Natural Language Processing

Prof. Pawan Goyal
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

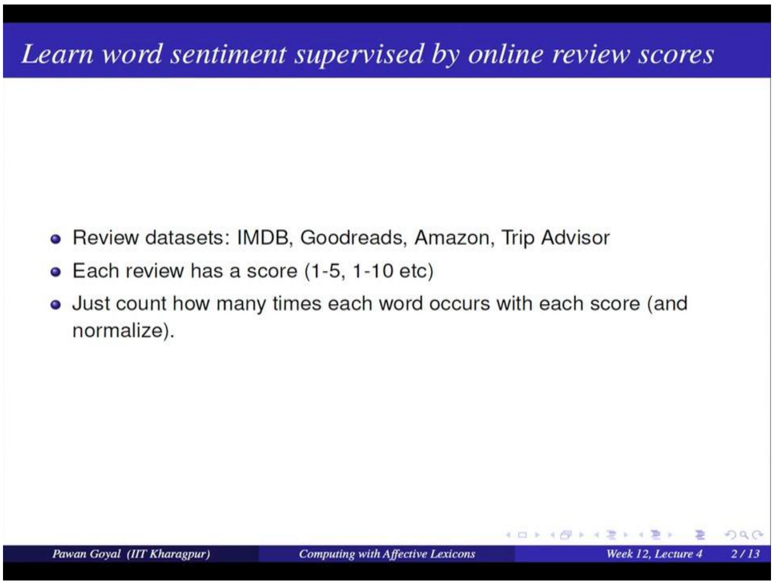
Lecture - 59
Computing with Affective Lexicons

Welcome back for the lecture four of this week. So, in the last lecture, we discussed how do you bootstrap your opinion sentiment lexicons from using some simple seed sets, from your corpus and from the WordNet.

So, in this lecture, what we will see, so we will see two things mainly, suppose you have your sentiment lexicons, and you can use that directly to up to find out the sentiment score of your different sentences, and all in your data. So, these are very standard methods for doing that you can take average and all. But what are certain things that you have to keep in mind that is what we will talk about. And then from the different data sets we are use here abundance of sentiments like review data set over movies and all, what are that nice trends you see about different words.

So, what kind of major do you should use to be able to see those trends, also to be able to find out what words are getting what words are more prominent with positive and negative emotions in those.

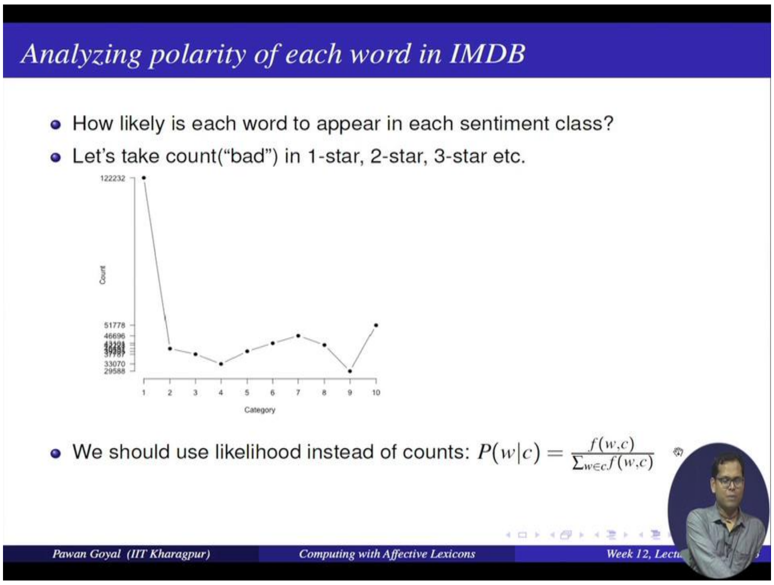
(Refer Slide Time: 01:24)



So, let us see. Suppose, I want to learn what is the sentiment score of a word by see how often does that occur in a review corpus. So, I take some corpus from online corpus say I IMDB I take all the reviews from different movies. And now I am going to see how often how much correlated a particular word is with a particular rating a score. So, how do I do that? So, it can start with any review data sets.

So, here are some examples, you can take from movies, you can take IMDB, you want to take reviews for books, you can take good reads, you want to take reviews for hotels, and all you can take trip adviser and for various products you can take Amazon or other websites. Now, on different pages, you will find different ratings, somewhere you will find 1 to 5, somewhere you find 1 to 10 etcetera. And now I want to find out how often a word occurs in a particular sentiment class a particular rating class. So, what can be a good measure for doing that?

