1

00:00:00,025 --> 00:00:04,628

[SOUND] This lecture is about,

2

00:00:04,628 --> 00:00:11,073

Opinion Mining and Sentiment Analysis,

3

00:00:11,073 --> 00:00:15,133

covering, Motivation.

4

00:00:15,133 --> 00:00:18,208

In this lecture,

we're going to start, talking about,

5

00:00:18,208 --> 00:00:21,170

mining a different kind of knowledge.

6

00:00:21,170 --> 00:00:27,150

Namely, knowledge about the observer or

humans that have generated the text data.

7

00:00:27,150 --> 00:00:31,070

In particular, we're going to talk about

the opinion mining and sentiment analysis.

8

00:00:32,660 --> 00:00:37,540

As we discussed earlier, text data

can be regarded as data generated

9

00:00:37,540 --> 00:00:41,573

from humans as subjective sensors.

10

00:00:43,450 --> 00:00:50,880

In contrast, we have other devices such

as video recorder that can report what's

11

00:00:50,880 --> 00:00:56,670

happening in the real world objective to

generate the viewer data for example.

12

00:00:58,290 --> 00:01:03,495

Now the main difference between test

data and other data, like video data,

13

00:01:03,495 --> 00:01:10,440

is that it has rich opinions,

14

00:01:10,440 --> 00:01:15,720

and the content tends to be subjective

because it's generated from humans.

15

00:01:16,730 --> 00:01:22,170

Now, this is actually a unique advantaged

of text data, as compared with other data,

16

00:01:22,170 --> 00:01:28,190

because the office is a great

opportunity to understand the observers.

17

00:01:28,190 --> 00:01:31,740

We can mine text data to

understand their opinions.

18

00:01:31,740 --> 00:01:35,024

Understand people's preferences,

how people think about something.

19

00:01:37,105 --> 00:01:43,155

So this lecture and the following lectures

will be mainly about how we can mine and

20

00:01:43,155 --> 00:01:47,775

analyze opinions buried

in a lot of text data.

21

00:01:49,600 --> 00:01:53,036

So let's start with

the concept of opinion.

22

00:01:53,036 --> 00:01:57,580

It's not that easy to

formally define opinion, but

23

00:01:57,580 --> 00:02:02,340

mostly we would define

opinion as a subjective

24

00:02:02,340 --> 00:02:07,000

statement describing what a person

believes or thinks about something.

25

00:02:08,770 --> 00:02:11,140

Now, I highlighted quite a few words here.

26

00:02:11,140 --> 00:02:15,480

And that's because it's worth thinking

a little bit more about these words.

27

00:02:15,480 --> 00:02:20,160

And that will help us better

understand what's in an opinion.

28

00:02:20,160 --> 00:02:25,370

And this further helps us to

define opinion more formally.

29

00:02:25,370 --> 00:02:30,990

Which is always needed to computation to

resolve the problem of opinion mining.

30

00:02:30,990 --> 00:02:35,390

So let's first look at the key

word of subjective here.

31

00:02:35,390 --> 00:02:39,579

This is in contrast with objective

statement or factual statement.

32

00:02:40,780 --> 00:02:44,130

Those statements can be proved right or

wrong.

33

00:02:45,180 --> 00:02:49,470

And this is a key differentiating

factor from opinions

34

00:02:49,470 --> 00:02:53,520

which tends to be not

easy to prove wrong or

35

00:02:53,520 --> 00:02:56,960

right, because it reflects what

the person thinks about something.

36

00:02:59,010 --> 00:03:06,640

So in contrast, objective statement can

usually be proved wrong or correct.

37

00:03:07,660 --> 00:03:15,170

For example, you might say this

computer has a screen and a battery.

38

00:03:16,700 --> 00:03:18,600

Now that's something you can check.

39

00:03:18,600 --> 00:03:22,200

It's either having a battery or not.

