PART B UNIT 6
NATURAL LANGUAGE PROCESSING

CLASS 10
ARTIFICIAL INTELLIGENCE (417)
INDIAN SCHOOL AL WADI AL KABIR

Natural Language Processing

Natural Language Processing (commonly called NLP) takes in the data of Natural Languages which humans use in their daily lives and operates on this.

Natural Language Processing, or NLP, is the sub-field of AI that is focused on enabling computers to understand and process human languages.

Applications of Natural Language Processing:

Automatic Summarization:

Information overload is a real problem when we need to access a specific, important piece of information from a huge knowledge base. Automatic summarization is relevant not only for summarizing the meaning of documents and information, but also to understand the emotional meanings within the information, such as in collecting data from social media.

Sentiment Analysis:

The goal of sentiment analysis is to identify sentiment among several posts or even in the same post where emotion is not always explicitly expressed.

Companies use Natural Language Processing applications, such as sentiment analysis, to identify opinions and sentiment online to help them understand what customers think about their products and services

Text classification:

Text classification makes it possible to assign predefined categories to a document and organize it to help you find the information you need or simplify some activities. For example, an application of text categorization is spam filtering in email.

Virtual Assistants:

Accessing our data, helps us in keeping notes of our tasks, make calls for us, send messages and a lot more. With the help of speech recognition, these assistants can not only detect our speech but can also make sense out of it.

Ex: Google Assistant, Cortana, Siri, Alexa, etc

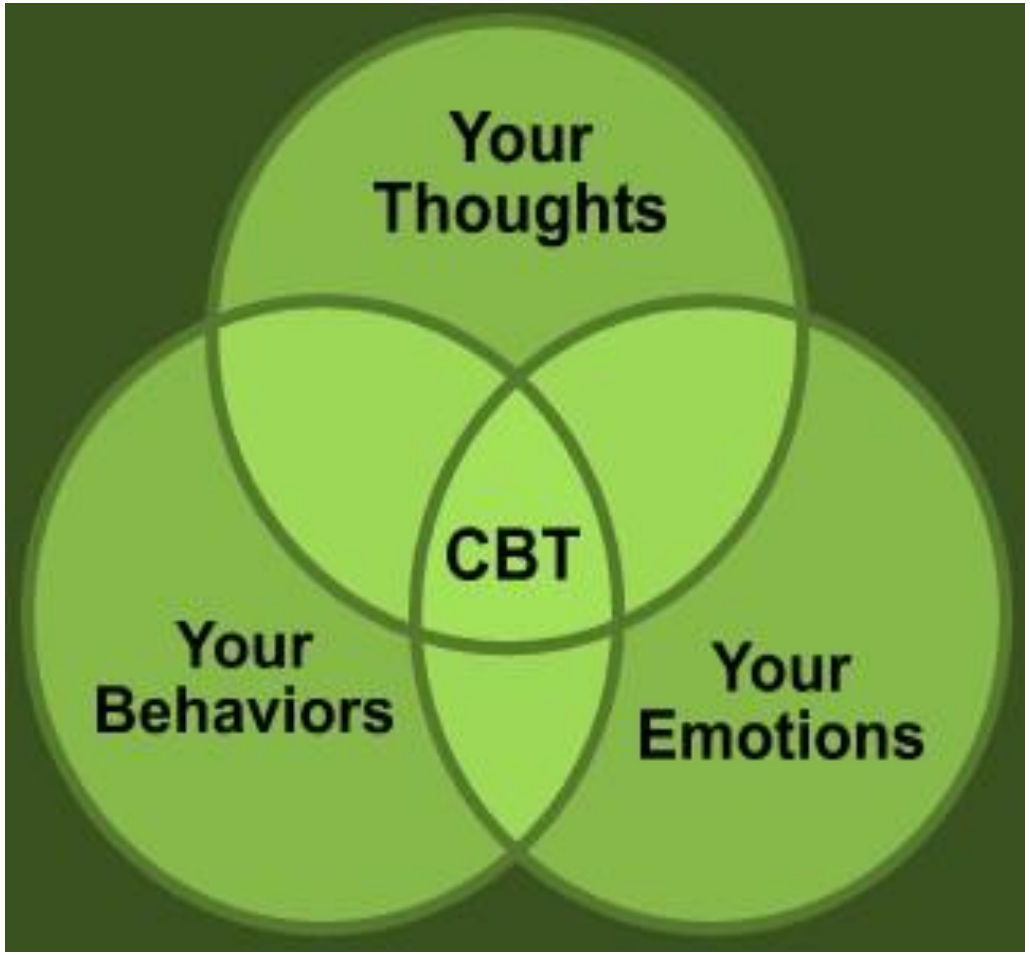
Wordtune (AI writing tool that rewrites, rephrases, and rewords your writing)

Revisiting the AI Project Cycle

Let us try to understand how we can develop a project in Natural Language processing with the help of an example.

