Affective Computing

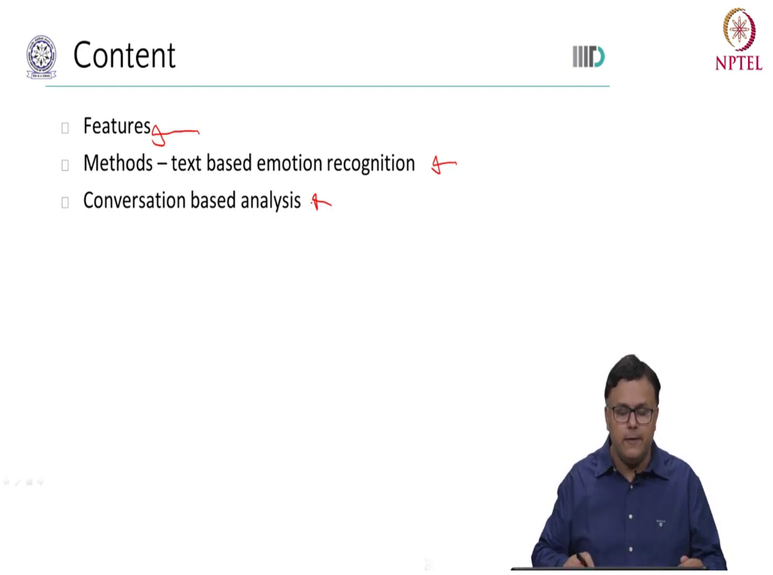
Dr. Abhinav Dhall
Department of Computer Science and Engineering
Indraprastha Institute of Information Technology, Delhi

Week - 06
Lecture - 02
Emotion Analysis with Text

Hello and welcome. I am Abhinav Dhall from the Indian Institute of Technology, Ropar and friends today we are going to discuss about Analysis of Emotion from Text. So, this is the 2nd lecture in the text analysis for emotion recognition in the Affective Computing series.

So, in the last lecture we discussed about what is the importance of analysis of the emotion from the text. The text could be in the form of a tweet, could be a document and then we discussed, how simple organization of text in the terms of the font, the curvature of the lines which are constituting the font, essentially topography, how that affects the perception of emotion.

(Refer Slide Time: 01:11)



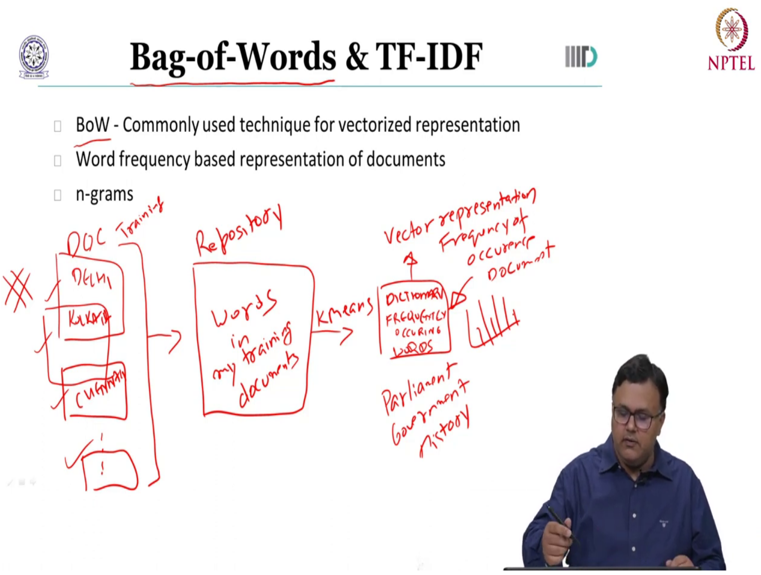
Content

- □ Features
- □ Methods – text based emotion recognition
- □ Conversation based analysis

And today first I am going to discuss with you a very important building block of a text based emotion recognition system which is the important features. How are we going to understand the context, the linguistics, how are we going to represent the text in a tweet or a document. Later on, I am going to discuss with you few methods, which are proposed in the academic community which are predicting emotion from text.

Later on we are going to switch the gears a bit and we are going to discuss about how emotion recognition can be performed when there is a conversation happening. When two people are let us say conversing through text, the perception of the emotion and the understanding of the emotional state of the user that can be very dynamic. So, how the emotion changes, how the perception changes, over the conversation flow. So, we are going to look into that.

(Refer Slide Time: 02:19)



Bag-of-Words & TF-IDF

- BoW - Commonly used technique for vectorized representation
- Word frequency based representation of documents
- n-grams

So, let us start with one of the most commonly used feature representation for text and to link it to how we were using audio and video for emotion recognition. Recall we discussed that we can have a bag of words based representation, right. So, we can represent parts of an image or parts of a speech sample as a word and then we can create a representation.

Similarly, when we are talking about your bag of words, typically what we are saying is let us say you have a document. Now, for the sake of the example, imagine the document contains information about the different cities in India ok. So, this one let us say talks about Delhi, then the second document talks about Kolkata and then the third one talks about Chennai and so forth, right. So, we have a long list of documents.

Now, the content in each document would be different that would mean the number of words in each document will be different. Now, what we want to do, in the plain vanilla bag of words representation friends is we want to take all the documents which are present during the training and we want to create a large repository, which contains all the words, so all the words in my training documents ok.

From this we can apply a simple clustering technique like k means, we have discussed this in facial expression based emotion recognition and that is going to give me a dictionary, which will contain the frequently occurring words. So, you can say these frequently occurring words are the building block of my documents.

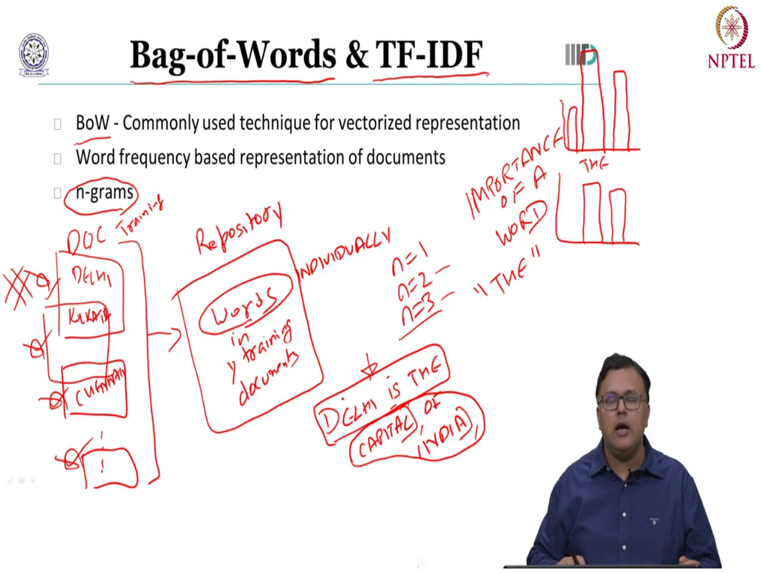
And once I have identified this frequently occurring words, I can then have a vector representation for each document separately, which is essentially the frequency of occurrence of these important words in my document, right. So, that is a vector representation.

So, for example, if you are talking about Delhi, then the word for example, parliament and government, history these might be more common within the Delhi based document as compared to let us say, document discussing another city which is not the state capital and is let us say New City, right.

