

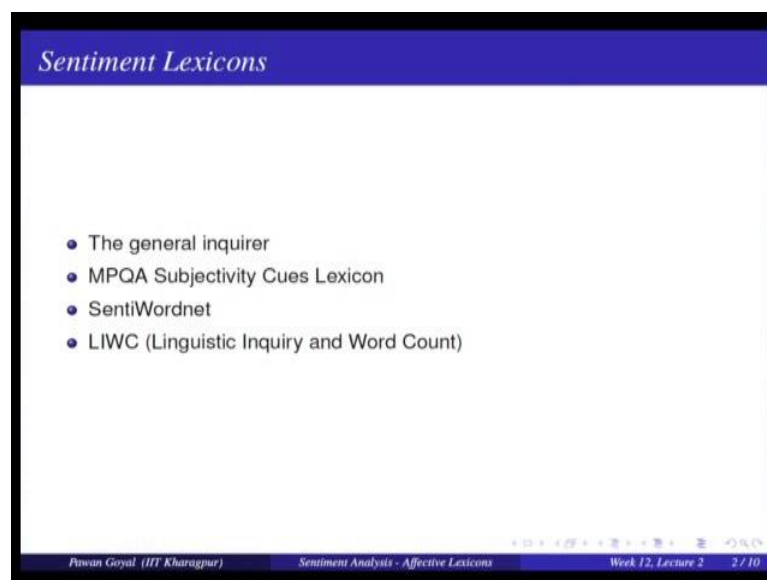
Natural Language Processing
Prof. Pawan Goyal
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture - 57
Sentiment Analysis - Affective Lexicons

Welcome back for the second lecture of this week. So, we talked about Sentiment Analysis. And we were saying that the one of the most important tools for doing sentiment analysis is using some sort of Lexicons. So, were some people have manually annotated some words whether they are positive or negative and so on, and given a new text you can use that to find out what is the major polarity and you can use some complicated models also to see how it changes by using communications and so on.

But important is finding out whether the some of the words are having positive or negative polarity. And that can be obtained by using sentiment lexicons. So, there are many different lexicons.

(Refer Slide Time: 00:59)



Sentiment Lexicons

- The general inquirer
- MPQA Subjectivity Cues Lexicon
- SentiWordnet
- LIWC (Linguistic Inquiry and Word Count)

Pawan Goyal (IIT Kharagpur) Sentiment Analysis - Affective Lexicons Week 12, Lecture 2 2 / 10

In this lecture I will just give some introduction to some of these and they are all available and you can go and download those in using your applications. So for example,

the general inquirer is one of the most more popular lexicons, then MPQ; multi prospective question answering subjectivity cues lexicon, SentiWordnet is something that has been used widely and then in the recent times the tool LIWC that is linguistic inquiry and word count has been popularly used for doing sentiment analysis along with some many other task. So, it has been very very popular.

So, let us have a look at these tools. Interestingly the tools I will be showing are mostly built for English there are efforts and on using these tools took good strap lexicons in other languages and many such efforts have actually been able to build such lexicons and such tools. But there is a lot of scope on for doing that in say Indian languages. Some of the tools available and many of these are not available. And if you have to process comments etcetera in Indian context you might want to use to build some of these also. So, we will also see how do you learn these in the next lecture.

(Refer Slide Time: 02:21)

The General Inquirer												
Categories												
<ul style="list-style-type: none"> Positive (1915 words) and Negative (2291) words Strong vs weak, active vs passive, overstated vs understated pleasure, pain, virtue, vice, motivation, cognitive orientation etc. 												
Entry	Source	Positive	Negative	Actv	Aff	Ngiv	Hostile	Strong	Power	Weak	Submit	Active
ABANDON	194,195		Negative			Ngiv				Weak		
ABANDONMENT	194		Negative							Weak		
ABATE	194,195		Negative									
ABATEMENT	194		Negative									
ABDICATE	194		Negative							Weak	Submit	
ABIDE	194	Positive			Aff		Hostile					Active
ABIDE	194	Positive						Strong		Weak	Submit	
ABJECT	194	Positive	Negative									
ABLE	194,195	Positive		Actv								
ABNORMAL	194,195		Negative			Ngiv						
ABORD	194,195		Negative									
ABOLISH	194,195		Negative			Ngiv	Hostile	Strong	Power			Active
ABOLITION	194		Negative					Strong				
ABOMINABLE	194		Negative									
ABORTIVE	194		Negative									
ABOUND	194	Positive										
ABOUT#1	194,195											
ABOUT#2	194,195											
ABOUT#3	194,195											
ABOUT#4	194,195											
ABOUT#5	194,195											
ABOUT#6	194,195											
ABOUT#7	194,195											
ABOVE#1	194,195											
ABOVE#2	194,195											
ABOVE#3	194,195											
ABOVE#4	194,195											