The Scenario

The world is competitive nowadays. People face competition in even the tiniest tasks and are expected to give their best at every point in time. When people are unable to meet these expectations, they get stressed and could even go into depression. We get to hear a lot of cases where people are depressed due to reasons like peer pressure, studies, family issues, relationships, etc. and they eventually get into something that is bad for them as well as for others. So, to overcome this, cognitive behavioural therapy (CBT) is considered to be one of the best methods to address stress as it is easy to implement on people and also gives good results. This therapy includes understanding the behaviour and mindset of a person in their normal life. With the help of CBT, therapists help people overcome their stress and live a happy life.



Problem Scoping

CBT is a technique used by most therapists to cure patients out of stress and depression. But it has been observed that people do not wish to seek the help of a psychiatrist willingly. They try to avoid such interactions as much as possible. Thus, there is a need to bridge the gap between a person who needs help and the psychiatrist. Let us look at various factors around this problem through the **4Ws problem canvas**.

Who Canvas – Who has the problem?

Who are the stakeholders?	o People who suffer from stress and are at the onset of depression.
What do we know about them?	o People who are going through stress are reluctant to consult a psychiatrist.

What Canvas – What is the nature of the problem?

What is the problem?	oPeople who need help are reluctant to consult a psychiatrist and hence live miserably.
How do you know it is a problem?	o Studies around mental stress and depression available on various authentic sources.

Where Canvas – *Where does the problem arise?*

What is the context/situation in which the stakeholders experience this problem?	When they are going through a stressful period of time Due to some unpleasant experiences

Why Canvas – *Why do you think it is a problem worth solving?*

What would be of key value to the stakeholders?	○ **People get a platform where they can talk and vent out their feelings anonymously** ○ **People get a medium that can interact with them and applies primitive CBT on them and can suggest help whenever needed**
How would it improve their situation?	○ **People would be able to vent out their stress** ○ **They would consider going to a psychiatrist whenever required**

The problem statement templates go as follows:

Our	People undergoing stress	Who?
Have a problem of	Not being able to share their feelings	What?
While	They need help in venting out their emotions	Where?
An ideal solution would	Provide them a platform to share their thoughts anonymously and suggest help whenever required	Why

Goal of our project which is:

"To create a chatbot which can interact with people, help them to vent out their feelings and take them through primitive CBT."

Data Acquisition

To understand the sentiments of people, we need to collect their conversational data so the machine can interpret the words that they use and understand their meaning. Such data can be collected from various means:

1. Surveys
2. Observing the therapist's sessions
3. Databases available on the internet
4. Interviews, etc.

Data Exploration

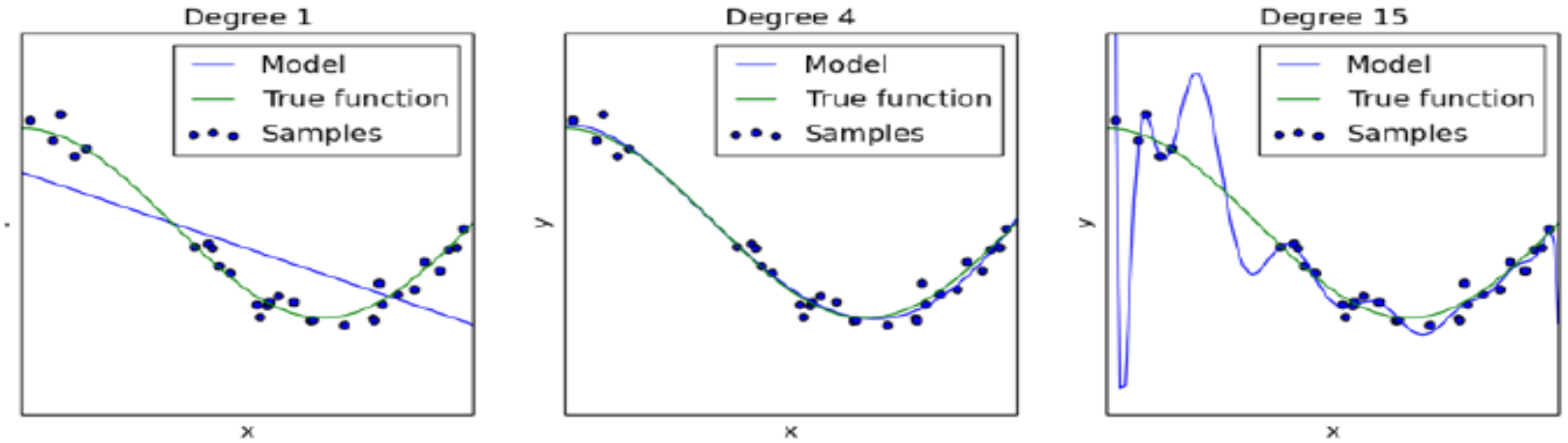
Once the textual data has been collected, it needs to be processed and cleaned so that an easier version can be sent to the machine. Thus, the text is normalized through various steps and is lowered to minimum vocabulary since the machine does not require grammatically correct statements but the essence of it.

Modelling

Once the text has been normalized, it is then fed to an NLP based AI model. Note that in NLP, modelling requires data pre-processing only after which the data is fed to the machine. Depending upon the type of chatbot we try to make, there are a lot of AI models available which help us build the foundation of our project.

Evaluation

The model trained is then evaluated and the accuracy for the same is generated on the basis of the relevance of the answers which the machine gives to the user's responses. To understand the efficiency of the model, the suggested answers by the chatbot are compared to the actual answers.



As you can see in the above diagram, the blue line talks about the model's output while the green one is the actual output along with the data samples.

Figure 1 The model's output does not match the true function at all. Hence the model is said to be underfitting and its accuracy is lower.

Figure 2 In the second one, the model's performance matches well with the true function which states that the model has optimum accuracy and the model is called a perfect fit.

Figure 3 In the third case, model performance is trying to cover all the data samples even if they are out of alignment to the true function. This model is said to be overfitting and this too has a lower accuracy.

Once the model is evaluated thoroughly, it is then deployed in the form of an app which people can use easily.

Chatbots

One of the most common applications of Natural Language Processing is a chatbot.

	• **Mitsuku Bot*** https://www.pandorabots.com/mitsuku/
	• **CleverBot*** https://www.cleverbot.com/
	• **Jabberwacky*** http://www.jabberwacky.com/
	• **Haptik*** https://haptik.ai/contact-us

	• **Rose*** http://ec2-54-215-197-164.us-west-1.compute.amazonaws.com/speech.php
	• **Ochatbot*** https://www.ometrics.com/blog/list-of-fun-chatbots/

CHAT BOTS

Script- bot and Smart-bot.

- **Chatbots** also known as conversational agents, are software applications that mimic written or spoken human speech for the purposes of simulating a conversation or interaction with a real person

- **Examples of script bot** may include the bots which are deployed in the customer care section of various companies. Their job is to answer some basic queries that they are coded for and connect them to human executives once they are unable to handle the conversation.

- **Examples of smart bot :** On the other hand, all the assistants like Google Assistant, Alexa, Cortana, Siri, etc. can be taken as smart bots as not only can they handle the conversations but can also manage to do other tasks which makes them smarter.

There are 2 types of chatbots around us: Script- bot and Smart-bot.

Script-bot	Smart-bot
Script bots are easy to make	Smart-bots are flexible and powerful
Script bots work around a script which is programmed in them	Smart bots work on bigger databases and other resources directly
Mostly they are free and are easy to integrate to a messaging platform	Smart bots learn with more data
No or little language processing skills	Coding is required to take this up on board
Limited functionality	Wide functionality

Language Differences

Human Language VS Computer Language

- **Human Language:** Language is a medium through which humans communicate and express their thoughts and feelings to each other. Language is unique to humans. Humans communicate through language, which we process all the time. Our brain keeps on processing the sounds that it hears around itself and tries to make sense out of them all the time.
- The sound reaches the brain through a long channel. As a person speaks, the sound travels from his mouth and goes to the listener's eardrum. The sound striking the eardrum is converted into neuron impulse, gets transported to the brain and then gets processed. After processing the signal, the brain gains understanding around the meaning of it. If it is clear, the signal gets stored. Otherwise, the listener asks for clarity to the speaker. This is how human languages are processed by humans

- On the other hand, the computer understands the language of numbers. Everything that is sent to the machine has to be converted to numbers. And while typing, if a single mistake is made, the computer throws an error and does not process that part. The communications made by the machines are very basic and simple.