So, you will have a histogram a vector lies the representation, then standard method you can train a machine learning model and predict the emotion, which let us say the text is conveying about the city, right. Now, as you would have figured out friends here, what we are simply doing is?

We are counting the number of words, the unique words in a document and then trying to see the frequently used words across all the documents, how many times they are occurring in my current document. Now, imagine when you were talking about the input stage you know the first step, wherein I said well you create a repository of all the words in all the documents.

(Refer Slide Time: 06:48)



Bag-of-Words & TF-IDF

- BoW - Commonly used technique for vectorized representation
- Word frequency based representation of documents
- n-grams

So, we were considering the words here individually. What will that mean? So, let us say a line in my document says Delhi is the capital of India. Now, you treat each word here separately Delhi is one word, is one word, the is one word and so forth, right. Now, very early on a concept was introduced specially for natural language processing, where we said well, let us also have combination of words which are occurring together. Now, that is referred to as n-grams.

So, now when n is equals to 1, you observe this scenario where Delhi is a separate word, is a separate word, the is a separate word, separate entities. When you have, n is equals to 2, notice what is going to happen? I am going to say in the same statement Delhi is the capital of India. I am going to take these two conjunctive words together.

So, Delhi is, now this is one unit, the capital that is another unit and of India that is another unit ok. One could move a step further and say I am also going to now take three neighbor's together as one unit, as one word. So, what will that be in our example statement? When n is equals to 3, a 3 gram a trigram you would say Delhi is the, is one word, is one unit.

And once you have established, the different combinations of words and the appropriate representation for n-gram, n is equals to 2 3 depending on the value, you can then repeat all the steps which we just discussed for bag of words. Wherein now these words before k means could be a combination of sequential words right, each word here could be for example Delhi is the.

Now, that is one unit, one word in the bag of words sense and one could then train the system. Why would that be useful? Simply because when let us say you encounter a individual unit like this where we say Delhi is the. So, from the words meaning perspective, what we understand is? That after this some information about Delhi is going to be presented, right.

So, here we say Delhi is the, and then for a trigram where n is equals to 3, you would have capital of India as a second unit, right. So, learning the relationship between sequentially occurring words as well, so that is another way of extracting information for our final goal which is understanding emotion from text analysis, right.

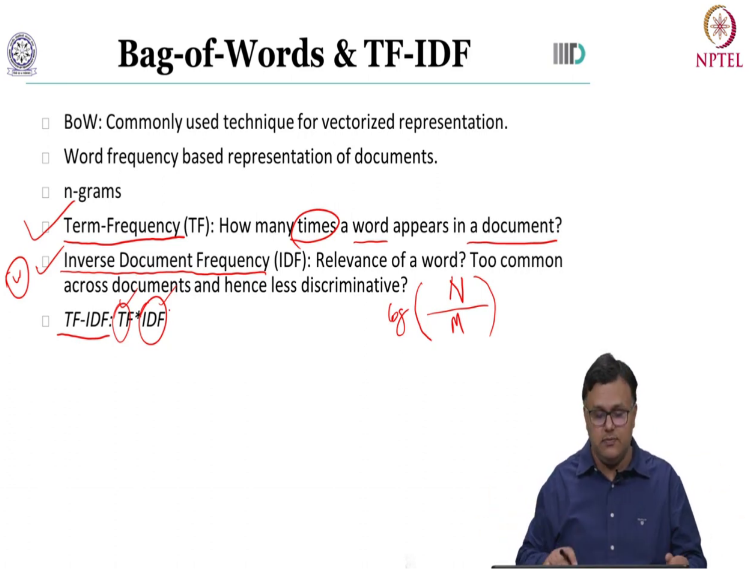
Now, similar to this you can also have another concept, wherein you can say well. I am looking at the words individually or sequential words as one unit how about I also take into consideration, the importance of a word ok. Now, what could be importance of a word there are different ways to discuss and describe this one could say well, if we have the word Delhi being repeated in a document several times, there is a high probability that the document is discussing about Delhi something in Delhi, something around Delhi, right.

The other word is let us say you have a word “the”. Now the word “the” is a common word and it would be repeated across different documents. So, even if let us see in a vector representation of a document, you know I am just going to draw histogram here, the word "the" could be repeated multiple times in a document. And other words are of you know similar or smaller frequency, you would find that in other document histogram the again could be very high in value.

So, large frequency, so "the" does not helps me much, in discriminating between the types of document and perhaps from the context of emotion recognition, it is not helping me much in discriminating between if let us say the perceived contents emotion from a document is class x and then in other document you know the perceived emotion is class y. Because the is happening so many times, it is not really helping me much, right.

So, how do we take care of the situation, how simply we can encode this information within our representation of the text? So, for that friends, we have the very popularly and commonly used TF-IDF representation.

(Refer Slide Time: 12:22)



Bag-of-Words & TF-IDF

- □ BoW: Commonly used technique for vectorized representation.
- □ Word frequency based representation of documents.
- □ n-grams
- □ Term-Frequency (TF): How many times a word appears in a document?
- □ Inverse Document Frequency (IDF): Relevance of a word? Too common across documents and hence less discriminative?
- □ *TF-IDF: TF*IDF*

$\log\left(\frac{N}{M}\right)$

Now, what is that? The first part of this is the term frequency TF simply that means, how many times a word is appearing in a document. Now, how many times for example, the word Delhi is occurring in a document. If that word is occurring multiple times, then it could be important ok. You have seen words like “the” and “a” they happen multiple times you know all the documents, right.

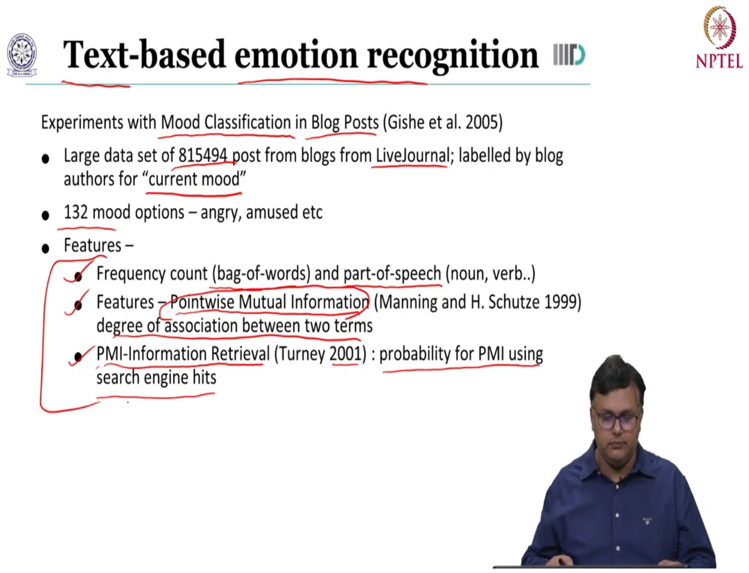
So, that could mean I cannot only depend on the term frequency, what I; what would I need? To balance it I would need a inverse document frequency it is called IDF. What it does is? It tells the importance relevance of a word, how is it doing it. Well, you say for a given word, I am going to compute the relevance by saying how many data points do I have in terms of documents. Let us say I have N data points.

Simply speaking and I am going divided by the number of times this particular word term happens in some of the documents ok. So, let us say M. The happen so many times. Now, I would like to use this as a weight ok and so that I do not do a very harsh treatment to the words which are happening multiple times in some of the documents, I am simply adding a logarithmic to it.