So, starting with the first (Refer Time: 02:23) general inquirer. As I was saying the label a lot of different words, it is not about some 10 or 50 words. So, they have 1915 positive words and 2291 negative words. And they have labeled them on various different dimensions. So, whether the word is positive, whether is negative, whether it is active, whether it is passive, overstated or understated. And then in the various dimensions like

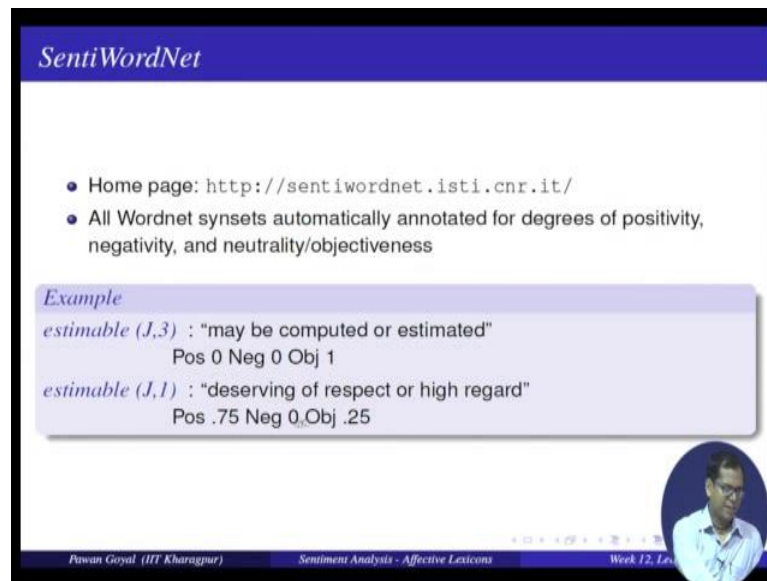
does it talk about pleasure, pain, virtue vice, motivation, cognitive, orientation etcetera. And this all comes niceliness spreadsheet, where you have different words and you can say whether it is positive negative and weak strong and so on.

So, let us see some examples; see here from the first few words in this spreadsheet, so you see what are negative words; abandon, abandonment, abate, abdicate, abject, abnormal. These are all negative words. And then you get some positive words like ability, able, abound, abide; these are all positive words. Then there are other dimensions like weak and strong. Is it a weak word or strong word? So, let us see weak versus strong. So, ability is a strong indicator, able strong, abolish is strong; but abdicate, abide, amendment are not very strong so they are weak.

So, this is also annotated this is weak sentiment this is an strong sentiment. And then they are other dimensions if you can see the other columns of this spreadsheet. And this to this comes readily available and you can use these for any you can see how to use that in your different tasks.

Then I was saying SentiWordnet is again very very popular. So, what was this effect? You know we talked about ordinary in this course. So, WordNet has lot of different synsets. So, what was done in SentiWordnet each of this synsets were taken and people, so by some way of finally doing manual supervisor supervision they try to label what is the sentiment score of this particular synset. Is it positive, negative? What is the score? What is the objectivity, subjectivity? So, they try to manually label that.

(Refer Slide Time: 04:45)

A presentation slide titled "SentiWordNet" with a blue header. It contains two bullet points: "Home page: http://sentiwordnet.isti.cnr.it/" and "All Wordnet synsets automatically annotated for degrees of positivity, negativity, and neutrality/objectiveness". Below this is a light purple box labeled "Example" containing two entries for the word "estimable": one for sense (J,3) with scores Pos 0, Neg 0, Obj 1, and another for sense (J,1) with scores Pos .75, Neg 0, Obj .25. The slide footer includes the name "Pawan Goyal (IIT Kharagpur)", the course "Sentiment Analysis - Affective Lexicons", and "Week 12, Lecture 12". A small circular portrait of the speaker is in the bottom right corner.

SentiWordNet

- Home page: <http://sentiwordnet.isti.cnr.it/>
- All Wordnet synsets automatically annotated for degrees of positivity, negativity, and neutrality/objectiveness

Example

estimable (J,3) : "may be computed or estimated"
Pos 0 Neg 0 Obj 1

estimable (J,1) : "deserving of respect or high regard"
Pos .75 Neg 0 Obj .25

Pawan Goyal (IIT Kharagpur) Sentiment Analysis - Affective Lexicons Week 12, Lecture 12

So, they took all WordNet synsets and they were automatically annotated for degrees of positive, negativity and neutrality and objectiveness; so examples are like that. So, you have two different synsets of the word estimable, so 3 and 1. So, J 3 is may be computed or estimated you can see that this is like a more objective does not have any sentiment, so it has positive 0 negative 0 objective 1.