(Refer Slide Time: 02:34)



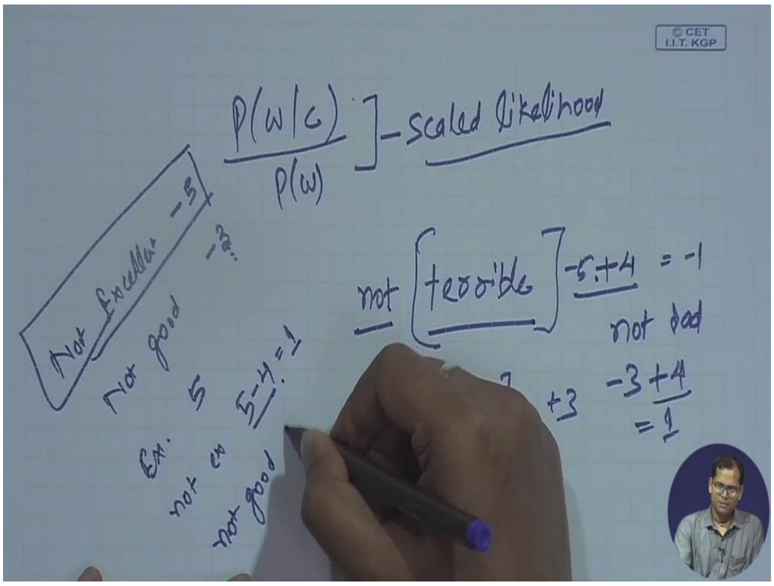
So, let us see. So, we are taking the examples from IMDB, and I want to find out I want to analyze polarity of each word in IMDB, so that is how often does that occur with different, different ratings. So, what will be the simplest measure that will come to your mind? You can see I have readings from 1 to 10, and I want to find out how often this word occurs with each of the rating 1, 2, 3, 4, 5, 6, 7, 8 times and so on and I can do like that k. So, let us count the word bad in 1 star, 2 stars, and 3 star etcetera. So, on x-axis, you can see different ratings 1 to 10; and on y-axis, you can see that count how many times a word occurs.

So, this can be a approach that can give you how many what are the counts of a word, but this will not help you in finding out some sort of seeing a some sort of normalized picture. So, for example, does this word occur in one rating more with more probability than others other ratings, so what would be a criteria. So, you want to see that it might happen that in my corpus, there are more reviews with rating 1 than rating 5, so by that simple statistics the count of this word in 1 will be higher than count in 5. So, I should be able to do some normalization.

So, first normalization I can do is what is the count of the word in a particular rating divide by count of all the words in that rating to some sort of probability of a word in a particular rating that can be the first measure that I can take. So, what is the probability of a word given a particular rating? So, we will see is noting a rating and that you can obtain by seeing the number of times the word occurs with that rating divide by all the words that occur in that rating whatever times or you can say this is the size of that rating how many different words occur. So, this will give you the probability of the word occurring in that rating. So, this can help you normalize across the ratings. So, what is the probability of this word rating 1, rating 2, rating 3 and you can see does it have a higher probability rating one and so on.

But suppose now you want to compare across words. So, this is ok for comparing across the ratings, but if you want to compare across words this may not be a good measure. And why is that because it might again happen that this word is very, very common in lexicon so that means the probability of this word is actually very high. So, by that logic again the probability of this word occurring with this review might this rating might also be high. So, I want to do another normalization that takes into account what is the probability of this word.

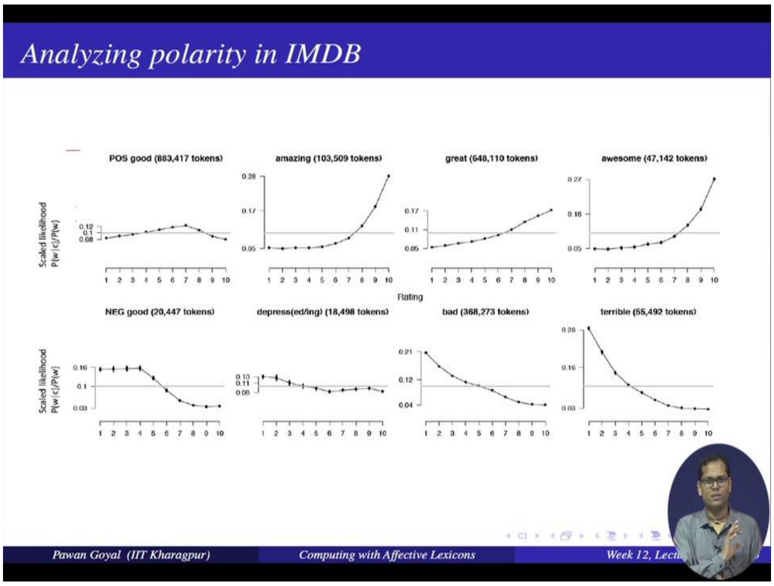
(Refer Slide Time: 05:28)