40

00:03:23,430 --> 00:03:28,820

But in contrast with this, think about

the sentence such as, this laptop has

41

00:03:28,820 --> 00:03:34,950

the best battery or

this laptop has a nice screen.

42

00:03:34,950 --> 00:03:38,629

Now these statements

are more subjective and

43

00:03:38,629 --> 00:03:43,290

it's very hard to prove

whether it's wrong or correct.

44

00:03:45,680 --> 00:03:49,000

So opinion, is a subjective statement.

45

00:03:50,260 --> 00:03:54,450

And next lets look at

the keyword person here.

46

00:03:54,450 --> 00:03:56,940

And that indicates that

is an opinion holder.

47

00:03:56,940 --> 00:04:02,310

Because when we talk about opinion,

it's about an opinion held by someone.

48

00:04:02,310 --> 00:04:04,660

And then we notice that

there is something here.

49

00:04:04,660 --> 00:04:06,860

So that is the target of the opinion.

50

00:04:06,860 --> 00:04:09,810

The opinion is expressed

on this something.

51

00:04:11,440 --> 00:04:15,780

And now, of course, believes or

thinks implies that

52

00:04:15,780 --> 00:04:21,010

an opinion will depend on the culture or

background and the context in general.

53

00:04:21,010 --> 00:04:24,760

Because a person might think

different in a different context.

54

00:04:24,760 --> 00:04:29,690

People from different background

may also think in different ways.

55

00:04:29,690 --> 00:04:34,240

So this analysis shows that there are

multiple elements that we need to include

56

00:04:34,240 --> 00:04:36,580

in order to characterize opinion.

57

00:04:38,160 --> 00:04:42,470

So, what's a basic opinion

representation like?

58

00:04:42,470 --> 00:04:46,830

Well, it should include at

least three elements, right?

59

00:04:46,830 --> 00:04:49,520

Firstly, it has to specify

what's the opinion holder.

60

00:04:49,520 --> 00:04:51,220

So whose opinion is this?

61

00:04:51,220 --> 00:04:55,720

Second, it must also specify the target,

what's this opinion about?

62

00:04:57,230 --> 00:05:00,830

And third, of course,

we want opinion content.

63

00:05:00,830 --> 00:05:03,830

And so what exactly is opinion?

64

00:05:03,830 --> 00:05:05,160

If you can identify these,

65

00:05:05,160 --> 00:05:10,750

we get a basic understanding of opinion

and can already be useful sometimes.

66

00:05:10,750 --> 00:05:14,130

You want to understand further,

we want enriched opinion representation.

67

00:05:15,210 --> 00:05:17,500

And that means we also want to

understand that, for example,

68

00:05:17,500 --> 00:05:22,740

the context of the opinion and

what situation was the opinion expressed.

69

00:05:22,740 --> 00:05:26,290

For example, what time was it expressed?

70

00:05:26,290 --> 00:05:32,310

We, also, would like to, people understand

the opinion sentiment, and this is

71

00:05:32,310 --> 00:05:39,370

to understand that what the opinion tells

us about the opinion holder's feeling.

72

00:05:39,370 --> 00:05:43,660

For example, is this opinion positive,

or negative?

73

00:05:43,660 --> 00:05:47,911

Or perhaps the opinion holder was happy or

was sad, and

74

00:05:47,911 --> 00:05:52,813

so such understanding obvious

to those beyond just Extracting

75

00:05:52,813 --> 00:05:56,527

the opinion content,

it needs some analysis.

76

00:06:00,484 --> 00:06:03,860

So let's take a simple

example of a product review.

77

00:06:03,860 --> 00:06:09,350

In this case, this actually expressed the

opinion holder, and expressed the target.

78

00:06:09,350 --> 00:06:12,140

So its obviously whats opinion holder and

79

00:06:12,140 --> 00:06:16,860

that's just reviewer and its also often

very clear whats the opinion target and

80

00:06:16,860 --> 00:06:21,340

that's the product review for

example iPhone 6.

