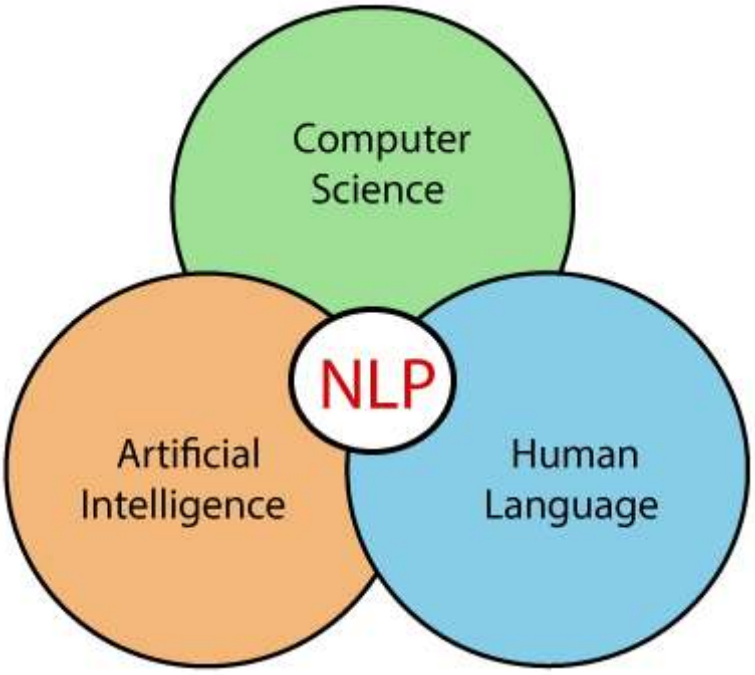
1. List various applications of NLP and discuss any one application in detail.
2. What is Natural language processing (NLP) ? Discuss various stages involved in NLP process with suitable example.
3. Describe various challenges in Natural language processing.
4. How to build an NLP pipline.
5. What's the need for text summarization? Discuss different approaches.
6. Explain text summarization with neat diagram.
7. How does a text summarization algorithm work?

What is NLP?

NLP stands for **Natural Language Processing**, which is a part of **Computer Science, Human language,** and **Artificial Intelligence.** It is the technology that is used by machines to understand, analyse, manipulate, and interpret human's languages. It helps developers to organize knowledge for performing tasks such as **translation, automatic summarization, Named Entity Recognition (NER), speech recognition, relationship extraction,** and **topic segmentation.**



Major Challenges of Natural Language Processing (NLP)

Natural Language Processing limitations and problems:

- **Contextual words and phrases and homonyms**
- **Synonyms**
- **Irony and sarcasm**
- **Ambiguity**

- **Errors in text or speech**
- **Colloquialisms and slang**
- **Domain-specific language**

1. Contextual words and phrases and homonyms

   - The same words and phrases can have different meanings according the context of a sentence and many words – especially in English – have the exact same pronunciation but totally different meanings.
   - Eg. I ran to the store because we ran out of milk.
   - Homonyms – two or more words that are pronounced the same but have different definitions – can be problematic for question answering and speech-to-text applications because they aren't written in text form. Usage of their and there, for example, is even a common problem for humans.

2. Synonyms
   - Synonyms can lead to issues similar to contextual understanding because we use many different words to express the same idea.
   - for building NLP systems, it's important to include all of a word's possible meanings and all possible synonyms.

3. Irony and sarcasm
   - Irony and sarcasm present problems for machine learning models because they generally use words and phrases that, strictly by definition, may be positive or negative, but actually connote the opposite.

4. Ambiguity
5. Errors in text and speech
   - Misspelled or misused words can create problems for text analysis. Autocorrect and grammar correction applications can handle common mistakes, but don't always understand the writer's intention.

6. Colloquialisms and slang
   - Informal phrases, expressions, idioms, and culture-specific lingo present a number of problems for NLP – especially for models intended for broad use.

7. Domain-specific language
   - Different businesses and industries often use very different language. An NLP processing model needed for healthcare, for example, would be very different than one used to process legal documents.

Advantages of NLP

- NLP helps users to ask questions about any subject and get a direct response within seconds.
- NLP offers exact answers to the question means it does not offer unnecessary and unwanted information.

- NLP helps computers to communicate with humans in their languages.
- It is very time efficient.
- Most of the companies use NLP to improve the efficiency of documentation processes, accuracy of documentation, and identify the information from large databases.

Disadvantages of NLP

A list of disadvantages of NLP is given below:

- NLP may not show context.
- NLP is unpredictable
- NLP may require more keystrokes.
- NLP is unable to adapt to the new domain, and it has a limited function that's why NLP is built for a single and specific task only.

Components of NLP

There are the following two components of NLP -

1. Natural Language Understanding (NLU)

Natural Language Understanding (NLU) helps the machine to understand and analyse human language by extracting the metadata from content such as concepts, entities, keywords, emotion, relations, and semantic roles.

NLU mainly used in Business applications to understand the customer's problem in both spoken and written language.

NLU involves the following tasks -

- It is used to map the given input into useful representation.
- It is used to analyze different aspects of the language.

2. Natural Language Generation (NLG)

Natural Language Generation (NLG) acts as a translator that converts the computerized data into natural language representation. It mainly involves Text planning, Sentence planning, and Text Realization.

Note: The NLU is difficult than NLG.