So, what can be a alternative measure. So, I already have probability of the word given this class, see there is a particular rating class. Now, I can divide it by the probability of the word is k itself and that will be a normalized measure. So, now it will be comparable across different words. So, I am seeing what is the probability of the word in the corpus divide in the particular rating divide by what is the probability of word overall. So, you can also think it like how much does that depart from its actual probability; if this is very different from its actual probability that means, yes, there is a very, very it is a nice indicative of a particular class.

So, this is called likelihood and this together is called as scaled likelihood. So, we will be trying to compare words across categories and different words by using their scaled likelihood. So, how do they occur in different ratings with its scaled likelihood? So, by that I am making them comparable across words by taking probability word given the class divide by probability word.

(Refer Slide Time: 06:45)



Now, let us take some example from IMDB data set. So, what we are seen on x-axis, we have different ratings 1 to 10; and on y-axis, we have the scaled likelihood. And it is a scale, so it will not go beyond 1, so it has to be between 0 to 1. So, what are you seeing here. So, let us see the first column positive good and negative good. So, what we are seeing as you go from rating 1 to 7 positive good is increasing so that means we are talking good without a negation it is increasing, but when you are from 7 to 10, it is decreasing. Negative good, it is high initially and then it is decreasing.

So, now can you make sense of that? So, when you are going to higher reviews initially good in increasing and then it is decreasing. And this is actually this might be the case why because yes when you go to in review of rating 1, you might not have much good return, but when you are going to higher review, yes, there is good. But when you go to even higher reviews, you might not be using a word like good, you might be using a better adjective like amazing and so on, fantastic.

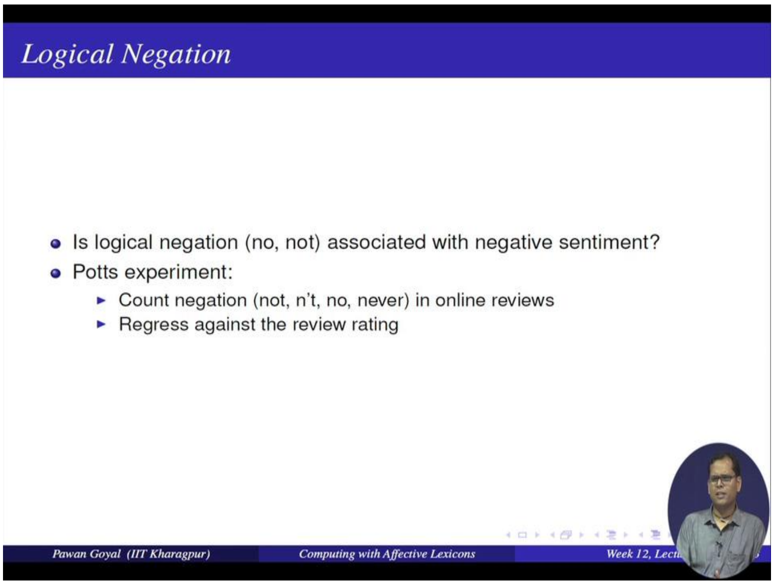
So, you are using adjectives like that, so that is what you will see in the in the next columns. So, you take a adjective like amazing and you will see in reviews 1 to 5, it is nearly very low 0.05. And then start shooting up when it goes to 0.28 in the reviews of rating time that means, the word amazing occurs a lot in the reviews of rating time by the scaled likelihood and this give a nice picture.

Now, you talk about negative good adverse. So, adverse, decreasing, it was decreasing as you go from high reviews. So, in review of 10 they want be much occurrence of not good. But now if you talk about terrible, or depressing; so, you have the word depressing depressive. So, you see that initially in rating 1, it was occurring with a high likelihood - the scaled likelihood; and as you go down, it keeps on decreasing. So, again this looks this pattern looks interesting.

