Institute of Policy Studies

ENGAGING MINDS, EXCHANGING IDEAS

Digital Frontiers Seminar: Man and Machine (2): Online and Offline Sensing of Public Sentiments on National Day Rally 2016

Thursday, 23 February 2017

Seminar Room 3-1, Level 3, Manasseh Meyer Building

Lee Kuan Yew School of Public Policy

National University of Singapore

Digital Frontiers Seminar:
Man and Machine (2): Online and Offline Sensing of Public Sentiments on National Day Rally 2016

Part I
Big Data and Online Sensing

Associate Professor Jiang Jing
Living Analytics Research Centre
School of Information Systems
Singapore Management University

Digital Frontiers Seminar:
Man and Machine (2): Online and Offline Sensing of Public Sentiments on National Day Rally 2016

Part I
Big Data and Online Sensing

Dr Palakorn Achananuparp
Research Scientist
Living Analytics Research Centre
School of Information Systems
Singapore Management University

Lee Kuan Yew School of Public Policy
National University of Singapore

Institute of Policy Studies

Engaging Minds, Exchanging Ideas

Analytics for Consumer & Social Insights
Computational Social Science
LARC
Living Analytics Research Centre
Network-Centric
Experiment-Driven
Closed-Loop

Singapore Management University
School of Information Systems

Online Sensing for NDR 2016

Background

- How did Singaporeans respond to the topics covered in NDR 2016?
  - Traditional survey – scale is limited
  - Analyzing social media – data is public, number of online users is big
- LARC's capabilities in social sensing
  - Continuous crawling of social media content (Twitter, blogs, forums, Facebook, etc.) since 2011
  - Development of advanced data analytics techniques

Objectives of the Study

- By collecting and analyzing online content related to NDR 2016, we aim to understand the general public's interests and sentiments on NDR topics.

- By comparing the findings of online sensing with those from the survey done by IPS, we aim to gauge the representativeness of public interests and sentiments harnessed from social media.

Research Questions

- Which online platforms were the most popular for discussing NDR 2016?
- Which NDR topics gained the most attention online?
- What were the online sentiments towards NDR topics?
- Which NDR topics gained the most divided sentiments?
- How do topics differ across different online platforms?
- How do sentiments toward various topics differ across different online platforms?
- How close are online sensing results to survey results?

Data Sources

- Online content between Aug 7, 2016 and Sep 10, 2016. (Two weeks before and two weeks after NDR 2016)
- Platforms:
  - Twitter: ~150K Singapore-based users
  - Facebook: public pages identified by IPS
  - Blogs: 243 blogging sites (Mothership, Online Citizen, etc.)
  - Forums: Hardwarezone, AsiaOne, Sammy's Alfresco Coffee Forum, Reddit Singapore
  - Websites of mainstream media: Straits Times, Today, etc.
- Two types of content: posts and comments
  - Posts: content published independent of any previous content
  - Comments: content published in response to a post

Methodology

collected online posts and comments

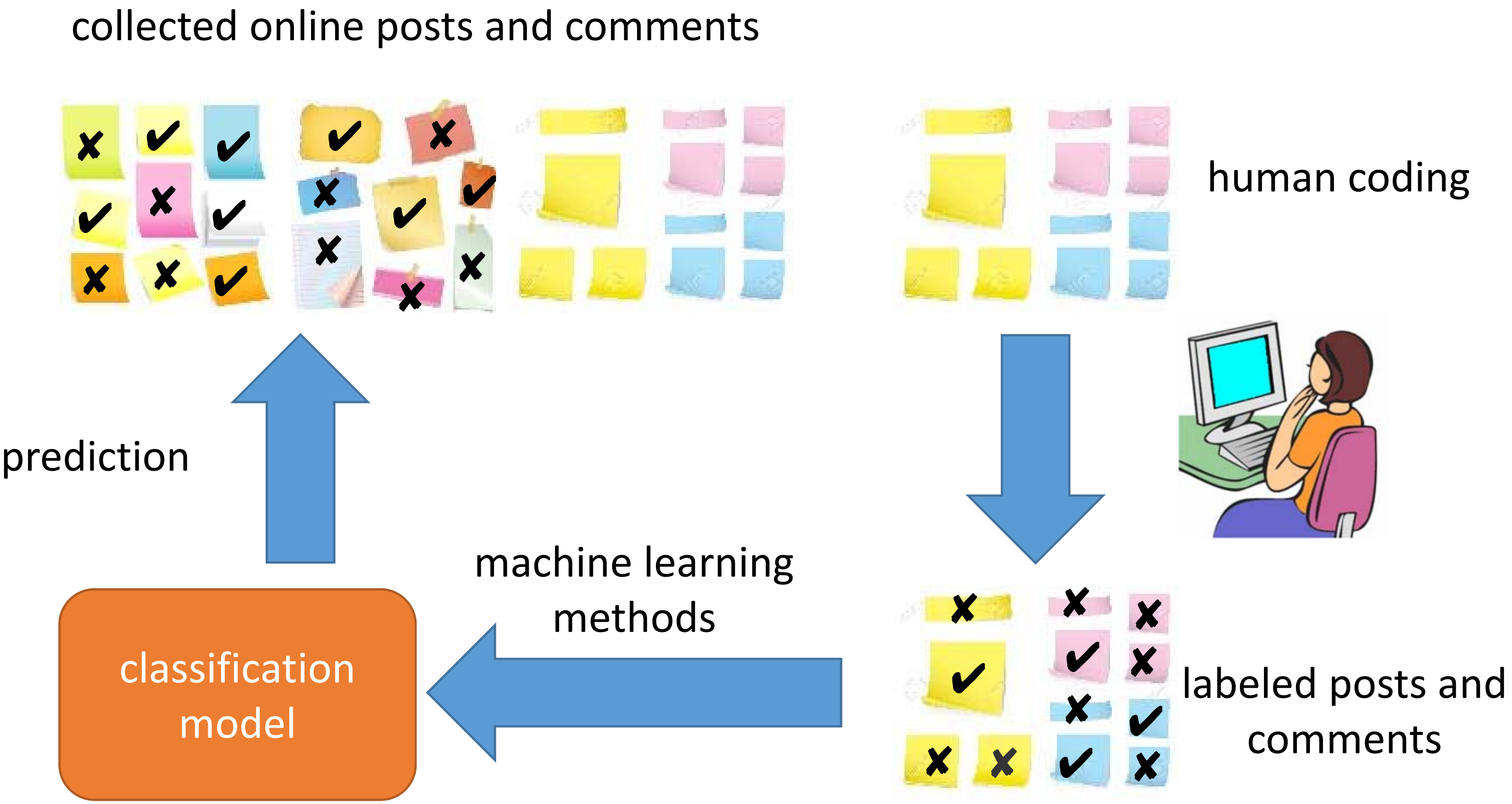
Step 1: Identify NDR-related content

Step 2: Categorize posts and comments by topics

Step 3: Classify sentiments

We use a combination of human coding and machine learning to perform the tasks above.

Methodology



collected online posts and comments

human coding

labeled posts and comments

machine learning methods

classification model

prediction

Methodology

- Step 1: Identify NDR-related content
  - Labels: relevant and irrelevant
- Step 2: Categorize posts and comments into topics
  - Labels: 26 topics identified by IPS
- Step 3: Classify sentiments
  - Labels: positive and negative

Topics

Elected Presidency	Eligibility Criteria	Minority Representation	Social Cohesion
Foreign Policy	South China Sea Dispute	Bilateral Relations	Terrorism
Leadership Succession	ARS	Good Politics	Progress Together
Entrepreneurship and Innovation	Disruption to Economy	Economy Performance	SkillsFuture
Heng Swee Keat	Lawrence Wong	Minister for Finance	Eldershield
PM Lee	Unwell	Schooling	Pokemon
Delivery of Speech	Others		

Classification Accuracy

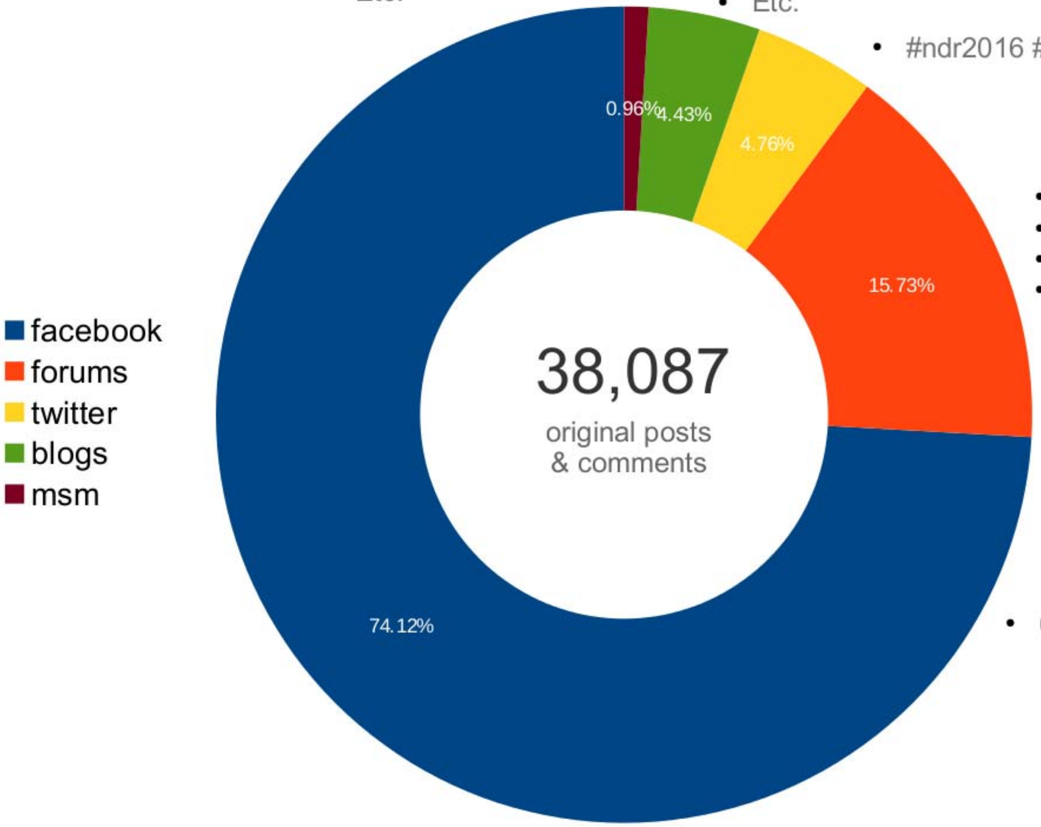
- Step 1 (relevant content): 0.8469
- Step 2 (topic categorization): 0.7351 (on blogs), 0.6573 (on tweets)
- Step 3 (sentiment classification): 0.84

- Typically a classification accuracy of 0.75 – 0.8 is considered good.
- Accuracy goes down when there are many categories and/or there is little training data.