81

00:06:21,340 --> 00:06:26,210

When the review is posted usually

you can't such information easier.

82

00:06:27,350 --> 00:06:30,540

Now the content, of course,

is a review text that's, in general,

83

00:06:30,540 --> 00:06:32,290

also easy to obtain.

84

00:06:32,290 --> 00:06:35,920

So you can see product reviews are fairly

85

00:06:35,920 --> 00:06:40,640

easy to analyze in terms of obtaining

a basic opinion of representation.

86

00:06:40,640 --> 00:06:45,730

But of course, if you want to get more

information, you might know the Context,

87

00:06:45,730 --> 00:06:47,300

for example.

88

00:06:47,300 --> 00:06:51,170

The review was written in 2015.

89

00:06:51,170 --> 00:06:57,190

Or, we want to know that the sentiment

of this review is positive.

90

00:06:57,190 --> 00:07:02,499

So, this additional understanding of

course adds value to mining the opinions.

91

00:07:04,340 --> 00:07:09,060

Now, you can see in this case the task

is relatively easy and that's

92

00:07:09,060 --> 00:07:13,789

because the opinion holder and the opinion

target have already been identified.

93

00:07:14,820 --> 00:07:17,230

Now let's take a look at

the sentence in the news.

94

00:07:17,230 --> 00:07:21,410

In this case, we have a implicit

holder and a implicit target.

95

00:07:21,410 --> 00:07:24,990

And the tasker is in general harder.

96

00:07:24,990 --> 00:07:31,700

So, we can identify opinion holder here,

and that's the governor of Connecticut.

97

00:07:32,790 --> 00:07:35,710

We can also identify the target.

98

00:07:35,710 --> 00:07:39,160

So one target is Hurricane Sandy, but

99

00:07:39,160 --> 00:07:43,920

there is also another target

mentioned which is hurricane of 1938.

100

00:07:43,920 --> 00:07:45,850

So what's the opinion?

101

00:07:45,850 --> 00:07:49,660

Well, there's a negative sentiment here

102

00:07:49,660 --> 00:07:52,029

that's indicated by words like bad and

worst.

103

00:07:53,360 --> 00:07:59,270

And we can also, then, identify context,

New England in this case.

104

00:08:00,420 --> 00:08:03,300

Now, unlike in the playoff review,

105

00:08:03,300 --> 00:08:08,370

all these elements must be extracted by

using natural RAM processing techniques.

106

00:08:08,370 --> 00:08:11,100

So, the task Is much harder.

107

00:08:11,100 --> 00:08:13,020

And we need a deeper natural

language processing.

108

00:08:14,860 --> 00:08:16,200

And these examples also

109

00:08:17,550 --> 00:08:22,260

suggest that a lot of work can be

easy to done for product reviews.

110

00:08:22,260 --> 00:08:23,896

That's indeed what has happened.

111

00:08:23,896 --> 00:08:29,170

Analyzing and

assembling news is still quite difficult,

112

00:08:29,170 --> 00:08:34,150

it's more difficult than the analysis

of opinions in product reviews.

113

00:08:36,810 --> 00:08:39,330

Now there are also some other

interesting variations.

114

00:08:39,330 --> 00:08:43,152

In fact, here we're going to

examine the variations of opinions,

115

00:08:43,152 --> 00:08:44,080

more systematically.

116

00:08:44,080 --> 00:08:46,150

First, let's think about

the opinion holder.

117

00:08:47,300 --> 00:08:50,870

The holder could be an individual or

it could be group of people.

118

00:08:50,870 --> 00:08:53,190

Sometimes, the opinion

was from a committee.

119

00:08:53,190 --> 00:08:55,060

Or from a whole country of people.

120

00:08:56,450 --> 00:08:58,770

Opinion target accounts will vary a lot.

121

00:08:58,770 --> 00:09:02,870

It can be about one entity,

a particular person, a particular product,

122

00:09:02,870 --> 00:09:04,830

a particular policy, ect.

