

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

Lecture - 56
Sentiment Analysis – Introduction




Welcome back for week-12 of this course. So, from last two weeks, we have been talking about lot of different applications in NLP and what kind of methods you can use for solving those applications. And we have discussed applications about text classification, text summarization, entity linking and functionality extraction. There are many other applications where of NLP that you can think of. And once you have learnt all the basic concepts, so you have to just see how what kind of algorithms you can apply for newer applications.

So, this week we will focus on one of the very, very interesting application that is of sentiment analysis and we will cover that from many different angles. So, sentiment analysis is a field that has been become very, very popular in the last decade; and because of the social media coming in and lot of opinions and comments that that people are putting online. So, it becomes important to understand what are peoples sentiment towards a particular product; a particular movie, a particular candidate in politics and so on. And can one do that by simply analyzing, the sort of text data that I find online and also on various different resources.

So, can I use that to do some sort of sentiment analysis. And this can later use for various predictions that how people sentiments will move in future and may be what will the market of particular product or a movie and so on. You can also do some sort of comparisons between products, people and so on. So, we will see what this field of sentiment analysis enters? And what are some of the things that will help you on taking up any new challenging problems sentiment analysis? So, this is what can be some sort of introduction to sentiment analysis that you have some sort of movie reviews and each reviews either saying ok this movie is very positive talking positively about the movie or the negatively about the movie.

(Refer Slide Time: 02:28)

Example: Positive or negative movie review?

-  • unbelievably disappointing
-  • Full of zany characters and richly applied satire, and some great plot twists
-  • this is the greatest screwball comedy ever filmed
-  • It was pathetic. The worst part about it was the boxing scenes.

Pawan Goyal (IIT Kharagpur) Sentiment Analysis - Introduction Week 12, Lecture 1 2 / 16

So, let us see some of the movie reviews here. So, it says unbelievably disappointing, and you may say it is this is negative review. Another way is ones is full of zany characters and richly applied satire, and some great plot twist, and you will say this is a positive review. This is the greatest screwball comedy ever filmed, you will say again as a positive review. It was pathetic. So, you see this becomes a negative review. So, you are having lot of different reviews and you are seeing which one is positive, which one is negative and so on.

(Refer Slide Time: 03:04)

The slide is titled "Where is Sentiment Analysis Used?" in a blue header. It lists five applications of sentiment analysis, each with a domain name in italics and a question in a standard font:

- Movie* Is this review positive or negative?
- Products* What do people think about the new iPhone?
- Public Sentiment* How is consumer confidence? Is despair increasing?
- Politics* What do people think about this candidate or issue?
- Prediction* Predict election outcomes or marked trends from sentiment

At the bottom of the slide, there is a footer with the text "Pawan Goyal (IIT Kharagpur)", "Sentiment Analysis - Introduction", "Week 12, Lecture 1", and "3 / 16".

What are the different places where you can use sentiment analysis in general? So, we took some examples from the movie domain. So, from the way people write reviews about the movie, you can say that whether the movie was the reviews are positive or negative, but you can talk in general about many different things. For example, you can say in terms of products. So, when you buy products online, there is also a field where you can put your reviews, you can put your reviews about the product.

And there are again one can capture whether about the product are using something positive or using something negative and so on. You can also go for the down saying this product has multiple aspects. So, this product like camera, you can talk about whether it is very heavy camera, whether or how good image, what is the image quality and so on. So, you can talk different aspects about the product; and each aspect you can say whether the opinion of the sentiment of the people is positive or negative.

Then you can enjoy talk about the public sentiment that is it can be over a particular brand name. So, on that brand, what is the consumer's confidence, what is the consumers reactions, it can be in general about the policies that government is making that what is the confidence of the people in that policy and what is their sentiment and so on. You can do it for the politics that is whenever there are some elections, some candidates are there.

So, what are the people's opinions and sentiments about different sort of different candidates? And this is again a very, very interesting field where you can say ok is it just positive negative or you can you can go further say ok. What positive things are people talking about a particular and candidate what negative things people are talking about this candidate and you can also do some lot of summarization on top of that. So, there is a lot of different things that you can do about this field of sentiment analysis.

And by doing all that you can also do some sort of prediction that is once I know the peoples sentiments about different political parties or candidates, can I predict the outcome of the election in advance. And you will see in the recent years, this has being done in a lot and lot to by many different sorts of researchers and companies. So, they are trying to predict the outcomes of elections, whether it is US election, whether it is elections in India and so on.