Which online platforms were the most **popular**?

Content Volume = Original Posts + Comments

Content Volume By Platforms



6 mainstream media sites

- The Straits Times
- Today
- Business Times
- Etc.

32 blog sites

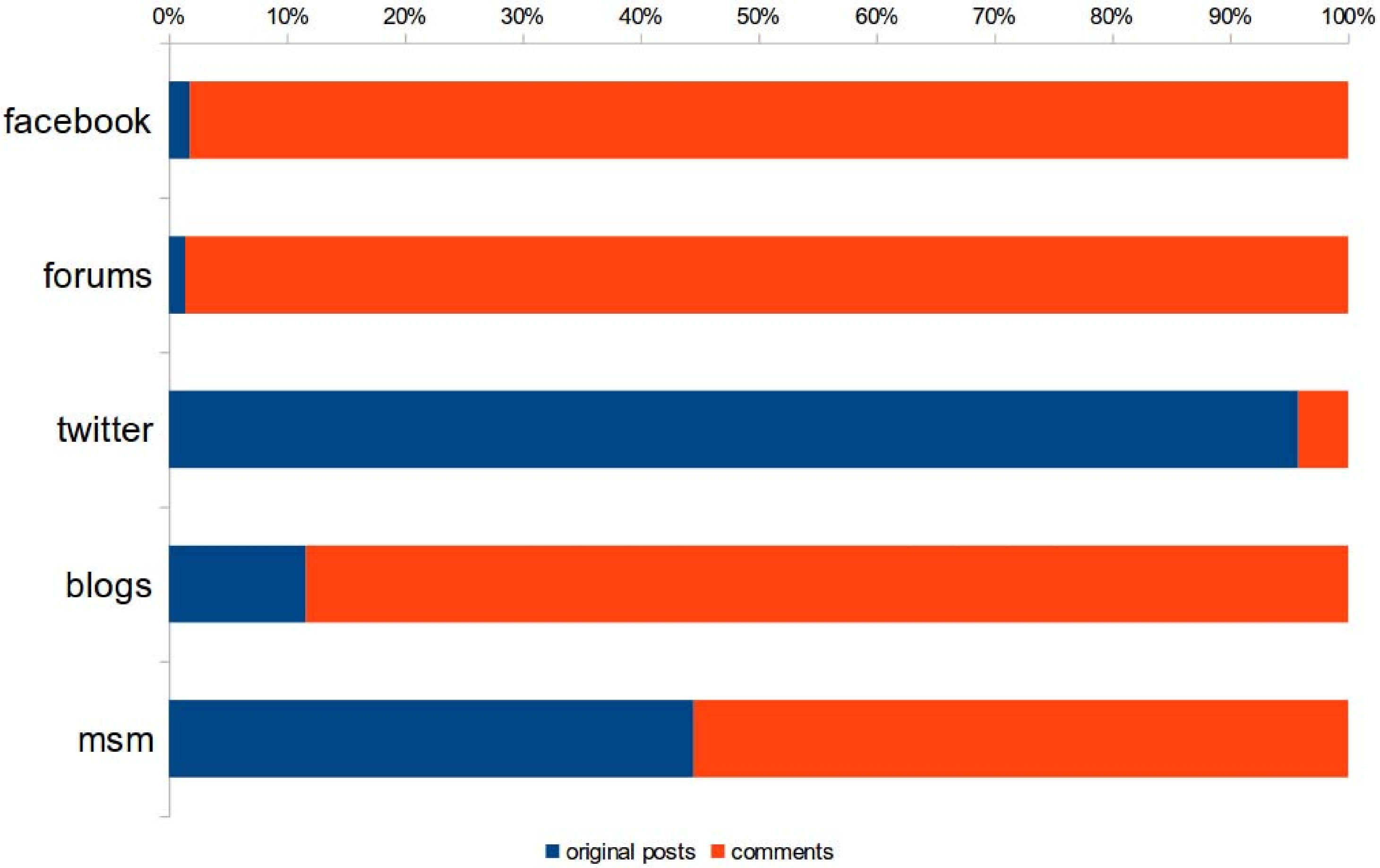
- Mothership
- The Online Citizen
- The Middle Ground
- Etc.

- #ndr2016 #ndrsg tweets

- Hardwarezone forum
- Sam's Alfresco Coffee forum
- AsiaOne forum
- Reddit /r/singapore

