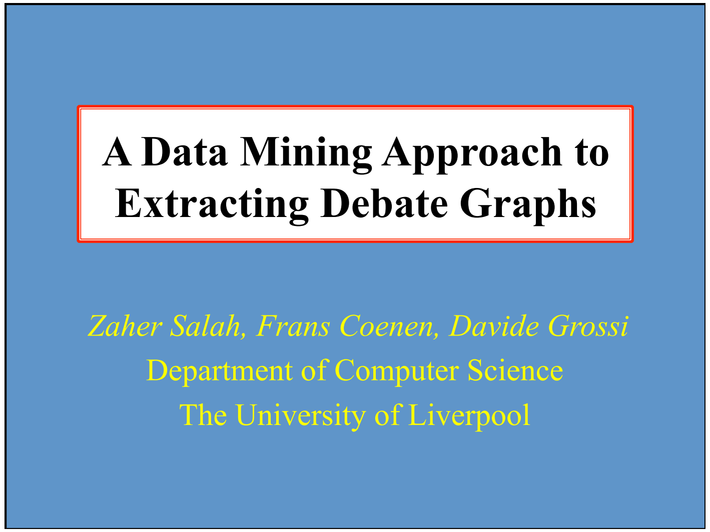
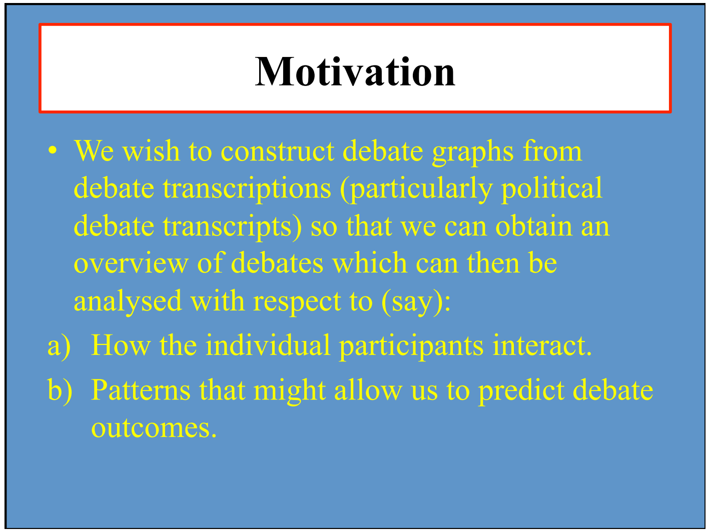


A Data Mining Approach to Extracting Debate Graphs

Zaher Salah, Frans Coenen, Davide Grossi
Department of Computer Science
The University of Liverpool

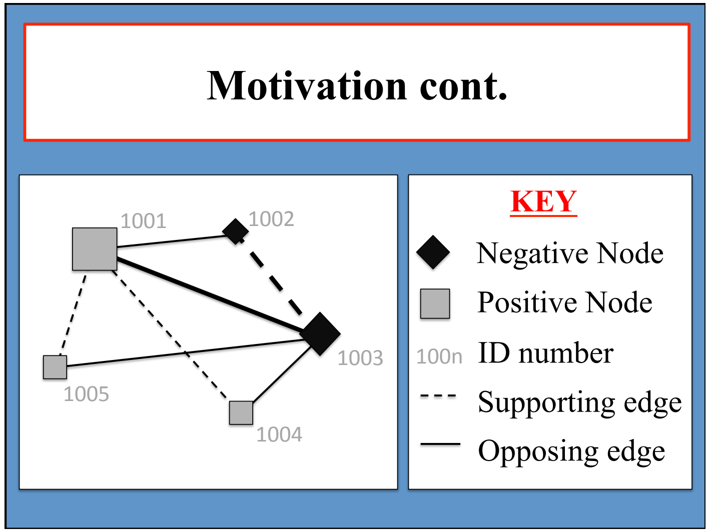


Motivation

- We wish to construct debate graphs from debate transcriptions (particularly political debate transcripts) so that we can obtain an overview of debates which can then be analysed with respect to (say):

a) How the individual participants interact.

b) Patterns that might allow us to predict debate outcomes.



