Course Name:Business Intelligence and Analytics
Professor Name:Prof. Saji.K.Mathew
Department Name:Department of Management Studies
Institute Name:Indian Institute of Technology Madras
Week:12
Lecture:44

INTRODUCTION TEXT MINING | BI&A | Prof. Saji K Mathew

Hello and welcome back. So, this session is on text mining. Text mining. There is a major difference between what we are used to so far and what text mining is. And that difference is in the type of data that you deal with. So far in all the techniques, be it descriptive or explanatory or predictive techniques that we learnt, we used numeric data.

Even if the data were categorical, we somehow made it numeric right, we numerically coded that data. So, because algorithms work on numeric data and unless data is in numeric format or in numeric type, the general algorithms that we learnt cannot function. So, the difficulty is that not all data are numeric, not all data are numeric. What are the data that you can think of which are not numeric? Feedback.

Feedback, ok. What feedback? Faculty feedback? Customer feedback. Customer feedback that is in text format. So, there are data type, but they are also data. We generally tend to think that data means numbers, that is something that you have to take out of your mind.

Data does not necessarily mean numbers. Is a experience data, customer experience data? An experience can be captured and documented in so many different ways. It could be videos, it could be textual description, it could be phenomenological account or case studies about experiences. It is all data.

Only thing is the data is not in numeric format. So, as you said, you go to customer feedback in Flipkart or Amazon. So, people talk a lot about their buying experience, about the products they bought, about the vendors and about the delivery experience and so on. So, there is a lot of information there, a lot of useful information there, but unfortunately not in numeric format. So, see one of the problems since you reminded about it, one of the problems my PhD student addressed was, how do we use feedback, customer feedback in e-commerce sites.

Use that feedback to discover attributes, customer defined attributes of products and services. So, generally any product has a list of attributes that the manufacturer specifies. For example, you take a water bottle, if you have to buy it from Amazon, they will talk about the color as an attribute, the height, the cap, the material, these are all attributes of the product. These are I would say the engineering attributes, the physical attributes of the product, but customers may also be interested in what others are talking about this product. So, for example, how easy it is to open the cap, if it takes a lot of time to open, it is not very easy to use.

So, there is some hidden attribute there which the customers are trying to define, this is good product, but I found it difficult to open. And if, say if 1000 people people gave right, wrote feedback and say many of them are repeatedly talking about some experience in using the product, there is some hidden attribute which the manufacturer has not specified, but customers are trying to say. So, you know the selfie as a feature in phones, arguably evolved through customers feedback or customers experience. Selfie is a feature that customers want or people want.

So, then if you look at the specs of a phone , now the selfie itself has become a feature or a feature in a detail. So, if you delve into customer feedback, you may find a lot of attributes that they define in textual format, but how do you measure it, how do you actually put a value to it. Let me give another side of e-commerce, you go to online for buying products, because you have something called convenience there, right, you can sit at home and browse through different look at different items, browse. And also unlike the brick and mortar stores, where the sales guy will tell you how good the product is, how bad the product is, you can read what others have written about the product. For example, if you are buying a phone , you know we found in our research that thousands of reviews not just one or two, which can become a book in itself.