What are the possible difficulties a machine would face in processing natural language?

1. Arrangement of the words and meaning

There are rules in human language. There are nouns, verbs, adverbs, adjectives. A word can be a noun at one time and an adjective some other time. Eg: Life and Life Insurance

There are rules to provide structure to a language. This is the issue related to the syntax of the language. Syntax refers to the grammatical structure of a sentence. When the structure is present, we can start interpreting the message. Now we also want to have the computer do this. One way to do this is to use the part-of-speech tagging. This allows the computer to identify the different parts of a speech.

Besides the matter of arrangement, there's also meaning behind the language we use. Human communication is complex. There are multiple characteristics of the human language that might be easy for a human to understand but extremely difficult for a computer to understand.

Analogy with programming language:

Different syntax, same semantics: 2+3 = 3+2

Here the way these statements are written is different, but their meanings are the same that is 5.

Different semantics, same syntax: 2/3 (Python 2.7) ≠ 2/3 (Python 3)

Here the statements written have the same syntax but their meanings are different.

In Python 2.7, this statement would result in 1 while in Python 3, it would give an output of 1.5.

What are the possible difficulties a machine would face in processing natural language?

2. Multiple Meanings of a word

Let's consider these three sentences:

His face turned red after he found out that he took the wrong bag

What does this mean? Is he feeling ashamed because he took another person's bag instead of his? Is he feeling angry because he did not manage to steal the bag that he has been targeting?

The red car zoomed past his nose

Probably talking about the color of the car

His face turns red after consuming the medicine

Is he having an allergic reaction? Or is he not able to bear the taste of that medicine?

Here we can see that context is important. We understand a sentence almost intuitively, depending on our history of using the language, and the memories that have been built within. In all three sentences, the word red has been used in three different ways which according to the context of the statement changes its meaning completely. **Thus, in natural language, it is important to understand that a word can have multiple meanings and the meanings fit into the statement according to the context of it.**

What are the possible difficulties a machine would face in processing natural language?

3. Perfect Syntax, no Meaning

Sometimes, a statement can have a perfectly correct syntax but it does not mean anything. For example, take a look at this statement:

Chickens feed extravagantly while the moon drinks tea.

This statement is correct grammatically but does this make any sense? In Human language, a perfect balance of syntax and semantics is important for better understanding.

These are some of the challenges we might have to face if we try to teach computers how to understand and interact in human language. So Natural Language Processing do this magic.

Data Processing

- Humans interact with each other very easily. For us, the natural languages that we use are so convenient that we speak them easily and understand them well too. But for computers, our languages are very complex. Natural Language Processing makes it possible for the machines to understand and speak in the Natural Languages just like humans.
- The language of computers is Numerical. So the very first step is to convert our language to numbers. This conversion takes a few steps to happen. **The first step to it is Text Normalisation**.
- Since human languages are complex, we need to first of all simplify them in order to make sure that the understanding becomes possible.

Data Processing

Text Normalisation : In Text Normalisation, we undergo several steps to normalise the text to a lower level. Before we begin, we need to understand that in this section, we will be working on a collection of written text. That is, we will be working on text from multiple documents and **the term used for the whole textual data from all the documents altogether is known as corpus**. Not only would we go through all the steps of Text Normalisation, we would also work them out on a corpus.

Text Normalisation helps in cleaning up the textual data in such a way that it comes down to a level where its complexity is lower than the actual data.

In Text Normalisation, we undergo several steps to normalise the text to a lower level.

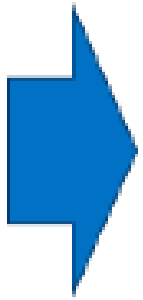
The Steps are:

1. Sentence Segmentation
2. Tokenisation
3. Removing Stopwords, Special Characters and Numbers
4. Converting text to a common case
5. Stemming
6. Lemmatization

1.Sentence Segmentation

Under sentence segmentation, the whole corpus is divided into sentences. Each sentence is taken as a different data so now the whole corpus gets reduced to sentences.