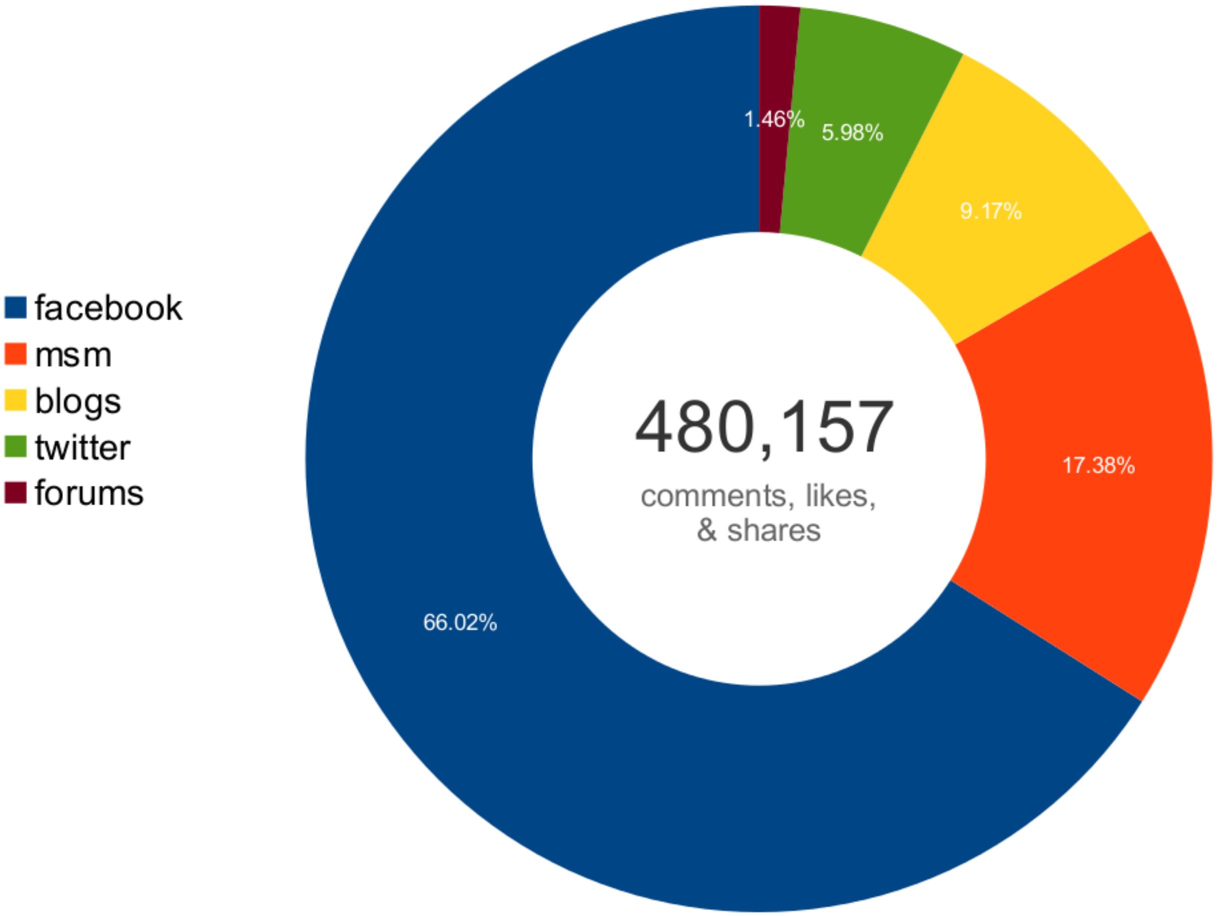
- 69 Facebook pages

Proportions of Content Types By Platforms

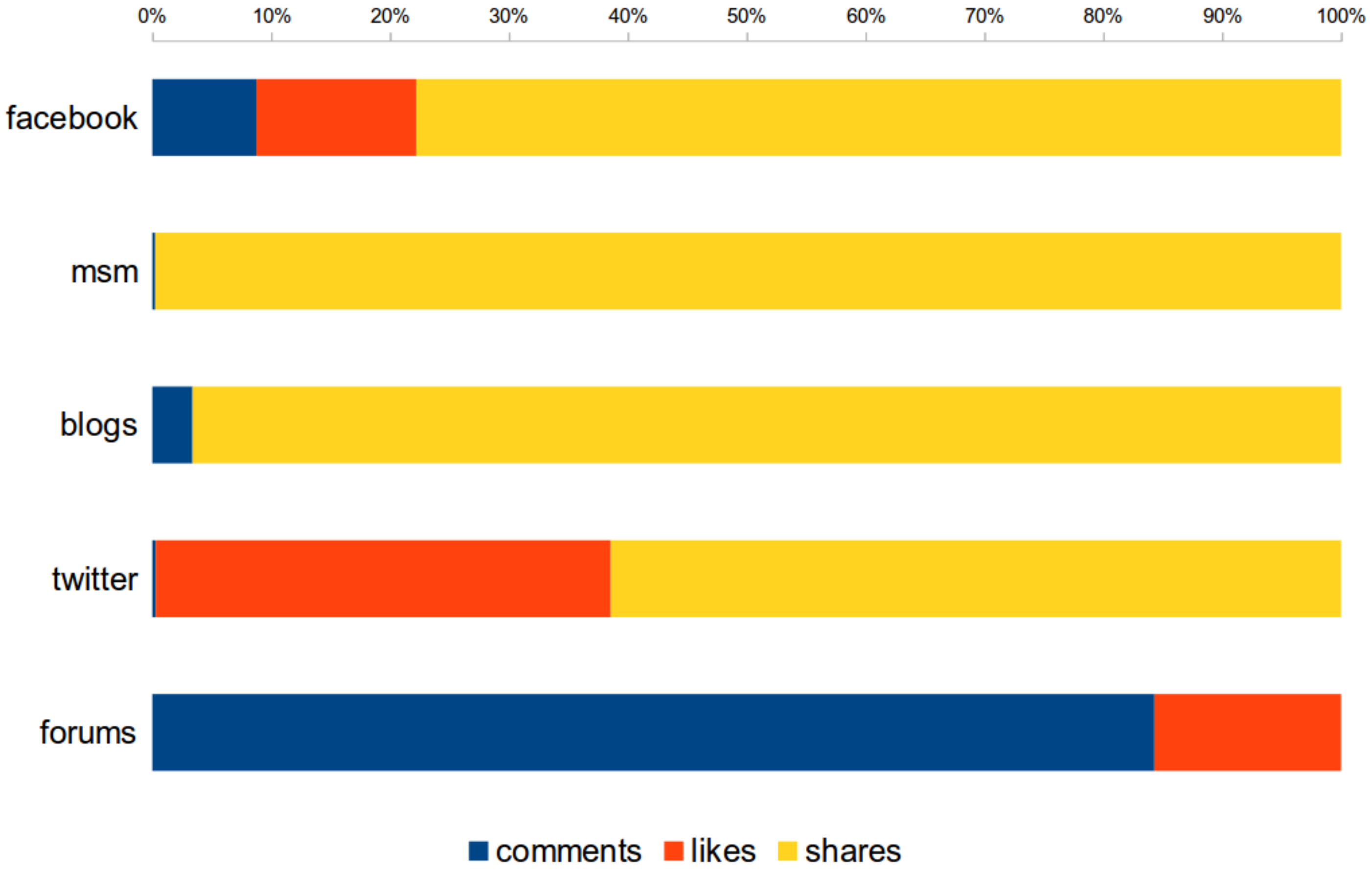


Engagement Volume = Comments + Likes + Shares

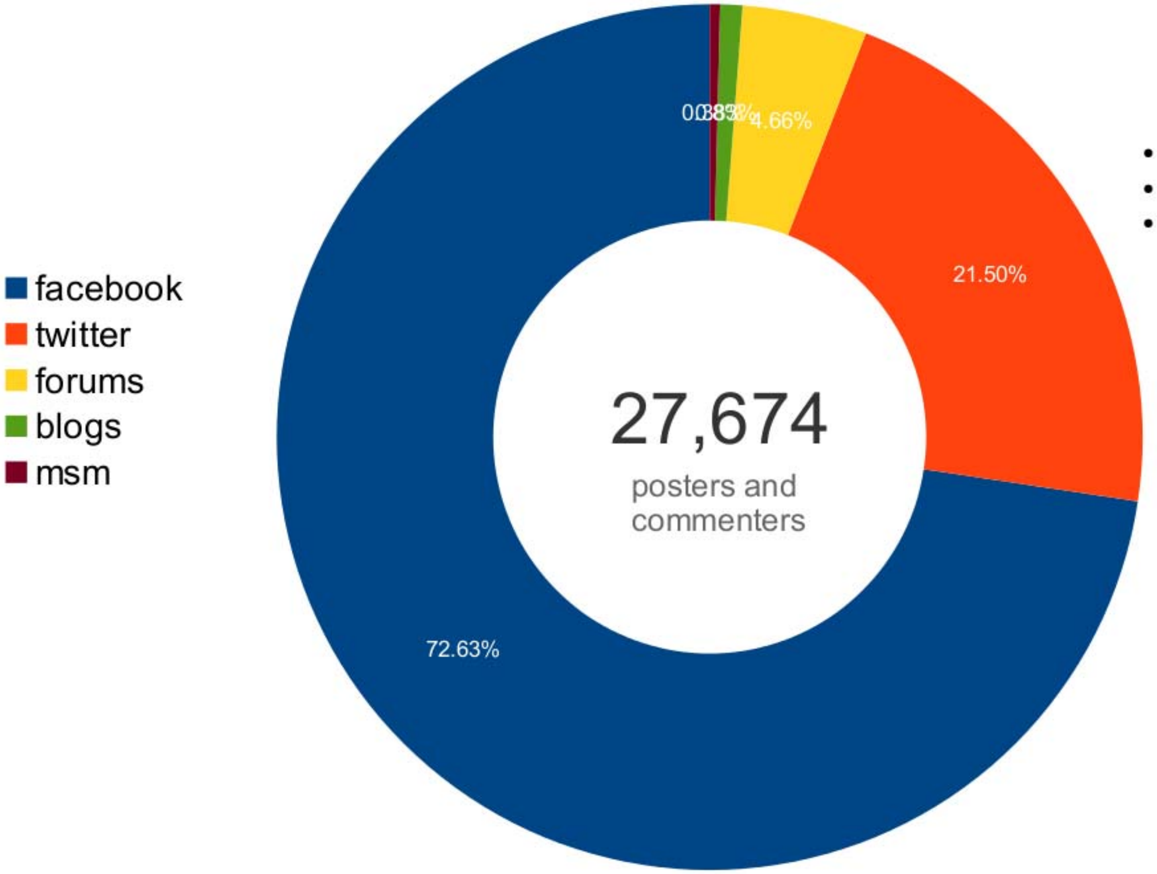
Engagement Volume By Platforms



Proportions of Engagement Types By Platforms