Difference between NLU and NLG

NLU	NLG
NLU is the process of reading and interpreting language.	NLG is the process of writing or generating language.
It produces non-linguistic outputs from natural language inputs.	It produces constructing natural language outputs from non-linguistic inputs.

Applications of NLP

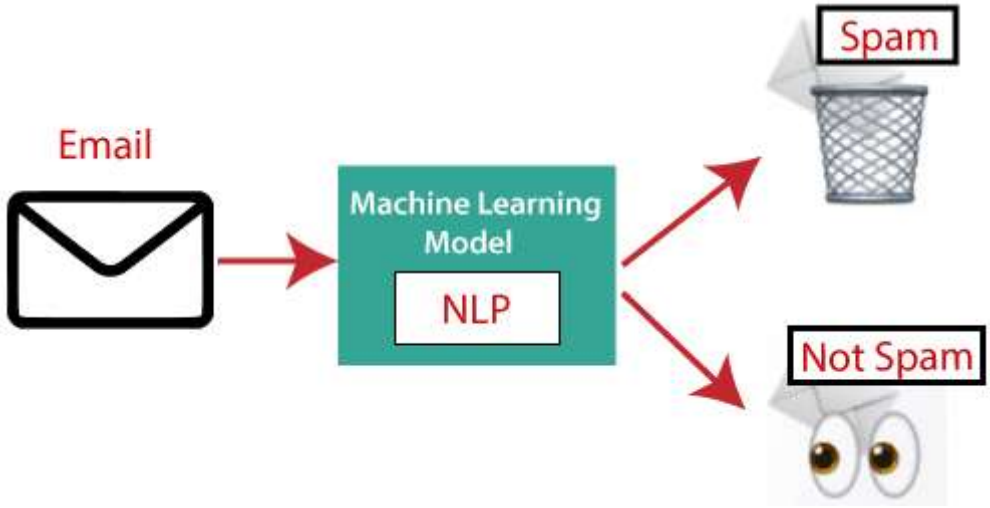
There are the following applications of NLP -

1. Question Answering

Question Answering focuses on building systems that automatically answer the questions asked by humans in a natural language.

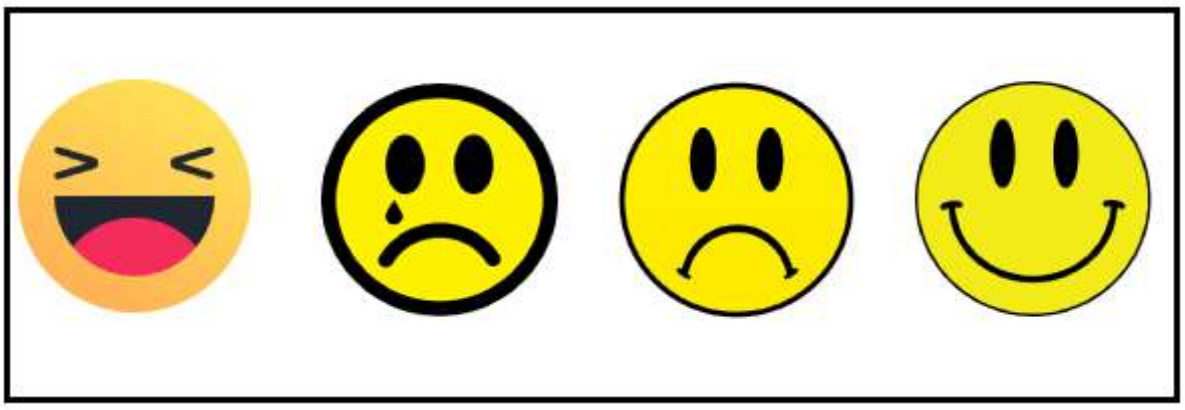
2. Spam Detection

Spam detection is used to detect unwanted e-mails getting to a user's inbox.



3. Sentiment Analysis

Sentiment Analysis is also known as **opinion mining**. It is used on the web to analyse the attitude, behaviour, and emotional state of the sender. This application is implemented through a combination of NLP (Natural Language Processing) and statistics by assigning the values to the text (positive, negative, or natural), identify the mood of the context (happy, sad, angry, etc.)



4. Machine Translation

Machine translation is used to translate text or speech from one natural language to another natural language.

Example: Google Translator

5. Spelling correction

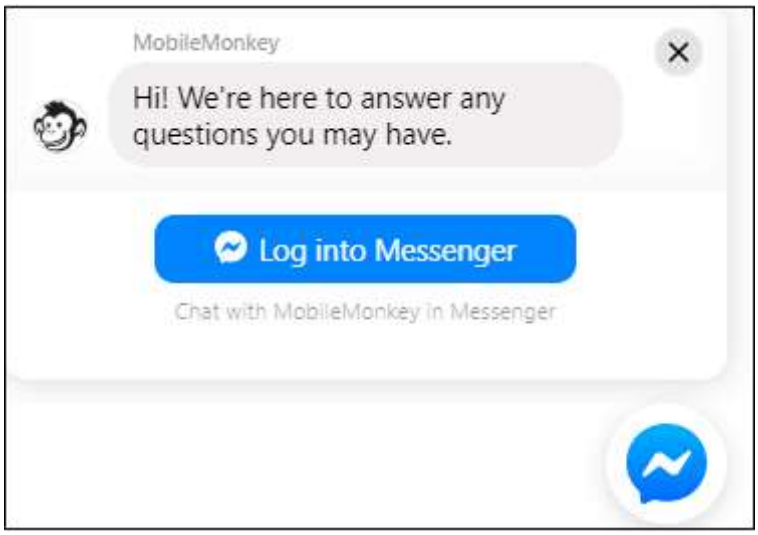
Microsoft Corporation provides word processor software like MS-word, PowerPoint for the spelling correction.