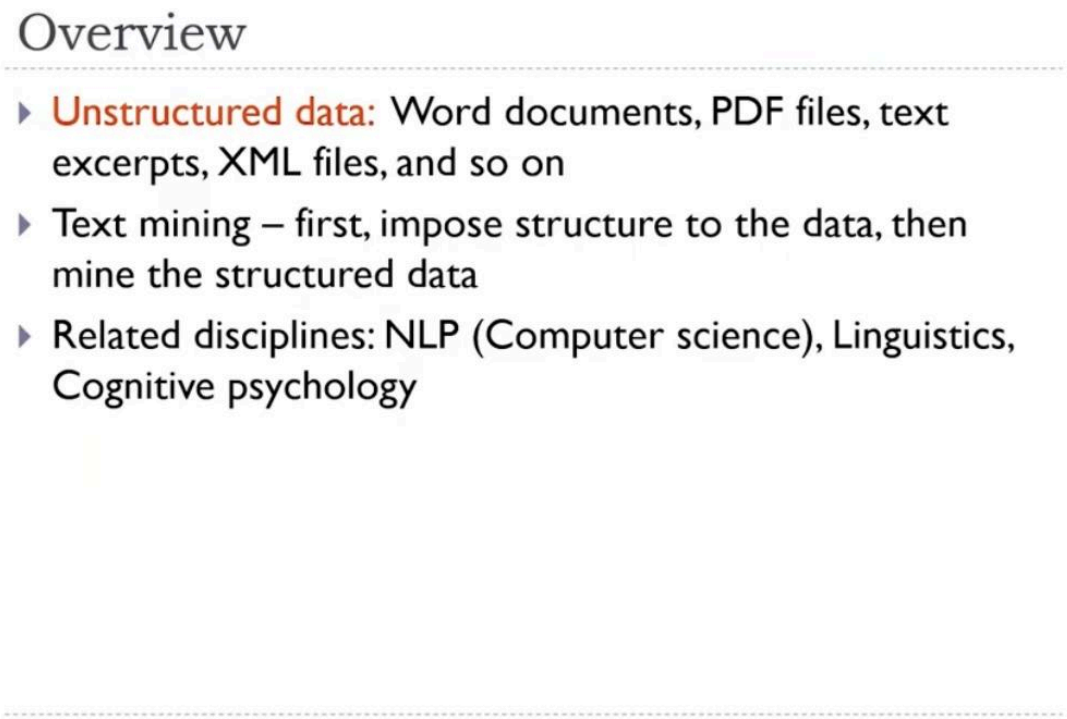
Customer reviews for popular products are so much of, in data in textual form that reading it is very inefficient. You read one feedback that is positive, then somebody else write which is totally opposite of what is first written, then you read one, you read the other, and when you read 10, 15, 20, 50, 100 you are lost, because it is voluminous data. And therefore, there was, we found there is a scope or there is a value in understanding this text and see if the text data can be converted into certain attributes. And there is something known as point of, sorry, POS tagging, not point of sales, POS, parts of text which tagging in text mining, which was used to tag, you know to use nouns as attributes and see what qualifies the nouns as customer comments, you know whether it is positive or negative. Based on that we developed a scale and we actually qualified these user defined attributes.

For example, something like the voice or a feature like selfie, how customers scale it or how customers experience it as scale. We did that using text mining. For example, selfie, in a scale of 0 to 5, sorry 1 to 5 or voice quality, 1 to 5. So, there are many attributes we derived from the data set, scaled it using the sentiment scores of the words used to talk about it. And gave this feedback to customers, do you see if you go to Amazon you get only the star rating, right whether the product is, you know a 4 star or a 3 star or, you know between them etc. But that is an overall feedback. When you buy a product, more than overall rating you are interested in certain features of that product.

So, the feature based measures to qualify the product is something that is possible, if you actually look into the details of the text and extract this kind of information from the text. And we did that, we have filed a patent also. So this is, these are the sort of opportunities that you get if you get into the details of the text, meaning text is not numeric data, but a lot of data is in textual format. If you neglect it you lose opportunity, perhaps to understand your customers, perhaps to understand your employees or perhaps to understand what people are talking about your products or services. So, therefore, you can see increasingly organizations are employing text mining techniques to develop better sense or understanding about how their products or services are being perceived by people.

So, when you look at the text. So, text mining that way has lot of potential for application in business and management. So as a whole, what is text mining? Text mining uses text data, textual data, data is in textual form, it could be a book, it could be customer feedback, it could be user manuals, it could be any form of text, but you want to make some sense of from that text. For example, what are people talking about a movie, what are the words that people use to describe a director or your favorite cine star, are they positive words or are they negative words, are they more positive words than negative words. A lot of ways at which you can actually look at the textual data, ask questions as you ask databases and if there are algorithms that can do this, you are actually, you are actually able to use that data and fortunately yes, today there are text mining techniques.

