verb
Noun Phrase
drive
Determiner
Noun
Prepositional Phrase
a
car
Preposition
Noun Phrase
to
Determiner
Noun
my
college

Why NLP?

- As we humans communicate with each other in a way that we call Natural Language which is easy for us to interpret but it’s much more complicated and messy if we really look into it.
- Because, there are billions of people and they have their own style of communicating, i.e. a lot of tiny variations are added to the language and a lot of sentiments are attached to it which is easy for us to interpret but it becomes a challenge for the machines.
- This is why we need a process that makes the computers understand the Natural Language as we humans do, and this is what we call Natural Language Processing(NLP).

Tools: Basic Python Libraries

- 1. Pandas – library for data analysis and data manipulation
  2. Matplotlib – library used for data visualization
  3. Seaborn – a library based on matplotlib and it provides a high-level interface for data visualization
  4. WordCloud – library to visualize text data
  5. re – provides functions to pre-process the strings as per the given regular expression

Tools: Natural Language Processing

- 1. nltk – Natural Language Toolkit is a collection of libraries for natural language processing
- 2. stopwords – a collection of words that don't provide any meaning to a sentence
- 3. WordNetLemmatizer – used to convert different forms of words into a single item but still keeping the context intact.

Tools: Scikit-Learn (ML Library)

- 1. CountVectorizer – transform text to vectors
- 2. GridSearchCV – for hyperparameter tuning
- 3. Decision Tree Classifier – machine learning algorithm for classification