6. Speech Recognition

Speech recognition is used for converting spoken words into text. It is used in applications, such as mobile, home automation, video recovery, dictating to Microsoft Word, voice biometrics, voice user interface, and so on.

7. Chatbot

Implementing the Chatbot is one of the important applications of NLP. It is used by many companies to provide the customer's chat services.



8. Information extraction

Information extraction is one of the most important applications of NLP. It is used for extracting structured information from unstructured or semi-structured machine-readable documents.

9. Natural Language Understanding (NLU)

It converts a large set of text into more formal representations such as first-order logic structures that are easier for the computer programs to manipulate notations of the natural language processing.

How to build an NLP pipeline

There are the following steps to build an NLP pipeline -

Step1: Sentence Segmentation

Sentence Segment is the first step for building the NLP pipeline. It breaks the paragraph into separate sentences.

Example: Consider the following paragraph -

Independence Day is one of the important festivals for every Indian citizen. It is celebrated on the 15th of August each year ever since India got independence from the British rule. The day celebrates independence in the true sense.

Sentence Segment produces the following result:

1. "Independence Day is one of the important festivals for every Indian citizen."
2. "It is celebrated on the 15th of August each year ever since India got independence from the British rule."
3. "This day celebrates independence in the true sense."

Step2: Word Tokenization

Word Tokenizer is used to break the sentence into separate words or tokens.

Example:

JavaTpoint offers Corporate Training, Summer Training, Online Training, and Winter Training.

Word Tokenizer generates the following result:

"JavaTpoint", "offers", "Corporate", "Training", "Summer", "Training", "Online", "Training", "and", "Winter", "Training", "."

Step3: Stemming

Stemming is used to normalize words into its base form or root form. For example, celebrates, celebrated and celebrating, all these words are originated with a single root word "celebrate." The big problem with stemming is that sometimes it produces the root word which may not have any meaning.

For Example, intelligence, intelligent, and intelligently, all these words are originated with a single root word "intelligen." In English, the word "intelligen" do not have any meaning.

Step 4: Lemmatization

Lemmatization is quite similar to the Stamming. It is used to group different inflected forms of the word, called Lemma. The main difference between Stemming and lemmatization is that it produces the root word, which has a meaning.

For example: In lemmatization, the words intelligence, intelligent, and intelligently has a root word intelligent, which has a meaning.

Step 5: Identifying Stop Words

In English, there are a lot of words that appear very frequently like "is", "and", "the", and "a". NLP pipelines will flag these words as stop words. **Stop words** might be filtered out before doing any statistical analysis.

Example: He **is a** good boy.

Step 6: Dependency Parsing

Dependency Parsing is used to find that how all the words in the sentence are related to each other.

Step 7: POS tags

POS stands for parts of speech, which includes Noun, verb, adverb, and Adjective. It indicates that how a word functions with its meaning as well as grammatically within the sentences. A word has one or more parts of speech based on the context in which it is used.

Example: "Google" something on the Internet.

In the above example, Google is used as a verb, although it is a proper noun.

Step 8: Named Entity Recognition (NER)

Named Entity Recognition (NER) is the process of detecting the named entity such as person name, movie name, organization name, or location.

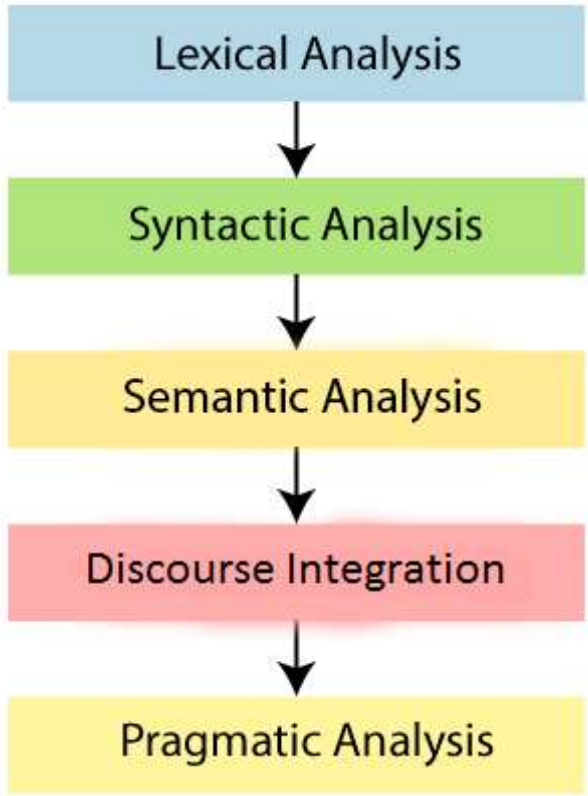
Example: Steve Jobs introduced iPhone at the Macworld Conference in San Francisco, California.

Step 9: Chunking

Chunking is used to collect the individual piece of information and grouping them into bigger pieces of sentences.