Then you see the word like great it is not as steep as amazing, but again this is nice trend; it is starts increasing from 1 to 10, from 0.05 it goes up to 0.17 that similarly starts from 0.21 very high likelihood in reading one it goes up to 0.04 as you go to rating time. And similarly if you see awesome and terrible, they have similar trends as amazing. So, awesome is very similar to amazing, yes, occurs a lot in rating 10, you see the numbers also nearly same 0.28, 0.27.

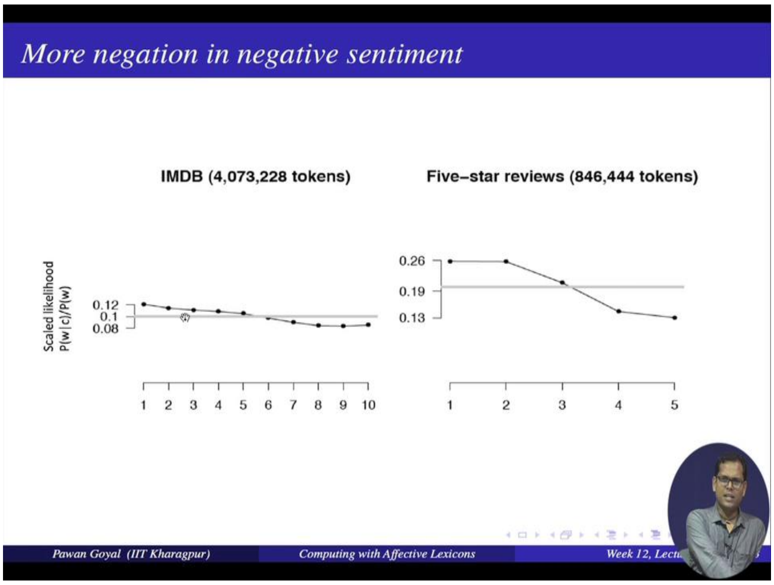
And terrible is like very, very high in rating 1, 0.28 and goes to 0.03 at rating trend. So, this is how you can now compare different words that how likely do they occur with reviews of rating 1 2 and 10 and you can try to compare which words behave in a very similar manner. And you might even be able to use this method to find out by using a review corpus which words are actually positive, which words are negative. So, you can use it for the task that we discussed in the last lecture also.

(Refer Slide Time: 10:01)



Now, you can see some other trends also. So, how often, so negation is a big problem, so is it that whenever you had negation in the sentence does it always conveys a negative sentiment. So, is negative below is logical negation associate with the negative sentiment. So, Potts said this experiment where they found out how many times negation like not, no, never were occurring in online reviews, and then they were trying to see its occurrence with the review ratings. So, how often do they occur with different review ratings?

(Refer Slide Time: 10:40)



Again they did took the scaled likelihood for IMDB and five-star reviews. And they found a very hyper sensitive trend here that in review of rating 1, it had a very high scaled likelihood 0.12. And as you go down up to 10, it had a 0.08, has not as different as you were seeing in terrible or awesome, but you can instant say that the negation is occurring more in the reviews with rating 1 than reviews in rating 10. And same thing they saw with other reviews were that were on five-star rating. So, it was occurring with 0.26 likelihood with review 1, review of rating 1 and a likelihood of 0.13 with the review of rating 5.

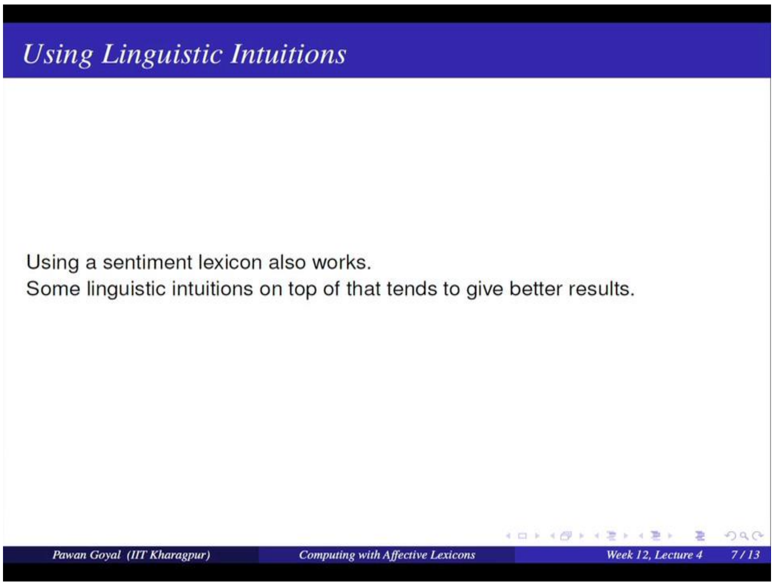
So, now, suppose I want to use this lexicon, I have a sentiment lexicon, I want to find use that to find out what is the sentimental score of a sentence in or a paragraph in my corpus. So, this will help I can just take the sentiment score of each of the words that is a very simple based on algorithm, you find out the sentiment of each word, add, take an average things like that and that will give you is score to the whole sentence. So, what I am trying to show here is that if you some linguistic intuition on top of that that might give you a better result.

