School of
Information Systems

SMU
SINGAPORE MANAGEMENT
UNIVERSITY

Introduction to Text Analytics

Data for Analytics

Structured

Product		
Name	Data Type	Nullable?
PRODUCT_ID	VARCHAR	NO
CATEGORY	VARCHAR	NO
LIST_PRICE	DECIMAL	NO

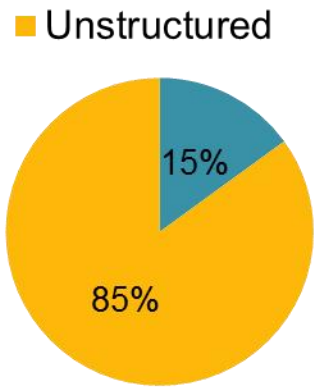
Unstructured

The challenge is:

- Approximately 75-90% of data is unstructured (while IT is built for structured data)
- Unstructured data is growing at nearly 10x the rate of structured data
- Less than 5% of unstructured data is proactively managed

Source: Natasha DeKroon and Brian Karp

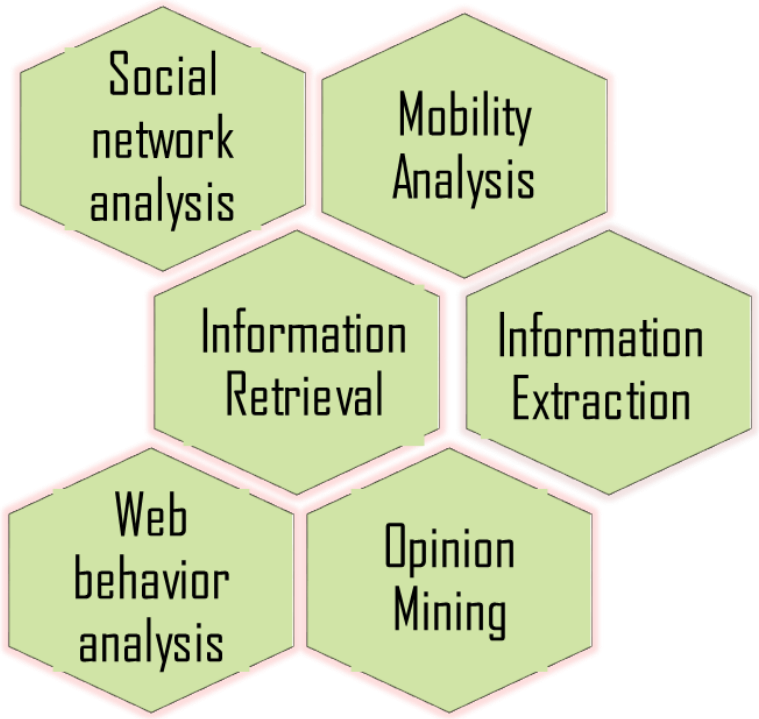
Why Need Text Analytics?



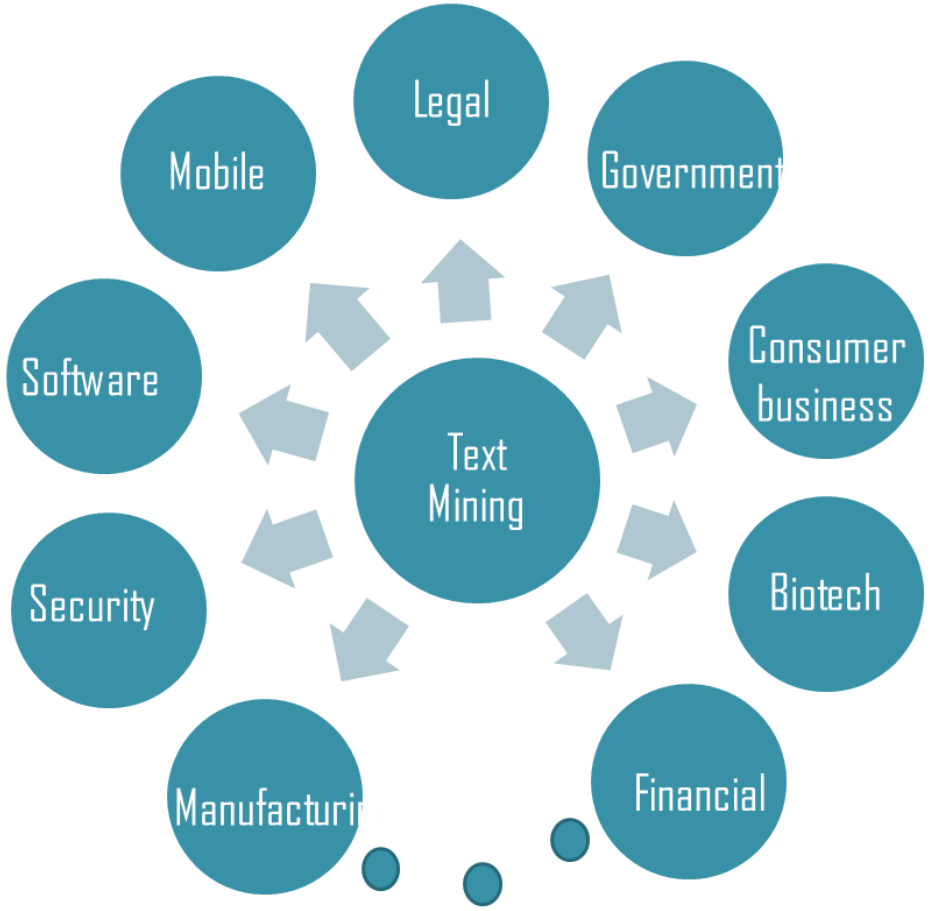
"Eight-five percent of data is **unstructured**, and you need **text analysis** and **text abstraction** along with a relational database to arrive at an integrated view," says Jerry Hill, vice president of manufacturing, for Teradata.