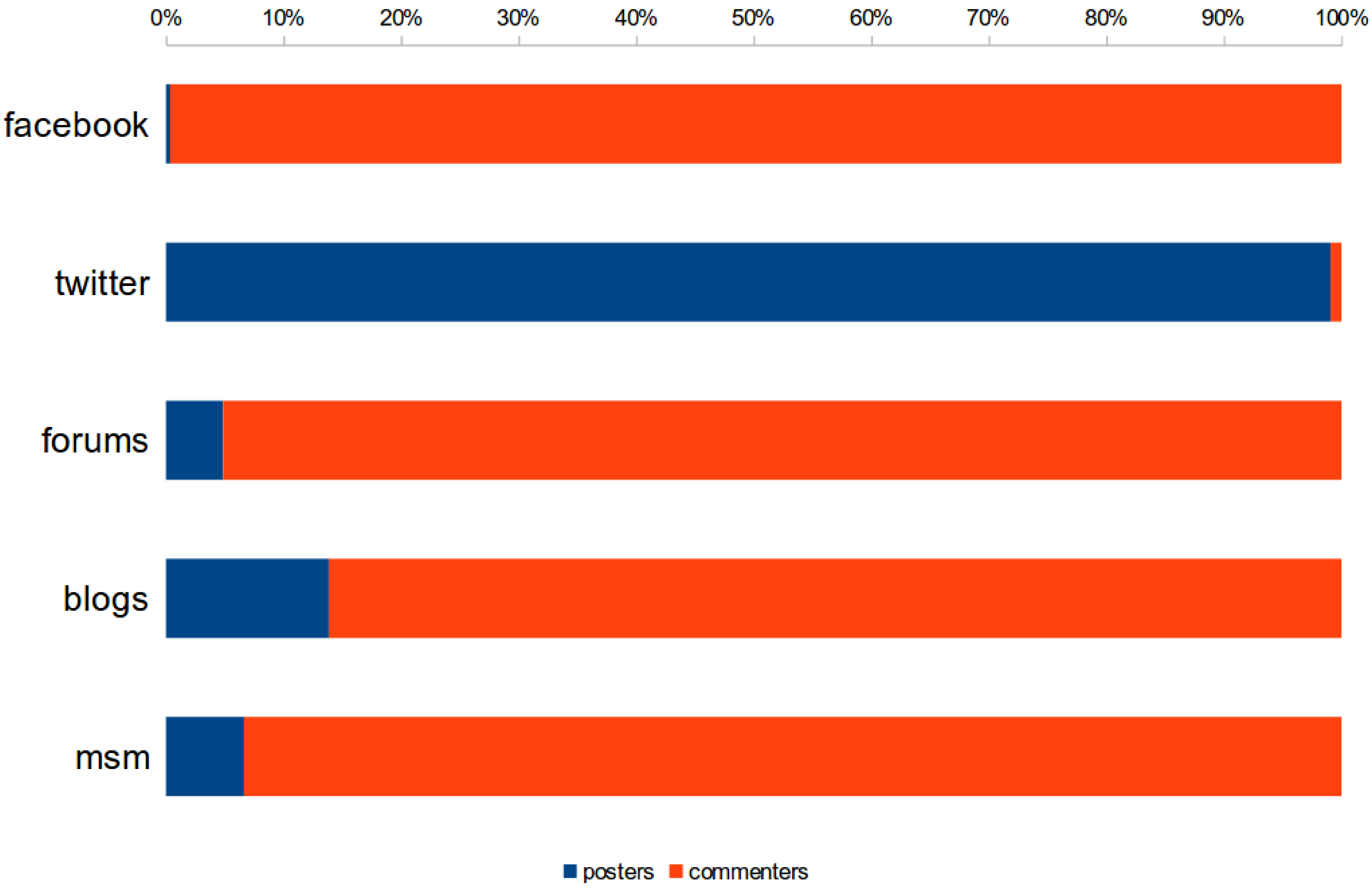


Numbers of Users By Platforms



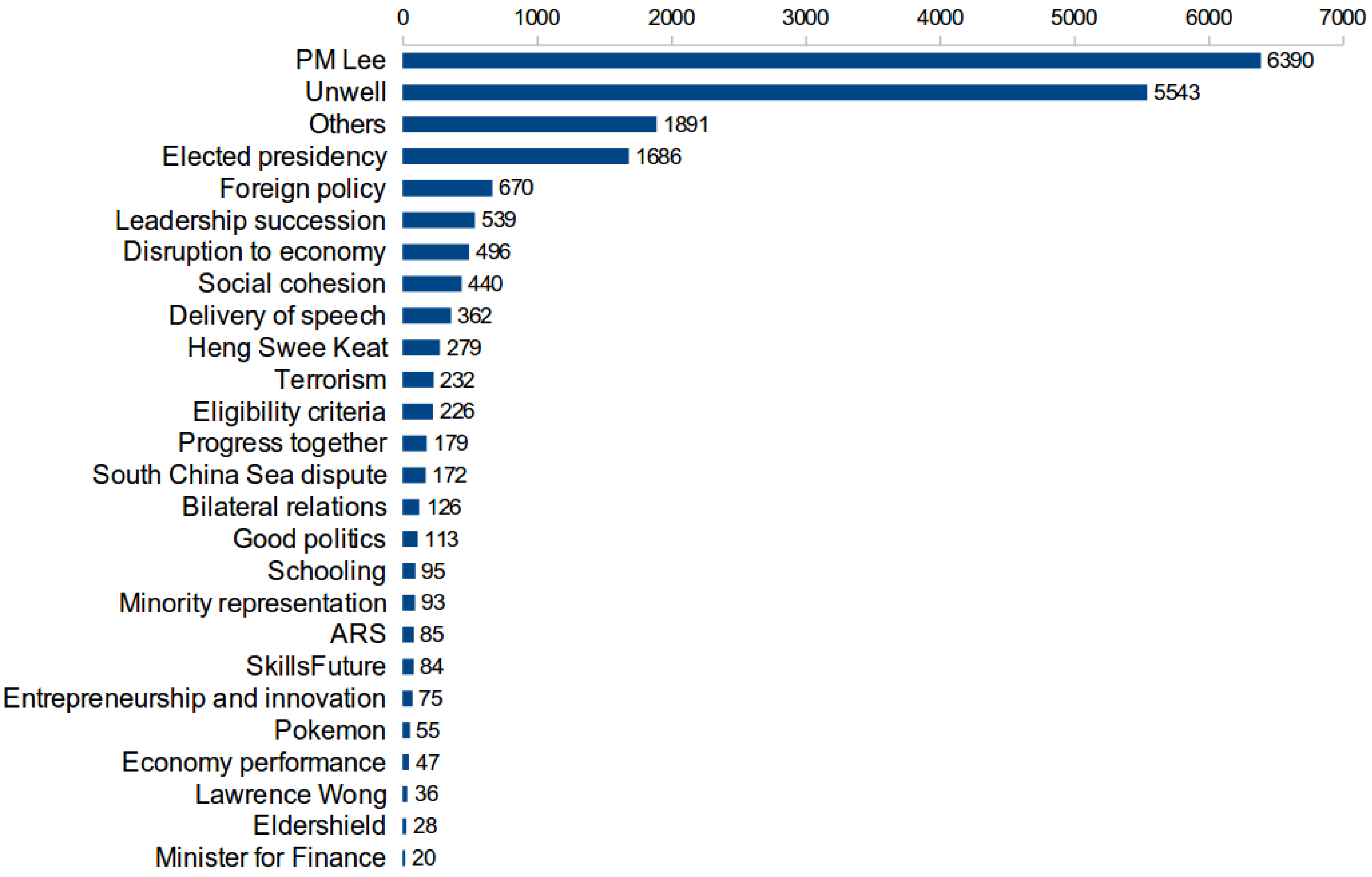
- 561 Posters
- 5334 Retweeters
- 55 Commenters

Proportions of Users Types By Platforms

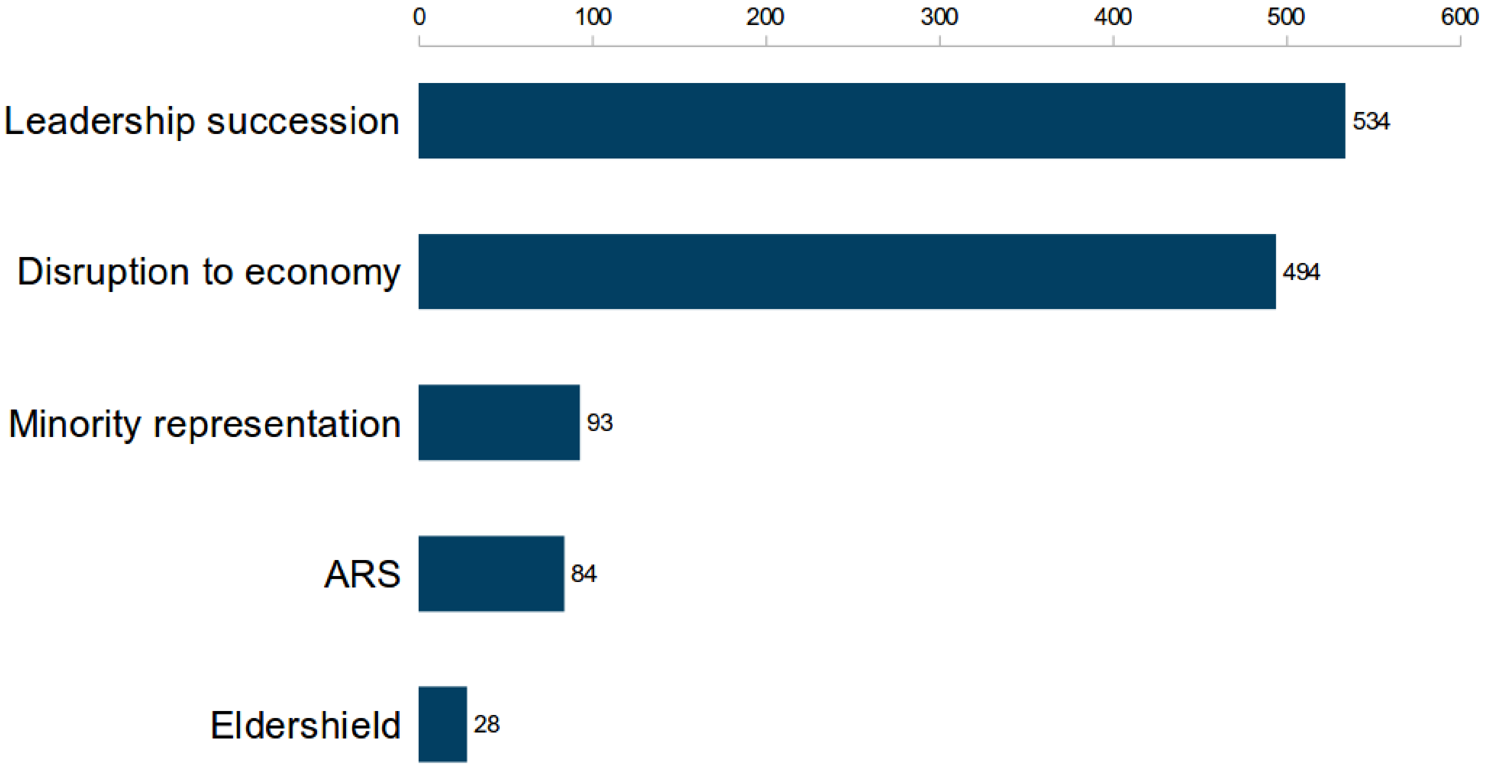


Which NDR topics gained the most attention online?