In CBT, we learn to decipher the lies we are undermining ourselves with— based on the bias embedded in the things we say. For example, "I'm never going to make any friends" is an example of all-or-nothing thinking and we feel bad because we buy into this thought.



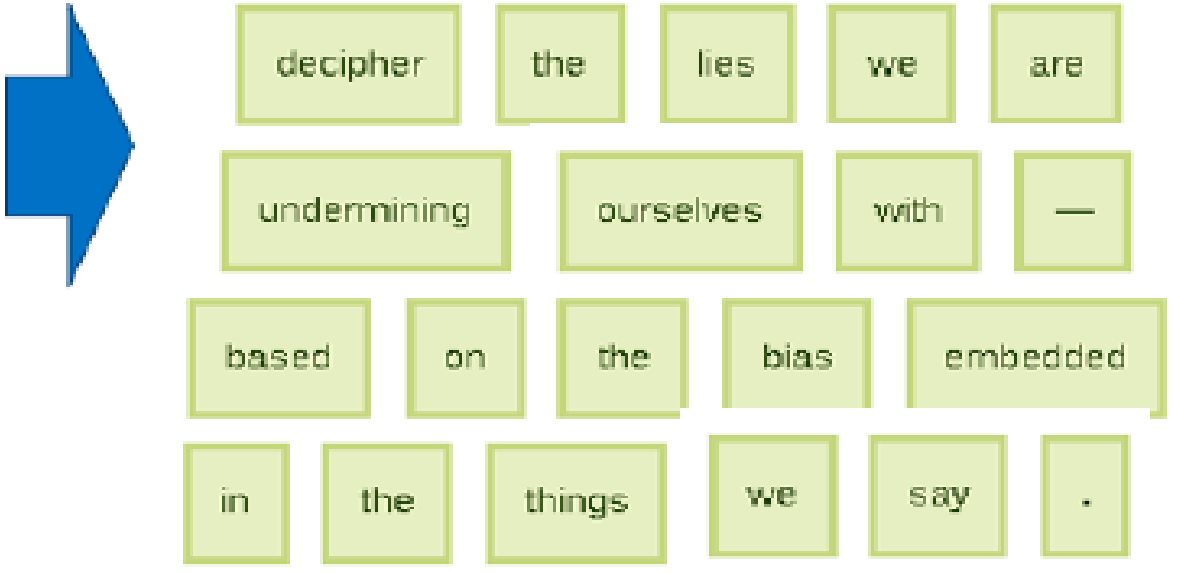
1. In CBT, we learn to decipher the lies we are undermining ourselves with— based on the bias embedded in the things we say.

2. For example, "I'm never going to make any friends" is an example of all-or-nothing thinking and we feel bad because we buy into this thought.

2. Tokenisation

After segmenting the sentences, each sentence is then further divided into tokens. Tokens is a term used for any word or number or special character occurring in a sentence. Under tokenisation, every word, number and special character is considered separately and each of them is now a separate token.