So, for example, one particular problem with this is negation. So, what do I do, I have in my lexicon word like terrible and word like good; I know terrible is very bad like minus

5 and good is say plus 3. If I get a negation here, not terrible, not good, what do I do I add the polarity of not with the polarity of terrible; similarly polarity of not with good that will not be a good approach. So, I should understand that these are some sort of function words they have some ten functions. So, can I take it as a function over this, a function over this? A simple function that you can think is just reverse the polarity whenever there is not reverse the polarity; terrible is minus 5, not terrible becomes plus 5, but is that a good approach, so that we will see.

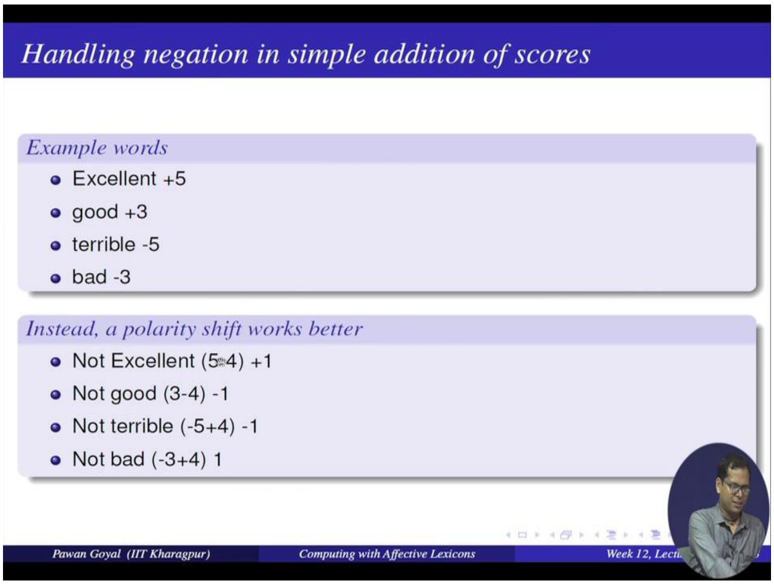
Secondly, there are some other words that are also like function words. So, you can say so well word like very. So, you say good has some polarity, now you attach very to that very good what be the polarity and some words can be somewhat so then what will the polarity somewhat good. So, good has some polarity how do you give a polarity to the somewhat good. So, like that you can use some linguistic conditions to make much more science from your lexicon and avoid doing some mistakes with all these different function words.

(Refer Slide Time: 13:42)



So, you can use some linguistic conditions on top and that can give you some better results.

(Refer Slide Time: 13:47)



So, let us see. So, I have these words like excellent with plus 5, good with plus 3, terrible with minus 5, and bad with minus 3. So, one as we said one simple thing as we can do is when you have negation is that is just can just reverse the polarity. So, if you reverse the polarity that is what we get not excellent minus 5, not good minus 3, not terrible plus 5, and not bad plus 3. Now, just have a look at that for a second and see if that make sense. So, we are saying not excellent is minus 5 and not good is minus 3.