Can you predict the market trends from sentiments? So, once you know the sentiment of the people, can you predict which sort of product will have a high market; you can also talk about the stock exchange about what stocks will go up, what stocks will go down and so on. So, there are lots of different interesting things that can be done using sentiment analysis.

(Refer Slide Time: 05:57)

Where is Sentiment Analysis Used?

- Frustration of callers to a help line
- Stress in drivers or pilots
- Depression and other medical conditions from social media
- Confusion in students talking to e-tutors

Pawan Goyal (IIT Kharagpur) Sentiment Analysis - Introduction Week 12, Lecture 1 4 / 16

Then there are other things like your help line and you want to see whether the scholars are somehow distressed or the frustrated and so on by doing sentiment analysis over their speeches. Then from the drivers and pilots, she want to find out the stress labels there that again it become important to avoid various accidents; so there also sentiment analysis used. Although their modality might be different, you might not be able to capture it by using simply the text data you might have look at their speech or some other sort of facial expression and all. So, in this course, and in this field we will only talk about taking it from the text data.

Then so that is why the other two kinds of applications become interesting that is the way people are using social media, so by the way they are writing over social media by the way they are making the connection with people, can we tell something about their health conditions like are they depressed, and they are works in recent times that are explore this area that is. Can you tell that by the way the tweeting behavior of a person, whether he is depressed, whether he is alcoholic and so on? And then when people are using various e-learning tools they are talking to e-tutors are they understanding the topic or are they having some confusions.

So, these sentiments can also be explored by using sentiment analysis. And you can think of hundred different scenarios, where this sentiment analysis will be used.

(Refer Slide Time: 07:51)

Affective States Typology

Emotion: angry, sad, joyful, fearful, ashamed, proud, elated

Mood: cheerful, gloomy, irritable, listless, depressed, buoyant

Interpersonal stances: friendly, flirtatious, distant, cold, warm, supportive, contemptuous

Attitudes: liking, loving, hating, valuing, desiring

Personality Traits: nervous, anxious, reckless, morose, hostile, jealous

Pawan Goyal (IIT Kharagpur) Sentiment Analysis - Introduction Week 12, Lecture 1 5 / 16

So, before going on to the how we do sentiment analysis let us discuss, what all comes in this broad field of sentiment, opinions, subjectivity what all can you think about. So, this is by this affective states typology. So, what are all the different things you can talk about this broad field? So, you can see talk about emotions right, emotions of something different. So, emotions will be are you angry, are you sad, joyful, fearful, ashamed, proud, elated all these are emotions.

Then you can top talk about the moods. So, whether the person is cheerful, he is gloomy, is it irritable, listless, depressed, buoyant. So, all these are different moods of people. Then one can talk about the interpersonal stances that are whether he is friendly, whether he is warm, supportive and so on. You can talk about attitudes – liking, loving, hating, valuing, desiring. And you can talk about the personality traits like being nervous, anxious, and reckless. And you will seen different, different applications, you might want to talk about different sort of states somewhere you want to talk about the emotions, somewhere only the interpersonal stances, we are talking about a group how people are behaving in a group.

So, if you talk about the social media and how people are reacting toward ho how peoples comments are, you will see we are mostly talking about the attitudes of the

people that is are they liking a product or they loving it or they hating it same about the movie or the candidate. So, when we talk about sentiment analysis, we mostly focus on these finding of this attitudes of people towards certain object or a person.

(Refer Slide Time: 09:28)

Sentiment Analysis

Sentiment Analysis is the detection of attitudes
enduring, affectively colored beliefs, dispositions towards objects or persons

The complete task

- Holder (source) of attitude
- Target (aspect) of attitude
- Type of attitude
 - From a set of types: *like, love, hate, value, desire*
 - Or simple weighted polarity: *positive, negative, neutral, together with strength*
- Text containing the attitude

Pawan Goyal (IIT Kharagpur) Sentiment Analysis - Introduction Week 12, Lecture 1 7 / 16

So, this is some sort of definition we can give that sentiment analysis, the detection of attitudes that is some sort of enduring our some sort of effectively colored beliefs, towards objects or person. So, now in general sentiment analysis can be a very, very complex task also. So, because when you are talking about sentiment analysis, there are many, many different things in picture. So, you are talking about the person who is the person who is holding the attitude, who is the person, who is liking it that is the holder of the attitude; and liking what, what is the target of this attitude. So, these are two important things who is the holder who is the target.