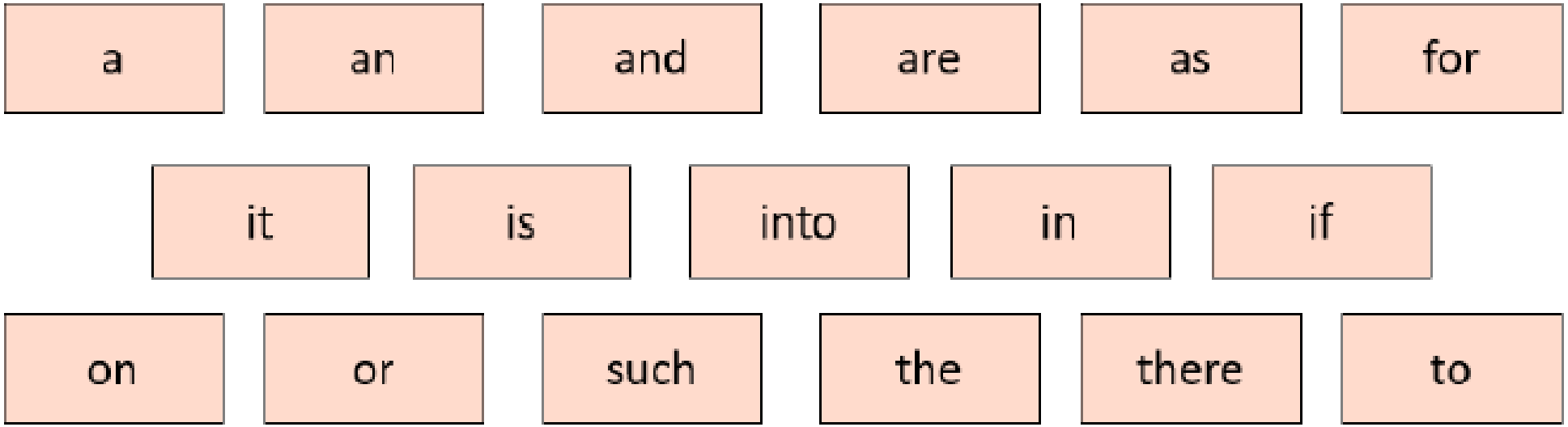
In CBT, we learn to decipher the lies we are undermining ourselves with— based on the bias embedded in the things we say.



3. Removing Stopwords, Special Characters and Numbers

In this step, the tokens which are not necessary are removed from the token list.

Stopwords are the words which occur very frequently in the corpus but do not add any value to it. Humans use grammar to make their sentences meaningful for the other person to understand. But grammatical words do not add any essence to the information which is to be transmitted through the statement hence they come under stopwords. Some examples of stopwords are:

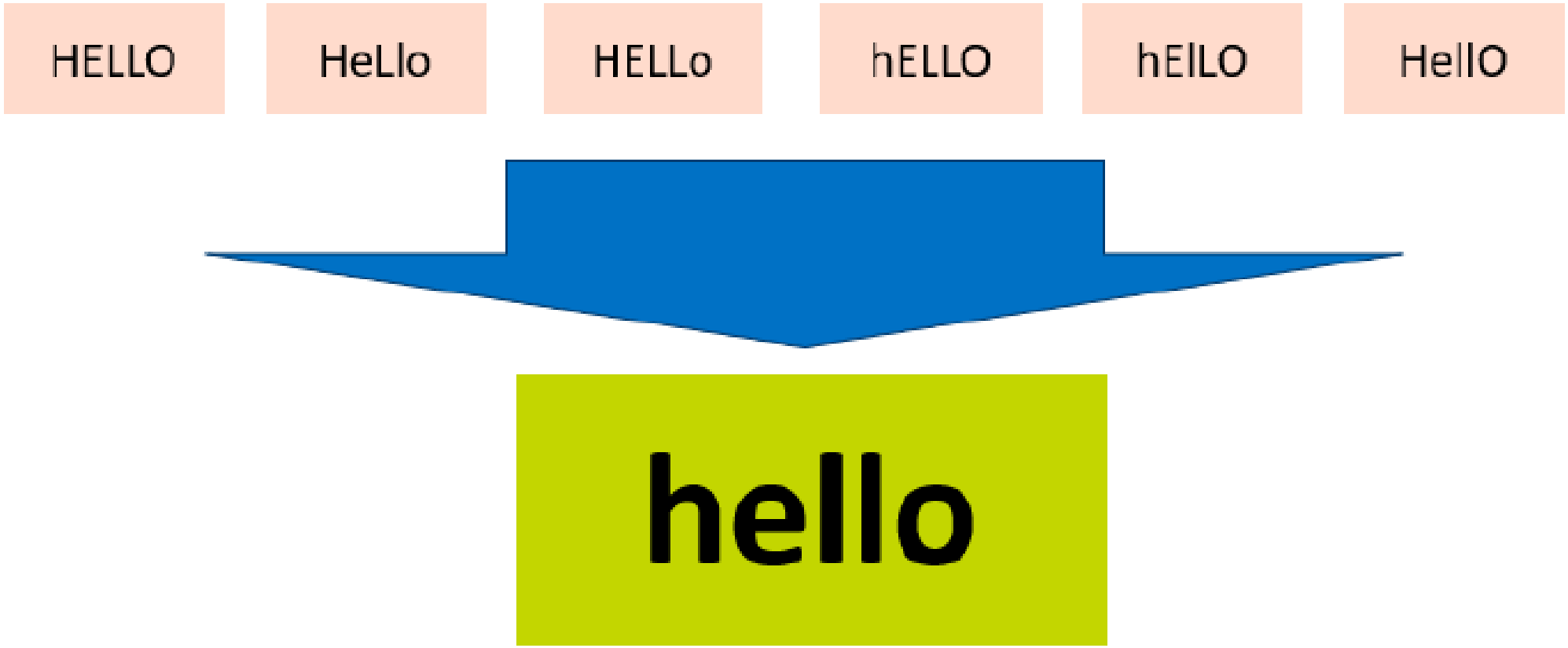


These words occur the most in any given corpus but talk very little or nothing about the context or the meaning of it. Hence, to make it easier for the computer to focus on meaningful terms, these words are removed.

Along with these words, a lot of times our corpus might have special characters and/or numbers. Now it depends on the type of corpus that we are working on whether we should keep them in it or not. For example, if you are working on a document containing email IDs, then you might not want to remove the special characters and numbers whereas in some other textual data if these characters do not make sense, then you can remove them along with the stopwords.

4. Converting text to a common case

After the stopwords removal, we convert the whole text into a similar case, preferably lower case. This ensures that the case-sensitivity of the machine does not consider same words as different just because of different cases.



5. Stemming

In this step, the remaining words are reduced to their root words. In other words, stemming is the process in which the affixes of words are removed and the words are converted to their base form.

Note that in stemming, the stemmed words (words which are we get after removing the affixes) might not be meaningful. Here in this example as you can see: healed, healing and healer all were reduced to heal but studies was reduced to studi after the affix removal which is not a meaningful word.

Word	Affixes	Stem
healed	-ed	heal
healing	-ing	heal
healer	-er	heal
studies	-es	studi
studying	-ing	study

Stemming does not take into account if the stemmed word is meaningful or not. It just removes the affixes hence it is faster.

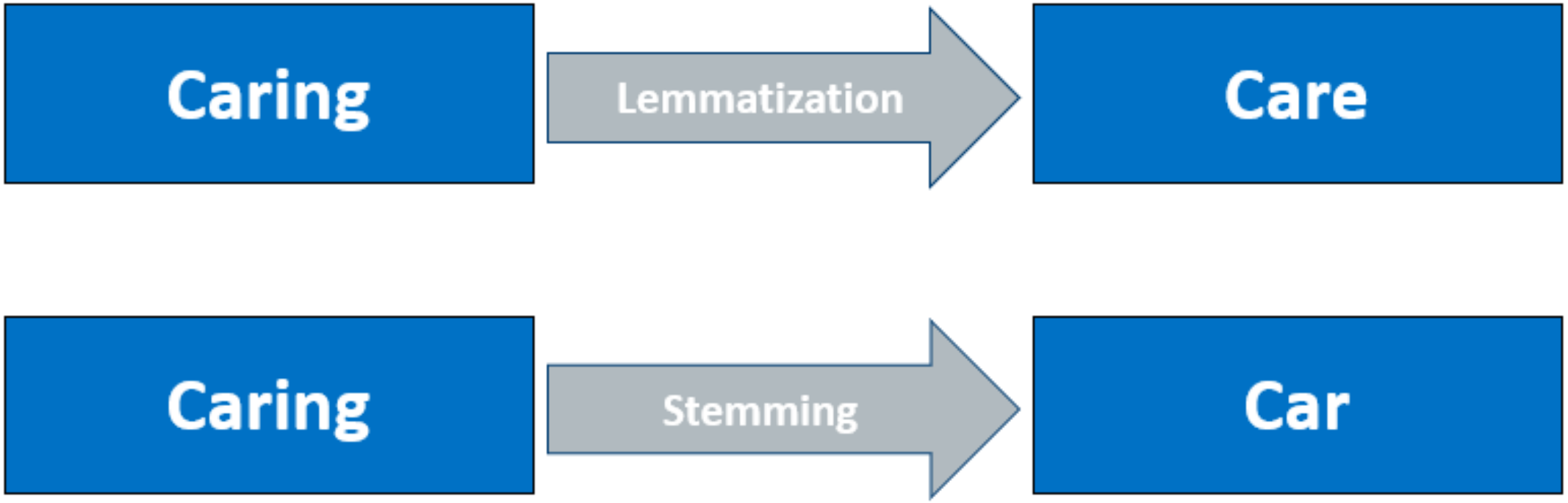
6. Lemmatization

Stemming and lemmatization both are alternative processes to each other as the role of both the processes is same – removal of affixes. But the difference between both of them is that in lemmatization, the word we get after affix removal (also known as lemma) is a meaningful one**. Lemmatization makes sure that lemma is a word with meaning and hence it takes a longer time to execute than stemming.**

As you can see in the same example, the output for studies after affix removal has become study instead of studi.