So, this fellow is going to give a weight, essentially if you were to understand in the importance saliency terms. So, for a given term its TF-IDF value would be its term frequency multiplied by its inverse document frequency. Now, Delhi may be a common keyword in some of the documents and it is repeated multiple times.

So, you will see a high value of TF, but it will not let us say be giving you a low value for IDF, so that means, salience. However, the word “the” would have a high TF, but a low IDF because it is happening all across multiple times in multiple documents.

(Refer Slide Time: 14:50)



Text-based emotion recognition

Experiments with Mood Classification in Blog Posts (Gishe et al. 2005)

- Large data set of 815494 post from blogs from LiveJournal; labelled by blog authors for "current mood"
- 132 mood options – angry, amused etc
- Features –
  - Frequency count (bag-of-words) and part-of-speech (noun, verb..)
  - Features – Pointwise Mutual Information (Manning and H. Schutze 1999) degree of association between two terms
  - PMI-Information Retrieval (Turney 2001) : probability for PMI using search engine hits

Now, friends let us look at a system proposed for a text-based emotion recognition. Now, we are going to discuss a classification task. So, this task proposed by Gishe and others is for classifying the perceived mood in blog posts. Now, the approach of the authors was as follows, they curated a large dataset of about 815000 posts from website called LiveJournal and then they asked the labellers to also indicate that current mood while they are were writing a particular blog.

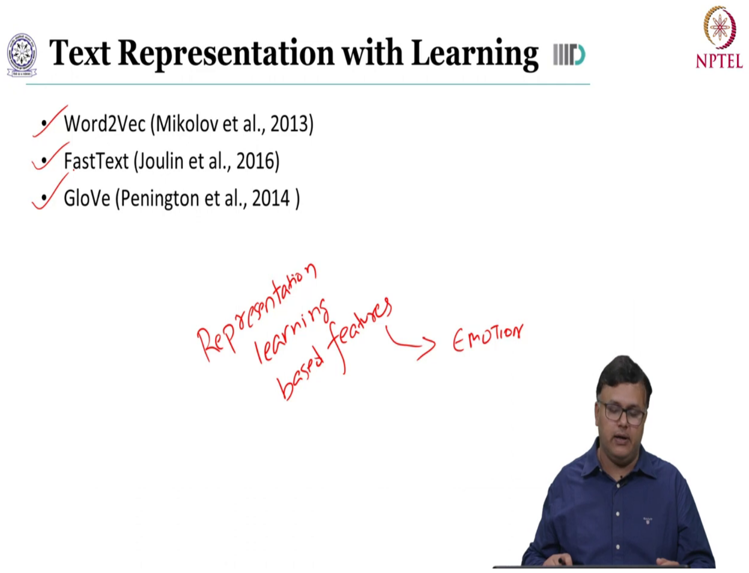
They had 132 categories of mood different keywords which the labellers, they gave to their particular blogs, how they were feeling. And now, let us look at the features which the authors used. The first is which we have already discussed looking at the bag of the words based representation. Now, what the authors also did is, they added what is referred to as part of speech?

Now, part-of-speech is essentially tagging the words in your data with them being a noun or an adjective or a verb. Because the number of nouns, the number of adjectives and you know the number of verbs in a particular document also gives us vital meta information about the task. Second, the authors used what is referred to as a point mutual information.

So, this is a feature which was proposed back in 99 and this point wise mutual information simply gives us the degree of association between two terms. How much two terms are related? Now, based on your point wise mutual information, the other feature which the authors used is the point wise mutual information hyphen information retrieval.

Now, this is a work from 2001 and this gives us the probability for PMI using search engine hits. So, you use the point wise mutual information and then you see, what is the probability of retrieving a certain result given certain keywords based on the PMI feature which essentially how close are two terms, you know what is the degree of association. Further they would then go forward combine this and then do classification.

(Refer Slide Time: 17:28)



Text Representation with Learning

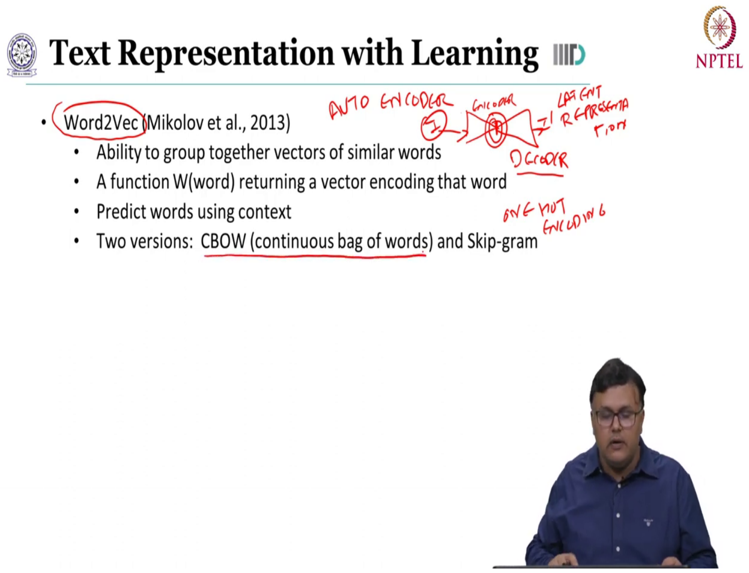
- ✓ Word2Vec (Mikolov et al., 2013)
- ✓ FastText (Joulin et al., 2016)
- ✓ GloVe (Penington et al., 2014)

Representation learning based features → Emotion

Now, we saw a paradigm shift right, as with any machine learning system, we saw earlier support vector machines, knife base kind of algorithms were used and then deep neural networks they kind of came back in the early 2012-2011. Which affected vision analysis, speech analysis and the same effect was felt in text analysis as well.

Now, to this end, we now are going to talk about representation learning based features. So, how representation learning based features can be used to map an input text into its emotion category. To mention one of the most popular ones which are used in the community there is the Word2Vec, then an extension from Word2Vec called FastText and later is the GloVe in feature representation. So, let us look at what is Word2Vec.

(Refer Slide Time: 18:42)



Text Representation with Learning

- Word2Vec (Mikolov et al., 2013)
    - Ability to group together vectors of similar words
    - A function W(word) returning a vector encoding that word
    - Predict words using context
    - Two versions: CBOW (continuous bag of words) and Skip-gram

Now, friends Word2Vec was proposed by Mikolov and others in 2013. Now, in this representation the authors leveraged the ability of a network to understand the presence of certain words together. So, if you look at a statement, words which are in a sequence are related to each other. So, can we use this observation this very simple observation and learn the vectorized representation.

