Artificial Intelligence for Economics

Prof. Adway Mitra

Artificial Intelligence

Indian Institute of Technology Kharagpur

Week – 08

Lecture - 39

Lecture 39 : Text Mining and NLP for Economics

Hello everyone. Welcome to this course on Artificial Intelligence for Economics. I am Adway Mitra, an Assistant Professor at Indian Institute of Technology, Kharagpur. So, like here we are at the last few lectures of this course. Today is the 39th lecture and topic of today's lecture is Text Mining and Natural Language Processing for Economics. So, like in the last lecture we discussed computer vision for economics.

So, computer vision and natural language processing these two are like you can say the core application areas of machine learning and artificial intelligence in which research has been going on for say 15, 20 years. Now, like in the last lecture we already discussed how computer vision can be useful in economics. Today we are going talking about natural language processing. So, we will first give you a very brief introduction to text mining and natural language processing.

We will also tell you the statistical and neural methods for natural language processing and after that we will go to the case studies where we will look at different research papers where natural language processing approaches have been used for various tasks in economics. So, like natural language processing it starts with text mining because I mean it is the data is usually textual in nature. It can also be speech, but for now we can focus on textual written data. So, text mining it starts with a large corpus of documents where each document we can consider it as a sequence of word tokens. There is every single individual word is called as a word token.

Now, from this corpus we can construct a vocabulary that is the list of words that is the unique word tokens. So, difference between words and word tokens is that a word is unique, but a word token I mean like that is each of these single words that are written on this document each of them is a word token and every word token maps to a particular word in the vocabulary. For example, this vocabulary is a word which has been used multiple times in this in this page. So, like here it has been used here it has been used and so on. So, these locations these are the word tokens and the vocabulary is a word which is

you being used in the word for both of these word tokens.

So, the vocabulary is the collection or the set of the unique words that are used in the corpus. Now, when we are considering the vocabulary, we tend to keep aside the stop words such as is, when, how, etcetera. These are also useful for certain tasks. For example, if we are trying to construct a new sentence, we say something like like natural language translation we are doing. In that case of course, the news we are constructing a new sentence or a new document which will of course, be meaningless unless we use these kinds of words, but for other kinds of semantic tasks of text mining or the or some of the more stand more classical tasks of natural language processing these are sometimes irrelevant.

So, these words they can be they are sometimes marked as stop words and they are not necessarily parts of the vocabulary. So, the main challenge in the text mining related tasks is the we need a mathematical representation of the semantic information and grammatic rules. Semantic information as in the meanings of the different words there are different tasks or typical tasks which we need in natural language processing or text mining we will discuss those tasks later. But many of those tasks it is like it is not enough to just have a collection of these words or the or it is not enough to represent a document as just a sequence of word tokens. Sometimes the meaning of the those words and which words have similar meanings or maybe opposite meanings or and so on or various other kinds of relations they have that also sometimes becomes important.

And apart from that like in many cases or many tasks these grammatical rules also somehow need to be taken into account. So, natural language processing is of course, like it if we consider any natural language let us say consisting of. when we consider any natural language such as like English, Bengali, Hindi and so on like it is based on three main premises. First of all the vocabulary, next the grammar rules and then the semantics. Vocabulary is of course the set of words, grammar rules is basically like some rules about which words can precede or succeed which other words to make a sentence meaningful and finally the semantics.

that is what are the meanings of the different words and so on. So, when we can like if we want to do natural language processing that is document is written in some natural language and we want to make sense of it, in that case we need to use like that is we need contributions on both of these or all three of these premises. So, without these we cannot do natural language processing. Now, some so what are some of the typical tasks of this text data mining or natural like. So, these are the you can say the if the easier or the lighter tasks of natural language processing which may not require detailed understanding of the meanings and so on.

So, like we so these are older tasks which have been typically known as the text data mining. So, the first of all there is part of speech tagging. So, what is part of speech tagging where we classify each word according to its part of speech. So, like here you can see a sentence I like to read books here every single word like I it is a pronoun like is a verb this. this read is again a verb, books is nouns and so on.