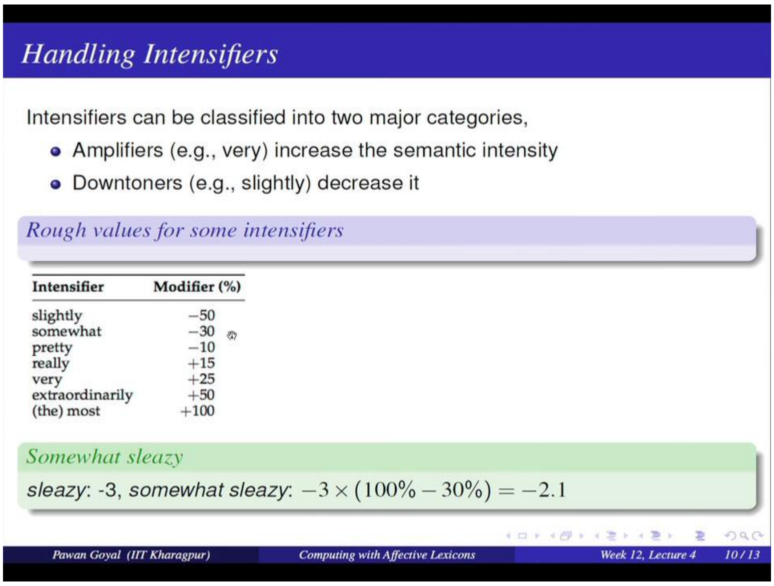
So, just think about it when I say excellent I mean this really good; when I say not excellent do I mean like terrible, if I say not excellent; that means, it is not excellent, but it may it is good, it is good, but not excellent. So, if I have to say it is not good, I will say not good. So, not excellent means something in the positive polarity. So, completely reversing the polarity will not be a good idea here. So, changing it from plus 5 to minus 5 will actually be a mistake.

Similarly, so not good is, but again you what you are doing not good is getting a better score than not excellent that is again not ideal. You want to give a lower score to not good than not excellent. Same thing you can think about to the other two words. So, you have not bad, not bad can mean something that is going towards good, but when you say not terrible you also you still mean that it is bad. So, not terrible should have a bad should have a negative score and much less than not bad, what is happening the reverse here. So, you seen not terrible is getting a higher score than not bad, this is not ideal.

So, you should have just reverse in the polarity, we can do something else and this is called you do something like a polarity shifting. So, if it is minus 5, you shift the polarity, so you add say plus 4 to bad. So, terrible is minus 5 add plus 4 you get minus 1, not terrible minus 5 minus 1 is ok. You do the same to the not bad. So, not bad will be minus 3; for bad plus 4 is equal to 1, so not bad starts getting good. So, you are getting a positive sense polarity.

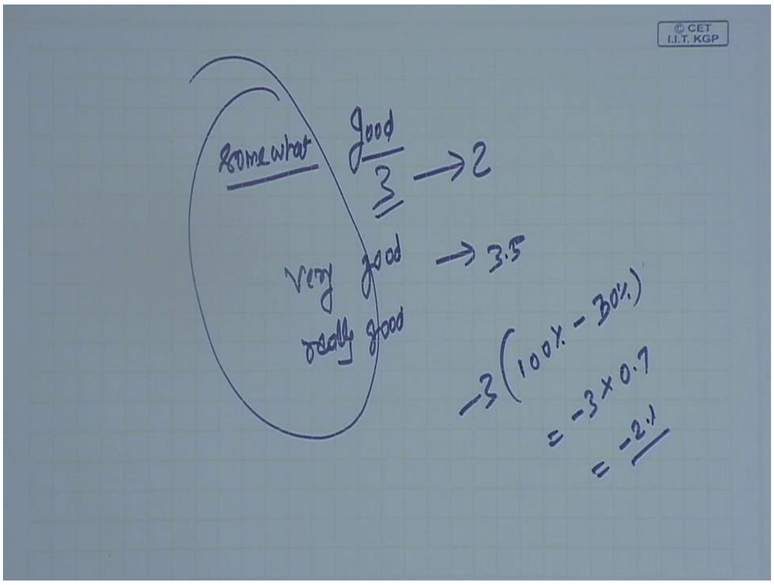
Same thing you do with excellent, excellent is 5, and you say not excellent shift polarity shift to so you say 5 minus 4 here you were doing plus four here you are doing minus four this is just a number you can change this and this will give you 1. So, you still have a positive score. And not good, good has 3, you shift the polarity, you get minus 1, you get a negative polarity and that would be a much better approach than simply reversing the polarity. So, this is like linguistic intuition that you can use. So, here is the polarity shift and something like that some 5 minus 4 gives you plus 1, 3 minus 4 gives you minus 1 and you can see not good has a more negative sentiment than not excellent same you can see with not terrible and not bad.