Motivation cont.

KEY

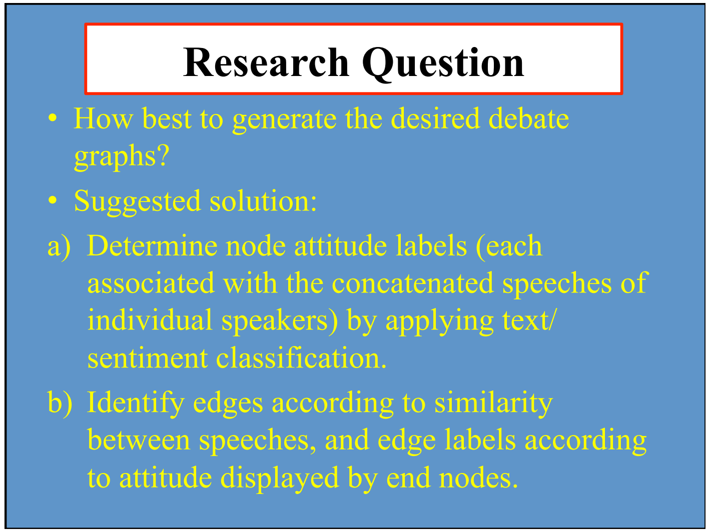
◆ Negative Node

■ Positive Node

100n ID number

- - - Supporting edge

— Opposing edge



Research Question

- How best to generate the desired debate graphs?
- Suggested solution:

a) Determine node attitude labels (each associated with the concatenated speeches of individual speakers) by applying text/ sentiment classification.

b) Identify edges according to similarity between speeches, and edge labels according to attitude displayed by end nodes.

Application Focus

- Verbatim transcripts of debates held within the UK House of Commons available in XML format from TheyWorkForYou.com.
- To the best knowledge of the authors, no one has attempted to describe House of Commons debates in this manner.

Example Debate (Start-up Loans)

Toby Perkins (Chesterfield, Labour)
….. The Government's failure to support small firms with access to finance cannot be camouflaged by this worthwhile scheme. Given that the Government have overseen a £14 billion reduction in lending to small business, will the Minister, at the same time as he celebrates his £50 million scheme, recognise that total lending from it is less than 1% of the shortfall in net lending that British business has experienced? *[Interruption.]*

John Bercow (Speaker)
Order. May I gently say to the hon. Gentleman that I think he is approaching his last sentence?

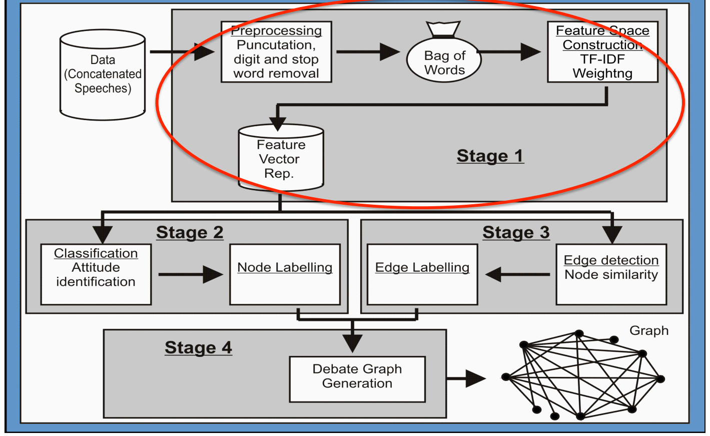
Toby Perkins (Chesterfield, Labour)
You are very wise, once again, Speaker, to notice that. Will the Minister make a statement on the real access to finance crisis that he has done so little about? Will he recognise the need for radical change to the banks through the Labour party's proposed network of local banks and support for challenger banks, which will lead to the desperately needed improvement in the position of small firms seeking access to finance?

