RSM Erasmus | Rotterdam School of Management
Erasmus University

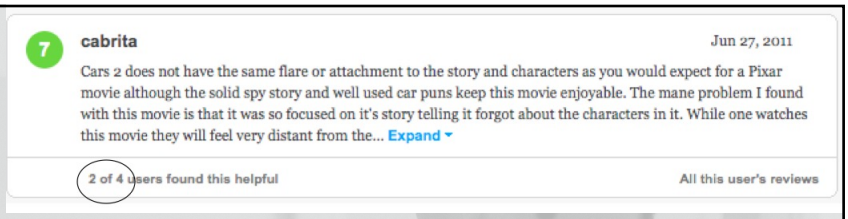
Session 3

NLP - Analysis & Testing

Prof. Gui Liberali

1

Variables Constructed from UGC



Review #	Stars or Rating (0…10)	Dim1: p(acting)	Dim2: p(storyline)	Dim3: p(Visual)	Sent. p(pos)	p (pos | acting)	p(pos| storyline)	p (pos | visual)	Daily Box office	Day
1	7	**0.6**	**0.1**	**0.3**	**0.6**	0.36	0.06	0.18	$6,958,490	Jun 27, 2011
2	8	**0.5**	**0.3**	**0.2**	**0.5**	0.25	0.15	0.1	$7,089,413	Jun 28, 2011
3	3	**0.4**	**0.5**	**0.1**	**0.3**	0.12	0.15	0.03	$5,540,193	Jun 29, 2011
..										
n	1	**0.1**	**0.8**	**0.1**	**0.2**	0.02	0.16	0.02	$5,540,193	Jun 29, 2011

5

5

Then what? How do I use these data ?

One possible way

Define the IV and DVs

- Definition
- Measurement
- Sentiment per topic?
- Extra variables of interest, e.g., temporal distance to first post & between posts, location, *sidedness*, thread, box office, sales, etc.

Define which descriptives are useful: e.g., mean and std deviation

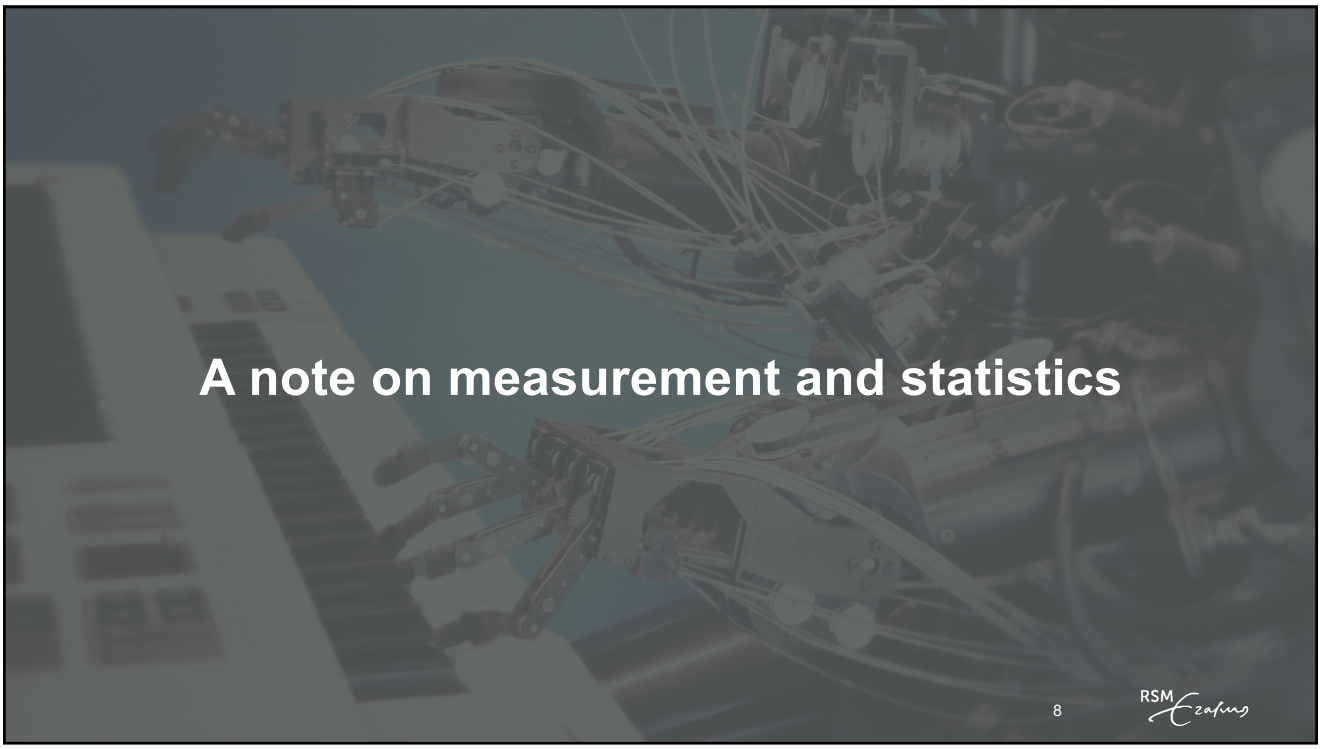
Define if the relationship between IV and DV is useful

Think how to analyze the relationship between IV and DV

- means?
- regression ?