(Refer Slide Time: 17:08)



Similarly, how do you handle various intensifiers like very, somewhat etcetera? So, what do you see here, so you can have some amplifiers that can increase the intensity, and then there are certain downtoners like slightly that slightly that can decrease it.

(Refer Slide Time: 17:28)

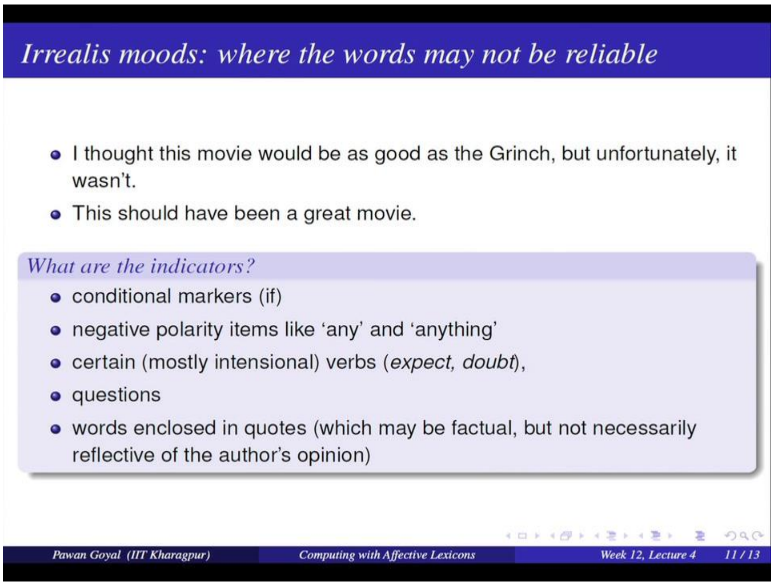


So, what do I want to do suppose you know the polarity for good, you know this score for good. Now, if it occurs with somewhat, somewhat good, suppose it is 3, you want to reduce this right somewhat good. So, you want to take from 3 to say 2. On the other hand, if it occurs with say very good or really good, so you want to increase that, so say 3.5 when we use 3, so very good might be 3.5.

So, again these are acting like some function words that can act as a function to modify your sentimental score of the main word. So, again in linguistic you can find outs and inform words in a paper. So, they had given some scores like slightly you do minus 50 percent, somewhat you do minus 30 percent, pretty minus 10 percent, really plus 15 percent. So, they are now an amplifier very is plus 25, extraordinary plus 50, and the most becomes plus 100.

So, you make these modifiers and how do you do computations. So, suppose you have a sentence like this somewhat sleazy. So, sleazy is minus 3. So, when you do somewhat sleazy, you will say ok, so what is this score of sleazy sorry somewhat minus 30 percent. So, you say minus 3, 100 percent minus 30 percent, so that will give you minus 3 into 0.7 it will give you minus 0.21, like that you can give a score with these downtoners or amplifiers. So, this is again some linguistic intuitions that can be used.

(Refer Slide Time: 19:02)



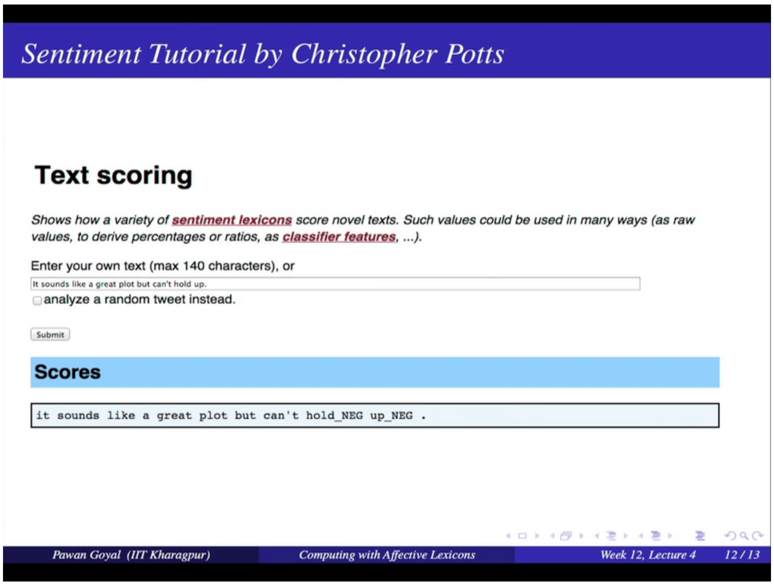
Then many times you have to be careful with specific sentences. So, like if people are using some irrealis moods. So, like I thought this movie would be as good as the Grinch, but unfortunately, it was not. So, here the author is saying I thought this movie would be as good as the Grinch. So, all the way there is a some positive words this say irrealis moods. So, you are saying I thought that this what happened, but this did not happen. So, this is again, so here you cannot just lie on your sentiment scores that are given in your lexicon.