Phases of NLP

There are the following five phases of NLP:



1. Lexical Analysis and Morphological

The first phase of NLP is the Lexical Analysis. This phase scans the source code as a stream of characters and converts it into meaningful lexemes. It divides the whole text into paragraphs, sentences, and words.

2. Syntactic Analysis (Parsing)

Syntactic Analysis is used to check grammar, word arrangements, and shows the relationship among the words.

Example: Agra goes to the Poonam

In the real world, Agra goes to the Poonam, does not make any sense, so this sentence is rejected by the Syntactic analyzer.

3. Semantic Analysis

Semantic analysis is concerned with the meaning representation. It mainly focuses on the literal meaning of words, phrases, and sentences.

4. Discourse Integration

Discourse Integration depends upon the sentences that proceeds it and also invokes the meaning of the sentences that follow it.

5. Pragmatic Analysis

Pragmatic is the fifth and last phase of NLP. It helps you to discover the intended effect by applying a set of rules that characterize cooperative dialogues.

For Example: "Open the door" is interpreted as a request instead of an order.

Why NLP is difficult?

NLP is difficult because Ambiguity and Uncertainty exist in the language.

Ambiguity

There are the following three ambiguity -

- **Lexical Ambiguity**

Lexical Ambiguity exists in the presence of two or more possible meanings of the sentence within a single word.

Example:

Manya is looking for a **match**.

In the above example, the word match refers to that either Manya is looking for a partner or Manya is looking for a match. (Cricket or other match)

- **Syntactic Ambiguity**

Syntactic Ambiguity exists in the presence of two or more possible meanings within the sentence.

Example:

I saw the girl with the binocular.

In the above example, did I have the binoculars? Or did the girl have the binoculars?

- **Referential Ambiguity**

Referential Ambiguity exists when you are referring to something using the pronoun.

Example: Kiran went to Sunita. She said, "I am hungry."

In the above sentence, you do not know that who is hungry, either Kiran or Sunita.

NLP Libraries

Scikit-learn: It provides a wide range of algorithms for building machine learning models in Python.

Natural language Toolkit (NLTK): NLTK is a complete toolkit for all NLP techniques.

Pattern: It is a web mining module for NLP and machine learning.

TextBlob: It provides an easy interface to learn basic NLP tasks like sentiment analysis, noun phrase extraction, or pos-tagging.

Quepy: Quepy is used to transform natural language questions into queries in a database query language.

SpaCy: SpaCy is an open-source NLP library which is used for Data Extraction, Data Analysis, Sentiment Analysis, and Text Summarization.

Gensim: Gensim works with large datasets and processes data streams.

Difference between Natural language and Computer Language

Natural Language	Computer Language
Natural language has a very large vocabulary.	Computer language has a very limited vocabulary.
Natural language is easily understood by humans.	Computer language is easily understood by the machines.
Natural language is ambiguous in nature.	Computer language is unambiguous.

https://www.analyticsvidhya.com/blog/2022/04/a-comprehensive-overview-of-sentiment-analysis/

A Comprehensive Overview of Sentiment Analysis

Introduction

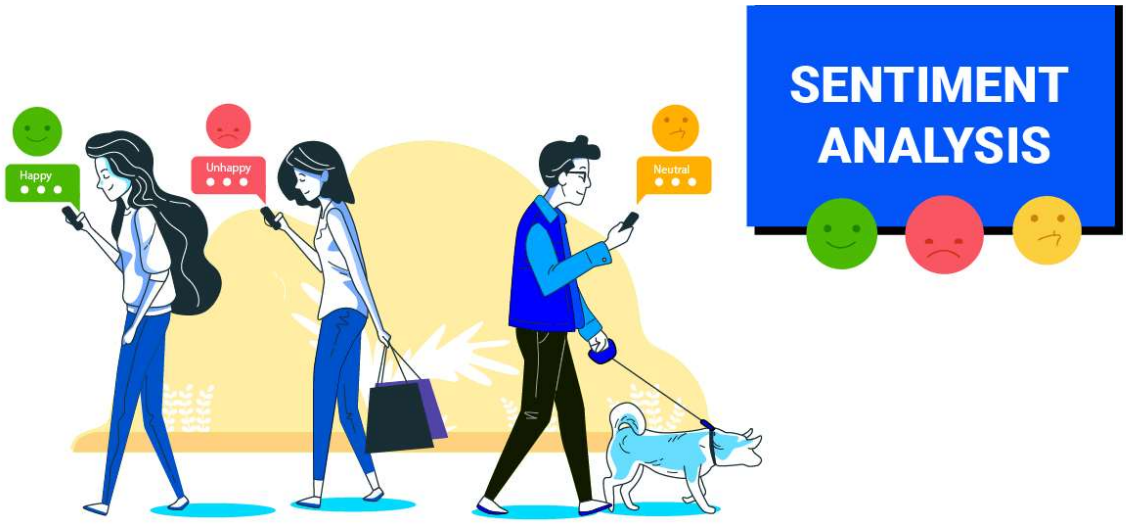
We can clearly see that sentiment analysis is becoming more popular as e-commerce, SaaS solutions, and digital technologies advance. We'll go through how this works and look at some of the most common corporate applications. We'll also discuss the analysis' existing issues and limitations.

Sentiment analysis examines how a text expresses emotion. Customer feedback, survey replies, and product reviews are all frequent uses. This can be useful in various situations, including social media monitoring, reputation management, and customer service. For example, Analyzing thousands of product reviews might provide important feedback on pricing and product features.