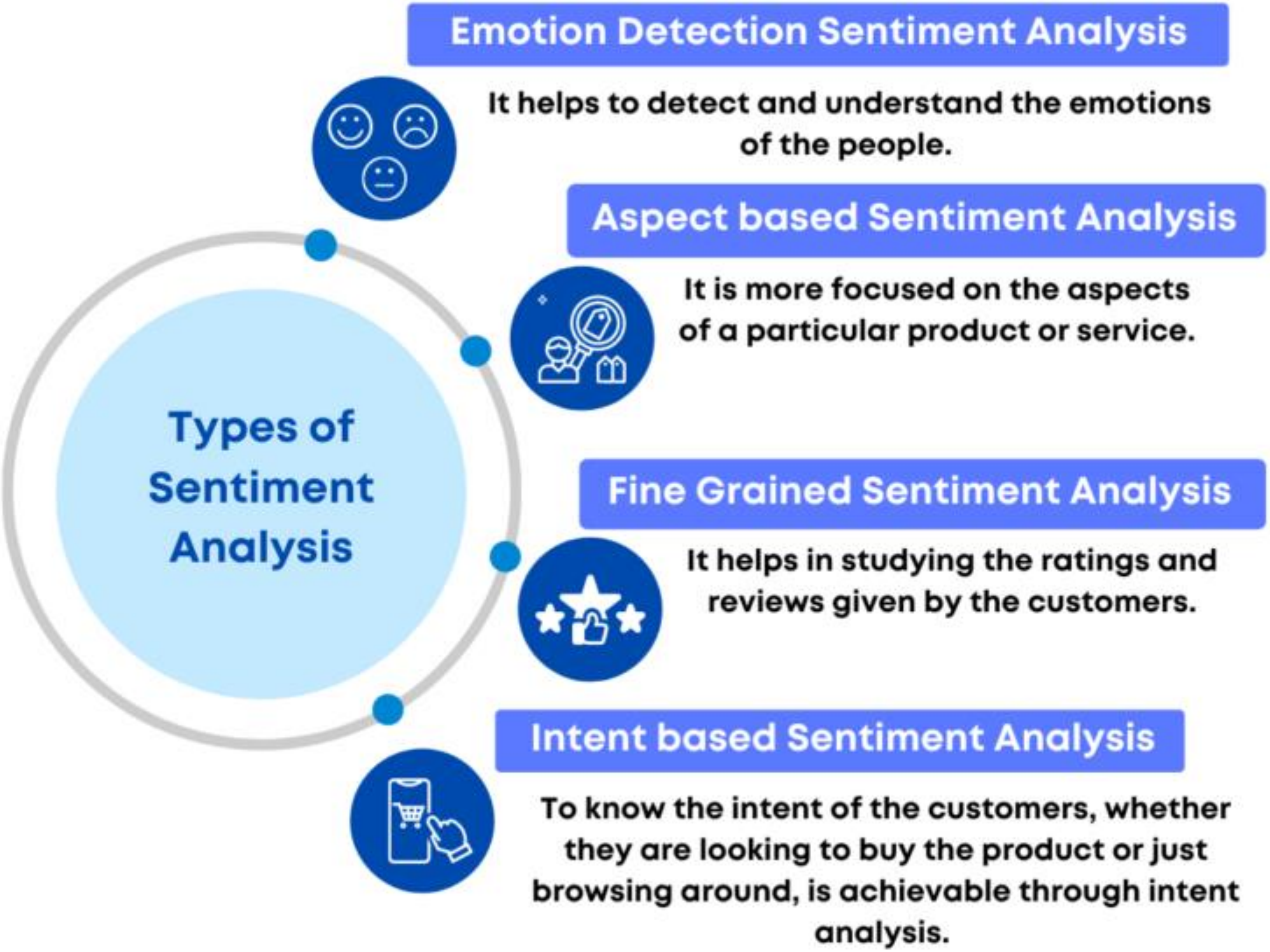
Tools: Evaluation Metrics

- 1. Accuracy Score – no. of correctly classified instances/total no. of instances
- 2. Precision Score – the ratio of correctly predicted instances over total positive instances
- 3. Recall Score – the ratio of correctly predicted instances over total instances in that class
- 4. Roc Curve – a plot of true positive rate against false positive rate
- 5. Classification Report – report of precision, recall and f1 score
- 6. Confusion Matrix – a table used to describe the classification models

Tools: ML Algorithms

- Naive Bayes Classifier: Based on Bayes’ theorem, it calculates the probability of a text belonging to a specific sentiment class.
- Support Vector Machines (SVM): A machine learning algorithm that separates data into different classes using hyperplanes.
- Recurrent Neural Networks (RNN): Particularly LSTM (Long Short-Term Memory) models, which capture sequential information in text data.
- Convolutional Neural Networks (CNN): Effective for capturing local patterns in text through convolutional filters.
- Decision Trees: Constructed based on features of the text to classify sentiments. The choice of algorithm depends on the specific requirements and characteristics of the sentiment analysis task.

Types of Sentiment Analysis



Types of Sentiment Analysis

Emotion Detection Sentiment Analysis

It helps to detect and understand the emotions of the people.

Aspect based Sentiment Analysis

It is more focused on the aspects of a particular product or service.

Fine Grained Sentiment Analysis

It helps in studying the ratings and reviews given by the customers.