Review #	Stars or Rating (0…10)	Dim1: p(acting)	Dim2: p(storyline)	Dim3: p(Visual)	Sent. p(pos)	p (pos | acting)	p(pos| storyline)	p (pos | visual)	Daily Box office	Day
1	7	**0.6**	**0.1**	**0.3**	**0.6**	0.36	0.06	0.18	$6,958,490	Jun 27, 2011
2	8	**0.5**	**0.3**	**0.2**	**0.5**	0.25	0.15	0.1	$7,089,413	Jun 28, 2011
3	3	**0.4**	**0.5**	**0.1**	**0.3**	0.12	0.15	0.03	$5,540,193	Jun 29, 2011
..										
n	1	**0.1**	**0.8**	**0.1**	**0.2**	0.02	0.16	0.02	$5,540,193	Jun 29, 2011

6



8

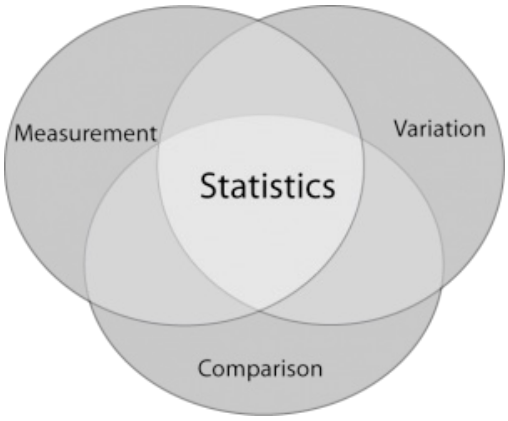
ROTTERDAM SCHOOL OF MANAGEMENT
ERASMUS UNIVERSITY

When does it make sense to use statistics?

The business school that thinks
and lives in the future

RSM Erasmus University

9



Source: A. Gelman's Notes at http://andrewgelman.com

Examples of latent variables: consideration set, topics being discussed in a text, propensity to switch, purchase intention.