The desire of people to interact with businesses, as well as the overall brand perception, is significantly influenced by public opinion. 93 percent of shoppers think online reviews influence their purchasing decisions, according to a Podium survey. After reading a few negative reviews, users may be less willing to give you a chance. They won't look into whether or not the feedback was genuine. They'll pick a different path. In this setting, companies that keep a close eye on their reputation can handle problems quickly and improve operations based on feedback. In the information era, such analysis enables the accurate measurement of people's attitudes toward a company.

What is Sentimental Analysis?

Sentiment analysis, also known as opinion mining, is a natural language processing (NLP) technique for determining the positivity, negativity, or neutrality of data. It is frequently used on textual data to assist organizations in tracking brand and product sentiment in consumer feedback, and better understanding customer demands.



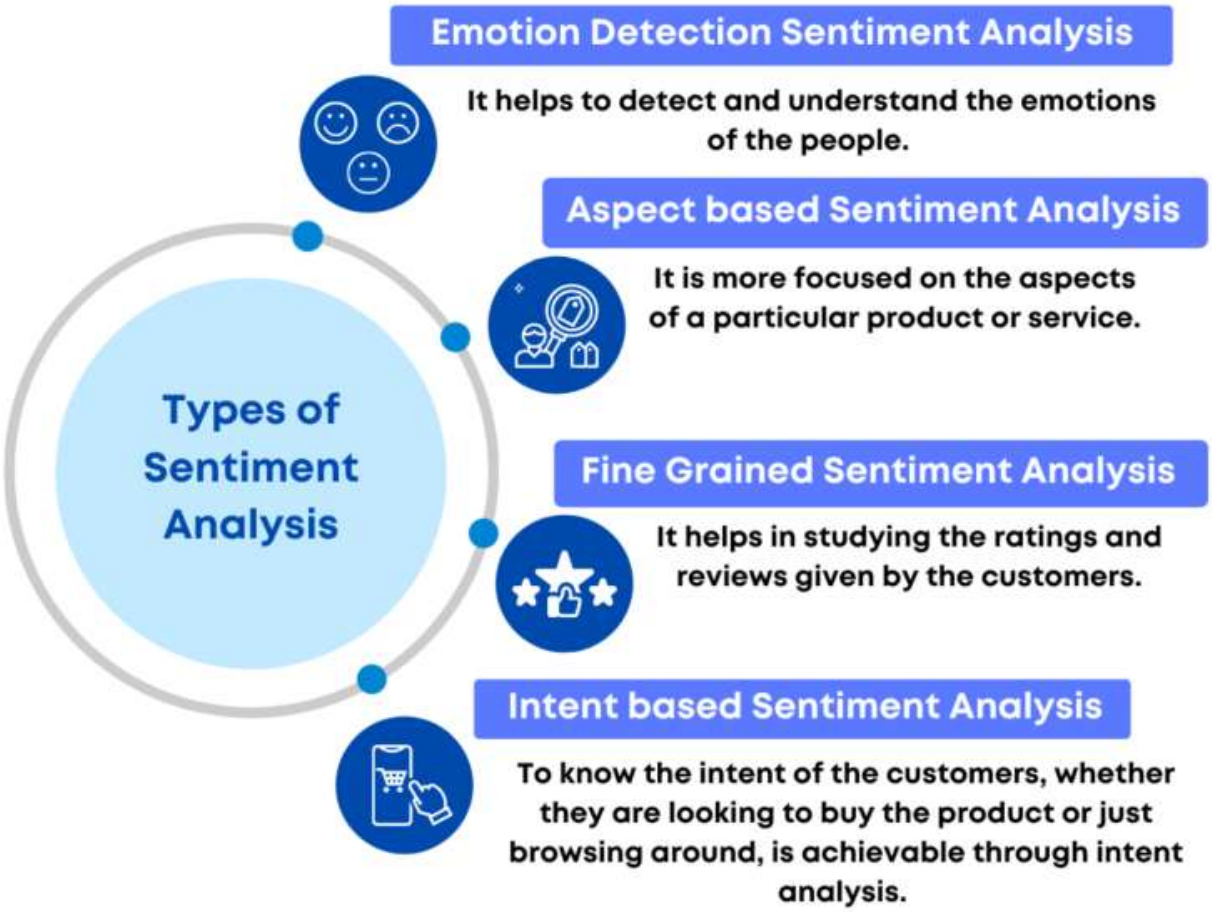
Source: Surveysensum.com

The tools assist businesses in extracting information from unstructured and unorganized text found on the internet, such as emails, blog posts, support tickets, webchats, social media channels, forums, and comments. To replace manual data processing, algorithms use rule-based, automatic, or hybrid techniques. Automatic systems learn from data using machine learning techniques, whereas rule-based systems execute sentiment analysis based on predetermined, lexicon-based rules. Both methodologies are combined in hybrid sentiment analysis.

While there are many different types of sentiment analysis techniques, fine-grained sentiment analysis, emotion detection, aspect-based sentiment analysis, and intent analysis are the most popular.

Types of Sentiment Analysis

Polarity categorization is an important part of sentiment analysis. The overall sentiment expressed by a paragraph, phrase, or word is referred to as polarity. This polarity can be measured using a “sentiment score,” which is a numerical rating. This score can be calculated for the complete text or for a single phrase.