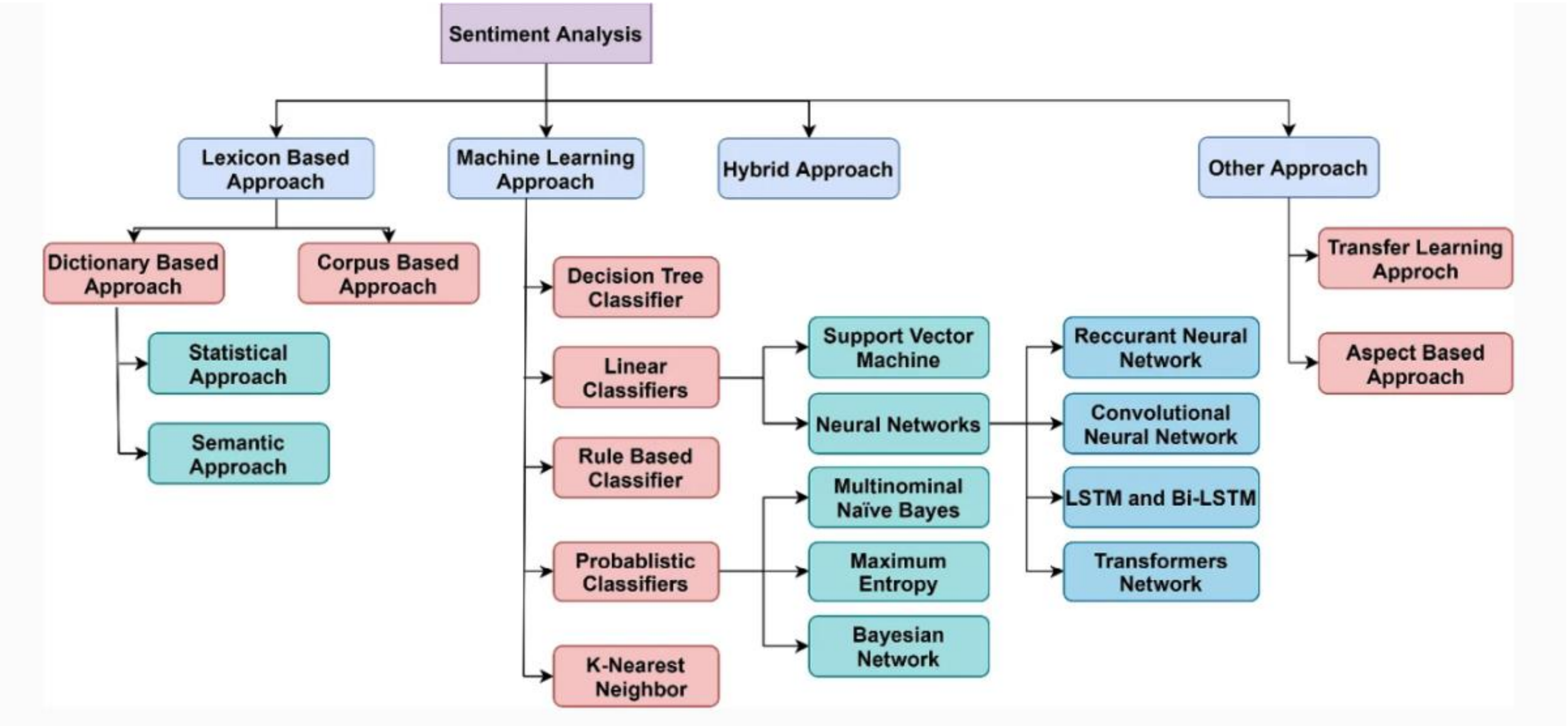
Intent based Sentiment Analysis

To know the intent of the customers, whether they are looking to buy the product or just browsing around, is achievable through intent analysis.