123

00:09:04,830 --> 00:09:08,020

But it could be about a group of products.

124

00:09:08,020 --> 00:09:10,630

Could be about the products

from a company in general.

125

00:09:11,860 --> 00:09:14,990

Could also be very specific

about one attribute, though.

126

00:09:14,990 --> 00:09:16,090

An attribute of the entity.

127

00:09:16,090 --> 00:09:21,220

For example,

it's just about the battery of iPhone.

128

00:09:21,220 --> 00:09:23,300

It could be someone else's opinion.

129

00:09:23,300 --> 00:09:27,160

And one person might comment on

another person's Opinion, etc.

130

00:09:27,160 --> 00:09:31,690

So, you can see there is a lot of

variation here that will cause

131

00:09:31,690 --> 00:09:34,080

the problem to vary a lot.

132

00:09:34,080 --> 00:09:38,680

Now, opinion content, of course,

can also vary a lot on the surface,

133

00:09:38,680 --> 00:09:42,050

you can identify one-sentence opinion or

one-phrase opinion.

134

00:09:42,050 --> 00:09:45,820

But you can also have longer

text to express an opinion,

135

00:09:45,820 --> 00:09:46,690

like the whole article.

136

00:09:48,390 --> 00:09:51,550

And furthermore we identify

the variation in the sentiment or

137

00:09:51,550 --> 00:09:56,520

emotion damage that's above

the feeding of the opinion holder.

138

00:09:56,520 --> 00:10:00,460

So, we can distinguish a positive

versus negative or mutual or

139

00:10:00,460 --> 00:10:02,420

happy versus sad, separate.

140

00:10:03,820 --> 00:10:05,930

Finally, the opinion

context can also vary.

141

00:10:05,930 --> 00:10:10,760

We can have a simple context, like

different time or different locations.

142

00:10:10,760 --> 00:10:13,360

But there could be also complex contexts,

143

00:10:13,360 --> 00:10:17,840

such as some background

of topic being discussed.

144

00:10:17,840 --> 00:10:22,420

So when opinion is expressed in

particular discourse context, it has to

145

00:10:22,420 --> 00:10:26,970

be interpreted in different ways than

when it's expressed in another context.

146

00:10:26,970 --> 00:10:32,230

So the context can be very [INAUDIBLE] to

entire discourse context of the opinion.

147

00:10:32,230 --> 00:10:34,170

From computational perspective,

148

00:10:34,170 --> 00:10:39,150

we're mostly interested in what opinions

can be extracted from text data.

149

00:10:39,150 --> 00:10:42,690

So, it turns out that we can

also differentiate, distinguish,

150

00:10:42,690 --> 00:10:46,810

different kinds of opinions in text

data from computation perspective.

151

00:10:46,810 --> 00:10:50,440

First, the observer might make

a comment about opinion targeting,

152

00:10:50,440 --> 00:10:54,250

observe the word So

in case we have the author's opinion.

153

00:10:54,250 --> 00:10:56,830

For example,

I don't like this phone at all.

154

00:10:56,830 --> 00:10:58,370

And that's an opinion of this author.

155

00:10:59,990 --> 00:11:07,940

In contrast, the text might also

report opinions about others.

156

00:11:07,940 --> 00:11:13,940

So the person could also Make observation

about another person's opinion and

157

00:11:13,940 --> 00:11:15,310

reported this opinion.

158

00:11:15,310 --> 00:11:19,740

So for example,

I believe he loves the painting.

159

00:11:19,740 --> 00:11:28,410

And that opinion is really about the It is

really expressed by another person here.

160

00:11:28,410 --> 00:11:32,250

So, it doesn't mean this

author loves that painting.

161

00:11:33,640 --> 00:11:38,970

So clearly, the two kinds of opinions

need to be analyzed in different ways,

162

00:11:38,970 --> 00:11:40,430

and sometimes in product reviews,

163

00:11:40,430 --> 00:11:45,220

you can see, although mostly the opinions

are false from this reviewer.

