RECAP + PEER-FEEDBACK

ADVANCED CRIME ANALYSIS

UCL

BENNETT KLEINBERG

11 MARCH 2019

UCL

Recap + peer-feedback

TODAY

- 2 case studies
- module recap
- your feedback
- peer-feedback

CASE STUDY 1 (URL)

Identifying the sentiment styles of YouTube's vloggers

Bennett Kleinberg
Department of Psychology
University of Amsterdam

Department of Security and Crime Science
University College London
b.a.r.kleinberg@uva.nl

Maximilian Mozes
Department of Informatics
Technical University of Munich
mozes@cs.tum.edu

Isabelle van der Vegt
Department of Security and Crime Science
University College London
isabelle.vegt.17@ucl.ac.uk

SENTIMENT STYLES: RQ

Are there patterns in sentiment usage in popular YouTube vlogs?

NEEDED

- vlog data
- sentiment “trajectories”
- cluster analysis

DATA PIPELINE

- retrieved list of most popular vloggers
- excluded those that were no real vloggers and non-English speaking
- scraped transcripts for each vlog
  - 27,333 transcripts
  - 40m tokens
  - 24b views

ANALYSIS

- unsupervised learning
- k-means method
- assigned each vlog to its cluster

FINDINGS

- 7 distinct sentiment styles
- preferance was moderated by gender

FINDINGS

Cluster	Family	Female	Male
Downhill from here	2.23	1.26	-2.88*
Mood swings	-2.31	1.96	1.25
Rags to riches	2.13	-1.95	-1.08
Riches to rags	-2.05	4.88*	-0.56
Bump in the road	1.69	-1.12	-1.08
End on a high note	-5.16*	-6.03*	8.32*
Twin peaks	3.83*	2.25	-4.99*

FINDINGS

- 7 distinct sentiment styles
- preferance was moderated by gender
- no effect on view count
- no effect on vlog length

CASE STUDY 2(URL)

Automatic Detection of Fake News

Verónica Pérez-Rosas[1], **Bennett Kleinberg**[2], **Alexandra Lefevre**[1]
Rada Mihalcea[1]
[1]Computer Science and Engineering, University of Michigan
[2]Department of Psychology, University of Amsterdam
vrncapr@umich.edu,b.a.r.kleinberg@uva.nl,mihalcea@umich.edu

FAKE NEWS PAPER: RQ

Can we detect fake news based on linguistic features in the news article?

NEEDED

- fake and real news
- different domains of news
- predictive analysis

DATA PIPELINE

- sourcing “real” news from mainstream news sites
- crowdsourcing approach for fake news
- news domain (sports, business, politics, technology, entertainment, education)

DATA PIPELINE

LEGITIMATE	FAKE
Nintendo Switch game console to launch in March for $299 The Nintendo Switch video game console will sell for about $260 in Japan, starting March 3, the same date as its global rollout in the U.S. and Europe. The Japanese company promises the device will be packed with fun features of all its past machines and more. Nintendo is promising a more immersive, interactive experience with the Switch, including online playing and using the remote controller in games that don't require players to be constantly staring at a display.	**New Nintendo Switch game console to launch in March for $99** Nintendo plans a promotional roll out of it's new Nintendo switch game console. For a limited time, the console will roll out for an introductory price of $99. Nintendo promises to pack the new console with fun features not present in past machines. The new console contains new features such as motion detectors and immerse and interactive gaming. The new introductory price will be available for two months to show the public the new advances in gaming.

Table 2: Sample legitimate and crowdsourced fake news in the Technology domain

DATA PIPELINE

LEGITIMATE	FAKE
Kim And Kanye Silence Divorce Rumors With Family Photo. Kanye took to Twitter on Tuesday to share a photo of his family, simply writing, "Happy Holidays." In the picture, seemingly taken at Kris Jenner's annual Christmas Eve party, Kim and a newly blond Kanye pose with their children, North, 3, and Saint, 1. After Kanyes hospitalization, reports that there was trouble in paradise with Kim started brewing. But E! News shut down the speculation with a family source denying the rumors and telling the site, "It's been a very hard couple of months."	**Kim Kardashian Reportedly Cheating With Marquette King as She Gears up for Divorce From Kanye West.** Kim Kardashian is ready to file for divorce from Kanye West but has she REALLY been cheating on him with Oakland Raiders punter Marquette King? The NFL star seemingly took to Twitter to address rumors that they've been getting close amid Kanye's mental breakdown, which were originally started by sports blogger Terez Owens. While he doesn't appear to confirm or deny an affair, her reps said there is "no truth whatsoever" to the reports and labeled the situation "fabricated."

Table 3: Sample legitimate and web fake news in the Celebrity domain

DATA PIPELINE

- sourcing “real” news from mainstream news sites
- crowdsourcing approach for fake news
- news domain (sports, business, politics, technology, entertainment, education)
- celebrity domain
  - Entertainment Weekly, People Magazine,RadarOnline
  - checked with GossipCop.com and other online sources

ANALYSIS

- extracted linguistic features
  - ngrams, punctuation, LIWC, readability, syntax
- supervised learning
  - SVM classifier
  - 5-fold CV
  - 50% baseline

FINDINGS

Features (# features)	Acc.	$F1_{Legit.}$	$F1_{Fake}$
Punctuation (12)	0.71	0.69	0.72
LIWC-Summ (7)	0.61	0.58	0.64
LIWC-LingProc. (21)	0.67	0.66	0.66
LIWC-PsyProc. (40)	0.56	0.56	0.55
LIWC (80)	0.70	0.70	0.70
Readability (26)	0.78	0.77	0.79
Ngrams (634)	0.62	0.62	0.62
CFG (1377)	0.65	0.64	0.65
All Features (2140)	0.74	0.74	0.74

Table 4: Classification results for the FakeNewsAMT dataset collected via crowdsourcing.

FINDINGS

Features (# features)	Acc.	$F1_{Legit.}$	$F1_{Fake}$
Punctuation (12)	0.69	0.69	0.69
LIWC-Summ. (7)	0.67	0.66	0.69
LIWC-LingProc (21)	0.72	0.72	0.71
LIWC-PsyProc (40)	0.67	0.68	0.66
LIWC (80)	0.74	0.74	0.74
Readability (28)	0.62	0.61	0.63
Ngrams (1317)	0.71	0.72	0.71
CFG (2599)	0.72	0.72	0.72
All Features (4048)	0.76	0.77	0.76

Table 5: Classification results for the Celebrity news dataset.

FINDINGS

	FakeNewsAMT	Celebrity
A1	0.71	0.80
A2	0.70	0.77
Sys	0.74	0.76

Table 9: Performance of two annotators (A1, A2) and the developed automatic system (Sys) on the fakenews datasets

MODULE RECAP

YOU'VE LEARNED A LOT:

- data generation/collection
- text analysis
- machine learning

DATA GENERATION/COLLECTION

- APIs
- custom-made webscraping

TEXT ANALYSIS

- sentiment
- ngrams
- LIWC
- TFIDF

MACHINE LEARNING

- supervised learning
- unsupervised learning
- cross-validation
- performance metrics

YOUR FEEDBACK?

PEER-FEEDBACK SESSION

NEXT WEEK

CLASS TEST

- Monday, 18 March 2019
- 1-3pm
- 60 min
- 10 questions (5 MC, 5 open)

END