Then you can talk about what is the type of attitude. So, again here you can choose from this wide range, you can say I have to choose one of the like, love, hate, value, desire and so on. Or simply you can say what is the polarity is it positive, is it negative, is it neutral or you can give some sort of strength is positive with that much of strength, negative with this much of strength and so on. So, again you see there is a lot of flexibility in how you can define your task. And you can some might be more complex than the other.

And you might also want to find out what is the text in this whole whatever attendance you have taken that contains that attitude. What part of text contains that attitude among this whole sentence or paragraph? So, this might be the complete task. Now, you might just want to use one of this task and not the other, one of the particular things like I want to just say where this whole sentence is positive or negative I do not worry what part is positive or negative. Or I want to say who is having the opinion and towards whom and that is all I want to do not want to go further into detail. So, you can choose which of what part of this whole task, you want to compute to for your particular application.

(Refer Slide Time: 11:29)

The slide is titled "Sentiment Analysis" in a blue header. It lists three levels of sentiment analysis tasks in light blue boxes:

- Simplest Task**: Is the attitude of this text positive or negative?
- More complex**: Rank the attitudes of this text from 1 to 5
- Advanced**: Detect the target, source, or complex attitude types

At the bottom, there is a footer with the text: "Pawan Goyal (IIT Kharagpur) Sentiment Analysis - Introduction Week 12, Lecture 1 8 / 16".

So, simplest task would be you are given a text and is the attitude positive or negative that is the simplest task. And you will see most of the works are done for this simple task. So, and if you want to make it more complex, you will say ok, all these attitudes are there can I give them some sort of score. So, I score them in the range of 1 to 5, 1 might mean very negative and 5 might mean very positive; it can be from minus 5 to 5.

So, they are different scales on which you can give the ratings 1 to 5, 1 to 10 and so on. You will see when you give rating to movies or the products on different sites; you have ratings from 1 to 5 or 1 to 10 and so on. So, again you might want to predict the ratings on that that is skilled from the text and yes, then it can be more advanced where you want

to detect who everything target, source, and the complex attitude types. So, you can see that definition can vary. It can be simple, more complex or advanced sentiment analysis.

(Refer Slide Time: 12:37)

Sentiment Analysis in Movie Reviews

Polarity detection
Is an IMDB movie review positive or negative?

| | |
|---|--|
| | |
| <p>when _star wars_ came out some twenty years ago , the image of traveling throughout the stars has become a commonplace image . [...] when han solo goes light speed , the stars change to bright lines , going towards the viewer in lines that converge at an invisible point . cool . _october sky_ offers a much simpler image—that of a single white dot , traveling horizontally across the night sky . [...]</p> | <p>“ snake eyes ” is the most aggravating kind of movie : the kind that shows so much potential then becomes unbelievably disappointing . it's not just because this is a brian depalma film , and since he's a great director and one who's films are always greeted with at least some fanfare . and it's not even because this was a film starring nicolas cage and since he gives a brauvara performance , this film is hardly worth his talents .</p> |

Pawan Goyal (IIT Kharagpur) Sentiment Analysis - Introduction Week 12, Lecture 1 9 / 16

So, let us see some example. So, you are given IDMB, you are having two different reviews and you want to find out whether they are positive or negative. So, here are two reviews. So, on the left hand side, so you have about when a star wars came out some twenty years ago, the image of traveling throughout the stars has become a image and then you find something like cool. And this is a movie October sky it offers a much simple simpler image white dot traveling horizontally across the night sky.

And this overall look gives a positive picture. We have the word like cool here and this gives a positive picture. And then if you see the one on the right hand side, and you will see immediately there is a word like this is unbelievably disappointing, and this film is hardly worth his talents and so on and you will see this become a negative review. So, from this whole text, you want to find out whether this is positive review or negative review.

(Refer Slide Time: 13:38)

Baseline Algorithm

- Tokenization
- Feature Extraction
- Classification using different classifiers
 - ▶ Naïve Bayes
 - ▶ MaxEnt
 - ▶ SVM

Pawan Goyal (IIT Kharagpur) Sentiment Analysis - Introduction Week 12, Lecture 1

So, now what are the some simple algorithm that you can apply. So, if you think about it for a while you can say this looks like a classification problem. So, we did classification just in the last week. So, you are having different text, and you want to classify them into either positive class or negative class; whether this is positive or negative. And we know we can use a lot of different models of classification, I can use SVMS, I can use logistic regression, I can use Naïve Bayes model that we talked about. So, I can use any of these models.