So, here given a sentence we have to tag every single word token as one of the parts of speech that is the post tagging. Secondly, we have word sense disambiguation. So, some words can have multiple meanings. Now, we have to like given a sentence we have to figure out which particular meaning is being intended in that for that particular word token. So, like for example, the word two examples are given here pine and cone both of them have like multiple meanings.

Now, if you come across the word pine in a sentence like is the is it this first meaning or is it this second meaning being used. The answer of course, lies in the context that is the surrounding word tokens in the sentence. they will give you a clue about which of these meaning these two senses is being used. So, that is basically the task of word sense disambiguation. Next there is named entity recognition.

So, like as you can see like here like a textual passage is given and now the aim is to find out the various specific entities or proper nouns of text documents. So, like an example of it you can see here. So, now what are some of the typical tasks in text data mining like we have already discussed a few. Let us continue for the few more task which may which are becoming progressively more complex. Next one is related to sentiment analysis.

So, now here we are interested in tagging every word token or sentence or passage to indicate whether it like whether it represents a positive sentiment or a negative sentiment. I mean sentiments are broadly of two polarities positive and negative it can be neutral also. So, it can be like it can be neutral also even among positive and negative there can be multiple degrees like strongly positive, strongly negative or weakly positive, weakly negative various polarities are possible for this sentiment analysis. So, like the basic task is that like given like. So, here you can see some examples.

So, this is this sentence it overall it represents a positive sentiment. Now, if you look at everything any individual word token in this sentence you may not be able to understand or like any like sentiment associated with it, but like there are some specific word tokens which may constitute a sentiment. Now, this is not a very easy task because often the like that is the languages can be quite the structures of sentences can be quite complex and it is possible that So, like some word tokens or some words viewed in isolation may seem to indicate one kind of sentiment, but put in the context of that sentiment of that sentence it may indicate a very different kind of sentiment. So, it is like sentiment analysis is not a

necessarily a very easy task and especially if some kind of sarcasm is involved it is very difficult to understand or represent such sarcasm. through mathematical models of text data.

So, that is why this sentiment analysis is a slightly more complex task than the things which we are earlier talking about such as post tagging or word sensitive simulation and so on. Another quite complex task is that of topic modeling. So, like let us say here is a we have a collection of documents and we know that every document is a like is like involves a few topics. Now, we like we want given this corpus of a large number of documents we want to find like that is we want to. So, that is make a list of the topics that are present in the corpus and furthermore we also want to assign like every single word token in each document to one of the topics.

So, that is the main aim. So, these topics can be something like say science, sports, economics, politics and so on and so forth. Now, these but this is an unsupervised problem we will not specify these topics beforehand like it is not a classification problem, but somehow we have to like it is we can say it is more like a clustering problem in which different word tokens from different documents will have to be grouped together like indicating that like they belong to the same topic. The task is further made difficult because the same word can occur in different context as parts of very different topics. And furthermore here we like it is quite complex to represent a topic, but in topic modeling every topic is represented as a probability distribution over the vocabulary. That is like say there is like if we consider one topic like say physics, then certain words in that vocabulary will be will have a higher probability of occurrence when I am considering that topic like say atom molecule or force or energy acceleration.

These words are quite likely under the topic of physics, but words like let us say energy a tree or socks or shoes these are less likely to be present on in a topic of physics. So, like a topic is can be represented as a probability distribution over the vocabulary. So, the task of topic modeling is basically this given a set of documents find out what are the major topics present in it and also assign every single word to one of those topics. Next, there is this task of segmentation or summarization where you have like you have a big text document which you want to divide into coherent parts and represent each of those parts concisely. So, that is the first one is segmentation and the next one is summarization.