This figure has been boosted by Social Media

What Text Analytics can do?

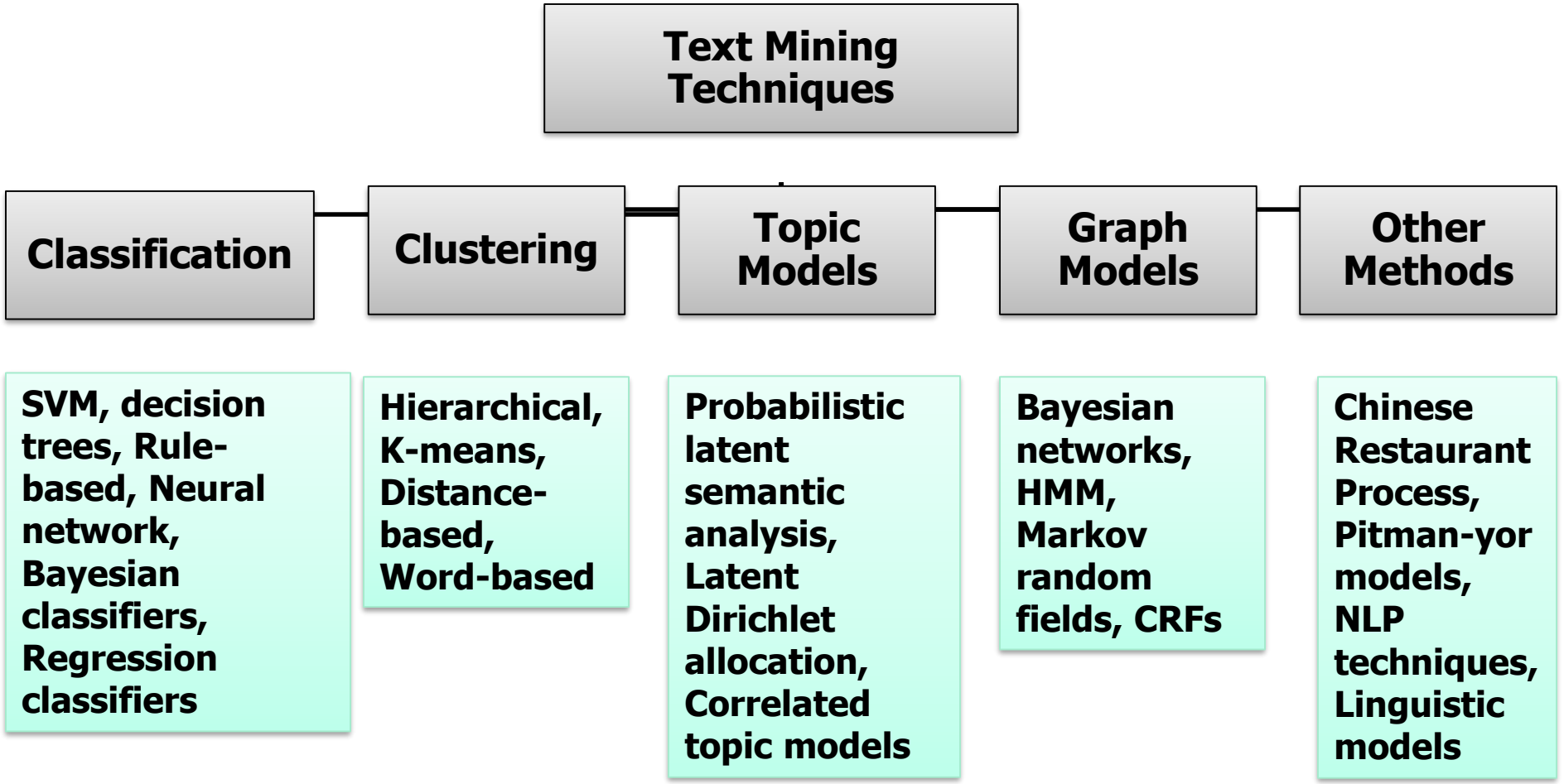


Text Analytics Functions

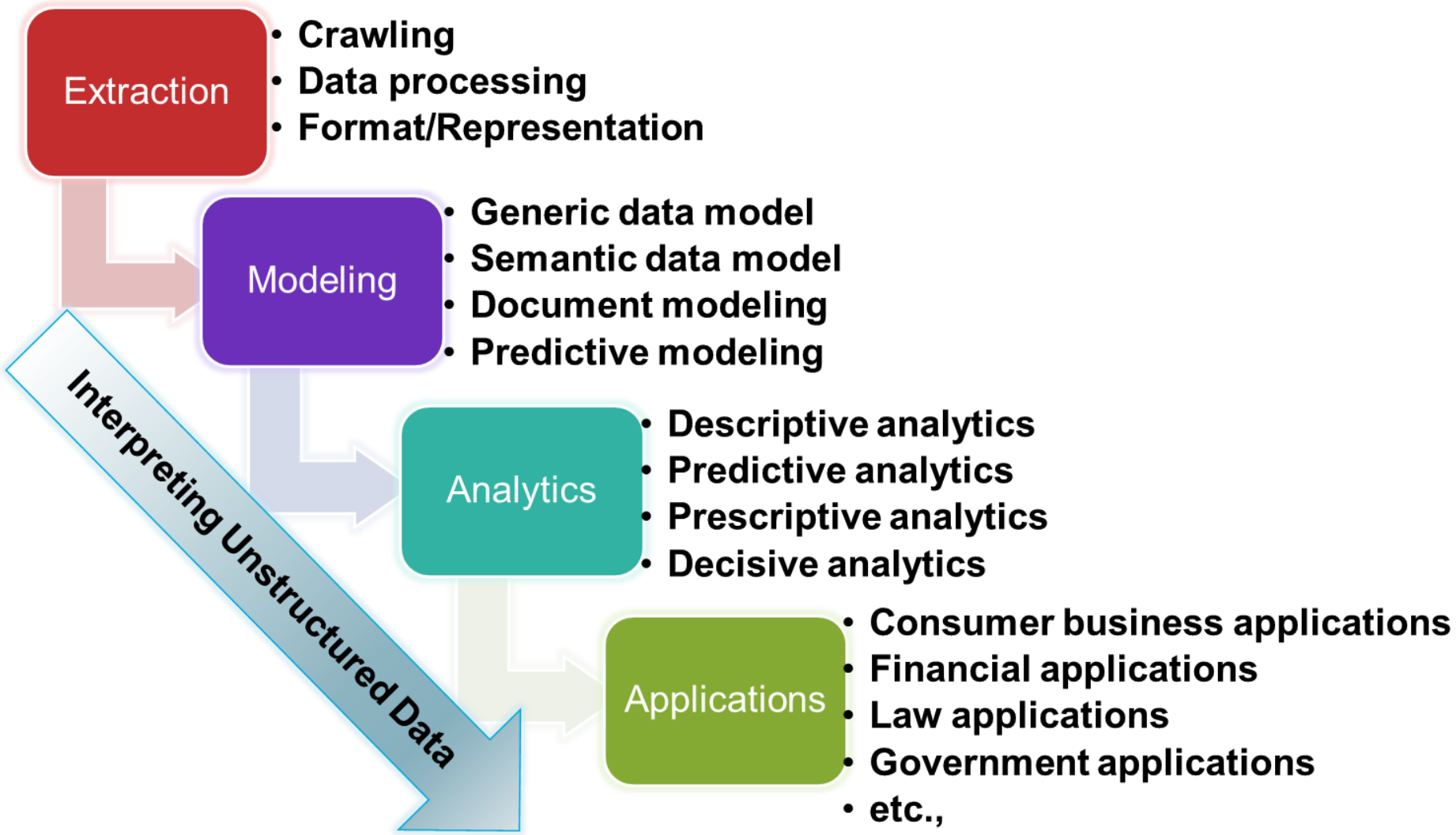


Text Analytics Domains

How- Text Mining Techniques

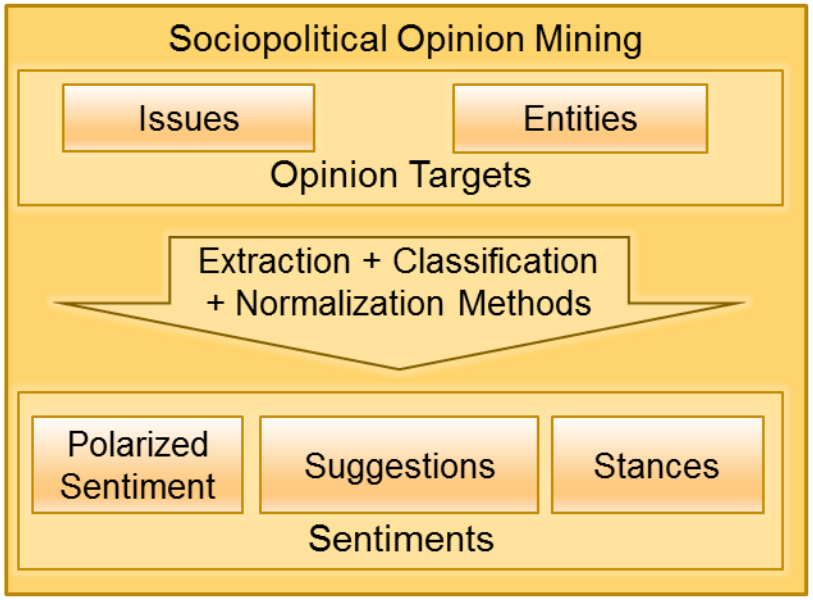


Stages in Applying Text Analytics



Example Text Analytics Application

Users' Feedback Channels:
Forums, Blogs, Political Sites, Facebook, Debate sites, etc.,



***Opinion Mining of Sociopolitical Comments from Social Media**, Swapna Gottipati, Thesis, 2014*

Example Comment from Social Media

- "My father was telling me in the past they worry abt 3 meals. The **government** that could give them 3 meals wins. Right now I worry abt my **retirement** and **housing**. The **gov** that can *lower **hdb** cost* and *discuss abt **job** and **retirement*** wins my vote. And I won't take 'no **retirement**' as an option".

What do we see?

1. Issues – Housing, Retirement, Job
2. Entities (People/Organizations)- Government
3. Suggestive opinions – Lower HDB cost
4. Valuable Comment

Results – Extracting Issues

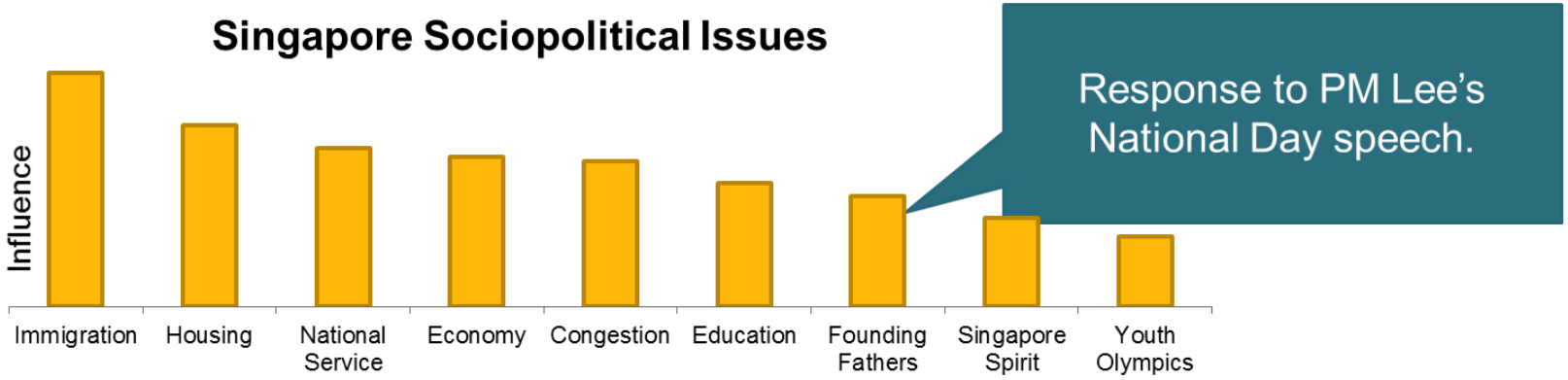
Table 1: Top words from PM's speech using JSC-LDA model