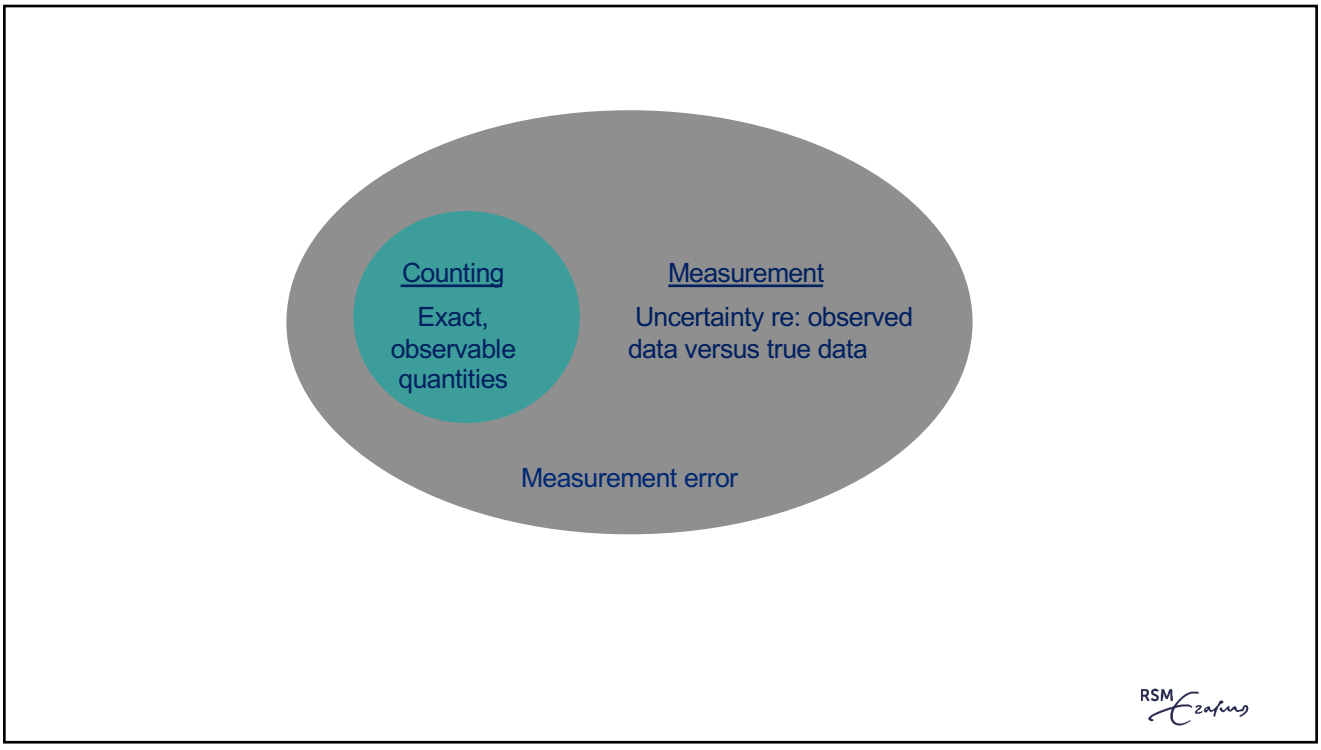
10

RSM Erasmus

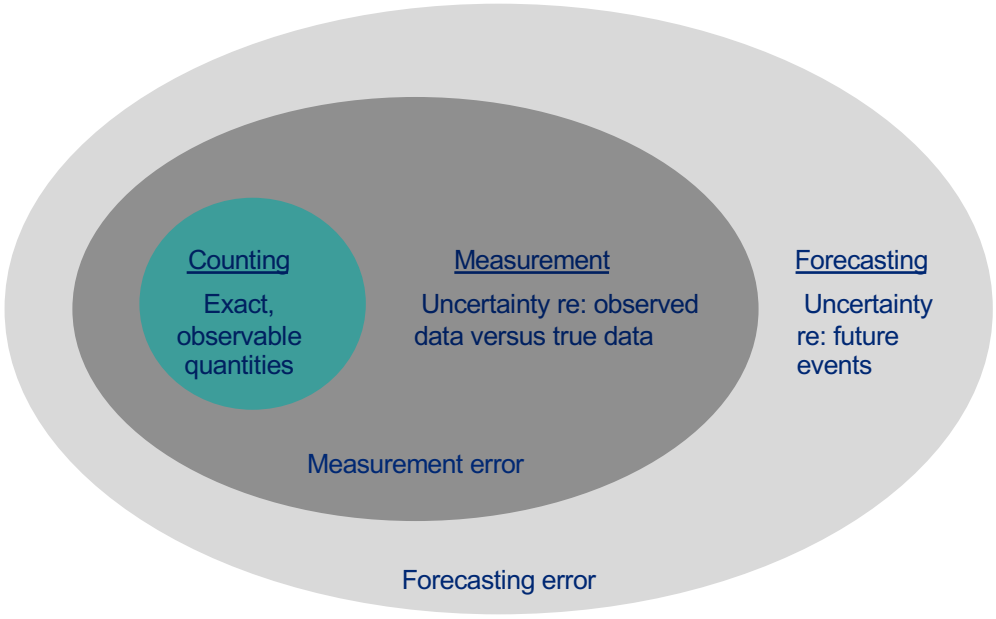
10


Counting
Exact, observable quantities

11


Counting
Exact, observable quantities
Measurement
Uncertainty re: observed data versus true data
Measurement error

12



13

13

The questions you want to answer define your analysis

Let's look at some examples

- **What type of questions do you have?**
- **What analyses were you planning to use?**

14

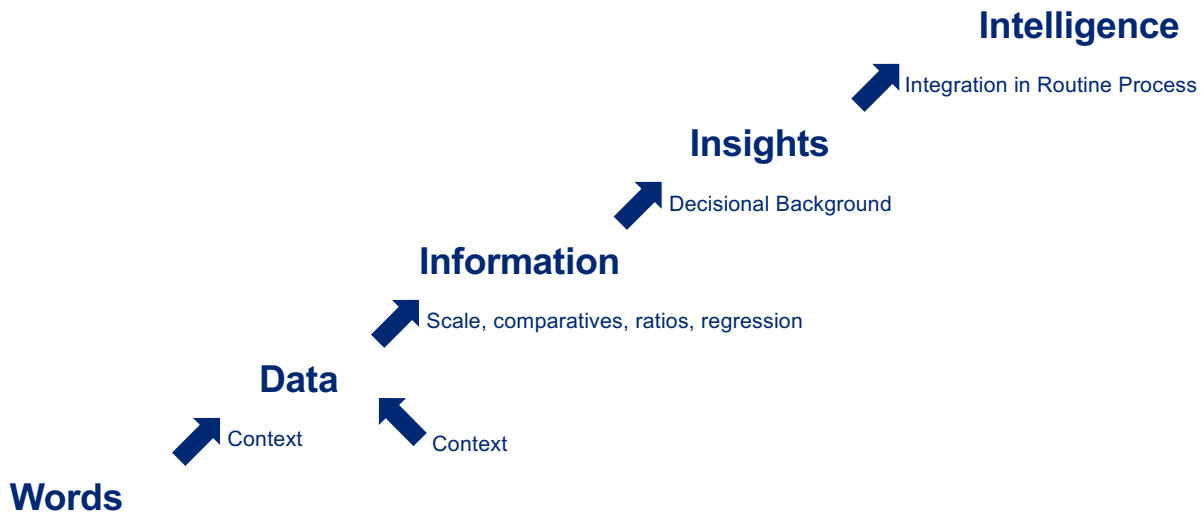
Analysis of Supervised Learning Data

1. All models are wrong. Some are useful.
2. What is the question you are trying to answer?
3. Keep in mind the data generating process
   - How was the (raw) UGC created?
   - Who created it? Why (from behavioral targeting? From a search?)
   - When? Under which circumstances (purchase?)
     What is the temporal difference between your IV and your DV?

RSM Erasmus

15

Context Matters



16

RSM Erasmus

16

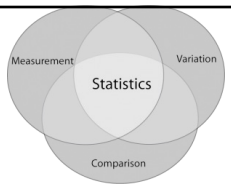
Some Types of UGC Analysis

Review #	Stars or Rating (0…10)	Dim1: p(acting)	Dim2: p(storyline)	Dim3: p(Visual)	Sent. p(pos)	p (pos | acting)	p(pos| storyline)	p (pos | visual)	Daily Box office	Day
1	7	0.6	0.1	0.3	0.6	0.36	0.06	0.18	$6,958,490	Jun 27, 2011
2	8	0.5	0.3	0.2	0.5	0.25	0.15	0.1	$7,089,413	Jun 28, 2011
3	3	0.4	0.5	0.1	0.3	0.12	0.15	0.03	$5,540,193	Jun 29, 2011
..										
n	1	0.1	0.8	0.1	0.2	0.02	0.16	0.02	$5,540,193	Jun 29, 2011