Similarly you see a sentence like this, this should have been a great movie right. Again if you just use your sentiment lexicon you will say it is a positive polarity, but this is again a irrealis moods and you cannot rely on the words directly. So, you should have some way of finding out its irrealis moods or it will either probably change my polarity or I use some something else. So, specifically you need to be careful about using conditional markers if that was the case, then this would have been good then and so on.

Then something like any or anything, certain intentional words like I expected, I doubt whenever this occurs, you might not rely on what follows. Similarly, if there are questions in the sentence then also you should not be able to fully rely on the words. And then this is important many a times when you do the sentiment analysis over the corpus like news corpus. So, what do you see you will find various quotes? So, he said something, so now you cannot assign a sentiment to a sentence by whatever is there in the quotes because he just reporting some other sentiment, but actually the author did not have any sentiment. So, you are reporting some sentiment.

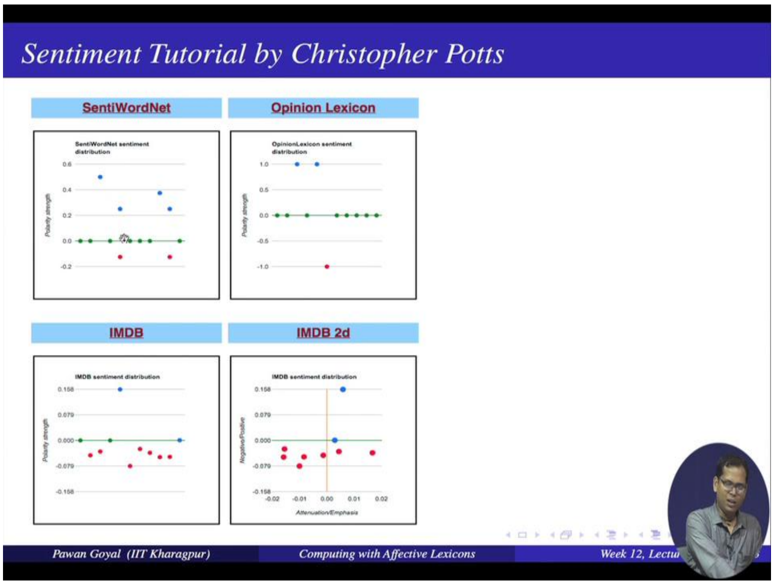
So, if you want to find out the sentiment of the sentence, it does not have any sentence. So, you should not use the sentiment of the quotes to give the sentiment of the sentence. So, the quotes also you should have to be careful with.

(Refer Slide Time: 21:03)



So, Christopher Potts has a very nice sentiment tutorial. So, you can just search for sentiment tutorial by Christopher Potts. And there you can also try different sentences and see how the tokenization is done. So, like it sounds like a great plot, but cannot hold up. So, you see the negation is coming with hold and up also. And further it tells you how different lexicons try to assign a score to this.

(Refer Slide Time: 21:28)



So, you will see how sentiwordnet deals with it, how opinion lexicon deals with it, how IMDB deals with it and so on. So, you see in sentiwordnet for different words, you find out what whether they are positive polarity or negative polarity. So, you can directly find out by using this tutorial website. So, and there are many other resources that will be helpful for you on that website itself. So, that was about this lecture that how do you compute with a affective lexicons what kind of nice trends you can see about the words occurring in the review data sets.

So, we will end this week, and also the course in the next lecture. So, we will take one another application. So, you just try to give hints on and what is the aspect based sentiment analysis. So, till now what we are doing, we are giving a score to the sentence this is positive or negative sentiment. But suppose it is not like the whole I am saying positive about the whole hotel as such, I am writing hotel review, I may not be saying positive about the hotel or negative about the hotel, I might be saying about certain aspects of that may be the service was good, but the room may be the food quality was good, but the room was dirty.

So, I might be saying some positive about one aspect, but negative about the another aspect. So, there some simple methods of capturing those that would you will see in the next lecture.

Thank you.