So in text mining, as I said the main challenge is that the data is unstructured, data is not structured, data is not numeric and data is not structured. Structured meaning you know you learn databases, relational databases, a table format you have a key and then you have other attributes. So data is organized into tuples and well defined structure for



the data, but text data is what, it is free flowing data. But in text mining the first effort in text mining is to convert this free flowing text into some structural format, impose some structure, give some structure to the text data and then use or then apply algorithms which work with this structured data. So, that could be your clustering, that could be any other analytical technique that can be used once the data is given a structure. So, text mining as we discuss it in this session has two steps that way, structuring the text and then analyzing the text using certain algorithms.

For example if you take your text book, how many chapters your text book has? Hopefully nobody will answer, right who cares, but there is something called a table of content, right a table of or, TOC. TOC gives you an overview of the book. For example, the book is divided into chapters, chapters are divided into sections, sections may have sub sections and so on. So there is some structure there, but that structure has to be formalized more for working with the overall text and then analyze it in a useful way. So, we are going to see how to structure the unstructured data and then work on it.

Now, foundations. So before going into any field, we should have some foundations of what is the basis for doing all this. So there are two sciences basically which informs text

mining. Text mining is more like a craft I would say and it only gives very basic insights, sometimes they can be very useful. Now, there are very advanced techniques in text mining or text analysis which I will give you a direction but we are not discussing that, we will be doing certain very fundamental steps or analysis in this session. So, related disciplines are number one, natural language processing NLP in computer science, that is pure algorithmic technique to analyze, not programming language, but natural language.

Natural language means language which humans speak. So natural language processing is about analyzing human languages. There is programming language like your C, Java etc. That is one aspect of CS, but CS also has developed a sub discipline known as NLP. Linguistics is what? Linguistics is the science of language. It is a philosophy of language. What is language at the end of the day? It is a medium for expression of our thoughts or for ideas.

That is a very advanced description of language. Generally the answer is language is for communication. But, you know humanity developed language at some point in time. Language is not very prevalent. Animals do not have the scripted languages, but they also communicate.

Even without language, suppose we do not have any language, I say the class has to communicate, but do not use any language, you will still communicate whatever I say. So, we have gestural language, we have sounds and so on. But when language developed, scholars looked at the philosophy of language and that is linguistics. Those who are very serious about learning text mining should actually credit courses on these two fundamentally. Cognitive psychology informs both, but NLP is a fundamental requirement for text mining.

And there are Coursera and NPTEL courses on this. I strongly recommend that you do that course, to advance your understanding of text mining. And linguistics, there is a really good course offered from our HSS faculty, I really benefited from that course. Since I talked about linguistics and NLP as foundations for text mining, there is a philosophy of language which is known as linguistics which developed I would say, in the last 70, 80 years. Linguistics is a fairly new discipline.

So, let me ask this question to the class. How do you learn language? How did you learn language? Two possibilities, one is rationalist other is empiricist meaning. One is you thought about the grammar, structure etc and you learned it through logical thinking, rational process. You developed your understanding about the language through thinking, that is one argument. The other is, you learned language from the environment.

For example, you observed how others speak. You learned, you developed your vocabulary, your grammar etc etc by observing what, how others speak. That is the empiricist argument or innateness versus behaviorism in the language of the linguistics expert. What do you think is the right answer? Is language learning rationalist or empiricist? What is it predominantly? What is the dominant approach humans follow in language learning? Empiricist, that means through observation you have learned a lot and then you reproduce that. Do not Google.

Linguistics foundations

- Philosophy of language
  - Mental representation of language, its expression in written form
  - Generative capacity of the mind