Source: Indiaai.com

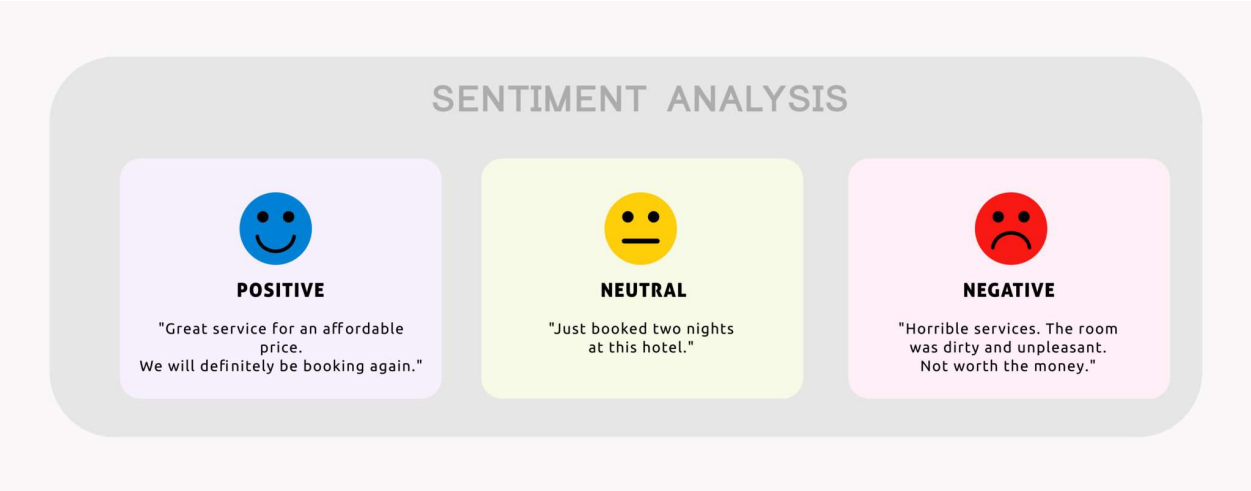
Depending on how you wish to interpret client feedback and inquiries, you can define and customize your categories to match your sentiment analysis needs. Meanwhile, these are some of the most common methodologies for sentiment analysis:

1. **Fine-grained Sentiment Analysis:** breaks down polarity into smaller groups, usually highly positive to very negative, to provide a more specific level of polarity. This can be compared to a 5-star rating system in terms of opinion.
2. **Aspect-based Sentiment Analysis (ABSA):** When it is related to a specific property or feature described in the text, it is most useful. ABSA is the process of discovering these traits or features and their sentiment. These features are referred to as “themes” at Thematic.

3. **Emotion detection:** Rather than detecting positive and negative emotions, emotion detection detects specific emotions. Happiness, frustration, shock, anger, and grief are only a few examples.
4. **Intent-based:** Intent-based analysis distinguishes between facts and opinions in a text. An online comment indicating dissatisfaction with changing a battery, for example, can motivate customer service to contact you to remedy the problem.

Why is it Important?

Sentiment analysis is snappily getting a pivotal tool for monitoring and understanding sentiment in all forms of data, as humans communicate their studies and passions more openly than ever ahead. Brands can discover what makes guests happy or unhappy by automatically assessing consumer input, similar to commentary in check replies and social media discourses. This enables them to knitter products and services to meet the requirements of their guests.



Source: Expressanalytics.com

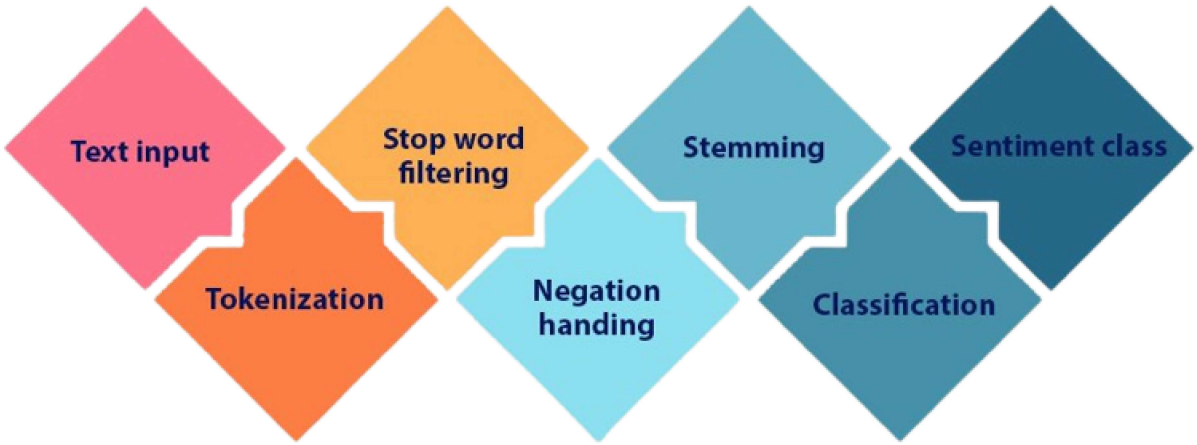
The following are some of the advantages:

1. **Sorting Data at Scale:** Can you imagine going through thousands of tweets, customer service discussions, or survey responses by hand? There is simply too much corporate data to process manually. Sentiment analysis aids firms in efficiently and cost-effectively processing large amounts of unstructured data.
2. **Real-Time Analysis:** Sentiment analysis can detect key concerns in real-time, such as whether a social media PR crisis is escalating. Is a disgruntled consumer about to leave? Sentiment analysis models can assist you in quickly identifying these types of circumstances so that you can take appropriate action.