So, now so we have discussed these problems in natural language processing, but how to solve it. So, since like about 2012 or so we have like ever since deep learning became very powerful most of the problems are solved by deep learning which we will discuss earlier, but prior to that like most of these problems used to be solved using statistical methods. So, like whenever we are talking about natural language processing like this language model is something of very great importance that is like the task of a language

model is given a sequence of word tokens you have to predict the next one. Somewhat like the this autofill or autocomplete feature which we have in Gmail, WhatsApp and so on. So, like you see like you have been provided with the word tokens wt minus 1 wt minus 2 and so on and so forth that is all the previous word token and your task is to predict the next word token or rather provide a probability distribution over the what word can be present as the next token.

So, now like so this kind of so the I mean a language model is basically defined as how it defines this kind of a probability distribution. So, some of the simplest probability simplest language models that was the Ngram model where this kind of probability distribution it like we had a frequentist construction of it that is we simply took the pattern like that is this this subsequence and search for the subsequence wrote the corpus and whenever and that is we saw how many times this particular subsequence occurs in the corpus and whatever and which word it is followed by. So, that way we know I mean which word is likely to or most likely to follow this particular subsequence accordingly we build a probability distribution. So, that is called the frequentist construction. Apart from this there were there are all some of the tasks which we mentioned earlier they used to be solved by latent variable models.

So, they are like in many cases the quantity which we do not know they are represented by latent variables. I mean when we I mean they are latent random variables say for example, the post tags associated with every word token or the word senses of every word token or some word importance score maybe. So these are like these are the these are represented as the latent variables and we build some kind of a probabilistic models to create conditional distributions over these latent variables. So some of these are well known probabilistic models like such as hidden Markov models, So, like here if you see this thing. So, let we have a sentence John can see will.

So, each of these words are to be associated with a post tag. So, the post tag is considered as a random variable and like what we need is a probability distribution over each of these random variables that and accordingly we will So, that is for that is like the post tag of this word can we will create a probability distribution over it and the most likely value will probably be will be considered as the post tag for that for this particular word token. but how do we construct the that probability distribution so there are so each of the these probabilistic models they like they have some their own way of representing these conditional probability distributions this is not the time to get into all that so the similarly there are also latent Dirichlet allocation this is a well known topic model. So in this case like it is considered like considered that first of all there are some topics as I said every topic is a distribution over the vocabulary then for every document there is a distribution over the topics indicating which topics are prominently present in that document and then within the document for every word token there is a single topic. and

the word is like the is a realization of that particular topic.

So, this is a complex probabilistic model. Now the task is to find their values of these latent variables which may be the post tags or the topics. So, we use probabilistic inference to estimate their latent variables. So, now let us come to a few use cases. So, like the here using natural language processing to read plans.

So, here the task is to study a 78 resilience plans from 100 resilient city network. The basic idea is that you have a large number of policy documents, public policy documents from which they are trying to figure out some like important information which is to be used for policy planning. so planners need to read plans to learn and adapt current practice planners may struggle to find time to read and study lengthy planning documents especially in emerging areas such as climate change and urban resilience recently natural language processing has shown promise in processing big textual data We asked whether planners could use NLP techniques to more efficiently extract useful and reliable information from planning documents by analyzing 78 resilience plans from 100 resilient city networks. We found that the results generated from topic modeling which is an NLP technique coincided to a large extent to those from conventional content analysis approach. So, like we just talked about topic modeling.

So, this paper is like you can say is an application of topic modeling. So, like all the documents which they are dealing with they are like policy documents related to urban planning. Now, within that they are looking searching for what are the prominent topics that have been discussed in these urban planning documents. So, these are some of the topics which they have identified.

So, disaster management goals etcetera. These are topic names which are probably given by themselves, but what the topic model really did is identified some frequent some words which seem to be occurring together frequently such as risk management, hazard mitigation, disaster recovery etcetera. So, the topic model found that these few words tend to be occurring frequently together. So, this was identified as a topic which can be given as a name of disaster management. Similarly, these words like public health, well-being etcetera like these words were found to occur together. So, they are another topic and so on and so forth.