On the other hand the second sense deserving of respect or high regard it looks like positive. So, positive is 0.75 negative is 0 and objectivity is 0.25. They have all these three scores.

(Refer Slide Time: 05:28)

Other Lexicons

MPQA Subjectivity Cues Lexicon

- Home page: http://www.cs.pitt.edu/mpqa/subj_lexicon.html
- 6885 words from 8221 lemmas: 2718 positive, 4912 negative
- Each word annotated for intensity (strong, weak)

Bing Liu Opinion Lexicon

- Bing Liu's Page on Opinion Mining
- <http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar>
- 6786 word: 2006 positive, 4780 negative

Pawan Goyal (IIT Kharagpur) Sentiment Analysis - Affective Lexicons Week 12, Lecture 2 5 / 10

Then another important lexicon is this MPQA subjectivity lexicon. Again you can go to this website where you can download this lexicon. So, it contains around 6885 words from 8000 plus lemmas and 2700 are positive and 4900 are negative. And each word has been labeled also for its intensity. So, in what are positive words, what are negative word, but you also know the intensity; is it strong is it weak and so on.

Then there is opinion lexicon by Bing Liu- that you can get from this the page on opinion by Bing Liu and there are again around refer the same words and some positive and some negative words. So, you can go and download these data sets and see whether this can be helpful for certain tasks.

(Refer Slide Time: 06:31)

LIWC (Linguistic Inquiry and Word Count)

- Home page: <http://www.liwc.net/>
- 2300 words, > 70 classes

Affective Processes

- Negative emotion (bad, weird, hate, problem, tough)
- Positive emotion (love, nice, sweet)

Cognitive Processes

Tentative (maybe, perhaps, guess), Inhibition (block, constraint)

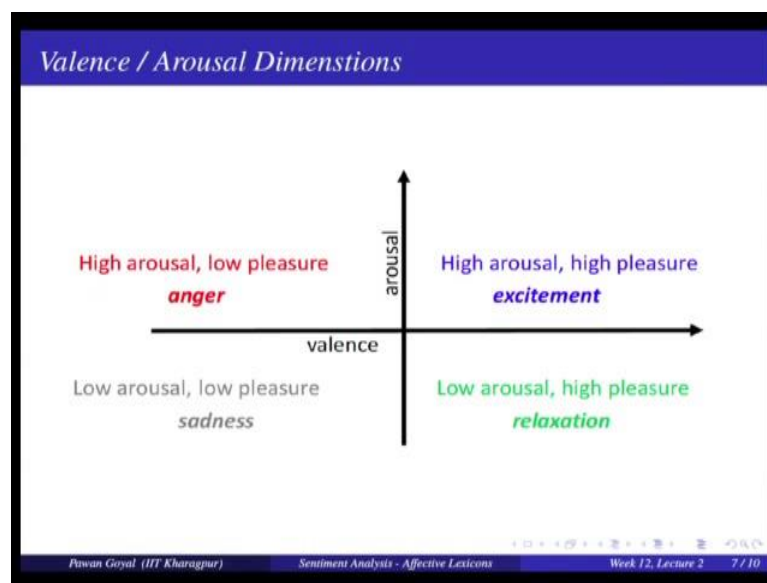
- Comes with a small fee

Pawan Goyal (IIT Kharagpur) Sentiment Analysis - Affective Lexicons Week 12, Lecture 2 6 / 10

And then I was saying in recent times another tool that has become very popular is the LIWC tool; linguistic enquiry word count. So, what is interesting is that. Again you can go to this webpage and explore what does this tool does. So, it takes 2300 different words and divides them into 70 different classes. And these classes range from simple positive negative to many different social things also. Like, so you are having negative emotions. So, you will say words like bad, weird, hate, problem, tough; they will be getting negative emotions. And love, nice, sweet are getting positive emotions.

But you will also see some cognitive processes like; some words denote some tentativeness like here, maybe, perhaps, guess; they might indicate in addition like block, constraint and so on. So, if you will see there are 70 different classes in which words are explored. And these classes make it much more richer representation that can be used for task like sentiment analysis also seeing other sort of effective states that we talked about earlier in the last week. So, we can talk about emotions, we can talk about inter personal things and so on by using this tool. Only thing is that it comes with a small fee, say if you have to use that you have to pay some small fee.

(Refer Slide Time: 07:51)



So, in general when we talk about sentiment you can also talk about two different aspects: one is called the valence another is called the arousal. So, by valence I mean positive or negative; how positive it is, how negative it is. And arousal is kind of strength. So, arouse means yes this is very very strong sentiment and low arousal means say it is not it is a weak sentiment.