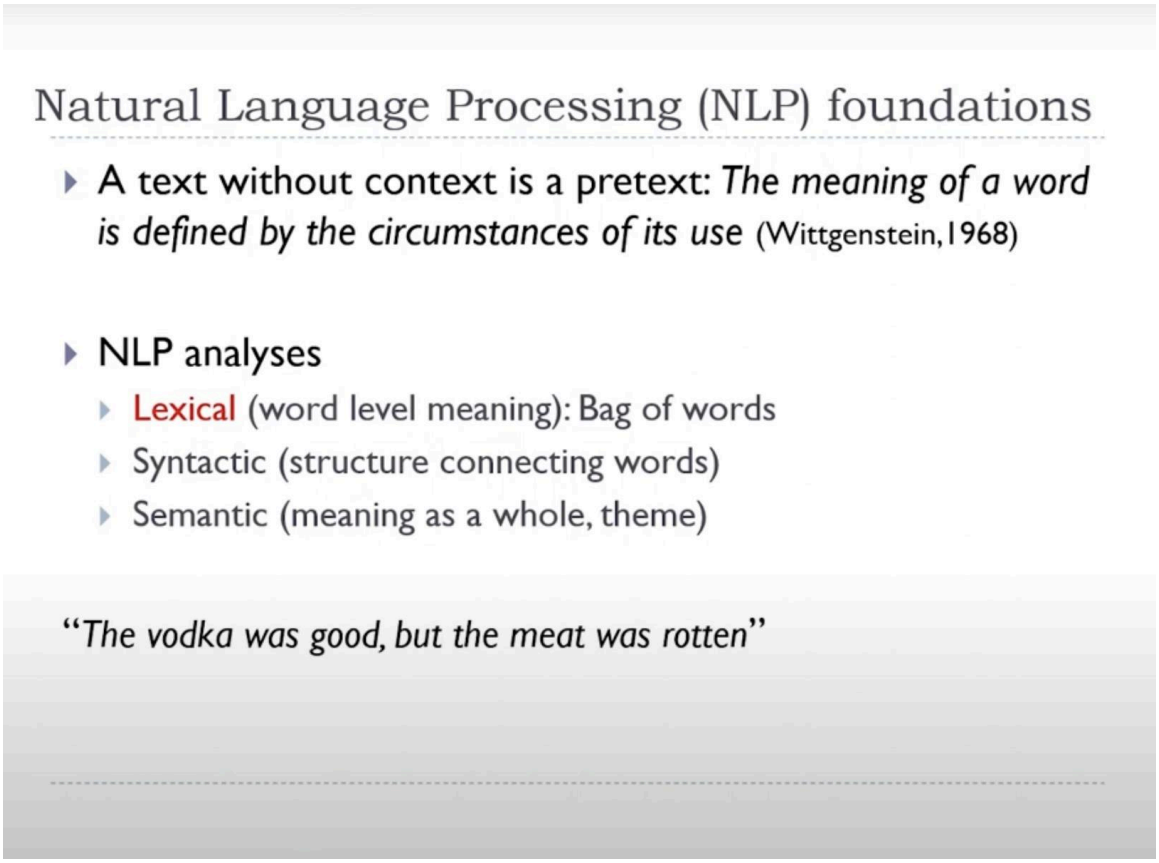
1. Rationalist approach (innateness)
   - Language formed in mind not by the senses but is fixed in advance, presumably by genetic inheritance (Chomsky, 1986)
2. Empiricist approach (behaviourism)
   - Language learning dominated by sensory inputs

Statistical NLP belongs to the second school

Now, let me ask you a question. How many sentences you can speak? How many sentences you can speak? Many. That is a good answer. Many means almost infinite. Does anyone restrict you? You are sitting in this class, I ask you to talk about something. You can construct, you construct, construct means you think, you think and create, articulate.

There is actually language learning and articulation is a rational process. If you are a writer for example, see the sort of imaginative or creative process involved in writing. So speaking, what do you speak in a given context? You think, right you basically have learned certain structure, vocabulary that aspect is there, but from that language is created. So, it is a rational process. So, that is the theory advanced by Chomsky

I suggest that you listen to Chomsky's lectures in UCLA and he is known as the father of linguistics. So he opposed the idea, I think there was some paper I think in the 1960s which suggested that language learning is empirical and he opposed that idea and built the theory of linguistics. So, but unfortunately this is about language learning among us, but when we get into this text mining processes, we are more going by the empirical approach where language, statistical NLP would actually learn language through empirical or observational data, not based on any logic that we impose on that. Yeah, rational approach became the dominant approach, post Chomsky.



So these are basic questions, how do you, how are you able to speak so much, so much if you were learning all by heart, you cannot actually do that, right. So they say that, those are the words I have used, you can go and read more about it or listen to it more, generative capacity of the mind, it is said to be infinite in sense. And today they talk about universal language and so on, because humanity as a whole has certain hidden structures for language learning and articulation which is common across languages and so on. The other aspect is the NLP. So NLP is a computer science discipline for working with natural language and there in NLP, there are three types of analysis, one is the lexical analysis which is word level, which is word level. For example, lexical analysis like dictionary.