So, like so this is a paper in which topic modeling has been used to like on a corpus of policy document to find out like mostly what kinds of topics these policy documents are talking about. And now if you are interested in one particular topic let us say I am I want to know the energy policies of the different cities. In that case like all the word tokens have already been assigned one topic by the topic model. So, like I will see which word tokens have been assigned to the energy topic.

So, I will read only those words. So, I am saved from having to read the entire document. So, in a sense it helps me to like it provides me an useful index for the document like it will which saves me the time of having to read entire documents if I am interested in very specific topics. Another interesting application of natural language processing in economics, this one is epidemiology of inflation expectations and internet search. So, here the broad idea is that like this paper is focusing on the like inflation and that is how much inflation is going to be in the coming days and what are the people's expectations about these about such inflations. So, when people have some expectations, they may form that expectation either on the basis of their day to day expenses experiences or on the basis of like what they read in social media and news and so on and so forth.

So, like here there this paper what it does is it looks into Google trends as an indicator of what kinds of things people are looking at what people are searching and as a result of that they are how their expectations about the inflation is changing. So, this paper investigates how inflation expectations of individuals are formed in India. We investigate if news on inflation plays a role in formation of inflation expectations following the epidemiology based work the standard literature on this topic. considers news coverage by the print and audio visual media as the sources of formation of inflation expectations. Instead we consider the internet as a potential common source of information based on which agents form their expectations about the future inflation.

Based on data extracted from Google trends Our results indicate that during the period 2006 to 18, the internet has indeed been a common source of information based on which agents have formed their expectations about future inflation. And the internet search sentiment has some impact on the inflation expectations. Once again we are seeing some connection with sentiment mining. Additionally based on the inflation expectation series derived from Google trends data we find that there is a presence of information stickiness in the system since only a small fraction of the population update their inflation based this period. So, one more important task in economics or one aspect of economics which is dependent on natural language processing is recommender systems and how they make like utilize the reviews by the users.

So, we all use recommender various types of recommender systems like starting from Amazon to booking.com and so on and so forth. So, whenever we make a transaction sometimes we are asked to write reviews and now the those recommender systems they ask us to they push us to write the reviews. So, that they can utilize these reviews to do better to give better recommendations in the future. So, recommender for the recommendation engine of course, they receive feedback from a very large number of people.

So, it is impossible for them to read and understand each of these reviews. So, instead they need some AI based algorithms or NLP based algorithms for that. So, some of the things, so like whenever we are dealing with these reviews, so like an important question is what are the major aspects which the customers are interested in. I mean the customers are provided with some products or services, but they all customers may not be interested in every single aspect. I may have some my own, I may have be interested in certain aspects of a product I buy.

If the same product is bought by someone else, they may be interested in some other aspects and so on and so forth. So, like the recommender system system wants to know from the reviews that what are the aspects which different people like and with what is and for different products whether certain particular aspects are considered to be good or not. So, if you look at this thing this example here is a review I loved the food the best sweet and sour soup ever everything was delicious and favorable delivery is pretty good nice people overall rating is 4 out of 5. So, now the like here different aspects have been found like one is food, the second is the people that is the staff of the restaurant or whatever they are talking about and so on. So, for each of these they have identified a sentiment score also that is because people is associated with nice.

So, it has a sentiment score of 4 and like because this soup is described as the best ever. So, it is given as 5. for is for the food in general it is written I love the food. So, the sentiment score is 4 and so on. So, here like in like so, they have in this study they have considered few recommender systems Yelp which is of course, a food recommender system TripAdvisor is for travel, Amazon is for like we all know for retail and so on.