Content Volume By Topics

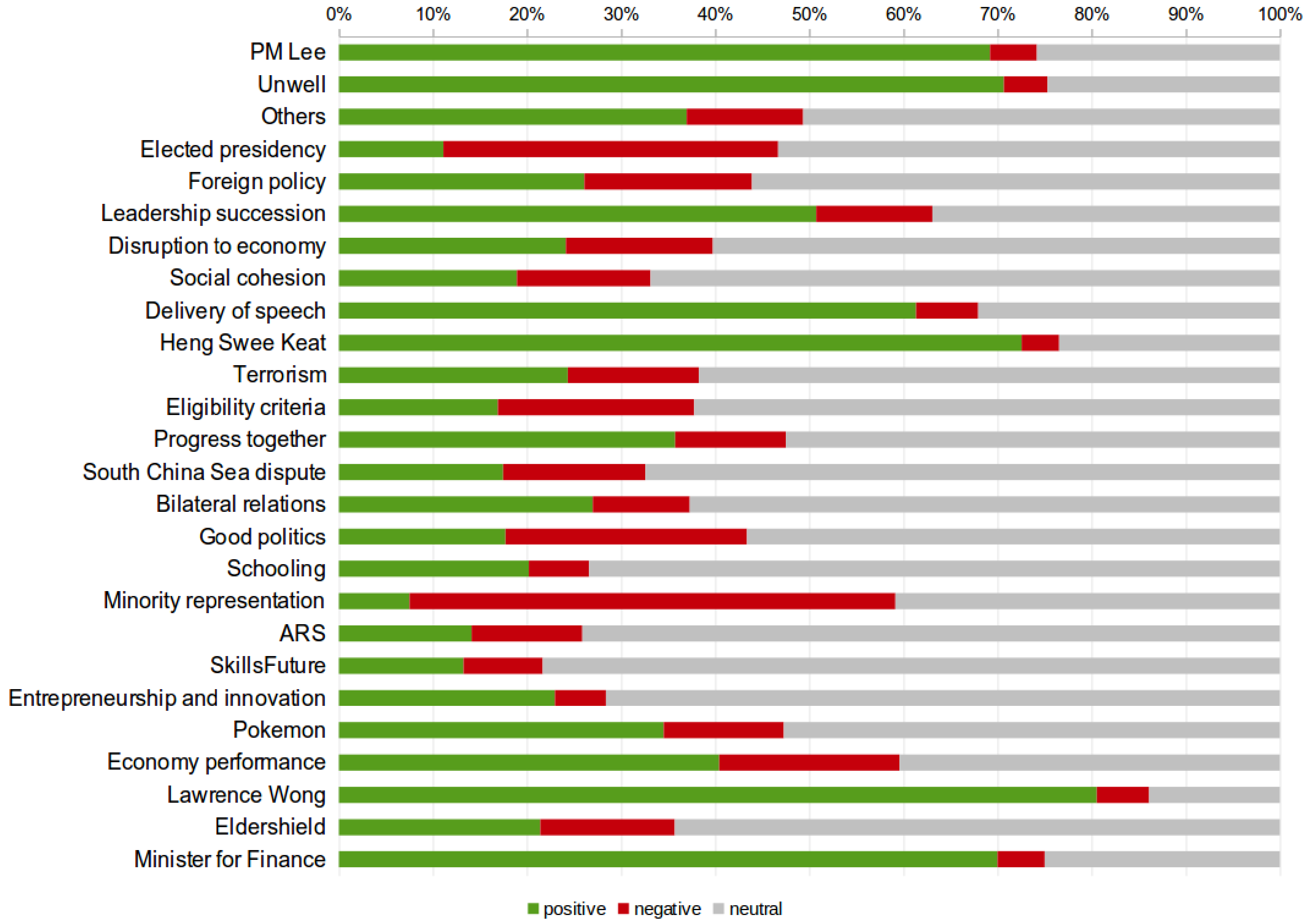


Content Volume By Survey Topics

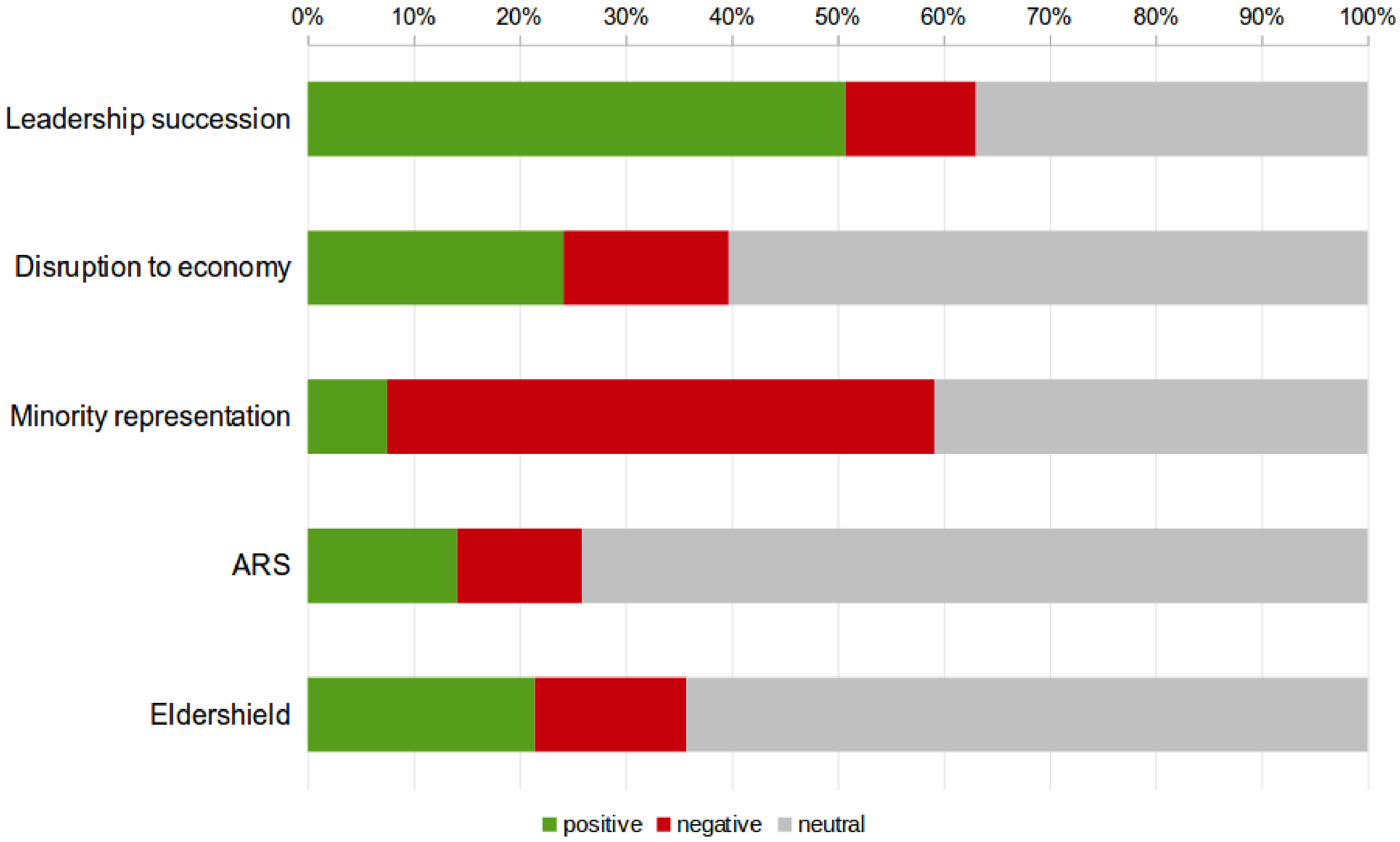


What were the online **sentiments** toward NDR topics?

Proportions of Sentiments By Topics



Proportions of Sentiments By Survey Topics

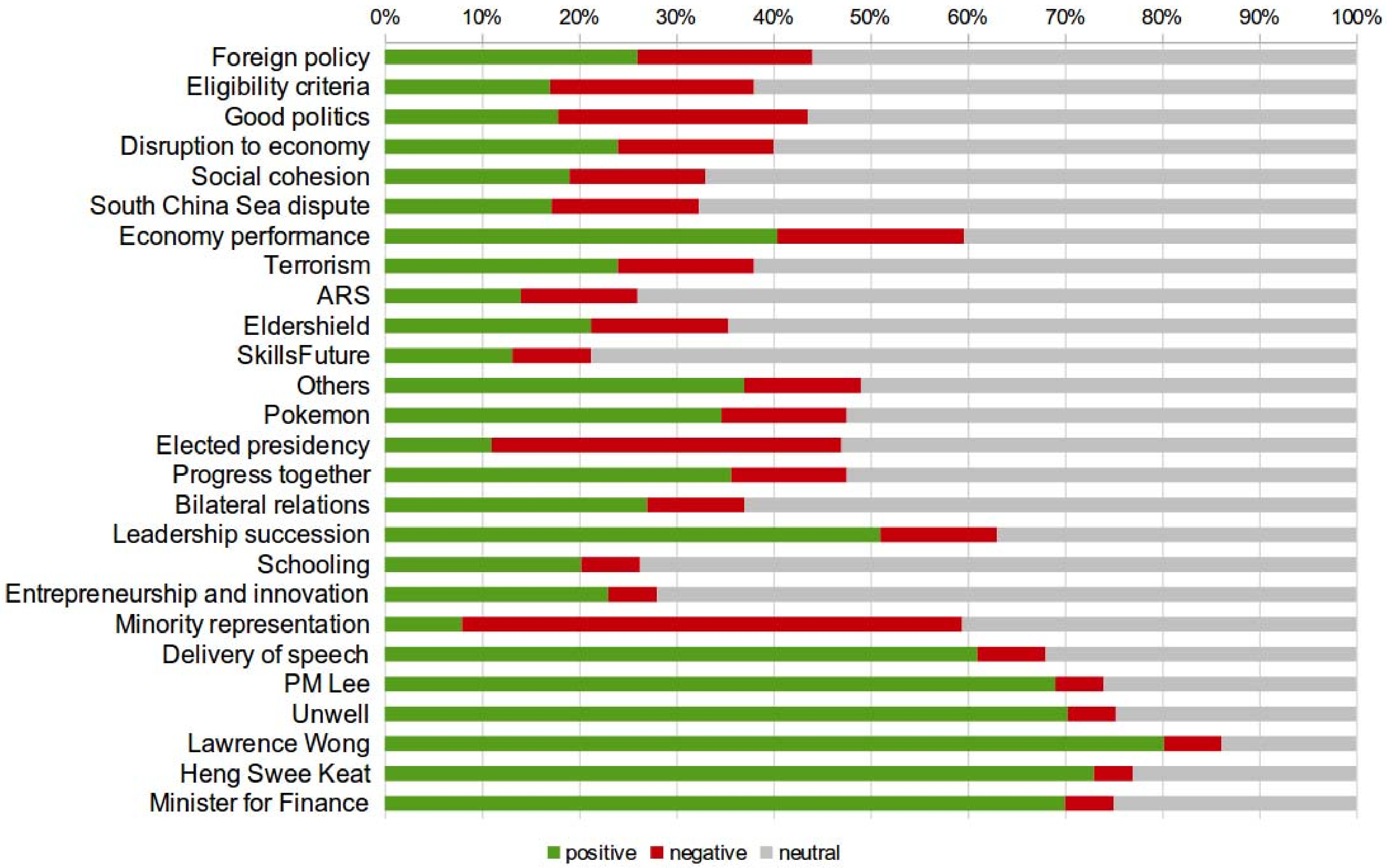


Which NDR topics gained the most **divided sentiments**?