So, this is also known as bag of words. So if you take a Shakespeare's Macbeth, suppose it is available in digital format, Macbeth is a bag of words, it is a collection of say 10 million words. That is how a text mining expert will look at it. So, only words matter but how words together create a meaning or a semantic, semantics to it is not analyzed in lexical analysis or bag of words approach. But bag of words is the dominant approach of, the starting analysis for text. So, it functions at the word level. Next is the structural level or syntactic level, where you also use grammar and third is the semantic analysis of text which is, what is most desirable and which is most difficult and today I will tell you a little more about it. Semantic is, suppose I am talking about text mining here for one hour or one and a half hours. At the end of the day, what is the meaning or you listen to someone's speech, the director of IIT Madras calls for a meeting. We do not know what he is going to talk about, but you listen to him for an hour and then what is the meaning, what is he trying to say, how do you actually read between the lines.

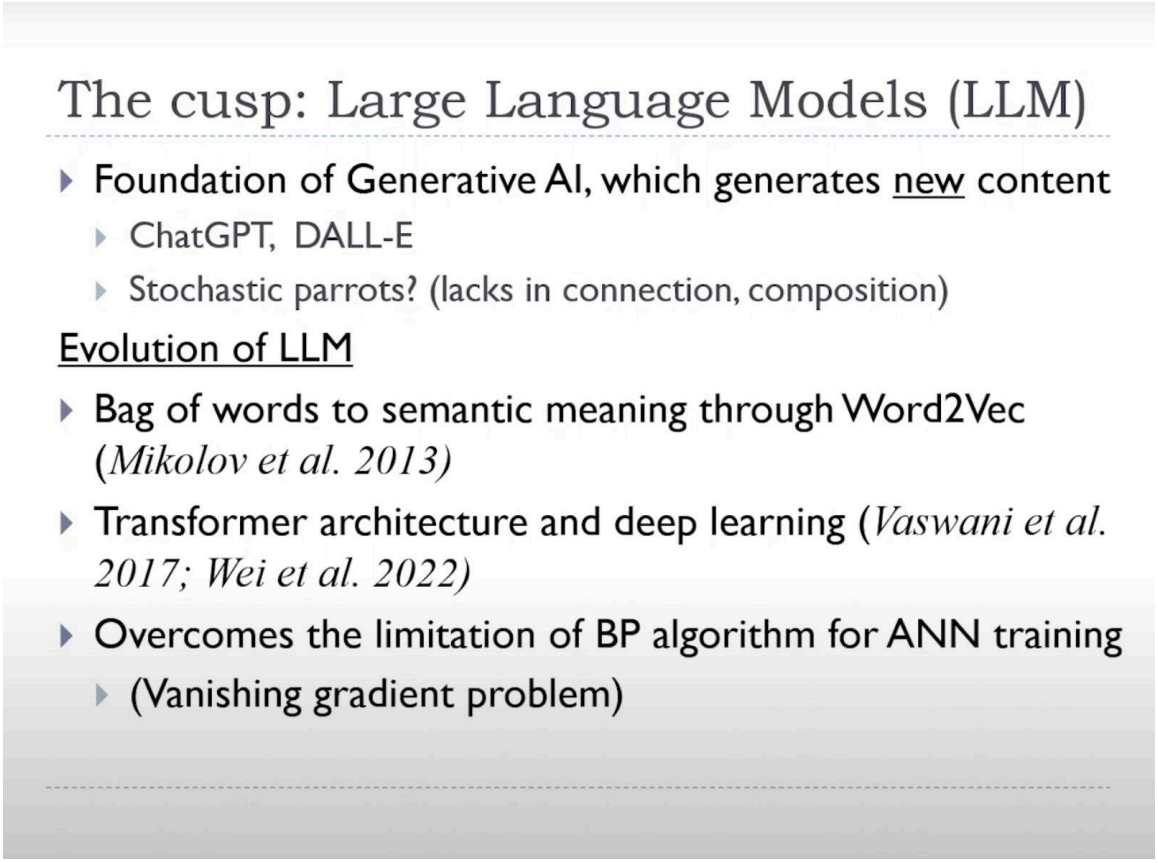
So, you create understanding about a collection of text. So, meaning as a whole, what are the themes in what someone spoke. This is semantic analysis. Semantic analysis is more complex, today being done, but lexical analysis is simple, that is what we are going to try unfortunately in this session, because this session is foundational. So, we will restrict ourselves to lexical analysis of, bag of words analysis of text. Now, another challenge with that kind of lexical analysis which you should be aware of is that, lot of meaning is contained in the context.