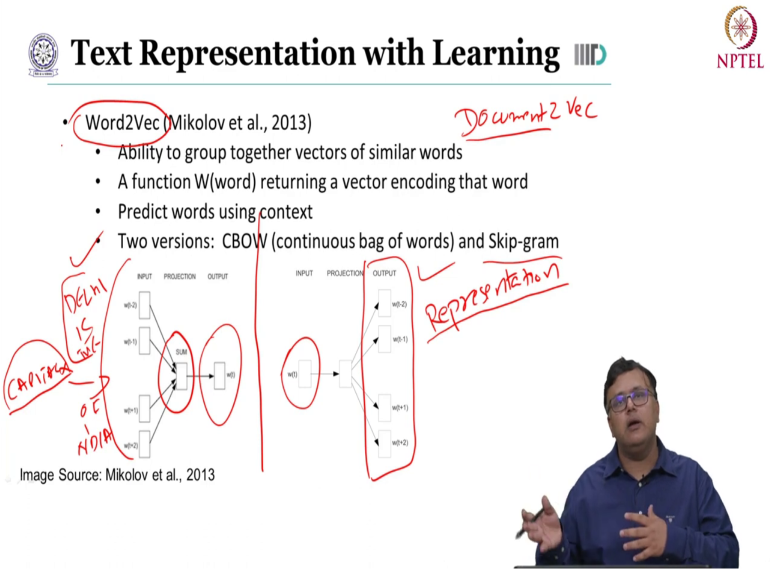
Now, what that means is? We would have an auto encoder to recall what was an auto encoder? You have a neural network, where the first part is your encoder and then the second part is a decoder. In between here, we have what is referred to as a latent representation which is essentially nothing but a compressed representation of the input and using this compressed representation which we input into the decoder, we get the reconstruction of the original input let me call it I dash.

Now, since we are talking about text and Word2Vec in specific, if we are inputting a series of words into my encoder then I would like to learn the relationship between them and I am going to use the latent representation, which is essentially going to give to me? The vectorized representation of the input words, so what we are doing is? We input into our network a one-hot encoding.

So, let us say the term is Delhi is the capital of India. Now, when you have the word Delhi under the focus, you keep the corresponding indexes 1, every other word gets a 0. Now, let us say in this case of Word2Vec you say Delhi is the capital of India, right. So, you could use Delhi is dash of India.

Now, the word capital is linked to the country which comes afterwards and it is linked to Delhi as well because Delhi is the capital of India. Therefore, the author they said well, let us have two types of representations two ways of computing the representation. The first one is referred to as your continuous bag of words. Now let us see what is that.

(Refer Slide Time: 21:32)



Text Representation with Learning

- Word2Vec (Mikolov et al., 2013)
  - Ability to group together vectors of similar words
  - A function W(word) returning a vector encoding that word
  - Predict words using context
  - Two versions: CBOW (continuous bag of words) and Skip-gram

Image Source: Mikolov et al., 2013

So, here you have your words as the input. So, you are saying Delhi is the I am not going to input capital here ok. So, this word is not input, this is the input here and later it is of India ok. Now, these are one-hot encoding vectors representation of the words, what we are saying is?

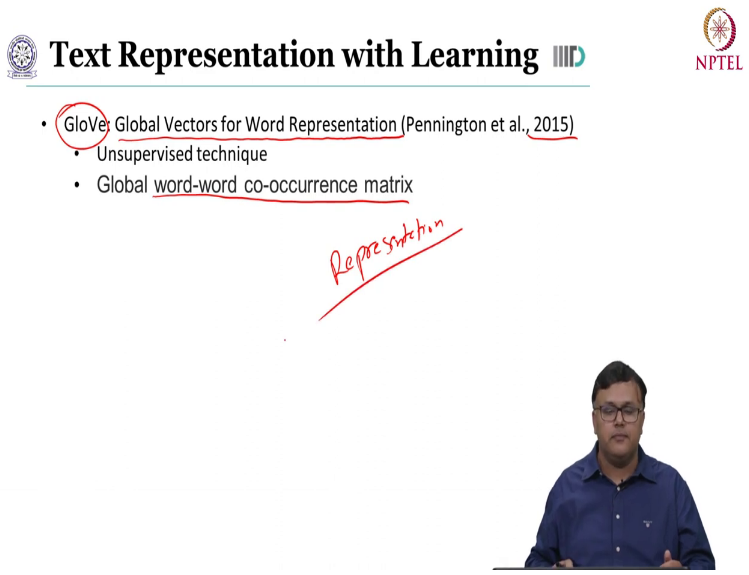
The network would not learn the presence of this word the representation for this word by submitting, the precursor and the content which comes afterwards the content which was before, the content which is afterwards and that is going to give me the output for representation of capital.

Another representation which was proposed in the same work is called your skip gram. Now, in that you are saying well you should be able to predict the neighboring words, right. So, if you had capital, capital city, capital of a country, city capital of a country, right. So, they the words before and afterwards have a relationship with the word which is being input so, how about trying to predict the neighborhood words.

And in both these independent pursues of your continuous bag of words and your skip grams we are learning the representation ok. So, this is a very powerful representation friends, which has been extensively used in predicting the vector representation for input words. Now, there have been some famous, widely used extensions of Word2Vec and that I will give an example to you is for example, your document 2 vector, right.

So, what is the vector representation for a document? So, again that is based on the concept of Word2Vec. Now, recall we wanted emotions, right. So, what will that mean? You take the vector representation for each word using Word2Vec pre-trained network, then you do the pooling as we have done discussed earlier and then you can learn a machine learning system to predict.

(Refer Slide Time: 23:52)



Text Representation with Learning

- GloVe: Global Vectors for Word Representation (Pennington et al., 2015)
  - Unsupervised technique
  - Global word-word co-occurrence matrix

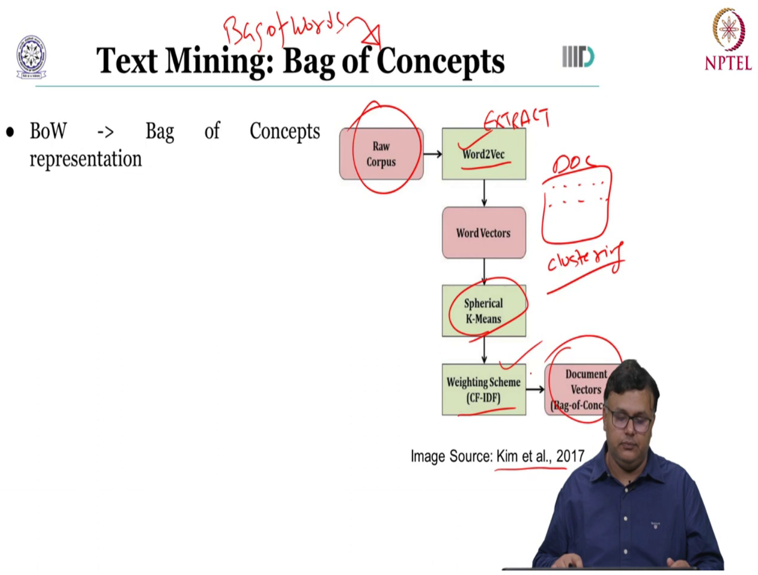
Representation

NPTEL

Now, another representation which I would if you like to mention to you is friends the global vector for word representation proposed in 2015 ok. In short popularly referred to as GloVe. Now, this is an unsupervised technique to learn the representation and simply it is based on creating and then little learning the word to word co occurrence matrix.

So, the authors created a matrix which is telling us, that what is the probability of occurrence of words together based on that, we learn a network and we are going to get a representation which I am going to be able to use later.