So, you can put them like a like an access here, so you have some quadrant this is like valence going from negative to positive and arousal going from low to high and you can see ok. So, here words are high arousal, high pleasure like excitement it has been having valence is positive for excitement and the arousal is also high. And you can accordingly think of a word that is positive, but arousal is low like getting relaxed, so relaxation. This is having again positive, but arousal is low.

Similarly negative you can take the contrast between anger and sadness. Anger and sadness both are having the negative valence, but in terms of arousals anger is in on high and sadness is low. There are again attempts that have put words in these two dimensions that what is the valence and what is the arousal of these words.

(Refer Slide Time: 09:14)

Lexicon of valence, arousal, and dominance

- Warriner, Amy Beth, Victor Kuperman, and Marc Brysbaert. "Norms of valence, arousal, and dominance for 13,915 English lemmas." *Behavior research methods* 45.4 (2013): 1191-1207.
- **Supplementary data:** This word is licenced under a Creative Commons.

Ratings for 14,000 words for emotional dimensions:

- **valence** (the pleasantness of the stimulus)
- **arousal** (the intensity of emotion provoked by the stimulus)
- **dominance** (the degree of control exerted by the stimulus)

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

So, this is the work; so lexicon of valence arousal and dominance; so norms of valence arousal dominance for 13915 English lemmas. And this word is licensed under a creative com commons license it is available for non commercial use.

So, what they have done? They have taken roughly 14000 words and they have labeled them for the valence that is how pleasant or unpleasant this is arousal. So, what is the intensity of the emotion? So whether it is highly arousal or low; and what is the dominance? So, what is the degree of control by the stimulus? So, these 14000 words are labeled with these three.

(Refer Slide Time: 09:59)

Lexicon of valence, arousal, and dominance

valence (the pleasantness of the stimulus)

- 9: happy, pleased, satisfied, contented, hopeful
- 1: unhappy, annoyed, unsatisfied, melancholic, despaired, or bored

arousal (the intensity of emotion provoked by the stimulus)

- 9: stimulated, excited, frenzied, jittery, wide-awake, or aroused
- 1: relaxed, calm, sluggish, dull, sleepy, or unaroused

dominance (the degree of control exerted by the stimulus)

- 9: in control, influential, important, dominant, autonomous, or controlling
- 1: controlled, influenced, cared-for, awed, submissive, or guided

Pawan Goyal (IIT Kharagpur) Sentiment Analysis - Affective Lexicons Week 12, Lecture 2 9 / 10

Let us see some example of the valence. So, how pleasant the stimulus is? So these are the words with valence of 9 and with the 1. And this you can easily see these are like what we are simply talking sentiment analysis. So, 9 words are; happy, pleased, satisfied, contented, hopeful, they are all 9. And 1 here; unhappy, annoyed, unsatisfied, despaired and bored and so on; so these are all in the valence.

Then if you go to arousal; again 9 they are like stimulated, excited, frenzied, and aroused; and 1 you will have words like relaxed, calm, sluggish, dull, and sleepy. So, you can see the degree of arousal also. Then there are; and you should not dominant. So, like in control, influential, important, dominant they are all having a reading of 9; and on 1 you will have controlled, influenced like controlled by someone else, cared-for, awed, submissive, (Refer Time: 11:02) low control.

(Refer Slide Time: 11:05)

Lexicon of valence, arousal, and dominance: Examples					
Valence		Arousal		Dominance	
vacation	8.53	rampage	7.56	self	7.74
happy	8.47	tornado	7.45	incredible	7.74
whistle	5.7	zucchini	4.18	skillet	5.33
conscious	5.53	dressy	4.15	concur	5.29
torture	1.4	dull	1.67	earthquake	2.14

So, here are some other examples with the words and the actual fraction with this they are labeled. So, vacation is getting a high valence. Yes happy gets a high valence, whistle gets (Refer Time: 11:18), conscious 5, torture gets 1.4. Rampage gets 7.5, arousal high tornado gets high, zucchini gets around 4, dressy and dull gets low. And similarly here self incredible get very high dominance; on the other hand words like earthquake get a very low dominance. So, these words are all labeled with all these dimensions.

With that there are many other sort of lexicons also that people have used and they also around on the web. There are some of the more important ones and most of these are available for download and you can use in your different tasks. But many of these lexicons were not created by manual labeling. Some of these were created automatically and someone then manually verified.

So, what we will see in the next lecture that is there is some way in which you can create these lexicons on your own by using some of the techniques and we talked about; by doing some sort of good strapping or by using some sort of (Refer Time: 12:23) can be learn these lexicons from the data. And this is again very interesting idea and many lexicons have been (Refer Time: 12:33) by using that method, so that you will talk about in the next lecture.

Thank you.