- Topic Analysis (content)
- Overall Sentiment Analysis
- Sentiment per Topic
- Time Analysis of Sentiment (overall and per topic)
- Time Analysis of Topic (overall and per topic)
- Co-Occurrence Analysis (e.g., lift)
- Network Analysis
- Social Influence Analysis

17

Text Analysis



Step 5 – Analysis

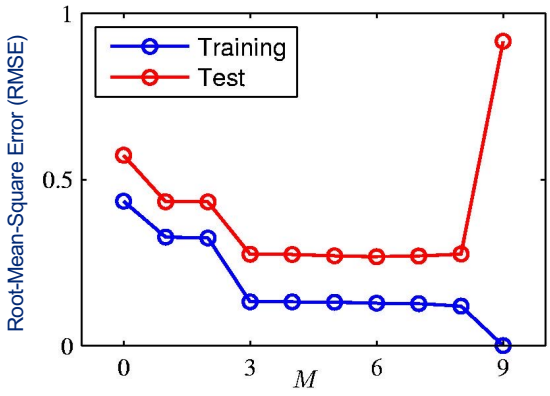
Computes distribution (mean, sd.) of sentiment or content across posts to understand variation over relevant groups such as

GENERAL	SUPPLY	DEMAND	COMPETITION	INSTITU-TION
Time	Products	Consumers	Competing firm	Markets
Reviewer	Product Categories	Geography (e.g., region, city)		
	Brands	RFM		
	Manager	CLV		
	Genre	Segments		
	Price	Cognitive Styles		
	Channels	Other characteristics		

18

20

Overfitting



21

Assessing Performance of a Learning Algorithm

Take out some of the training set

- Train on the remaining training set
- Test on the excluded instances

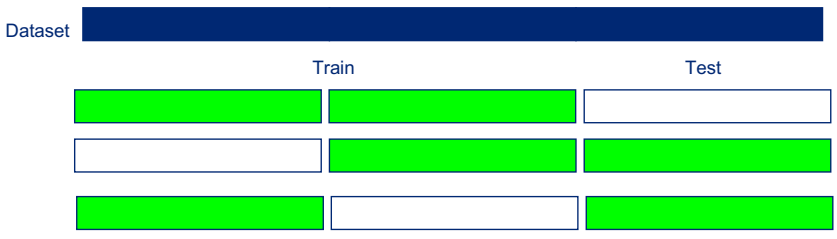
Report average results across all folds

Cross-validation

22

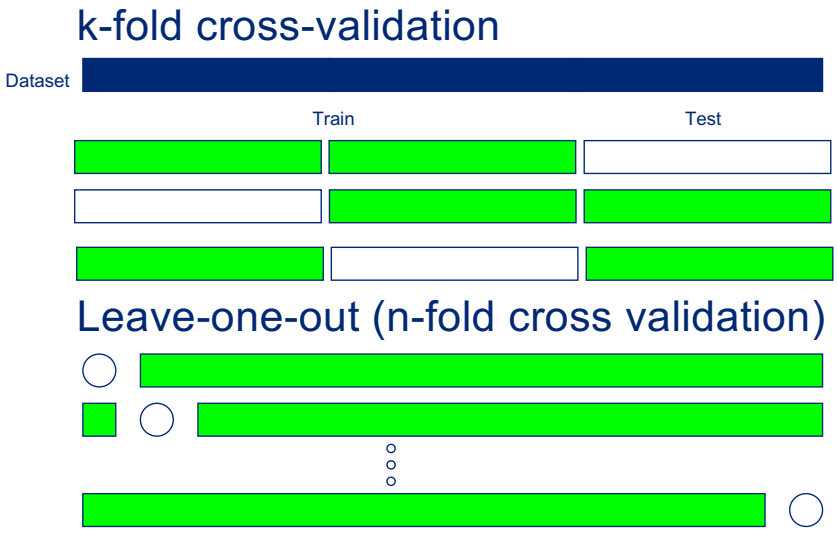
Common Data Splitting Strategies

k-fold cross-validation



23

Common Data Splitting Strategies



24

Confusion matrix

Say your model receive 13 images of cats and dogs and is asked to predict which images are of **cats.**

Here is how well it did:

		Actual class	
		Cat	Dog
Predicted class	Cat	5	2
	Dog	3	3

25

Confusion matrix

Say your model receive 13 images of cats and dogs and is asked to predict which images are of **cats.**

Here is how well it did:

		Actual class	
		Cat	Dog
Predicted class	Cat	5	2
	Dog	3	3

What is a false negative in this case?
And a false positive?