(Refer Slide Time: 24:42)



Text Mining: Bag of Concepts

- BoW -> Bag of Concepts representation

Raw Corpus

Word2Vec

Word Vectors

Spherical K-Means

Weighting Scheme (CF-IDF)

Document Vectors (Bag-of-Conc

Image Source: Kim et al., 2017

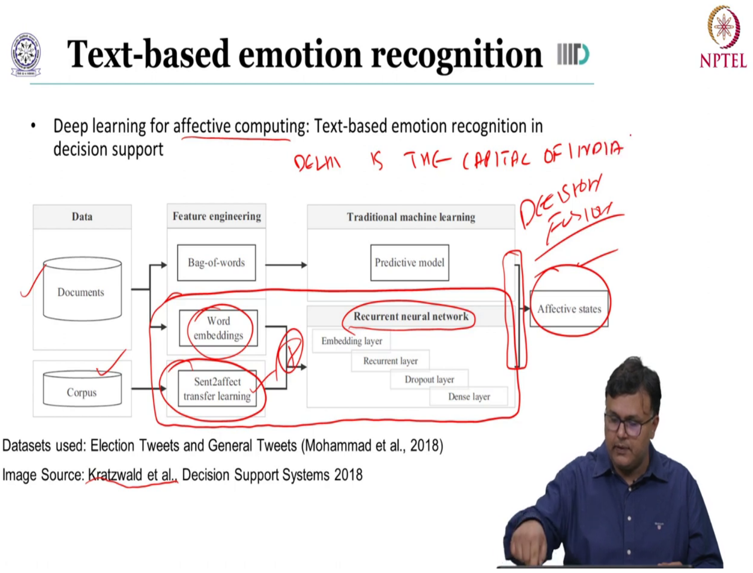
Now, let us look at one example where this representation learning is used. So, friends we have seen bag of words ok, we know this term bag of words now. So, from bag of words there is a an extension called bag of concepts. So, I do not want to input individual words or your bigrams or trigrams that is 2 words together, 3 words together as 1 unit, but I want to input into my system a vector which is created from a bag of concepts, right.

Now, let us look at one such work proposed by Kim and others, this you have the raw text data coming in you extract the representation using Word2Vec. Now, that gives you for a document the representation for each word. What we do is, we then compute a clustering on the representation which was extracted from Word2Vec and later we do this weighted scheme similar to TF-IDF which we have seen, right.

So, we are actually looking at the importance of each word. Now, notice in the plain vanilla bag of words, when you are doing vectorization of an input sample you would increment the bin in your histogram during the vectorization by let us say 1, when a particular new word arrives you have checked that word.

In this case we want the histogram to be a weighted accumulation of the content in a certain bin, right. So, you are doing your TF-IDF and then you are getting a representation which you can further use for predicting the emotion.

(Refer Slide Time: 26:44)



Text-based emotion recognition

- Deep learning for affective computing: Text-based emotion recognition in decision support

DELHI IS THE CAPITAL OF INDIA.

DECISION FUSION

Data

Documents

Corpus

Feature engineering

Bag-of-words

Word embeddings

Sent2affect transfer learning

Traditional machine learning

Predictive model

Recurrent neural network

Embedding layer

Recurrent layer

Dropout layer

Dense layer

Affective states

Datasets used: Election Tweets and General Tweets (Mohammad et al., 2018)

Image Source: Kratzwald et al., Decision Support Systems 2018

Here is another system friends on similar lines. Now, this work by Kratzwald and others that is deep learning for affective computing, text based emotion recognition in decision support.

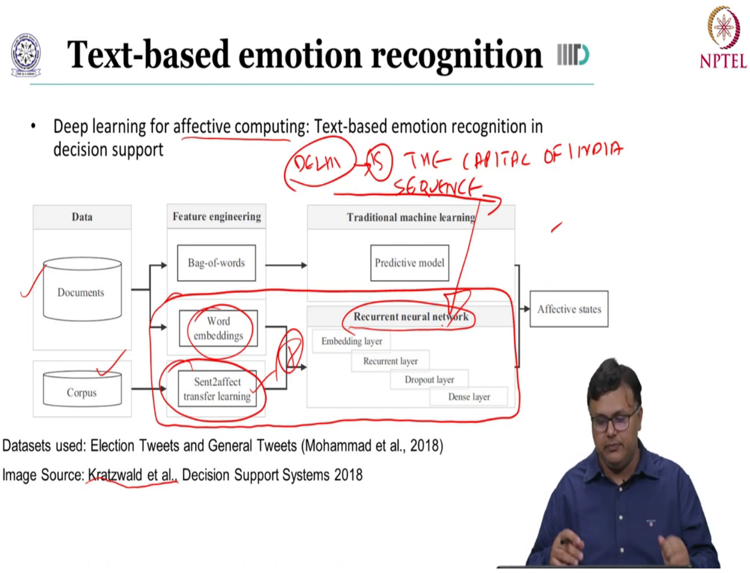
So, how we can use the affective states for predicting? The you know decisions which a machine could take and also assisting the user.

So, what do we have here? We have set of documents as training data. We have a pre-trained network which is learned, trained on a sentiment labelled corpus. Now, let us look at the features, the authors first are extracting bag of words and word embeddings from your pre-trained network.

Then you are predicting the network with the features you are doing feature fusion here for these two and then you are predicting the emotion. In parallel you use the word embeddings, which we have seen how they are extracted in the earlier slides. Then you extract the feature from this transfer learning based model concatenate and then you have a recurrent neural network.

You combine the outputs together, so you do the decision fusion and that gives you the affective states. Now, one thing to note here friends, I will like to draw your attention here is for the second part ok, for the second part here which is; when you have the word embedding and the features from transfer learning input into a recurrent neural network. What is typically happening in your RNN? Let us say the statement again which we have been using right; the example Delhi is the capital of India.

(Refer Slide Time: 28:41)



Text-based emotion recognition

- Deep learning for affective computing: Text-based emotion recognition in decision support

DELHI IS THE CAPITAL OF INDIA

SEQUENCE

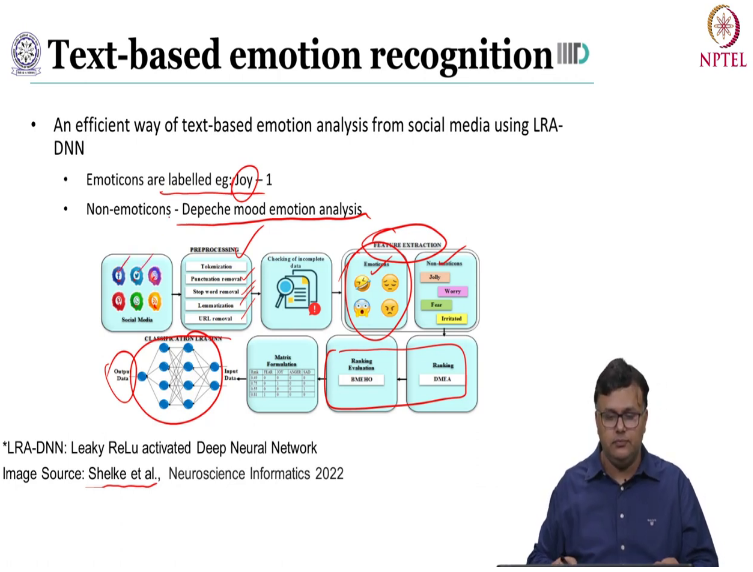
Datasets used: Election Tweets and General Tweets (Mohammad et al., 2018)

Image Source: Kratzwald et al., Decision Support Systems 2018

Now, when you are having a RNN what you are saying is essentially, you get a feature representation for this word you input it to the cell and along with that you have the feature representation of the second word available to the cell. So, there is a sequence which is being followed, right. So, this you could say that the sequence in which the words are arriving in my statement, I am learning that pattern in recurrent fashion in the recurrent neural network way, right.