But before doing that I will probably have to see given the text do some sort of tokenization, find out what are the words here, and I extract what are the important features that I can use for my classification tasks. So, this has to be done before running any of this algorithms. So, what is important is that can we do something that is specific to this task without about classification problem in general. And we know how to do all these for any given text. We will need to have some labels and we will learn Naïve Bayes classifier. But what is important for this task and that is something that you have to keep in mind when we are dealing with any application. So, there are some default set of methods that you can use for that application, but you have to see is the something is specific out of this application that you should use.

So, let us talk about tokenization right; we discussed tokenization in the very first week of this course. So, I need to find out what are the different words here. And there I do some sort of preprocessing. So, what are some sort of preprocessing. So, I do lower casing and I correct the spellings and so on. And if there are words that auto vocabulary and might also remove them. So, there are many things I can do. But when we are talking about sentiment analysis do you needs do something different, nothing about the tweets

(Refer Slide Time: 15:50)

Tokenization Issues

- Capitalization - preserve for word in all caps
- Word lengthening
- Handling emoticons


```

[<>]?           # optional hat/brow
[:]=8]          # eyes
[\\-o\\*\\'\\?  # optional nose
[\\)\\)\\(\\[dDpP/\\:~\\}\\{\\@|\\|\\] # mouth
|              ### reverse orientation
[\\)\\)\\(\\[dDpP/\\:~\\}\\{\\@|\\|\\] # mouth
[\\-o\\*\\'\\?      # optional nose
[:]=8]          # eyes
[<>]?           # optional hat/brow

```
- Handling negation
 - I **didn't** like this movie
 - I really like this movie

Add **NOT_** to every word between negation and following punctuation

 - didn't like this movie, but I ...

Pawan Goyal (IIT Khargpur) Sentiment Analysis - Introduction Week 12, Lecture 1 11 / 16

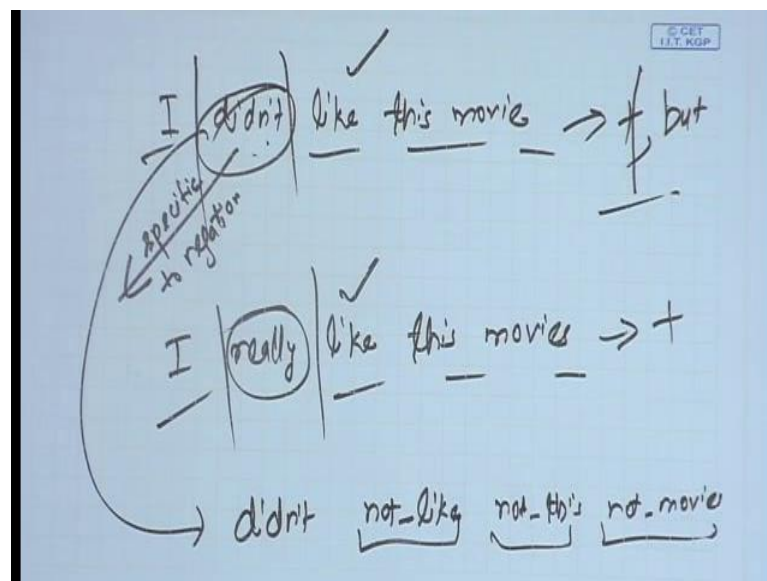
So, let us see what would a capitalization. So, suppose you are having a comment where the person is writing in all caps. So, what can you say immediately about that thing. So, whenever you are writing in all caps, in general that might indicate some sort of sentiment. So, you write cool in all caps, probably you are very, very happy about it or you are liking this. So, you want to preserve that this word occurred in all caps, you want to preserve that information, and you did not want to leave loose that information by doing lower casing. Then again you might have some lengthening of the words right. So, you are writing cool and you are writing five or six o in between c and l. So, you want to preserve that you are doing a lengthening here and this lengthening can again mean some sort of sentiment.

Then one interesting thing that comes is the emoticons that we never worried about, but

in sentiment analysis emoticons are a very good indicators of the sentiment of people emoticons can be either in comments when people are chatting to each other when they are tweeting. So, are they are they some easy way of caption the emoticons, so yes, they all come from some sort of regular expressions. So, if you can build some regular expressions and that are actually available with us

So, here are some example of regular expressions you can use and you can detect this contains a an eyes, optional nose, mouth, reverse orientation and so on, and each of this emoticon has some sort of sentiment. So, you can associate sentiments with each of these emoticons. Then another important thing is handling negation in sentiment analysis.