26

Confusion matrix

Say your model receive 13 images of cats and dogs and is asked to predict which images are of **cats.**

Here is how well it did:

		Actual class	
		Cat	Dog
Predicted class	Cat	5	2
	Dog	3	3

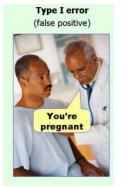
27

Confusion matrix

Say your model receive 13 images of cats and dogs and is asked to predict which images are of **cats.**

Here is how well it did:

		Actual class	
		Cat	Dog
Predicted class	Cat	5	2
	Dog	3	3



28

Confusion matrix

Say your model receive 13 images of cats and dogs and is asked to predict which images are of **cats.**

Here is how well it did:

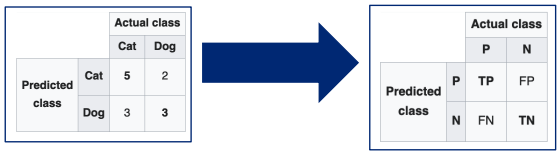
		Actual class	
		Cat	Dog
Predicted class	Cat	5	2
	Dog	3	3

		Actual class	
		P	N
Predicted class	P	TP	FP
	N	FN	TN

29

Confusion matrix – Accuracy and Precision

Say your model receive 13 images of cats and dogs and is asked to predict which images are of **cats.**



Accuracy – how correct are we overall (regardless the direction)
How many images did our model correctly label out of all images?

- Accuracy = (TP + TN) / (TP + FP + FN+TN)

Precision – how well do we do with positives?
How many of those we predicted as positive (e.g., as spam) are really positive (e.g., are spam) ?

- Precision = TP / (TP + FP)

There are many performance metrics to choose from including weighted scores (such as F1-score) and ROC curve.

- Your choice depends on the problem and what you want to minimize (FP? FN? Other?)
- Asymmetric cases: predicting whether a movie is okay for kids.
  - Best block a movie that would be okay instead of allowing a disturbing horror movie to be offered to a 6-year old kid.

30

Q. How can I know how good is my model? How can I compute the accuracy of my model?

31

Q. How can I know how good is my model? How can I compute the accuracy of my model?

A. The first part of the process is rather similar to obtaining training texts from judges.

You got this for free

1. Separate 20%-30% of your data for testing. The proportion can vary iteratively.
2. Break down your reviews in sentences
3. Give the sentences to a panel of judges.
4. Ask your judges to assign each sentence to a topic.
5. Compute, report the Cohen's Kappa **k** coefficient of agreement.
6. Compute the dimension of the sentences as given by argmax (posterior)

   e.g., arg max (0.05, 0.9, 0.05) = dimension 2

This is your job

7. For all chosen sentences, compare the dimensions agreed on by the judges and the dimensions given by your model **: FP, TP, TN, FN**
8. Report everything including the confusion matrix, accuracy and precision.

32

04

Examples

Classifiers in practice

33

33

What is Text Analysis Used for?

Collecting and using information on consumers, competition and the environment where commerce takes place to make better decisions.

34

What is Text Analysis Used for? Some examples

1. An on-going application using text analysis of *consumer forums, blogs and online press to observe how names of product categories evolve over time.*

   Application: radically new products

2. *text analysis of blogs to inspect how a brand is doing compared with others.*

   Application: pricing, product decisions, logistics decisions

3. Netzer et al., 2012 uses *text analysis of consumer forums to understand competition & market structure based on co-occurrence*
   - Who is competing against who
   - Problems and strong points of brands: ***pricing, product decisions, logistics decisions***

4. Liberali & Eliashberg (2021) use *text analysis of movie reviews to*
   - Forecast box office

35

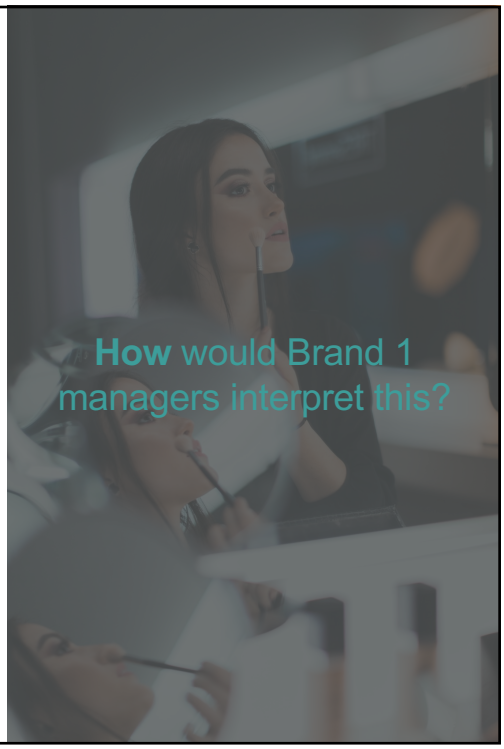
Example of a Project Using UGC:
Intelliseek

36

An UGC Application in the Make-Up Industry

Metric	Brand 1	Brand 2	Brand 3	Brand 4	Brand 5	Brand 6
Message Volume	585	174	1264	1071	1407	266
Sentiment						
Overall Score*	6.0	4.6	7.9	5.0	5.3	7.8
Positive	20%	16%	19%	17%	14%	18%
Negative	8%	6%	6%	8%	7%	5%
Neutral	70%	74%	74%	72%	77%	75%
Top Topics**						
Mascara	17%	18%	21%	35%	3%	5%
Lipcolor	15%	11%	9%	5%	15%	12%
Foundation	33%	14%	20%	30%	12%	1%
Total Boards	33	21	35	42	30	32
Total Authors	378	128	672	660	593	215

Sentiment is scored on a scale of 1 to 10, with 1 being negative and 10 being positive.% of Brand's Message Volume.*