So, like here are some of the main aspects that have been identified based on these reviews. These are like typically when people are utilizing these services, these are found to be the main aspects which they are looking at. And then for individual products, let us say a particular restaurant in Yelp or a particular let us say hotel in case of TripAdvisor etcetera. So, like for a particular hotel maybe all of like it they will try to find out the sentiment of people on each of these different aspects and accordingly they can give some form some overall rating for the for the hotel. or the overall rating is also sometimes not necessary they need may need some very aspect based ratings.

So, let us say that like I am interested in those hotels which let us say have very good bathrooms. So, in that case like they will look for the like this aspect called bathroom and then they will find which hotels have very positive sentiment in the reviews associated with them and so those are the ones which will be recommended to me and so on and so forth. So, this one is like another one somewhat similar to the inflation work. So, here we are stock market prediction analysis by incorporating social and news opinion and

sentiment. So, like here like just like in the previous work they focused on Google trends for like inflation here the task is like here the focus is stock market price prediction and they are trying to see if the social media and news.

they play some importance in like in the predicting the stock market price. So, here they have done their experiments and it shows that yes indeed that is the case. So, if you look at this figure. So, this black line these are the actual time series of prices of a particular company stock prices.

And, now they try to make predictions of the stock price. So, when they are considering or they are when they do the stock price prediction. Now, stock price prediction is a pretty well known problem in economics and people have been using various kinds of AI methods. In fact, when we earlier discussed about this problem also when we talked about the sequential data or sequential models LSTM etcetera. So, like if we use that kind of a model. So, it turns out that the predictions follow this dashed blue line which is which clearly deviates from the actual data that is the black line, but it is found that if like we when we are making the predictions we take into account the these social opinions which are taken from these news and social media and along with sentiment analysis it is found that like our prediction significantly improves.

So, in that case we have the red line which is found to closely align with the actual data that is the black lines. Now, so these as I said these are all these approaches are mostly related to the statistical methods for the NLP, but in recent times we have like we use neural network based methods. So, one main one of the main features of neural network based methods for NLP is word embedding where each word may be represented as a vector as an input to the ML model. So, like this like so one possible way is of course, to go for the one hot representation of a V dimensional vector.

So, where V is the vocabulary size. So, like if I am talking about the fifteenth word in the vocabulary then in the it will be a binary vector where only the fifteenth element will be 1 and all other elements will be 0. So, that can be a very silly represent vector representation for the word, but of course, it will be very high dimensional if my vocabulary has many words then accordingly the dimension of these vectors will also be small, but that is not what we want. So, now we want to make the and furthermore this considers all the words as different. Now different words may have similar may be related in various aspects, but since all of them are going to be represented by this binary vector that relation between the words cannot be captured. So, the idea of these word embeddings is to create is to represent every word using a vector in such a way that the vectors of similar words they should be close to each other with respect to Euclidean distance.

So, like for example, what you can see here that is the words cat and kitten since they are similar words we can see that they are like their vectors are also quite close to each other with respect to Euclidean distance. but the word dog that is slightly different from each other from those two. So, they it is slightly further away the words houses that is quite different from all of these. So, it is still further away and so on and so forth. Now also like if there are like if two pairs of words have similar relations like though they should be the line join or the vector connecting them those should be parallel to each other.

Like for example, the relation between man and woman is may be similar to the relation between king the words king and queen. So, the vectors connecting these two pairs they are like they are like just parallel to each other and their lengths are also similar. So, he like hereone word to vec is a popular model in using neural networks where every word can be mapped to a vector the idea being that similar words may occur frequently in each other others context. So, this fact is somehow taken into account while constructing the word vectors. Now, this concept of word vectors that can be like enhanced further specifically in the context of economics.

So, here for example, they have they are using word embeddings to repair to reveal monetary policy explanations changes. So, here like the input is the sequence of policy documents once again and from each of these policy documents they have the policy words. in the different type points of time t minus 1 t t plus 1 and so on. Now what they have observed is that the policy words keep evolving over the time. The same word can be the same thing can is often represented by different words at different points of time.