(Refer Slide Time: 17:54)



So, let us take these two examples. I did not like this movie versus I really like this movie. So, what will happen when you are doing the applying the standard Naïve Bayes model, so you will have this I did not like this movie, and then you have I really like this movie. So, what will happen in these sentences everything else is same except one word, and may be it can happen that because you are having a like you get a positive sentiment in both, yes, if you are doing a simple method you are applying a simple method by tokenizing taking each word and running a classifier.

So, question is can we do something a specific to this negation. And while there are many different ways people handle that one interesting idea is you apply not to all the words that are following this. So, you say ok convert that to I have did not then not like. So, it becomes a different token right like is one token and not like is another token, not this and not movie something like that. So, you append not to all the words that are following it. And you do that until there is punctuation. So, like here it can happen that after punctuation I am starting something. But so you do not want to carry forward this negation to after punctuation see you want to stop it here at punctuation. So, this is one way you can handle this.

So, what I want to convey is that although there are some standard things that we have talked about in this course, you do this lower case, you do this stemming, you do this and that, for a particular application you might have to deal with this differently. See you want to preserve here what are the upper cases so and what are the lengthening and what are the for negation you have to do something different find out the emoticons take them as an important features and so on. So, you have to do something a specific for this application. See, you have to look like this did not, not like, not this, not movie.

(Refer Slide Time: 20:05)

Naïve Bayes: Reminder

$$c_{NB} = \arg \max_{c_j \in C} P(c_j) \prod_{x_i} P(x_i | c_j)$$

$$\hat{P}(c_j) = \frac{\text{doc} - \text{count}(C = c_j) + 1}{N_{\text{doc}}}$$

$$\hat{P}(w_i | c_j) = \frac{\text{count}(w_i, c_j) + 1}{(\sum_{w \in V} (\text{count}(w, c_j)) + |V|)}$$

Pawan Goyal (IIT Kharagpur) Sentiment Analysis - Introduction Week 12, Lec 1

So, Naïve Bayes you remember we did in the last week itself. So, we will we will build a

classifier suppose our classes may be two or three positive negative and neutral and this is how you build a classifier. And you can find the probability of each class by from your training data you can find the probability of each feature in the class features can be ensemble words or something else again from trained data using this maximum likelihood estimate with add one smoothly.

(Refer Slide Time: 20:37)

Boolean Multinomial Naïve Bayes

- First remove all duplicate words from a test document d
- Then compute NB using the same equation

$$c_{NB} = \arg \max_{c_j \in C} P(c_j) \prod_{x_i} P(x_i | c_j)$$

Puwan Goyal (IIT Kharagpur) Sentiment Analysis - Introduction Week 12, Lecture 1 13 / 16

You can remove the duplicate words and you can use Naive-Bayes model to obtain what is the class, optimal class as per Naive-Bayes algorithm.

(Refer Slide Time: 20:51)

A piece of cake?

Is a given review on a known topic positive or negative?

"It may be a bit early to make such judgments, but Battlefield Earth may well turn out to be the worst movie of this century." (Elvis Mitchell, May 12, 2000)

don't we just need to look for "worst", "best", "love", "hate", etc.?

Pawan Goyal (IIT Kharagpur) Sentiment Analysis - Introduction Week 12, Lecture 1


So, this is a simple way, but what else people have tried and mind you this symbol Naive-Bayes works sometime, but it does not work with the very good precision, and there are many issues why it does not work. So, we will see some of those. So, as such if you talk about sentiment analysis, so it looks a simple problem right. I just have to so let us take this particular review it may be bit early to make such judgments, but Battlefield Earth may well turn out to be the worst movie of this century.

So, I read this review. And does that look very easy to find sentiment you say ok, yes, I see the worst movie and the worst is sufficient to tell me that this is a negative review. So, I can immediately assign it a polarity of negative, it looks easy. So is not it like I have to just see these words worst, best, love, hate in the reviews and then I can assign a sentiment is that as is.