Meaning is in the context. See what Wittgensteinh, Austrian philosopher said. The meaning of a word is identified by the circumstances of its use or as we generally say, a text without context is a pretext. So, I say some strong words standing here, you are not really doing good, get out, are you running out of the class? I said get out and nobody ran out of the class because there is a context. I said I am trying to test you out or I am saying there is a context and then in that context, I said get out. Are you taking it seriously? But the word, get out is there or the usage get out is there. So therefore, if you go by bag of word analysis and see the term get out, you may reach some other inference, if you do not know the context. So, the meaning is in the context. You can use the same word with positive meaning and negative meaning. You know that a lot of gali that we actually use, abusive language is not abusive when friends speak. It is rather very soothening . So words can be spoken in a context, are spoken in context.

So you take the meaning based on the context, you do not interpret it without context. Otherwise we will be all the time, you know facing a court if we are quoted out of context. Have have heard this? The vodka was good , but the meat was rotten. One application of text mining is in translation and in translation, you know that Google

translate and so on have become very advanced with training. So, there are two approaches, they say one is translate in context, other is translate word by word. So, think of Bible translation, this is in the context of Bible translation. So you can translate Bible word by word. You can also read a paragraph, understand the context and then translate it from say, Greek or Hebrew to English or vernacular language.

So, somebody attempted to translate. So some people argue it is God's word and therefore, the translator should not give meaning to it, simply translate the words as it is. So somebody did that. So to obey God and see the translation was this. The Bible was translated from the original language to Russian language. It is a, it is a Russian written in English of course, and it is a tragedy. So Jesus was about to be crucified and he was with the disciples and the disciples were sleeping and he found it very stressful. So he is going to be crucified next day, his friends are sleeping.



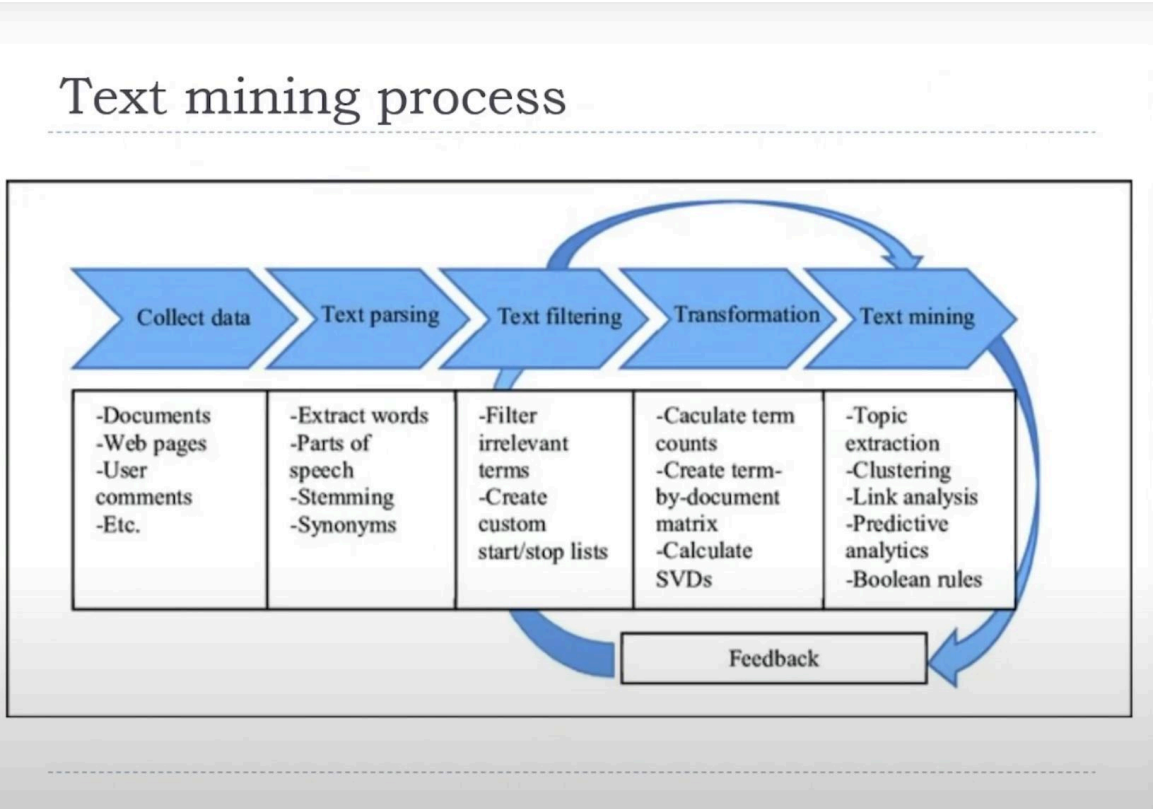
So, he goes to them and then one of them, Peter says, does anyone know what, that is what is translated here? The spirit is willing, but the body is weak, translation. The spirit is willing master, but the body is weak, you know. So translation such tragedies can happen and this is something you should be aware of when you do bag of words