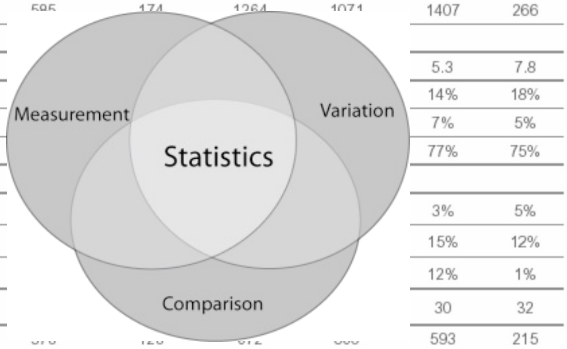


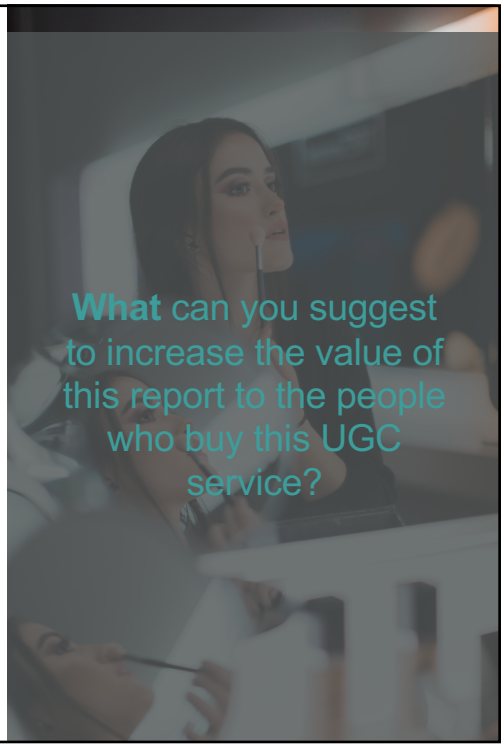
37

An UGC Application in the Make-Up Industry

Metric	Brand 1	Brand 2	Brand 3	Brand 4	Brand 5	Brand 6
Message Volume	595	174	1264	1071	1407	266
Sentiment						
Overall Score*					5.3	7.8
Positive					14%	18%
Negative					7%	5%
Neutral					77%	75%
Top Topics**						
Mascara					3%	5%
Lipcolor					15%	12%
Foundation					12%	1%
Total Boards					30	32
Total Authors	370	129	612	388	593	215

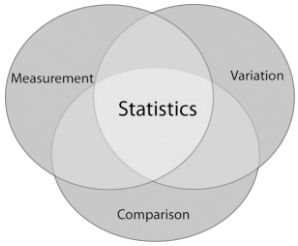
*Sentiment is scored on a scale of 1 to 10, with 1 being negative and 10 being positive.**% of Brand's Message Volume.





41

Analysis



It is often useful to check the distribution (mean and standard deviation) of sentiment or content across posts to understand variation over relevant groups.

Typical comparisons:

GENERAL	SUPPLY	DEMAND	COMPETITION	INSTITUTION
Time	Products	Consumers	Competing firm	Markets
Reviewer	Product Categories	Geography (e.g., region, city)		
	Brands	RFM		
	Manager	CLV		
	Genre	Segments		
	Price	Cognitive Styles		
	Channels	Other characteristics		

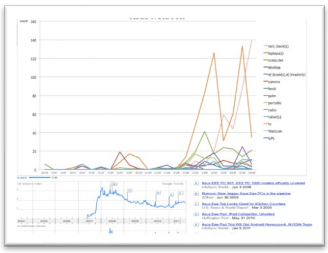
42

Some Applications of Text Analysis

Marketing is about collecting and using information on consumers, competition and the environment where commerce takes place to make better decisions.

:

1. An on-going application using text analysis of *consumer forums, blogs and online press to observe how names of product categories evolve over time.* ***Application:*** *launch of radically new products*

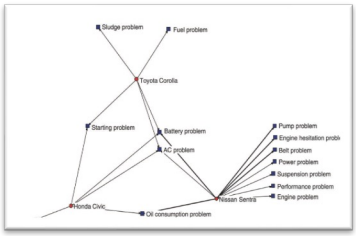


2. An application using *text analysis of blogs to inspect how a brand is doing compared with others.* ***b****: Improve 4Ps, detect early signals of issues*

Metric	Brand 1	Brand 2	Brand 3	Brand 4	Brand 5	Brand 6
Message Volume	585	174	1264	1071	1407	266
Sentiment						
Overall Score*	6.0	4.6	7.9	5.0	5.3	7.8
Positive	20%	16%	19%	17%	14%	18%
Negative	8%	6%	6%	8%	7%	5%
Neutral	70%	74%	74%	72%	77%	75%
Top Topics**						
Mascara	17%	18%	21%	35%	3%	5%
Lipcolor	15%	11%	9%	5%	15%	12%
Foundation	33%	14%	20%	30%	12%	1%
Total Boards	33	21	36	42	30	32
Total Authors	378	128	672	660	593	215

*Sentiment is scored on a scale of 1 to 10, with 1 being negative and 10 being positive.**% of Brand's Message Volume.

3. Netzer et al., 2012 uses *text analysis of consumer forums to understand competition & market structure based on co-occurrence*
- Who is competing against who
- Problems and strong points of brands.