164

00:11:45,220 --> 00:11:49,010

Sometimes, a reviewer might mention

opinions of his friend or her friend.

165

00:11:51,400 --> 00:11:56,240

Another complication is that

there may be indirect opinions or

166

00:11:56,240 --> 00:11:59,960

inferred opinions that can be obtained.

167

00:11:59,960 --> 00:12:02,500

By making inferences on

168

00:12:02,500 --> 00:12:06,520

what's expressed in the text that might

not necessarily look like opinion.

169

00:12:06,520 --> 00:12:10,190

For example, one statement that might be,

170

00:12:10,190 --> 00:12:14,260

this phone ran out of

battery in just one hour.

171

00:12:14,260 --> 00:12:23,250

Now, this is in a way a factual statement

because It's either true or false, right?

172

00:12:23,250 --> 00:12:27,120

You can even verify that,

but from this statement,

173

00:12:27,120 --> 00:12:31,980

one can also infer some negative opinions

about the quality of the battery of

174

00:12:31,980 --> 00:12:37,160

this phone, or the feeling of

the opinion holder about the battery.

175

00:12:37,160 --> 00:12:40,520

The opinion holder clearly wished

that the battery do last longer.

176

00:12:42,500 --> 00:12:46,460

So these are interesting variations

that we need to pay attention to when we

177

00:12:46,460 --> 00:12:47,880

extract opinions.

178

00:12:47,880 --> 00:12:52,180

Also, for

this reason about indirect opinions,

179

00:12:53,790 --> 00:12:58,300

it's often also very useful to extract

whatever the person has said about

180

00:12:58,300 --> 00:13:03,550

the product, and sometimes factual

sentences like these are also very useful.

181

00:13:03,550 --> 00:13:05,320

So, from a practical viewpoint,

182

00:13:05,320 --> 00:13:10,090

sometimes we don't necessarily

extract the subject of sentences.

183

00:13:10,090 --> 00:13:15,240

Instead, again, all the sentences that

are about the opinions are useful for

184

00:13:15,240 --> 00:13:18,420

understanding the person or

understanding the product that we commend.

185

00:13:19,520 --> 00:13:24,530

So the task of opinion mining can be

defined as taking textualized input

186

00:13:24,530 --> 00:13:27,500

to generate a set of

opinion representations.

187

00:13:27,500 --> 00:13:32,010

Each representation we should

identify opinion holder,

188

00:13:32,010 --> 00:13:34,750

target, content, and the context.

189

00:13:34,750 --> 00:13:39,070

Ideally we can also infer opinion

sentiment from the comment and

190

00:13:39,070 --> 00:13:41,260

the context to better understand.

191

00:13:43,170 --> 00:13:43,690

The opinion.

192

00:13:44,700 --> 00:13:48,270

Now often, some elements of

the representation are already known.

193

00:13:48,270 --> 00:13:52,140

I just gave a good example in

the case of product we'd use

194

00:13:52,140 --> 00:13:57,780

where the opinion holder and the opinion

target are often expressly identified.

195

00:13:57,780 --> 00:14:05,120

And that's not why this turns out to be

one of the simplest opinion mining tasks.

196

00:14:05,120 --> 00:14:09,920

Now, it's interesting to think about

the other tasks that might be also simple.

197

00:14:09,920 --> 00:14:12,880

Because those are the cases

where you can easily build

198

00:14:12,880 --> 00:14:15,791

applications by using

opinion mining techniques.

199

00:14:17,973 --> 00:14:23,810

So now that we have talked about what is

opinion mining, we have defined the task.

200

00:14:23,810 --> 00:14:29,260

Let's also just talk a little bit about

why opinion mining is very important and

201

00:14:29,260 --> 00:14:31,130

why it's very useful.

202

00:14:31,130 --> 00:14:35,790

So here, I identify three major reasons,

three broad reasons.