analysis, we are not looking at context. So if you go by words, you may actually do all this adventures. So that is why, so this slide is about the advance in NLP and text mining, that is the large language models.

Increasingly we hear about and some of you are using, of course chatGPT, chatDOC and Dall- E for painting, right. You you can paint, you give a textual description of what needs to be painted and you want this to be painted in the renaissance art, the tool actually gives you the painting and give. And so this mode, this advanced method of using text in large volumes in more semantic sense or in a more meaningful sense than lexical sense, is today having a lot of traction and that is known as large language models, large language models. So, the so called generative AI, in generative AI, it is not simple prediction. But it is also generative or it is generates something new, it paints a painting or it provides a summary or it writes something new in a format that you want it, it writes research papers. So there is a generative aspect to it. There is actually a meaning or a more, what you call semantic analysis of data, not just the lexical analysis of text that is involved and that is a turning point in the story line of text mining. We are starting, standing at that point. It is called the curse. I call it the curse for the transition.

So, if you look at text mining from the era of lexical analysis to large language models. So, some of the milestones I have given in this slide. This is based on the editorial of Information System Research, which recently gave an overview of this changing trend. So bag of words which we are going to use to semantic meaning through word to vector. So this was published in 2013, where words organized in the vector form could be analyzed semantically. And then the transformer architecture, which some of you must have read about, we studied feed forward networks ANN. We used a feed forward network.

So we, so there is no feedback. Transformer architecture is a more advanced architecture for neural networks. It is widely used in large language models. So transformer architecture was proposed by Vaswani and others in Google in 2017. Attention is all that matters, their paper, you know there is a famous paper. So it is in that paper, they actually proposed the transformer architecture and coupled with deep learning. So we talked about back propagation algorithms, but deep learning is an advance in ANN training and which overcomes a limitation of back propagation algorithm known as vanishing gradient problem. Vanishing gradient problem in summary is that when you try to train a neural network using a back propagation algorithm it does not become effective often times, because the gradient or the error which is used to calculate the weight vectors, when it is propagated from the output layer to the hidden layer, it becomes smaller and smaller.



So, the training is not effective, if you use the back propagation algorithms for ANN. So the advance is in deep learning techniques where this problem of vanishing gradient is addressed and you use a new architecture which is known as transformer architecture. Thanks to these two developments in neural networks, large language models which can be used to extract meaning from text, not just word to word correlation or descriptive understanding of the text, but more semantic or contextual meaning of the text, can be extracted by large scale training using textual data. That is the current state of text mining. So that is the current state of text mining. Unfortunately we are not going to LLM, we are limiting our analysis to a foundational level which is the lexical analysis or bag of words analysis.