4. Liberali & Eliashberg (2015) use *text analysis of movie reviews to*
- Find out what drives review helpfulness – useful for review platforms to become more relevant

43

Example of a Project Using UGC:

Understanding Market Structure and Detecting Problems

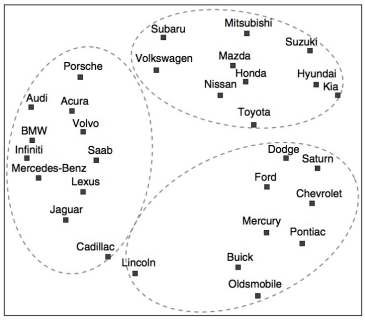
44

Application – Auto Industry

Netzer et al., 2012 uses text analysis of consumer forums to understand competition & market structure based on co-occurrence

- Who is your competitor in the eyes of the consumer
- Problems and strong points of brands.

Figure 3 MDS Map of Discussion of Car Brands



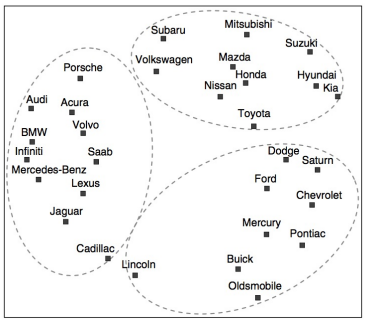
45

Application – Auto Industry

Netzer et al., 2012 uses text analysis of consumer forums to understand competition & market structure based on co-occurrence

- Who is your competitor in the eyes of the consumer
- Problems and strong points of brands.

Figure 3 MDS Map of Discussion of Car Brands



Consider a review website about cars and a simple case
P(toyota) = 2/100 = 0.02
P(break) = 3/100 = 0.03

P(toyota) x P(break) = 0.02 x 0.03 = 0.0006

If “Toyota” and “break” appear more often than that there is information in there

46

Application – Auto Industry

Netzer et al., 2012 uses text analysis of consumer forums to understand competition & market structure based on co-occurrence

- Who is your competitor in the eyes of the consumer
- Problems and strong points of brands.

Consider a review website about cars and a simple case
P(toyota) = 2/100 = 0.02
P(break) = 3/100 = 0.03

P(toyota) x P(break) = 0.02 x 0.03 = 0.0006

Figure 3 MDS Map of Discussion of Car Brands

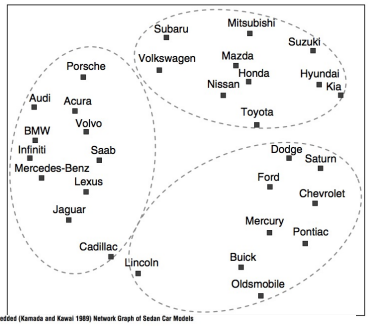
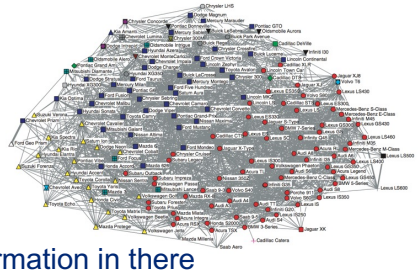


Figure 2 Spring-Embedded (Kamada and Kawai 1989) Network Graph of Sedan Car Models



If "Toyota" and "break" appear more often than that there is information in there

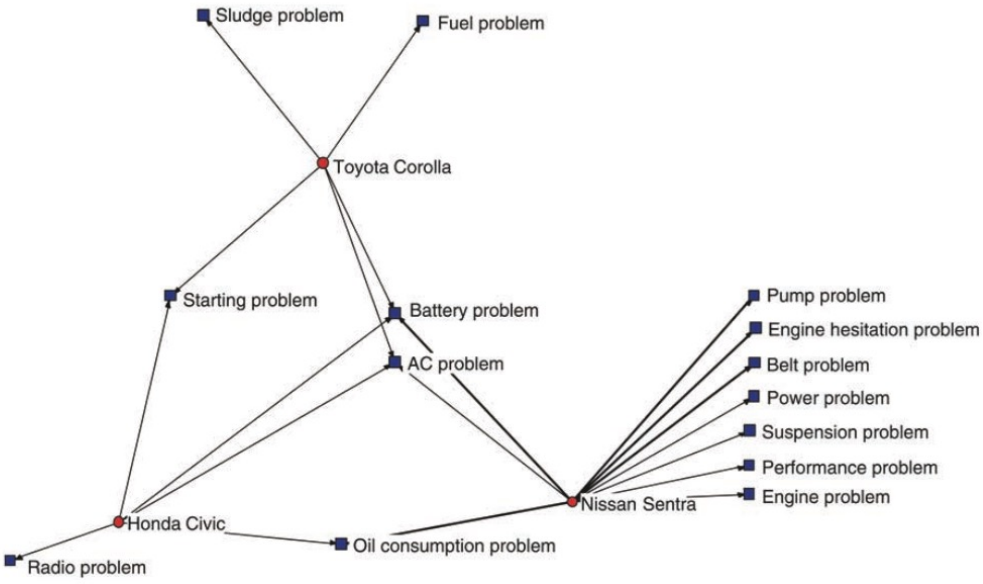
The shapes and color of the nodes (cars) in Figure 2 represent cluster membership.