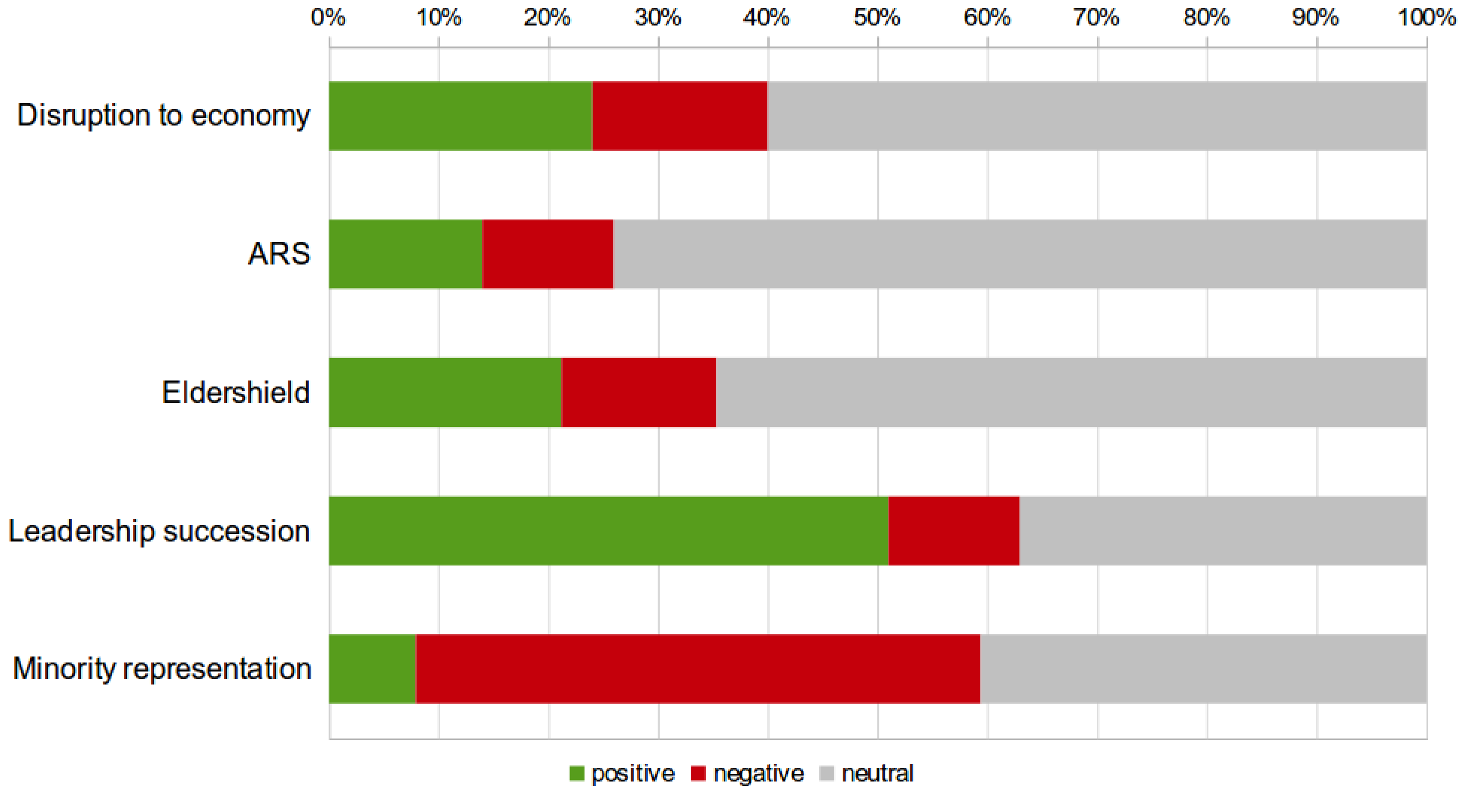
Divisiveness assumptions:

- High volume of subjective content (positive + negative)
- Equal portions of positive and negative sentiments

Most Divided Topics



Most Divided Survey Topics



How do **topics** differ across online platforms?

Distance Between Platforms

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Platform A	40%	30%	10%	10%	10%
Platform B	40%	30%	10%	10%	10%

distance(A, B) = 0

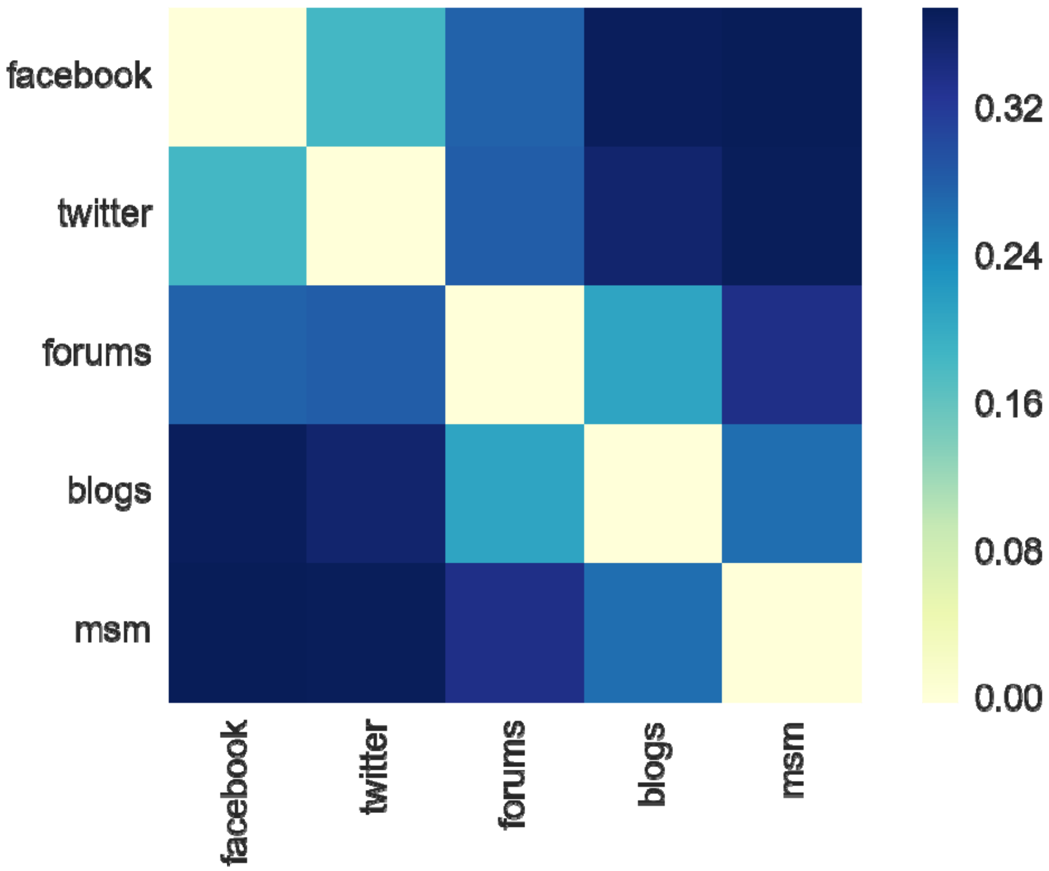
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Platform A	40%	30%	10%	10%	10%
Platform C	20%	20%	20%	20%	20%

distance(A, C) = 0.05

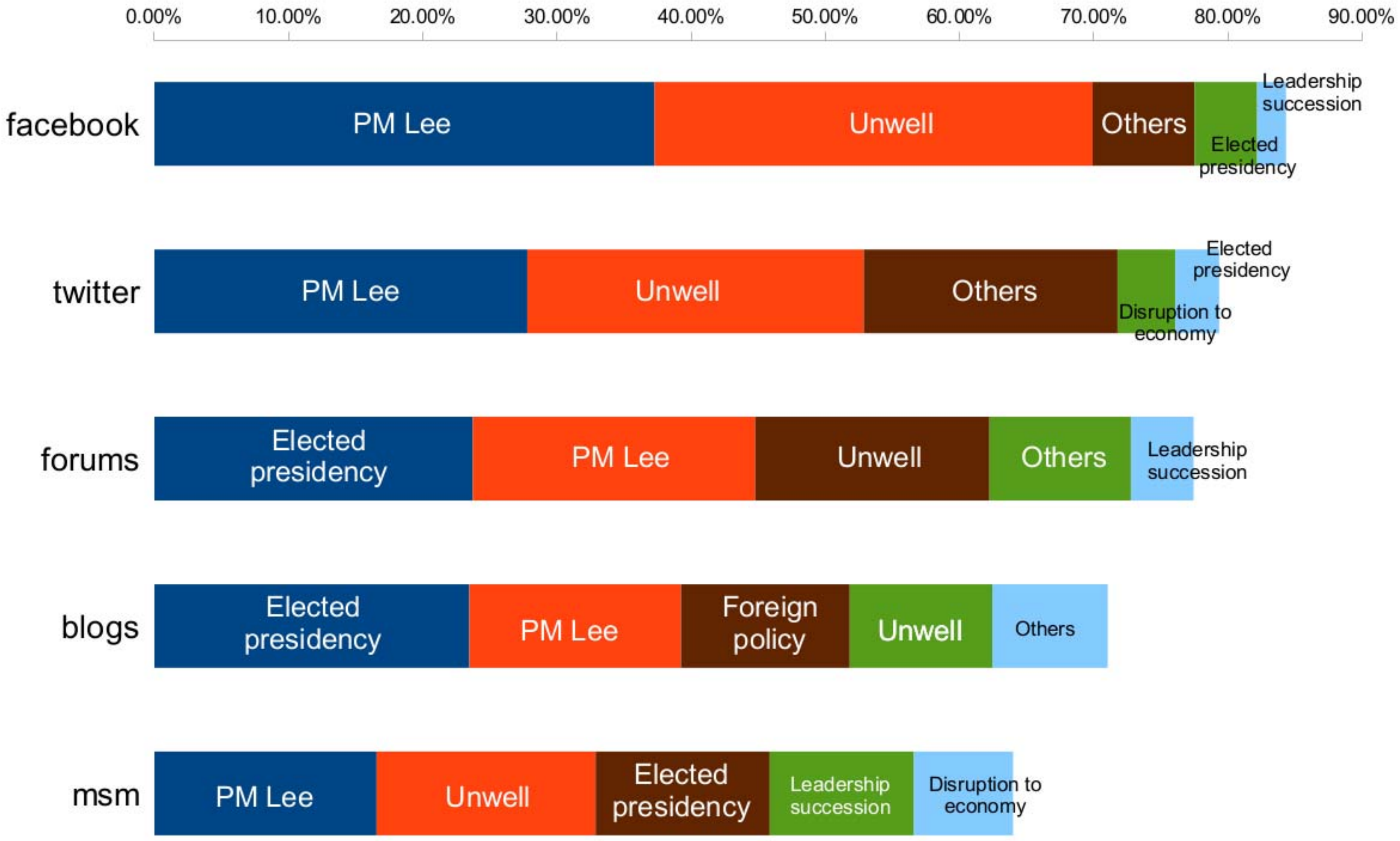
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Platform A	40%	30%	10%	10%	10%
Platform D	0%	0%	0%	0%	100%