So, what they are so like this fact they are like trying to capture with the help of word embeddings. that is even if the like if one particular thing is referred to using word w 1 at time t minus 1 and word w 2 in time t like then w 1 and w 2 their word vectors should be similar to each other because they are referring to essentially the same thing. So, these kinds of word embeddings are like they are trying to consider and this is another interesting application economic worth aware word embeddings. So, here what they are doing is they are trying to like that is apart from the semantic meaning of every word they are trying to attach an economic value also. and they are trying to when they are representing every word by a vector as we have already discussed the idea was that similar words with similar meanings should be should have their vectors close to each other.

But here the idea is that even words having similar worth they also should be close to each other in the embedding space. So, that is what that accordingly they have developed this word worth model or WWM to learn a word embeddings that capture the underlying economic word worths. So, accordingly like the like the I mean the word vectors they assign to the different words may be slightly different from the word the word vectors

from from word to vec. but they they explain that or they claim that from an economical point of view this is their their representation is better. And nowadays like I earlier talked about these language models and I mentioned the simplistic n-gram models and so on.

Nowadays we have this neural language models. So, the idea is that this probability that is the probability distribution over the next word w t given the sequence of previous words this is this probability distribution is represented by the latent representation Z of the RNN or an LSTM that is we consider that each of the previous words are sequentially input into a sequential deep learning model such as an RNN and LSTM as they we keep on inputting the previous words its latent states gets updated and at any given point What the next word is going to be like a probability distribution over that is created by the or can be represented by the this latent state of course, assuming that it has been trained appropriately. And we also have this in recent years in the past 5, 6 years one new concept has really grabbed the attention of this community and that is that word is attention itself. It basically quantifies the importance of the individual word tokens the this importance of course, learnable and there is exists a new kind of neural network architecture called transformers which are the state of the art neural network models for NLP and they are based on this concept of attentions. And there is this model called BERT, it is called the pre-training of deep bidirectional transformers for language understanding. So, like so, the it is called bidirectional the idea is quite simple it is the important thing here is the bidirectional.

So, where like the bidirectional basically indicates that like when we are constructing the context of a word that is the words nearby it. So, we are considering both the words before it as well as the words after it. So, that is why we are consider calling it as a bidirectional. And, this paper econ BERT towards robust extraction of named entities in economics. So, it tweaks the BERT model a little bit in the context of economics for this like for this task of named entity recognition.

We have already discussed the task of named entity recognition in economics. So, here I mean named entity recognition in general we have discussed. Specifically, in the context of economics, so like here they have considered three different kinds of named entities. One is outcome like in this case outcome means saving of money or it can be interventions like imposing participation requirements. and so like and so on and so forth. So, three different kinds of named entities they are looking for in this case and they have took the BERT model a little bit to match this requirement.

So, and like this last paper which we are discussing this is somewhat similar to the like the computer vision based approaches to economic problems which we discussed earlier. So earlier we had seen how satellite imagery etcetera can be used to estimate economic developments in different regions. Here the idea is we are predicting economic

developments using geolocated Wikipedia articles. So, what is a geolocated article? So, we are all familiar with Wikipedia. So, some Wikipedia pages especially the ones which are about a particular region they may be geotagged that is they may have some coordinates associated with them.

Now, like when we come across such a geotagged article we like from that article itself we can like if we do a natural language processing we can find out various attributes about that place I mean about what kind of economic development it has and we can also correlate that Wikipedia article with the satellite imagery of that place. So, like this is how we can say it is a multimodal approach it has two modes of input one is the Wikipedia articles as well as the satellite imagery especially the night time lights which we already had discussed in the previous lecture. So, like these two form inputs to two branches of a neural network which work together I mean which create a common representation of that of that region and from that common representation that the socioeconomic development of that region is somehow predicted. So, these are the references of the different papers which we discussed together. So, in summary natural language processing has a huge huge potential in solving many tasks of economics and so with that we come to the end of this lecture and so after this we will have just one lecture left. So, see you then stay well and take care bye.