47

Application – Auto Industry

Text analysis of consumer forums shows problems of brands

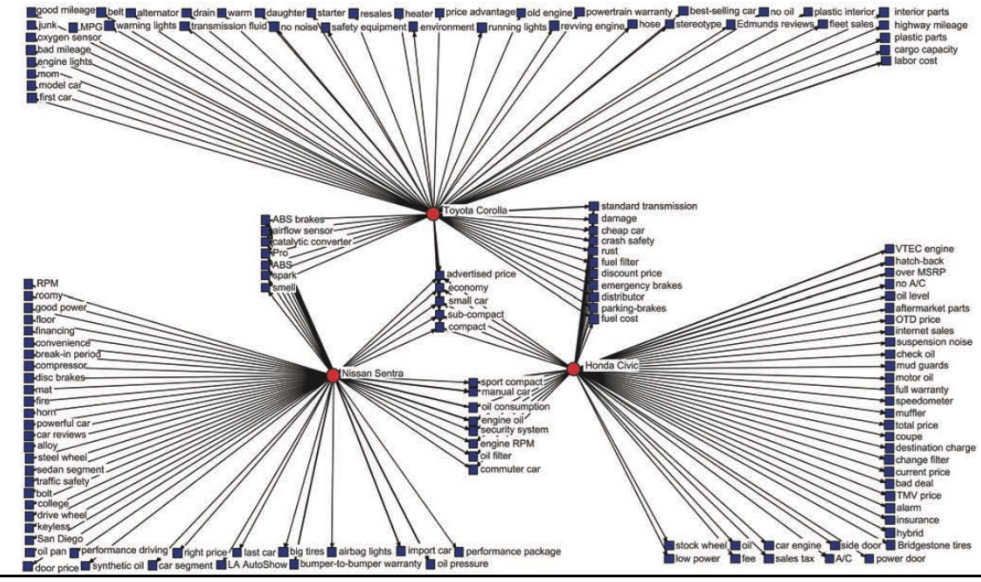
Figure 9 Problems Commonly Appearing with the Honda Civic, Nissan Sentra, and Toyota Corolla



48

Visualization of complex data can be handled

Figure 8 **Terms Commonly Appearing with the Honda Civic, Nissan Sentra, and Toyota Corolla**



Example of a Project Using UGC:

Understanding the Creation of New Categories

51

Products and Buzz

Category (not used as filter)	Pioneer	Another typical product
*Netbook	1.Palm Foleo (30/05/2007)	25. ASUS Eee PC 700 (16/10/2007)
*e-reader	2. LIBRIé ou PRS-500 Sony (09/2006)	26. Kindle (19/11/2007)
*Tablet	3. Ipad (03/04/2010)	27. Samsung Galaxy Tab (02/10/2010)
*GPS navigation device	4. TomTom GO (01/07/2004)	28. StreetPilot c330 (05/2005)
*Webcam / QuickCam	5. Connectix QuickCam (01/2005)	29. iSight (16/12/2006)
*Blu-ray Players	6. Sony BDP-S1 (08/2006)	30.Samsung BD-P1000 (09/2006)
*Portable Video Players	7. Creative Zen Vision (09/2005)	31.Archos AV500 Mobile DVR (11/2005)
*Motion Controller	8. Wii Remote (23/04/2006)	32. Playstation Move (02/06/2009)
*Bluetooth Headset	9. Jabra FreeSpeak BT250 (06/2004)	33. Motorola HS850 (05/2005)
Portable Speakers	10. Sony SRS-T57 (06/2004)	34. Altec Lansing inMotion iM4 (07/2005)
Portable Projector	11. HP ep9010 (08/2004)	35. Mitsubishi PK-10 (02/2005)
*Mobile TV	12. Nokia N92 (11/2005)	36. Samsung SGH-P910 (2006)
*OLED TV – Ultra Thin TV	13. Sony XEL-1 (12/2007)	37. Hitachi 1.5 (01/2008)
*LaserTV	14. LaserVue L65-A90 (2008)	38. Mitsubishi LaserVue L75-A91 (2010)
*Digital Cameras MFT	15. Panasonic Lumix DMC-G1 (10/2008)	39. Olympus PEN E-P1 (07/2009)
Digital Photo Frames	16. Philips 7FF1 (01/2006)	40. Kodak EasyShare EX-811 (05/2007)
*Digital Media Receivers	17. Linksys WMLS11B (05/2004)	41. ViewSonic WMA100 (11/2004)
*Personal Mobility/ Scooters	18. Segway i2 (07/2006)	42. Toyota Winglet (07/2008)
Travel Cooler	19. Black & Decker BDV212F (07/2007)	43. Vector 12V Mini Console (11/2007)
RGB LED design lamp	20. LivingColors Philips (11/2007)	44. Multi-Color E27 LED Light Bulb (2008)
Indoor Pet Containment	21. PetSafe PIF-275 (2004)	45. Innotek ZND-100 (2006)
Crocs	22. Croc (2004)	46. Cloggens (2007)
*Microblogging	23. Twitter (2006)	47. Pownce (01/2008)
*Social Buying	24. LivingSocial (2007)	48. Groupon (2008)

53

Asus Eee PC - Netbook

Co-occurrence of Name of the Product and its Category in Consumer Posts

Table 1. Word count and Density Index for **Consumer** Buzz in the last period of **ASUS**

Category i	Word count at last t	$D^{C}_{last\ t,i,j=Asus}$
net(-)book(s)	139	0.604
laptops(s)	35	0.152
computer	21	0.091
camera	11	0.048
tablet(s)	10	0.043
desktop	6	0.026
portable	4	0.017
radio	3	0.013
e(-)book(s),e(-)reader(s)	1	0.004
Total	230	1

RSM Erasmus

Your contact persons.

Gui Liberali
Professor of Digital Marketing

liberali@rsm.nl
+31 10 4082732

59