Word	Affixes	lemma
healed	-ed	heal
healing	-ing	heal
healer	-er	heal
studies	-es	**study**
studying	-ing	study

Difference between stemming and lemmatization can be summarized by this example:



With this we have normalised our text to tokens which are the simplest form of words present in the corpus.

Now it is time to convert the tokens into numbers. For this, we would use the Bag of Words algorithm

Bag of Words

Bag of Words is a Natural Language Processing model which helps in extracting features out of the text which can be helpful in machine learning algorithms. In bag of words, we get the occurrences of each word and construct the vocabulary for the corpus.

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

This image gives us a brief overview about how bag of words works. Let us assume that the text on the left in this image is the normalised corpus which we have got after going through all the steps of text processing. Now, as we put this text into the bag of words algorithm, the algorithm returns to us the unique words out of the corpus and their occurrences in it. As you can see at the right, it shows us a list of words appearing in the corpus and the numbers corresponding to it shows how many times the word has occurred in the text body. Thus, we can say that the bag of words gives us two things:

1. A vocabulary of words for the corpus
2. The frequency of these words (number of times it has occurred in the whole corpus).

Here calling this algorithm “bag” of words symbolises that the sequence of sentences or tokens does not matter in this case as all we need are the unique words and their frequency in it.

Here is the step-by-step approach to implement bag of words algorithm:
1. Text Normalisation: Collect data and pre-process it
2. Create Dictionary: Make a list of all the unique words occurring in the corpus. (Vocabulary)
3. Create document vectors: For each document in the corpus, find out how many times the word from the unique list of words has occurred.
4. Create document vectors for all the documents.

Let us go through all the steps with an example:
Step 1: Collecting data and pre-processing it.
Document 1: *Aman and Anil are stressed*
Document 2: *Aman went to a therapist*
Document 3: *Anil went to download a health chatbot*
Here are three documents having one sentence each. After text normalisation, the text becomes:
Document 1: [aman, and, anil, are, stressed]
Document 2: [aman, went, to, a, therapist]
Document 3: [anil, went, to, download, a, health, chatbot]
Note that no tokens have been removed in the stopwords removal step. It is because we have very little data and since the frequency of all the words is almost the same, no word can be said to have lesser value than the other.

Step 2: Create Dictionary

Go through all the steps and create a dictionary i.e., list down all the words which occur in all three documents:

Dictionary:

aman	and	anil	are	stressed	went
download	health	chatbot	therapist	a	to

Note that even though some words are repeated in different documents, they are all written just once as while creating the dictionary, we create the list of unique words.

Step 3: Create document vector

In this step, the vocabulary is written in the top row. Now, for each word in the document, if it matches with the vocabulary, put a 1 under it. If the same word appears again, increment the previous value by 1. And if the word does not occur in that document, put a 0 under it.

Since in the first document, we have words: aman, and, anil, are, stressed. So, all these words get a value of 1 and rest of the words get a 0 value.

aman	and	anil	are	stressed	went	to	a	therapist	download	health	chatbot
1	1	1	1	1	0	0	0	0	0	0	0

Step 4: Repeat for all documents

Same exercise has to be done for all the documents. Hence, the table becomes:

aman	and	anil	are	stressed	went	to	a	therapist	download	health	chatbot
1	1	1	1	1	0	0	0	0	0	0	0
1	0	0	0	0	1	1	1	1	0	0	0
0	0	1	0	0	1	1	1	0	1	1	1

In this table, the header row contains the vocabulary of the corpus and three rows correspond to three different documents. Take a look at this table and analyse the positioning of 0s and 1s in it.

Finally, this gives us the **document vector table** for our corpus. But the tokens have still not converted to numbers. This leads us to the final steps of our algorithm: TFIDF(Term Frequency & Inverse Document Frequency)