So, you have the cells you are inputting from the prior and to from the current representation. Now, this has been very commonly used in natural language processing, you know the recurrent neural networks, your LSTMs, by lSTMs and they are processing that information sequentially.

(Refer Slide Time: 29:23)



Text-based emotion recognition

- An efficient way of text-based emotion analysis from social media using LRA-DNN
  - Emoticons are labelled eg: Joy - 1
  - Non-emoticons - Depeche mood emotion analysis

*LRA-DNN: Leaky ReLu activated Deep Neural Network

Image Source: Shelke et al., Neuroscience Informatics 2022

Now, the community slowly has started moving to other approaches where; we are saying well we do not need sequential approach. But how about if I was able to parallelize and these are also referred to as your non-regressive models and after this I will show you few examples of that as well ok.

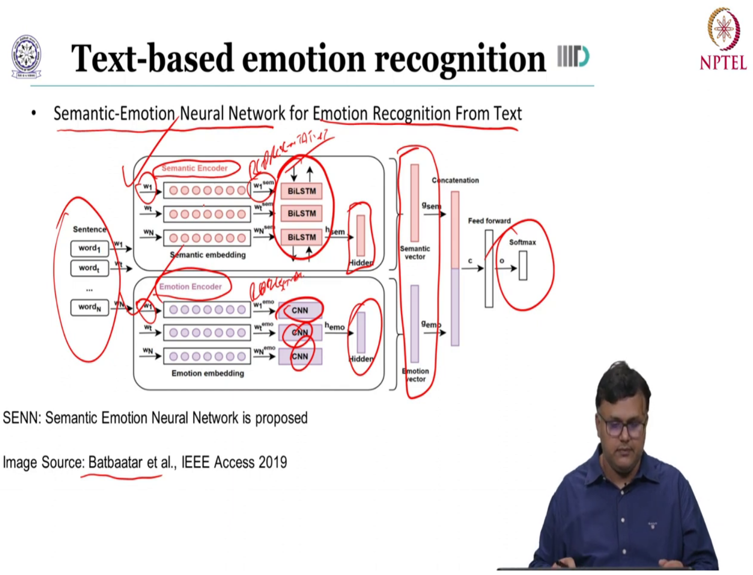
Now, let us look again at another system proposed by Shelke and others, which is using social media data as input to predict the emotion from the social media post. Now, notice in social media post, you are not only going to have the text or you could also have emoticons.

So, in this very interesting work the authors are first pre-processing the input text from social media post these are the different platforms from this they are fetching the information. You see they are doing the tokenization that is word level individual unit creation they are removing the punctuations; they are removing the stop word during the stop word removal, removing URLs and also doing lemmatization.

Now, for the emoticons, they are also adding labels to it as well. For example, they say well you see the joy emoticon, let us say you know the representation for that is a one. And for the text they are actually using another mood analysis system it is called the Depeche mood emotion analysis.

So, they combine, it extract features from the emoticons from the text and then they are doing a machine learning based ranking. They rank the presence of the emoticons along with the text and then input that into a deep neural network to predict what is the emotion, which is perceived from let us say post on Facebook or on Twitter and so forth.

(Refer Slide Time: 31:10)



Text-based emotion recognition

- Semantic-Emotion Neural Network for Emotion Recognition From Text

SENN: Semantic Emotion Neural Network is proposed

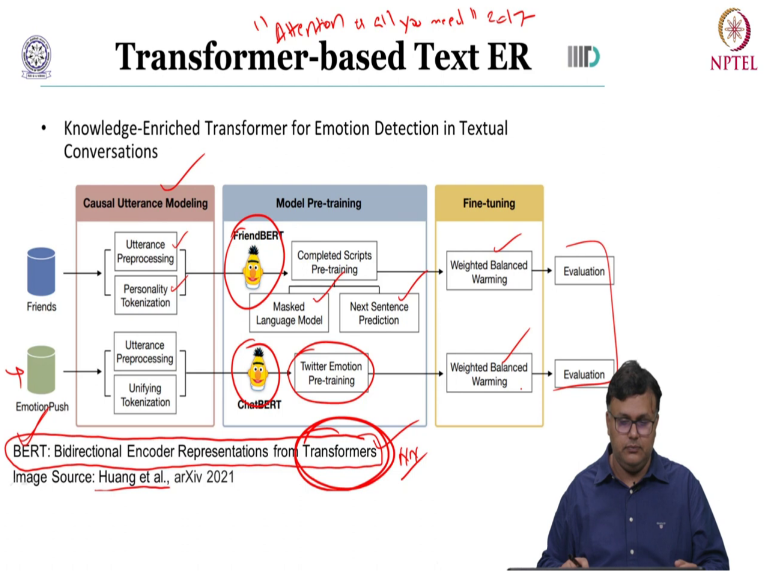
Image Source: Batbaatar et al., IEEE Access 2019

Now, let us look at another work friends, in this work the authors Batbaatar and others, they proposed a work called semantic emotion neural network for emotion recognition from text. What do we have in this work? We have the input words; we are extracting two types of embeddings from this work.

The first is a we have a pre-trained semantic encoder; it takes one word as an input at time and gives you the representation which is relating to the semantic information. In parallel the same word is input into an encoder which is giving us a representation which has the emotion representation. For the earlier one, there is a recurrent neural network and we get the fully connected layer.

In the emotion channel we are using a CNN based representation, we fuse it get the hidden network and then we are fusing this concatenating. So, what is happening here? We are getting your feature fusion and then we are predicting the emotion class. So, what did we achieve here? We had a semantic pre-trained representation, we had a emotion representation and then we are fusing it together.

(Refer Slide Time: 32:28)



"Attention is all you need" 2017

Transformer-based Text ER

- Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations

BERT: Bidirectional Encoder Representations from Transformers

Image Source: Huang et al., arXiv 2021

Now, as I was referring to you earlier, about these non-regressive techniques coming right, where we could input the data in parallel. After Word2Vec type of representations, the community has moved to attention based systems. Now, attention is simply saying an input statement comes in, what is important, what is not so important and how do you learn that well there is a whole mechanism and I am writing it here for you, to check out separately.

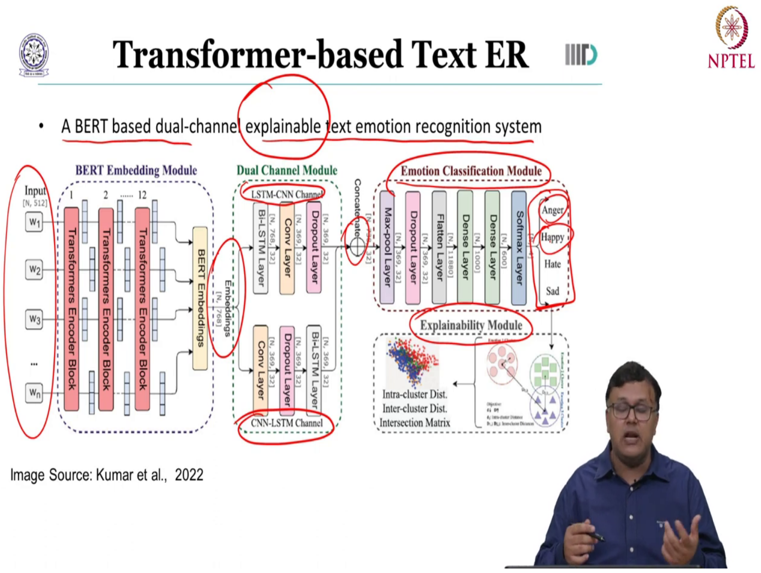
So, there is a paper called attention is all you need. This is a seminal work proposed to 2017. Will discusses how attention can be applied to a network for text task particularly in this. And this has given birth to neural networks, which are referred to as transformers. Now, these are non-regressive methods that is the tokens which are input into it, the words which are input into it are parallelly taken care of.