Applications of Sentiment Analysis

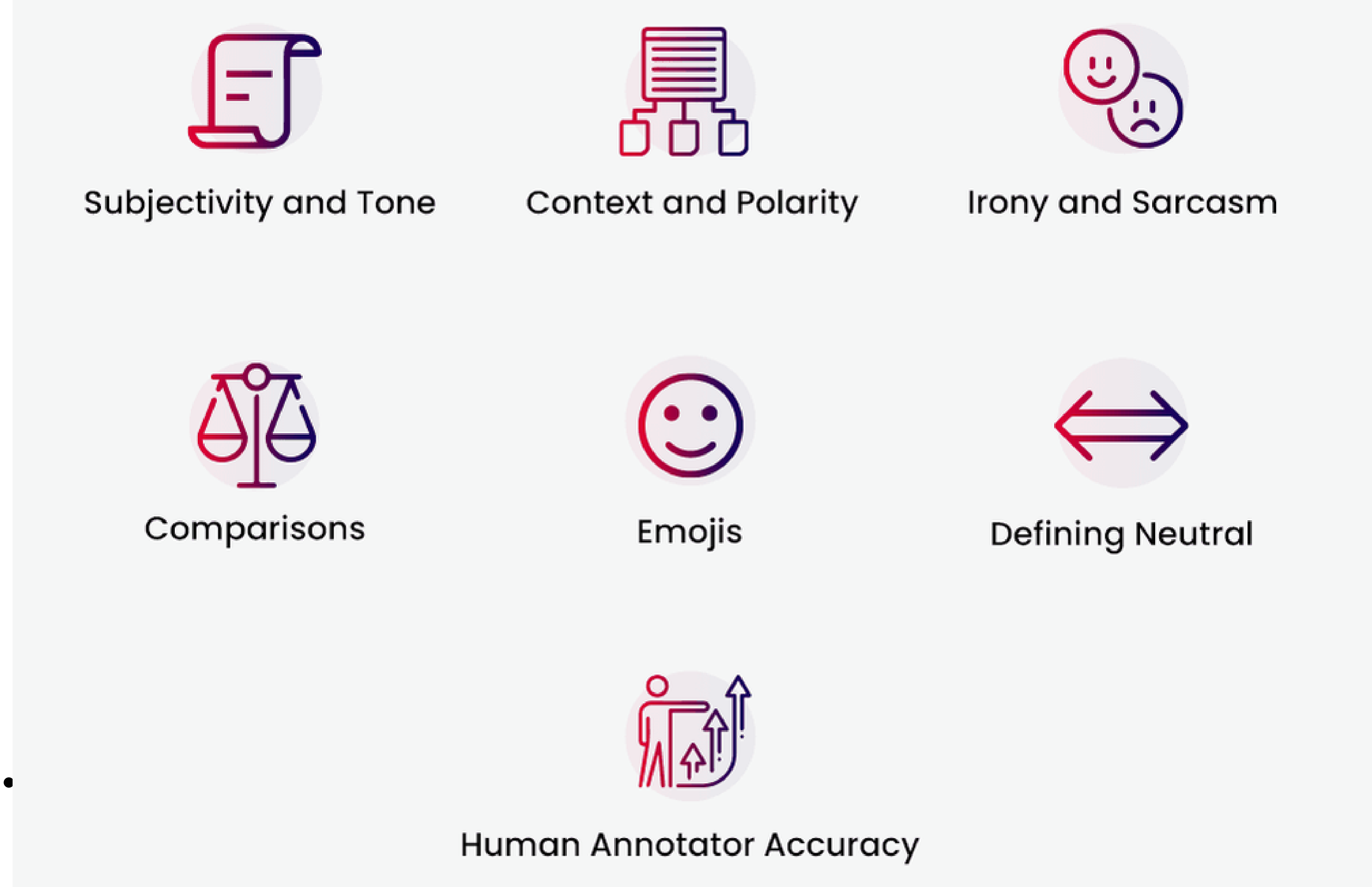
- Brand Monitoring
- Customer Service
- Finance and Stock Monitoring
- Business Intelligence Buildup
- Enhancing the Customer Experience
- Market Research and Analysis

Sentiment Analysis Methods



https://researchmultiple.com/sentiment-analysis-methods/.

Challenges Faced During Sentiment Analysis



Subjectivity and Tone

Context and Polarity

Irony and Sarcasm

Comparisons

Emojis

Defining Neutral

Human Annotator Accuracy

Challenges [Cont'd]

- **Identification of sarcasm:** sometimes without knowing the personality of the person, you do not know whether "bad" means bad or good.
- **Lack of text structure:** in the case of Twitter, for example, it may contain abbreviations, and there may be a lack of capitals, poor spelling, poor punctuation, and poor grammar, all of which make it difficult to analyze the text.
- **Many possible sentiment categories and degrees:** positive and negative is a simple analysis, one would like to identify the amount of hate there is inside the opinion, how much happiness, how much sadness, etc.
- **Identification of the object of analysis:** many concepts can appear in text, and how to detect the object that the opinion is positive for and the object that the opinion is negative for is an open issue. For example, if you say "She won him!", this means a positive sentiment for her and a negative sentiment for him, at the same time.
- **Subjective text:** another open challenge is how to analyze very subjective sentences or paragraphs. Sometimes, even for humans it is very hard to agree on the sentiment of these highly subjective texts.

THANK YOU