blue – feedback terns

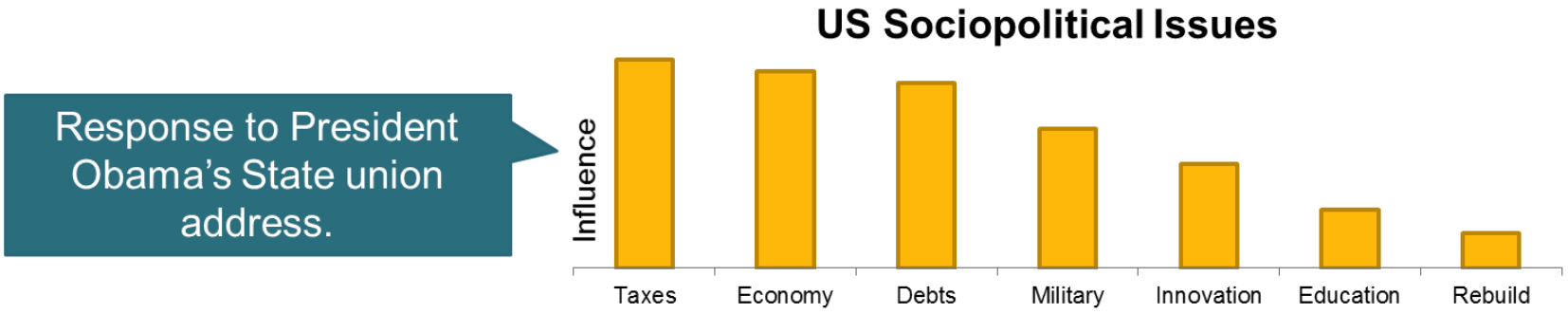
Issue	Top Words
Economy	ft, government, good, jobs, job, money, time, pay, working, bad, oil, country, workers, employers, simple
Immigration	foreign, workers, jobs, citizens, foreigners, chinese, talent, economy, immigrants, understand, world, local, foreigner
National Service	foreigners, salary, country, govt, people, ns, nsmen, vote, election, pay, policies, send, lower, family, private, sporeans, service
Congestion	time, job, change, hours, line, work, problem, place, talented, trains, people, working, foreigners, coming, run, bad
Housing	hdb, flats, live, foreigners, people, local, housing, property, long, high, poor, time, work, clear, afford, fw, stop, population
Education	good, students, school, schools, education, programme, poly, universities, work, government, academic, normal, university, overseas

What's public feedback? Linking high quality feedback to social issues using social media. Swapna Gottipati, Jing Jiang. In Proceedings of the International Conference on Social Computing (SocialCom'12) (poster), pages 546-551, 2012

Results - Comment Linking



1 REACH has received close to 1,700 comments on the National Day Rally (NDR) 2010 Feedback Exercise. This has surpassed the number of inputs REACH received last year on NDR by more than 50%. Immigration, Housing and Education top the list of topics with the highest number of inputs.



Ranked issues by the influence on public

Results - Extracting Entities and Suggestions

Input – Example comments from Social Media:

"The government should lift diplomatic immunity of the ambassador."
"Govt must inform the romanian government of what happened."
"SG government needs to cooperate closely with romania."
"Hope the government help the victims by at least paying the legal fees."
"I believe that goverment will help the victims for legal expenses"

Output:

Entity	Suggestive opinion (Sentiment)
government	lift diplomatic immunity of the ambassador and get him
government	inform the romanian government of what happened
government	cooperate closely with romania
government	help victims by at least paying the legal fees

Extracting and normalizing entity-actions from users' comments. Swapna Gottipati and Jing Jiang. In Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)

Embedded Analytics in Business Process in Healthcare

- Prediction from clinical notes using NLP
- Recommendation by Question-Answering System