3. **Consistent criteria:** When it comes to determining the sentiment of a text, it's estimated that just 60-65 percent of the time, people agree. Text sentiment tagging is a highly subjective process impacted by human experiences, thoughts, and beliefs. Companies can apply the same criteria to all of their data by adopting a centralized sentiment analysis system, which helps them enhance accuracy and generate better insights.

How does Sentiment Analysis Work?

The artificially intelligent bots are programmed to detect whether a message is favorable, negative, or neutral based on millions of pieces of text. Sentiment analysis divides communication into topic chunks and assigns each one a sentiment score.



Source: Medium.com

The process for basic sentiment analysis of text documents is simple:

1. Break down each text document into its individual components (sentences, phrases, tokens, and parts of speech).
2. Identify each phrase and component that carries a sentiment.
3. Each phrase and component should be given a sentiment score (from -1 to +1).
4. For multi-layered sentiment analysis, combine scores (Optional).

Deep Learning & Sentiment Analysis

Deep learning is worth investigating further since it produces the most accurate sentiment analysis. Traditional machine learning techniques, which involve manual

work to define categorization features, dominated the area until recently. They also frequently overlook the significance of word order, and NLP has been changed by deep learning and artificial neural networks.

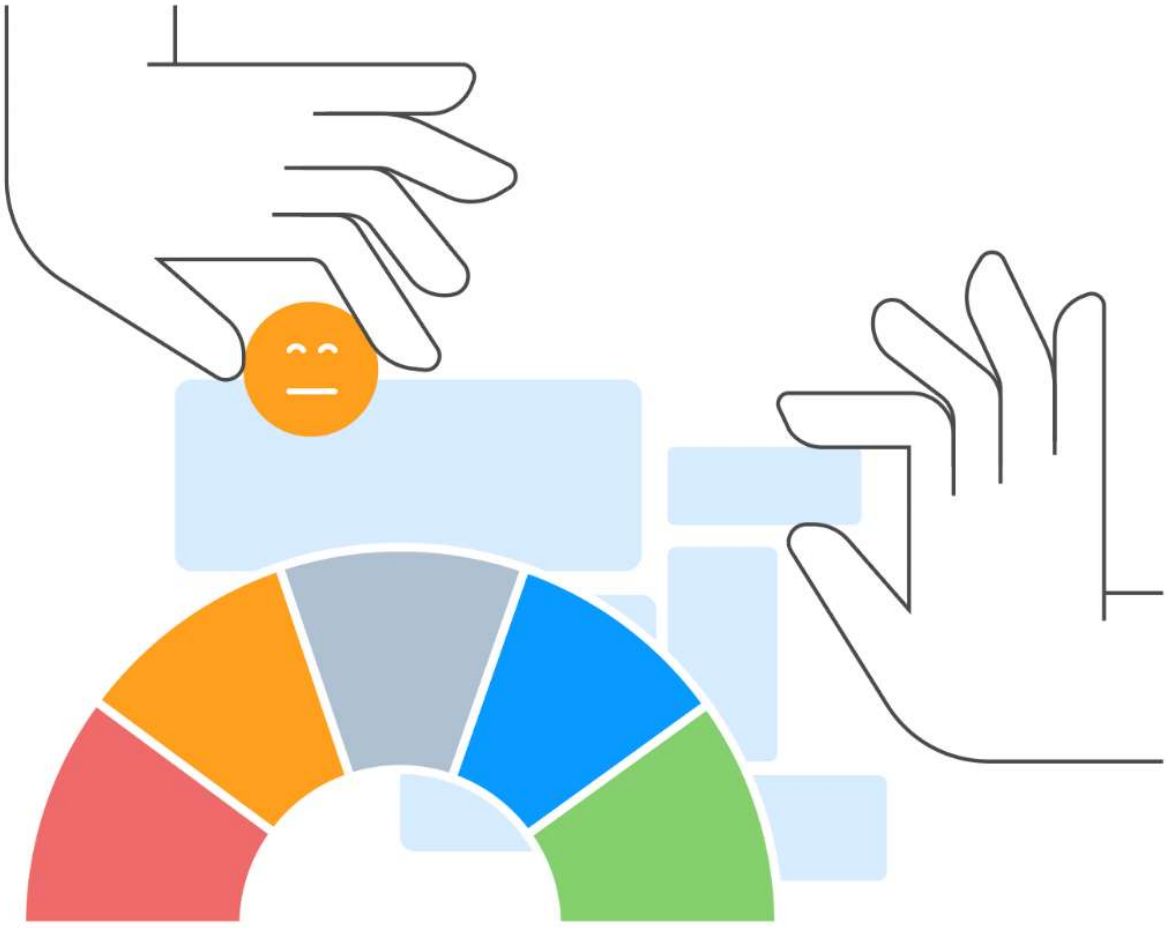
Source: Monkeylearn.com

Well, the structure and function of the human brain-inspired deep learning systems. The accuracy and efficiency of sentiment analysis improved due to this technique. When using deep learning, a neural network can learn to self-correct when it makes a mistake. Errors in traditional machine learning require human involvement to correct.