And similar to Word2Vec, the transformer based representation which is extremely common now in natural language processing and hence for natural language processing based emotion prediction is called BERT. So, this is your Bidirectional Encoder Representation which is based on a transformer architecture.

So, here is a work by Huang and others, what they are doing here is? They are modeling the utterances ok, so they are doing utterance pre-processing and they are looking at the personality tokenization. So, this is the pre-processing part, they input it into a pre-trained BERT model, which is called a friend BERT. And then the second one from the emotion corpora, they are input into a chat BERT. So, these are of course, you know as the name suggest, have been trained on different types of data.

The representation which they get, they are actually using that to do a pre-training on the input data and they are using a masked language model and a next sentence prediction what would come afterwards and for the output coming from the chat BERT, they are using it for a emotion pre-training for the Twitter data. Data, they do fine tuning and then they are doing the evaluation ok. So, you can use these BERT base representation nowadays as well.

(Refer Slide Time: 34:51)



Transformer-based Text ER

- A BERT based dual-channel explainable text emotion recognition system

Image Source: Kumar et al., 2022

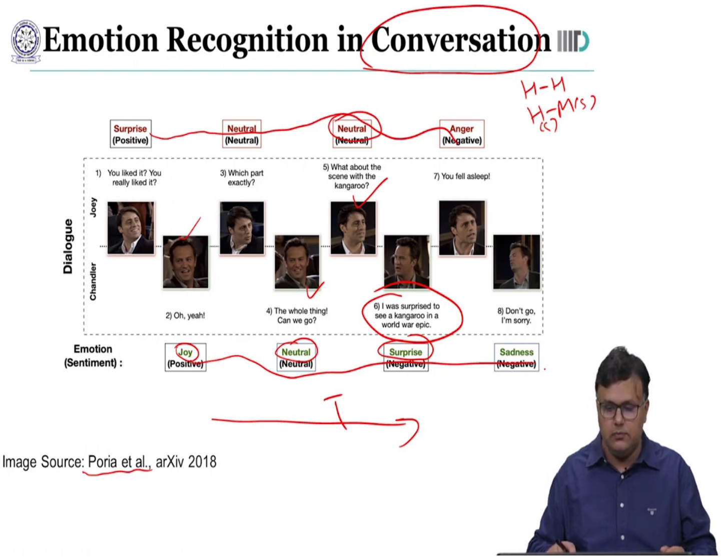
Now, here is another work. So, this is again by Kumar and others, it is called a BERT base dual channel explainable text emotion recognition system ok. Now, notice it is explainable ok. Now, you have the input representation for each word. So, these are the tokens which you are inputting into your BERT module.

You get the feature embeddings from the BERT module, but do we have data we have a RNN-CNN model and a CNN-RNN model this is just to exploit the temporal relationship, we can catenate the feature output and then we have a classifier. Now, this classifier is telling us one of these four states anger, happy, hate or sad.

Further the authors also propose an explainability module, which is looking at the intra clustered distance, inter clustered distance of the outputs to explain, why we have anger or happy as a particular label which is predicted. So, notice how the transition has happened for text base emotion recognition systems.

In the beginning we were using bag of words, bag of concepts base representation. And now we are in 2023 we are using systems such as BERT which are based on transformers.

(Refer Slide Time: 36:08)



Emotion Recognition in Conversation

H-H
H-M(s)
(s)

1) You liked it? You really liked it?

2) Oh, yeah!

3) Which part exactly?

4) The whole thing! Can we go?

5) What about the scene with the kangaroo?

6) I was surprised to see a kangaroo in a world war epic.

7) You fell asleep!

8) Don't go, I'm sorry.

Image Source: Poria et al., arXiv 2018

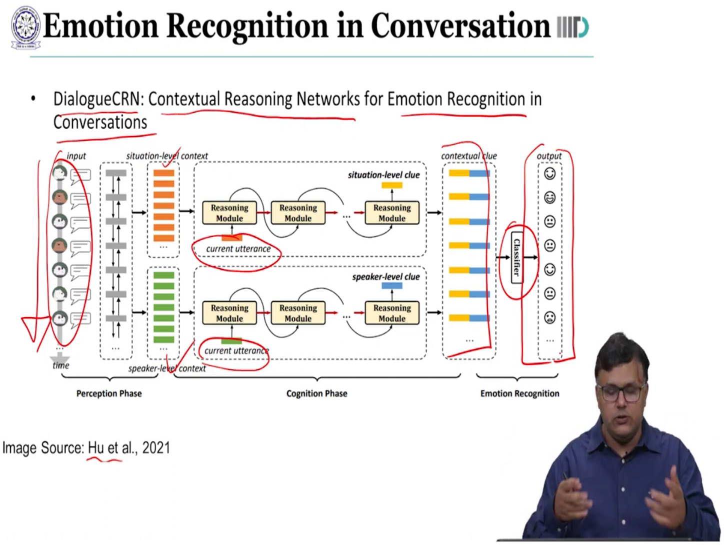
Now, let us change the conversation a bit friends, we have been talking about emotion recognition in text when the text is coming from let us say a blog, a document, a tweet or a post on social media. There is another dimension rather very important dimension to emotion analysis that is during conversations. Conversation could be between two humans. So, a human conversation, it could be a human machine conversation, it could be human machines conversation or humans and machines conversation.

So, as the conversation happen, the emotion changes the intensity changes. Now, what we have here on the screen is an example from Poria and others ok. What you see is a dialogue between two characters, now this is coming from a very popular sitcom, when what you see as the dialogue proceeds. So, here is the time axis you would notice that the subject is showing different emotions, right.

Here this fellow Chandler is showing joy, then there is a neutral, then there is a surprise after, then input coming from this person so this he said something, then this say this guy Joe says something, now emotion is elicited and you know this is actually showing surprise for Chandler and so forth.

So, you see how for both the subjects the emotion based on the text that is varying here is right. So, that means, for emotion recognition when conversation is happening, we would require a more dynamic approach. We require an approach which would be utilizing, what the other person said and what the user under focus replied back, right. And you could use a time series information as well to get the context the long term context.