203

00:14:35,790 --> 00:14:39,290

The first is it can help decision support.

204

00:14:39,290 --> 00:14:41,912

It can help us optimize our decisions.

205

00:14:41,912 --> 00:14:47,180

We often look at other people's opinions,

look at read the reviews

206

00:14:47,180 --> 00:14:51,100

in order to make a decisions like

buying a product or using a service.

207

00:14:52,160 --> 00:14:56,770

We also would be interested

in others opinions

208

00:14:56,770 --> 00:14:59,330

when we decide whom to vote for example.

209

00:15:00,978 --> 00:15:02,030

And policy makers,

210

00:15:02,030 --> 00:15:07,130

may also want to know people's

opinions when designing a new policy.

211

00:15:07,130 --> 00:15:10,460

So that's one general,

kind of, applications.

212

00:15:10,460 --> 00:15:12,540

And it's very broad, of course.

213

00:15:12,540 --> 00:15:17,830

The second application is to understand

people, and this is also very important.

214

00:15:17,830 --> 00:15:22,200

For example, it could help

understand people's preferences.

215

00:15:22,200 --> 00:15:24,460

And this could help us

better serve people.

216

00:15:24,460 --> 00:15:30,010

For example, we optimize a product search

engine or optimize a recommender system

217

00:15:30,010 --> 00:15:34,290

if we know what people are interested in,

what people think about product.

218

00:15:35,770 --> 00:15:39,870

It can also help with advertising,

of course, and we can have targeted

219

00:15:39,870 --> 00:15:47,170

advertising if we know what kind of

people tend to like what kind of plot.

220

00:15:48,560 --> 00:15:53,800

Now the third kind of application

can be called voluntary survey.

221

00:15:53,800 --> 00:15:59,360

Now this is most important research

that used to be done by doing surveys,

222

00:15:59,360 --> 00:16:00,248

doing manual surveys.

223

00:16:00,248 --> 00:16:03,670

Question, answer it.

224

00:16:03,670 --> 00:16:07,680

People need to feel informs

to answer their questions.

225

00:16:07,680 --> 00:16:13,610

Now this is directly related to humans

as sensors, and we can usually aggregate

226

00:16:13,610 --> 00:16:18,680

opinions from a lot of humans through

kind of assess the general opinion.

227

00:16:18,680 --> 00:16:24,490

Now this would be very useful for

business intelligence where manufacturers

228

00:16:24,490 --> 00:16:29,890

want to know where their products

have advantages over others.

229

00:16:31,120 --> 00:16:33,700

What are the winning

features of their products,

230

00:16:33,700 --> 00:16:35,650

winning features of competitive products.

231

00:16:37,090 --> 00:16:40,590

Market research has to do with

understanding consumers oppinions.

232

00:16:40,590 --> 00:16:43,535

And this create very useful directive for

that.

233

00:16:43,535 --> 00:16:48,840

Data-driven social science research

can benefit from this because they can

234

00:16:48,840 --> 00:16:52,290

do text mining to understand

the people's opinions.

235

00:16:52,290 --> 00:16:56,990

And if you can aggregate a lot of opinions

from social media, from a lot of, popular

236

00:16:58,290 --> 00:17:04,540

information then you can actually

do some study of some questions.

237

00:17:04,540 --> 00:17:12,310

For example, we can study the behavior of

people on social media on social networks.

238

00:17:12,310 --> 00:17:17,890

And these can be regarded as voluntary

survey done by those people.

239

00:17:19,350 --> 00:17:24,130

In general, we can gain a lot of advantage

in any prediction task because we can

240

00:17:24,130 --> 00:17:29,350

leverage the text data as

extra data above any problem.

241

00:17:29,350 --> 00:17:34,371

And so we can use text based

prediction techniques to help you

242

00:17:34,371 --> 00:17:39,203

make predictions or

improve the accuracy of prediction.

243

00:17:39,203 --> 00:17:49,203

[MUSIC]