distance(A, D) = 0.53

Topic Distance Between Platforms

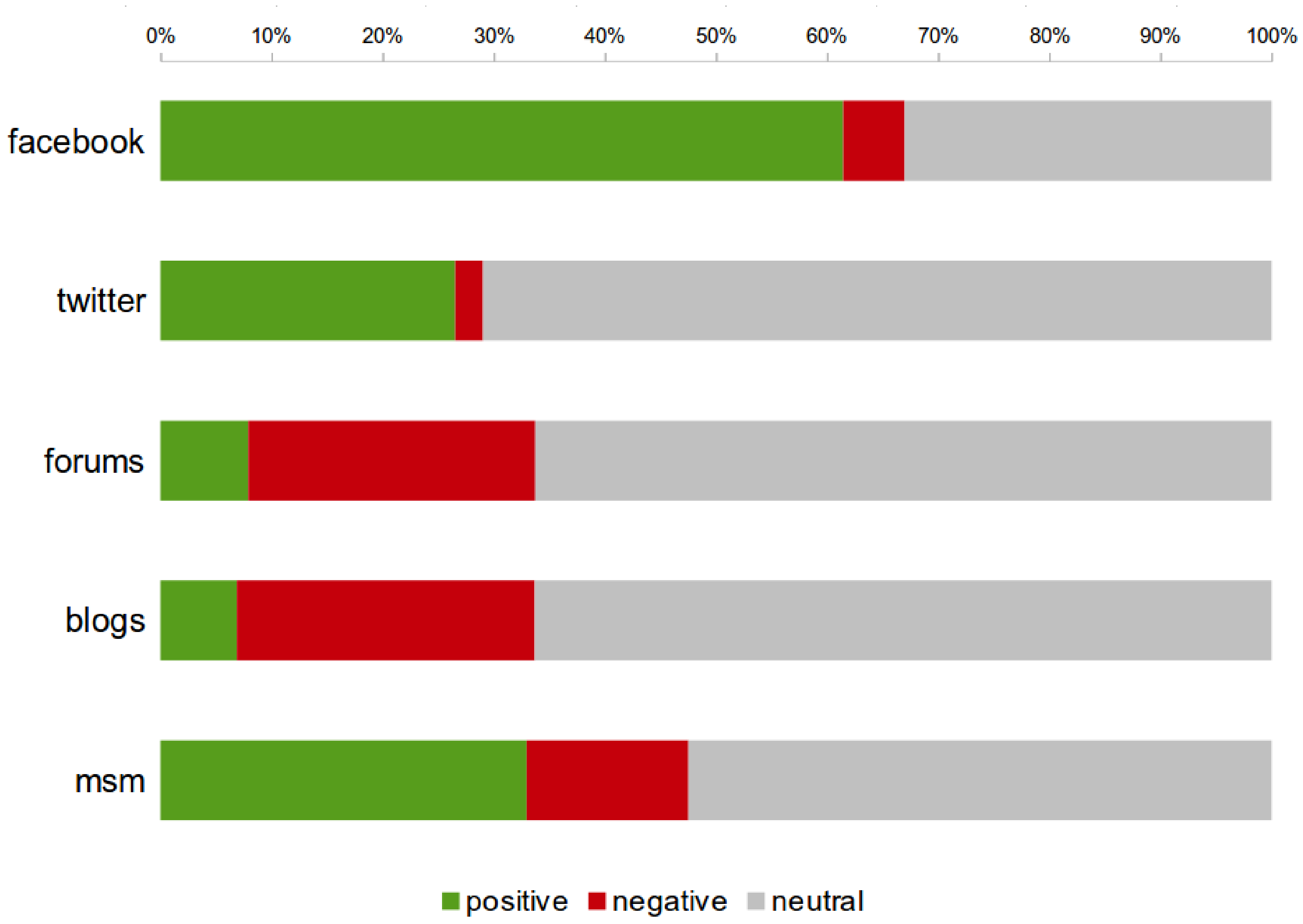


Top-5 Topics By Platforms



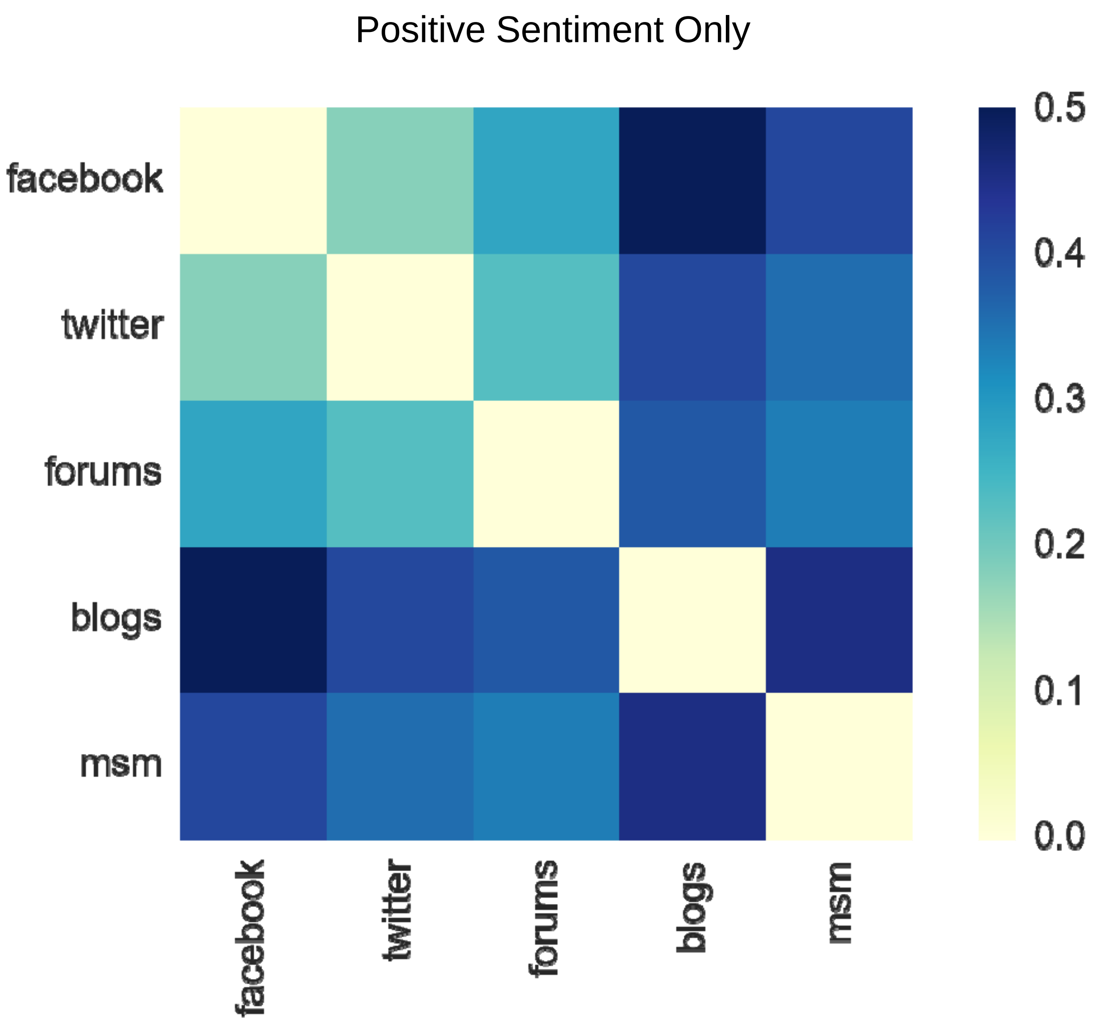
How do **sentiments** differ across platforms?