(Refer Slide Time: 37:58)


Emotion Recognition in Conversation

- DialogueCRN: Contextual Reasoning Networks for Emotion Recognition in Conversations

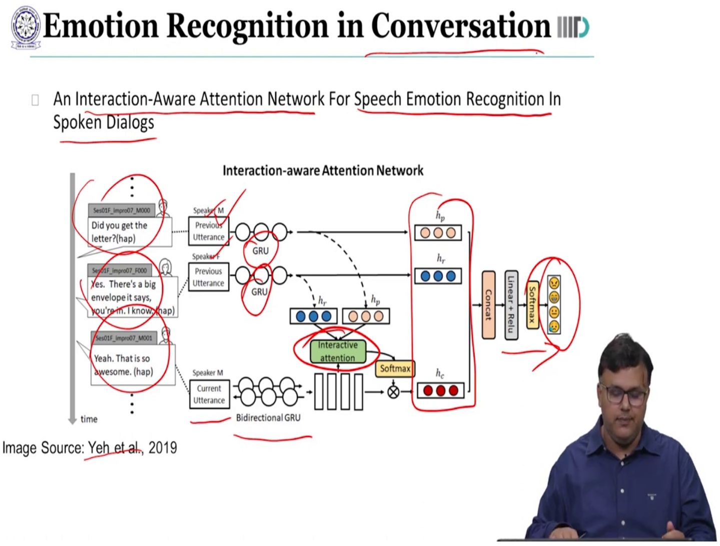
Image Source: Hu et al., 2021

Now, here is an example of a work, where emotion recognition is performed during a conversation. This work by Hu and others is titled as dialogue CRN, the contextual reasoning networks for emotion recognition during conversations. What we have friends here is the timeline and the input conversation between individuals.

What we are doing is, we are extracting the feature representations, which are getting us the situation level context and also the speaker level context again these are the feature representations. Then the authors they are proposing utterance module, you know this is extracting the situation in which the conversation is happening again these are using the neural networks.

In parallel we are analysing the speaker level clue as well, that is individually what a person said. This information is fused, then there is a classifier to predict what is the emotion during a conversation, right. So, two takeaways from this analyse the situation, analyse the content which a particular current person is speaking. Situation is going to vary, based on how the conversation is happening and the conversation is also going to be affected by the situation and the context where a person is what are they speaking and so forth.

(Refer Slide Time: 39:22)



Emotion Recognition in Conversation

- An Interaction-Aware Attention Network For Speech Emotion Recognition In Spoken Dialogs

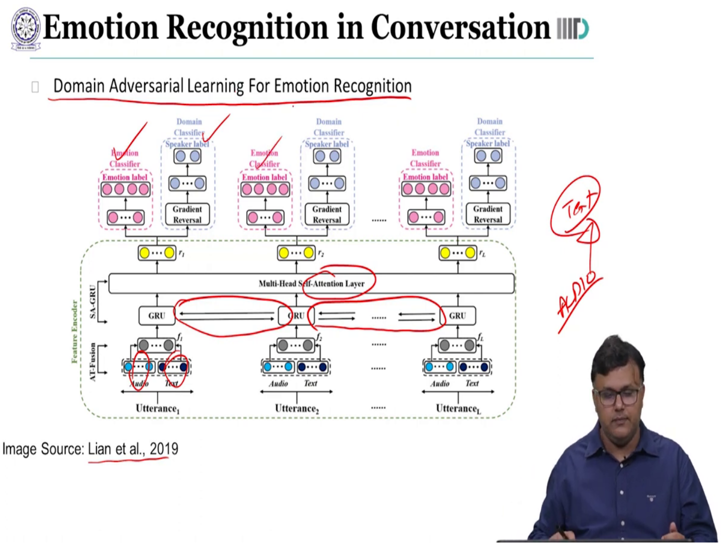
Image Source: Yeh et al., 2019

Now, here is another work by Yeh and others, this work is an interaction aware network for speech emotion recognition in spoken dialogues, alright. Now, in this what do you see here some conversation text which is there over time? So, you are extracting the utterances, here you have a GRU you know you are a current networks.

And you are also adding attention again this is coming from the work which I was referring to attention is all you need. Then you are looking at the utterances of the speakers in parallel. So, you know you have M and F let us say male speaker and female speaker and then again, a bi-directional GRU extracting the representation concatenating and then predicting the emotion.

Notice how as compared to text based emotion recognition the architectures are changing when there is a conversation. Because one person speaks, then the other person speaks, right. So, the system needs to not only analyse the content spoken by one subject at a given time, but in parallel also look at the conversation happening together as well.

(Refer Slide Time: 40:34)



Emotion Recognition in Conversation

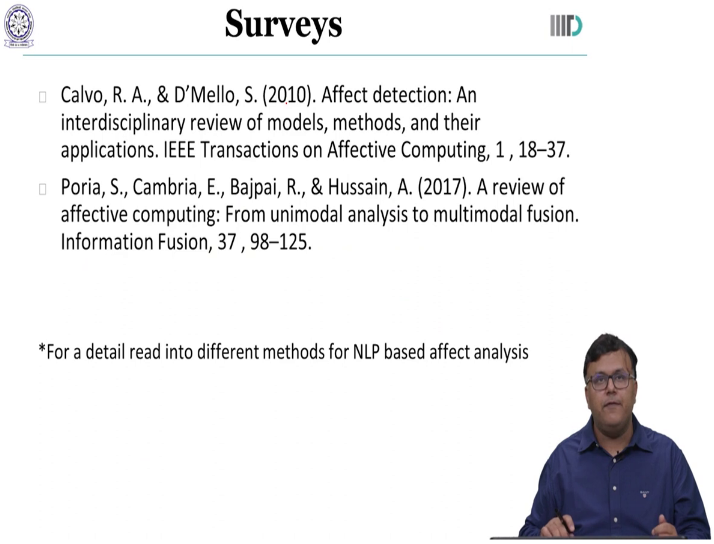
□ Domain Adversarial Learning For Emotion Recognition

Image Source: Lian et al., 2019

Here is another work friends Lian and others 2019 which is called domain adversarial learning for emotion recognition ok. What you have here is you have the utterances as input, now this is text and audio, which are coming in now you would wonder why text and audio. Well, it is possible that the conversation is happening in the voice modality you do speech to text and you get the text which was being spoken ok.

Now, notice how for the difference utterances, the conversation is being mapped for its dynamic through a GRU then you are adding again a attention layer and what we are doing is at different times, we are predicting the emotion and the speaker level was speaking emotion and speaker level and so forth.

(Refer Slide Time: 41:21)



Surveys

- Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. IEEE Transactions on Affective Computing, 1 , 18–37.
- Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion, 37 , 98–125.

*For a detail read into different methods for NLP based affect analysis

Now, friends this was a brief introduction to the very wide variety of works, which have been proposed in the literature for emotion recognition during conversations. I also invite you to look at two of the survey works. Now these are very detailed survey works which are looking at the different aspects of affect prediction using text, if you are interested in going deeper into this area.

So, friends with this, we come to the end of the today's lecture, we discussed about the different features which have been proposed in the literature for analysing the text which is essentially to create a vector representation, we started with the bag of words with representation, we talked about what is n-gram, then we looked at what is the concept of the TF-IDF and later on how we can have a bag of concept.

From this we moved on to, that how with the progress in deep neural network, we are using your representation learning in the form of pre-trained networks such as Word2Vec. From that the community has moved to attention based systems. Wherein now we are using transformer like architectures for predicting the emotion. And in the same context we moved a bit forward when let us say a conversation is happening.

When you see a machine and a person interacting or two human beings interacting, there is a very dynamic play of emotion which is happening. Therefore, the system needs to not only understand the individual persons utterances what a person said, but also look at how the relationship between the utterances that is being analysed.

Thank you.