(Refer Slide Time: 22:01)

In a small scale experiment (Pang et al., 2002)

| | Proposed word lists | Accuracy |
|-------------------------|--|----------|
| Human 1 | Positive: dazzling, brilliant, phenomenal, excellent, fantastic Negative: suck, terrible, awful, unwatchable, hideous | 58% |
| Human 2 | Positive: gripping, mesmerizing, riveting, spectacular, cool, awesome, thrilling, badass, excellent, moving, exciting Negative: bad, cliched, sucks, boring, stupid, slow | 64% |
| Statistics-based | Positive: love, wonderful, best, great, superb, beautiful, still Negative: bad, worst, stupid, waste, boring, ?, ! | 69% |



Pawan Goyal (IIT Kharagpur) Sentiment Analysis - Introduction Week 12, Lecture 1

So, it turns out it is not that easy. So, then experiment was reported in Pang et al 2002. So, what they did, so they took a corpus of again I think some reviews, and they asked the people before showing the corpus, if they can come up with a list of positive words and negative words. And I will label a review as positive, if it contains a word positive; negative if it contains positive negative of the majority of both. And people then came up with different, different lists.

So, for example, so human one came up with this list like dazzling, brilliant, phenomenal, excellent, fantastic and so on; and negative suck, terrible, awful, unwatchable, hideous. Human two came up with this list of gripping, mesmerizing, riveting, spectacular, cool; and negative as bad, clichéd, sucks, boring, stupid so and on and so forth. So, they came up with different, different words in positive and negative class, but when they tried to see what will be the accuracy of using this word list that are created by humans on this task, they found that human 1 had an accuracy of 58 percent and human two had an accuracy of 64 percent.

On the other hand, if we try to obtain this from statistics, you get words like love, wonderful, best, great, superb, beautiful, words like a still also come up, because this is obtained using statistics; and negative we have words like bad, worst, stupid, waste,

boring and some question mark exclamation etcetera. And this is still does better than the other two. So, this gives a performance of 69 percent than the other two. So, you see simply by using the words may not be sufficient, and why they that might be the case. So, why using these positive, negative words are not sufficient. So, mind you they are very important for sentiment analysis.

So, next lecture, we will complete devote on that that, what are different sort of lexicons or dictionaries about sentiment that you can use for this task, but what is the problem one problem with them.

(Refer Slide Time: 24:07)

Why can't we just look for words like "great" and "terrible"?

- This laptop is *a great deal*.
- *A great deal* of media attention surrounded the release of the new laptop.
- This laptop is *a great deal* ... and I've got a nice bridge you might be interested in.
- This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can't hold up.

Pawan Goyal (IIT Khargpur) Sentiment Analysis - Introduction Week 12, Lec 1

So, for that you have to see how people write about in their comments. So, this is one example, this laptop is a great deal. If you see the sentence you will say yes this looks like a positive review. But now let us see it in different like, a great deal of media attention surrounded the release of the new laptop. I am using the same words right great and deal, and this sentence also contains great and deal also contains the laptop. Would you say this is a positive sentiment? the sentence looks like neutral. So, you say ok. There is a lot of media attention, but says nothing about the laptop that whether the laptop is good or bad while the previous laptop was doing that. So, a case does not a stop here.

People will also write like that this laptop is a great deal and I have got a nice bridge you might be interested in right. So, this is some sort of phrase people use to say to use sarcasm, this is if this nice deal nice deal (Refer Time: 25:11) that you might be interested in so that means so this is like a really bad laptop. If you are talking is nice still, this is like a it is like a joke. So, then the text can also contains some humor or sarcasm. So, how do you find that this is not a positive sentiment, although there is no negative word in this sentence?

Similarly, here the film should be brilliant, it sounds like a great plot, that the actors are first grade; supporting cast is good as well. And Stallone is attempting to deliver a good performance everything positive. However, it cannot hold up; and the last sentence immediately turns all the polarity to negative. So, it says although it contains lot of positive words because of this is all negative. So, this that is what makes sentiment analysis much more difficult that you cannot do it easily by using simple features.

So, this was about the introduction. So, now, what we will do we will talk about different method that can be used. Again the most in common method is using some sort of known words that are having positive or negative polarity. And you have some sort of large list of these words not manually created some 10 or 12 words, it can be some 1000 of words that people have annotated with positive or negative sentiments, and this is called the affective lexicons. So, you can use these lexicons and they are readily available. So, we will see those. And you can use that for your task.

And then we will go further into how do you learn these lexicons and so on. So that is for this lecture, and then we will talk about this in the next lecture.

Thank you.