Challenges with Sentiment Analysis

Inaccuracies in training models are usually the source of problems with sentiment analysis. Objectivity, or neutral-sentiment comments, are an issue for systems and are frequently mistaken. For example, if a consumer received the wrong color item and left a review like “The product was blue,” it would be categorized as neutral rather than negative.

Detecting sentiment might be difficult when systems can’t understand the context or tone. When the context is not provided, answers to polls or survey questions like “**nothing**” or “**everything**” are difficult to categorize. They could be characterized as positive or negative depending on the question. Similarly, irony and sarcasm are difficult to teach and often result in mislabeled emotions.



Source: Medium.com

People's statements can be conflicting. The majority of evaluations will include both good and negative feedback, which may be managed by analyzing sentences one at a time. However, the more informal the medium, the more likely people are to mix diverse points of view in a single sentence, making it harder for a computer to comprehend.

Use Cases

Organizations can utilize sentiment analysis technologies for a variety of purposes, including:

1. Brand reputation management
2. Customer feedback

3. Crisis prevention
4. Market research
5. United Airlines
6. Politics

https://www.analyticsvidhya.com/blog/2023/02/how-to-build-a-chatbot-using-natural-language-processing/

Text summarization refers to the technique of shortening long pieces of text. The intention is to create a coherent and fluent summary having only the main points outlined in the document.

Automatic text summarization is a common problem in machine learning and natural language processing (NLP).

What's the need for text summarization?

Propelled by the modern technological innovations, data is to this century what oil was to the previous one. Today, our world is parachuted by the gathering and dissemination of huge amounts of data.

In fact, the International Data Corporation (IDC) projects that the total amount of digital data circulating annually around the world would sprout from 4.4 zettabytes in 2013 to hit 180 zettabytes in 2025. That's a lot of data!

With such a big amount of data circulating in the digital space, there is need to develop machine learning algorithms that can automatically shorten longer texts and deliver accurate summaries that can fluently pass the intended messages.

Furthermore, applying text summarization reduces reading time, accelerates the process of researching for information, and increases the amount of information that can fit in an area.

What are the main approaches to automatic summarization?

There are two main types of how to summarize text in NLP:

- **Extraction-based summarization**

The extractive text summarization technique involves pulling keyphrases from the source document and combining them to make a summary. The extraction is made according to the defined metric without making any changes to the texts.

Here is an example:

Source text: ***Joseph and Mary*** *rode on a donkey to* ***attend*** *the annual* ***event*** *in* ***Jerusalem****. In the city,* ***Mary*** *gave* ***birth*** *to a child named* ***Jesus****.*

Extractive summary: *Joseph and Mary attend event Jerusalem. Mary birth Jesus.*

As you can see above, the words in bold have been extracted and joined to create a summary — although sometimes the summary can be grammatically strange.

- **Abstraction-based summarization**

The abstraction technique entails paraphrasing and shortening parts of the source document. When abstraction is applied for text summarization in deep learning problems, it can overcome the grammar inconsistencies of the extractive method.

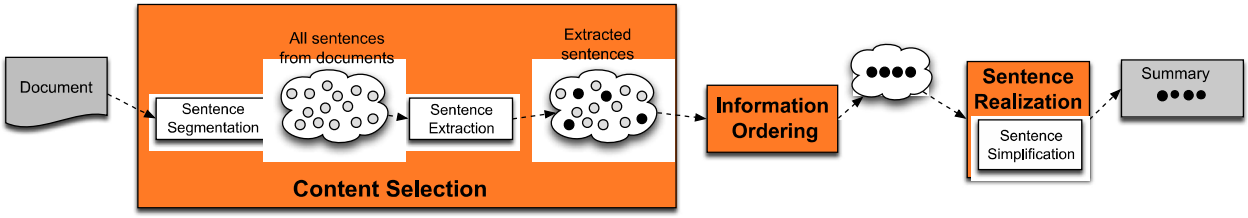
The abstractive text summarization algorithms create new phrases and sentences that relay the most useful information from the original text — just like humans do.

Therefore, abstraction performs better than extraction. However, the text summarization algorithms required to do abstraction are more difficult to develop; that's why the use of extraction is still popular.

Here is an example:

Abstractive summary: *Joseph and Mary came to Jerusalem where Jesus was born.*

Explain text summarization with neat diagram.



How does a text summarization algorithm work?

Usually, text summarization in NLP is treated as a supervised machine learning problem (where future outcomes are predicted based on provided data).

Typically, here is how using the extraction-based approach to summarize texts can work:

1. Introduce a method to extract the merited keyphrases from the source document. For example, you can use part-of-speech tagging, words sequences, or other linguistic patterns to identify the keyphrases.

2. Gather text documents with positively-labeled keyphrases. The keyphrases should be compatible to the stipulated extraction technique. To increase accuracy, you can also create negatively-labeled keyphrases.

3. Train a binary machine learning classifier to make the text summarization. Some of the features you can use include:

- Length of the keyphrase
- Frequency of the keyphrase
- The most recurring word in the keyphrase
- Number of characters in the keyphrase

4. Finally, in the test phrase, create all the keyphrase words and sentences and carry out classification for them.

Summary

Text summarization is an interesting machine learning field that is increasingly gaining traction. As research in this area continues, we can expect to see breakthroughs that will assist in fluently and accurately shortening long text documents.