Proportions of Sentiments By Platforms

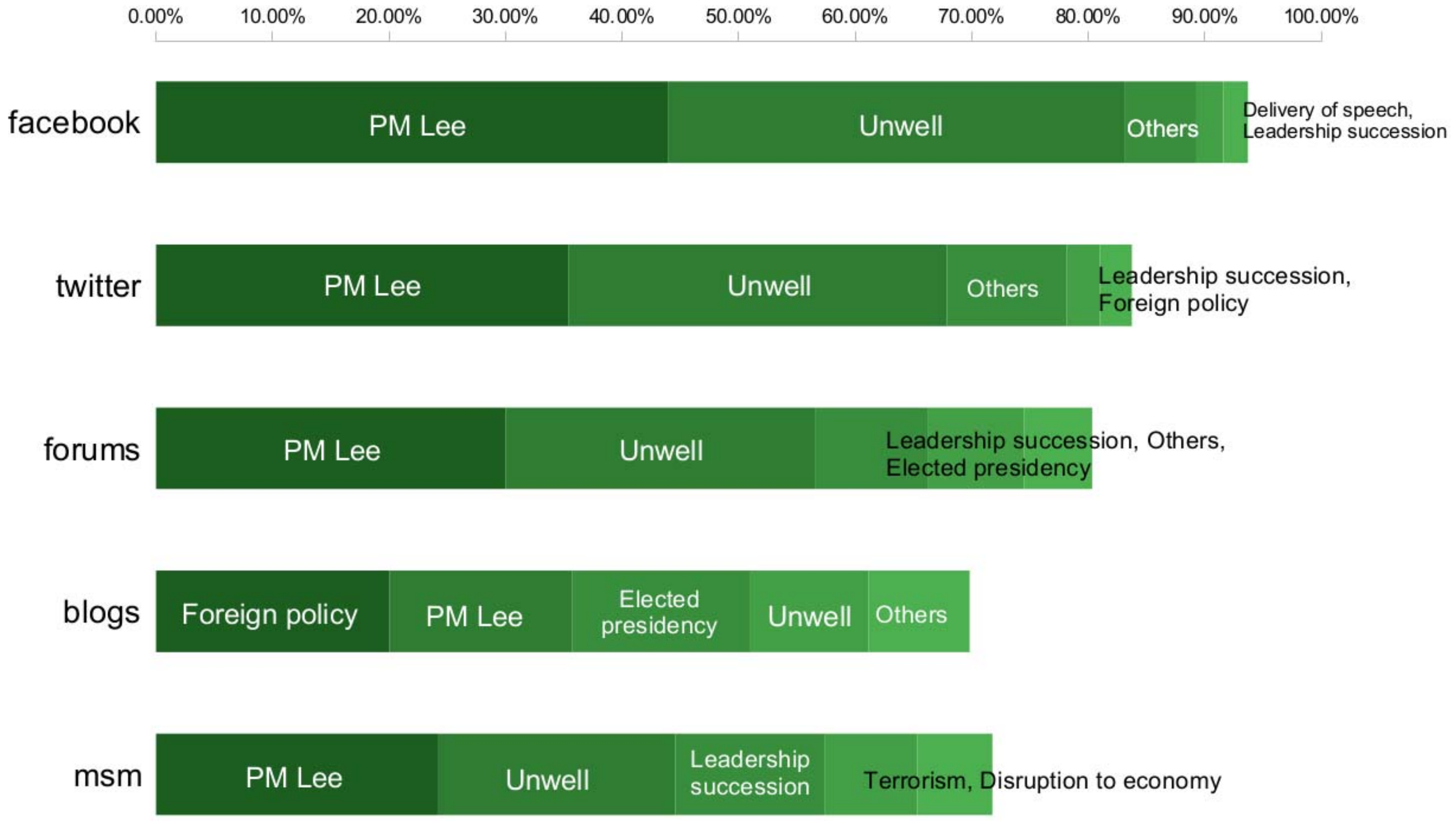


How do **sentiments** toward various **topics** differ across online platforms?

Topic-Sentiment Distance Between Platforms

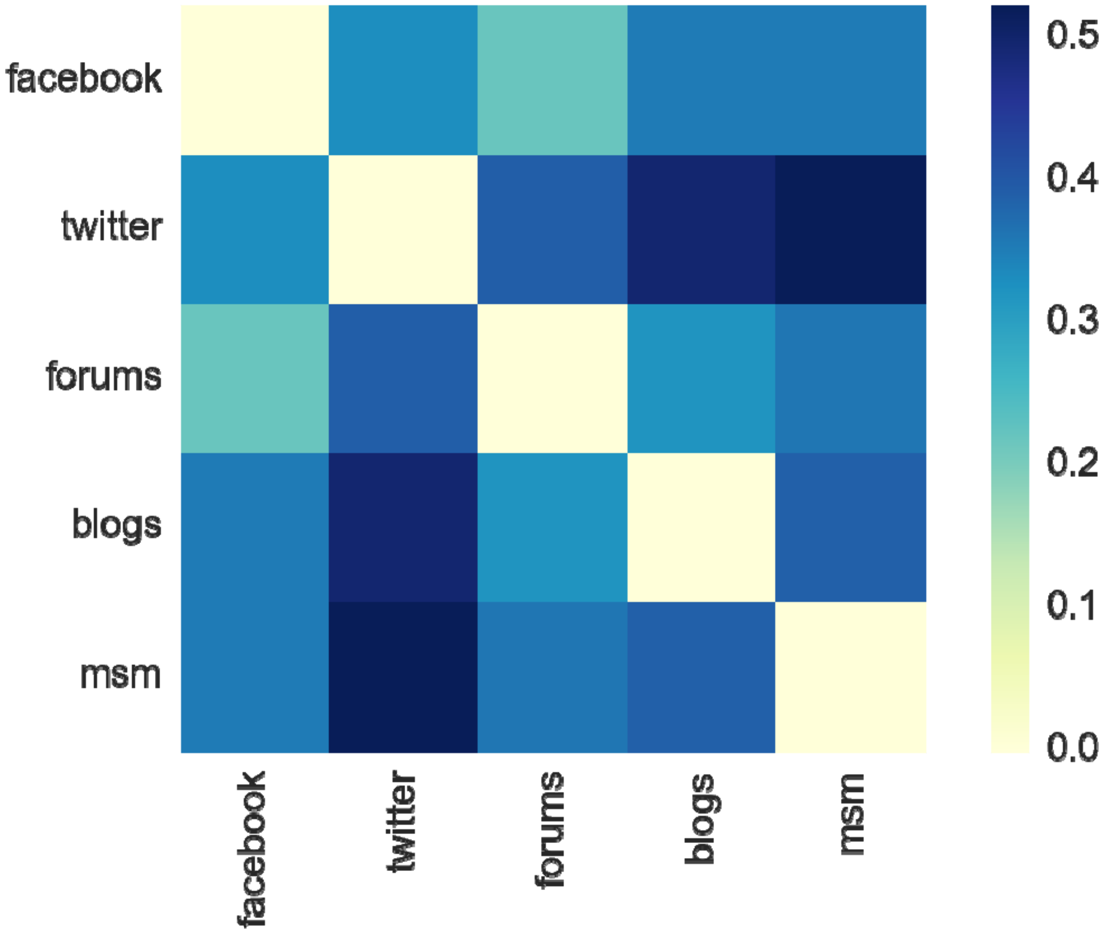


Top-5 Positive Topics By Platforms

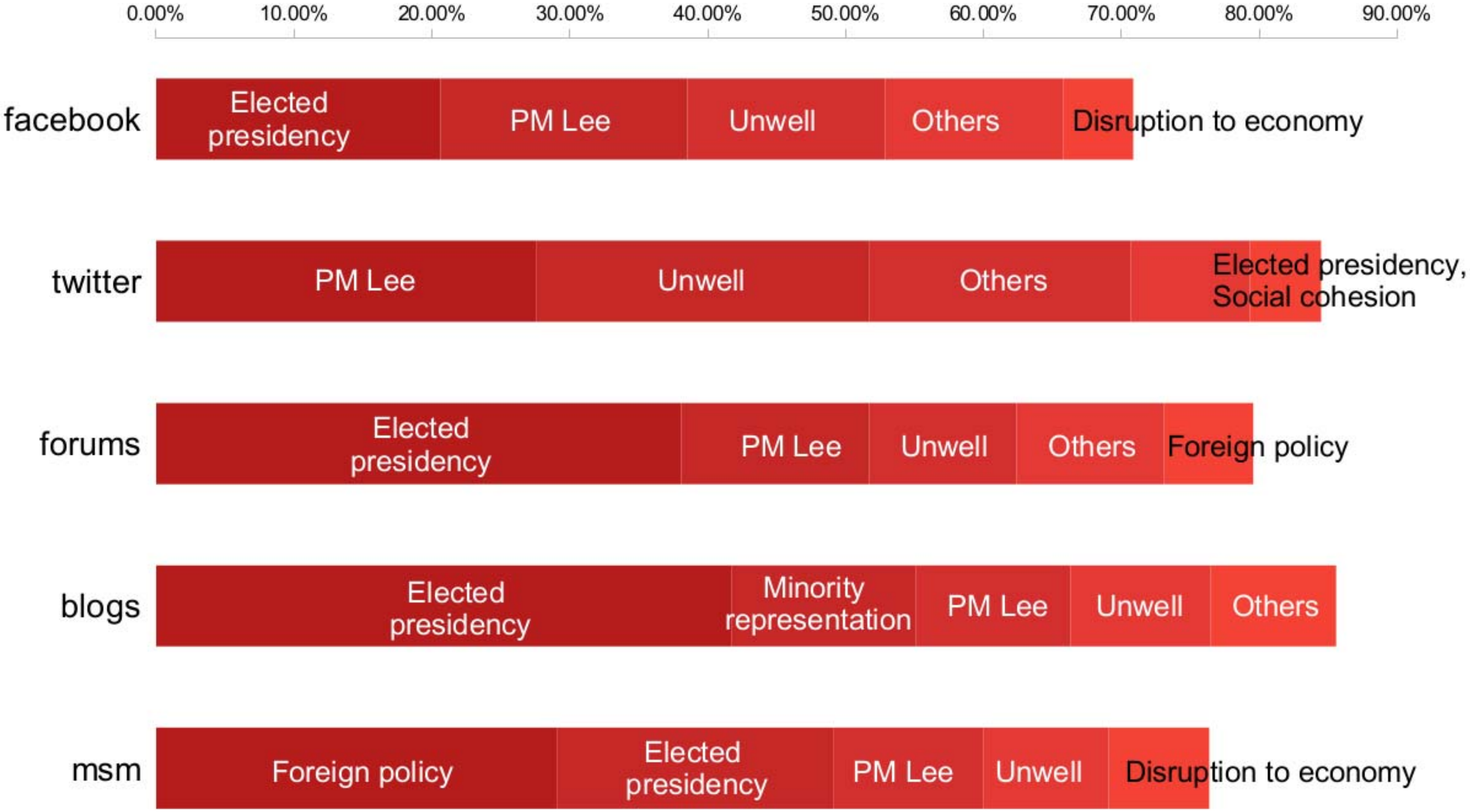


Topic-Sentiment Distance Between Platforms

Negative Sentiment Only



Top-5 Negative Topics By Platforms



Key Findings

- Facebook has more online users and more content generated than other platforms.
- The topic distributions on Facebook and Twitter are clearly different from those on blogs, forums.
- Sentiments are more positive on Facebook, Twitter and mainstream media than on blogs and forums.
- Online discussions (especially on Facebook and Twitter) are dominated by the topic on PM's health.
    - The event may have shifted the public's attention from the key NDR topics.
    - The event may have stimulated more interest and discussions on NDR.

THANK YOU!