Matthew Hancock (West Suffolk, Conservative)
Mr James Caan, who runs the start-up loans scheme on our behalf and to whom I pay tribute, is absolutely right to say how important mentoring is—and I think we have just seen why. What a pity that the Labour party cannot be enthusiastic about and supportive of a scheme that has done so much: …

Data Set (UKHCD2)

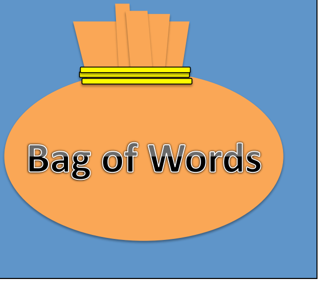
- We extracted transcripts associated with 100 debates from TheyWorkForYou.com conducted in 2012/13.
- Speeches associated with the same MP were concatenated together.
- Concatenated speeches with less than 50 words were ignored.
- 9473 concatenated speeches (4581 speeches made by speakers who voted Aye and 4892 who voted No) associated with 617 distinct Members of Parliament (MPs).

Framework


Data (Concatenated Speeches)
Preprocessing Puncutation, digit and stop word removal
Bag of Words
Feature Space Construction TF-IDF Weighting
Feature Vector Rep.
Stage 1
Stage 2
Classification Attitude identification
Node Labelling
Stage 3
Edge Labelling
Edge detection Node similarity
Stage 4
Debate Graph Generation
Graph

Preprocessing

- Upper to lower case alphabetic character conversion.
- Punctuation mark and numeric digit removal.
- Stop word removal (including domain specific stop words).
- (Snowball) stemming.



Feature Vector Construction

$$W_{ij} = TFIDF(i,j) = tf(i,j).\left(log\frac{N}{df(j)}\right)$$

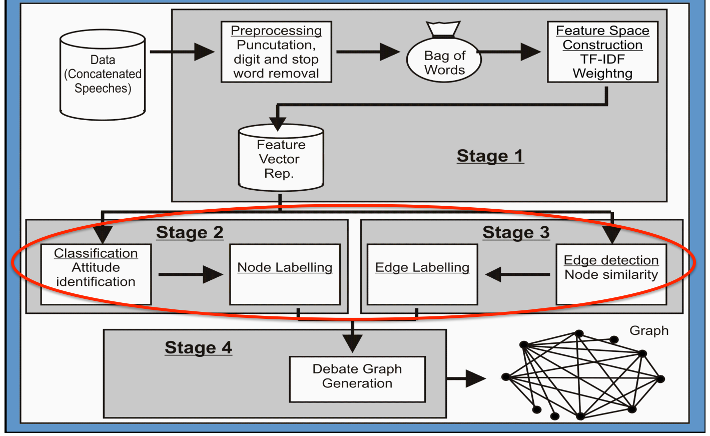
$$CosSim(d_i,d_j) = \frac{d_i \times d_j}{|d_i| \times |d_j|} = \frac{\sum_{k=1}^{k=z} w_{ik} \times w_{jk}}{\sqrt{\sum_{k=1}^{k=z} w_{ik}^2 \times \sum_{k=1}^{k=z} w_{jk}^2}}$$

Feature Vector Construction

- Bag of words used to define a feature space from which sets of feature vectors can be generated (one per concatenated speech).
- Feature vector elements hold term weightings (generated using TF-IDF).

Term	DF (Aye)	DF (No)	DF (Total)	Diff.
cuts	87	38	125	49
timetable	23	23	47	0
european	59	105	164	-46

Framework



Data (Concatenated Speeches)

Preprocessing Puncutation, digit and stop word removal

Bag of Words

Feature Space Construction TF-IDF Weightng

Feature Vector Rep.

Stage 1

Stage 2

Classification Attitude identification

Node Labelling

Stage 3

Edge Labelling

Edge detection Node similarity

Stage 4

Debate Graph Generation

Graph

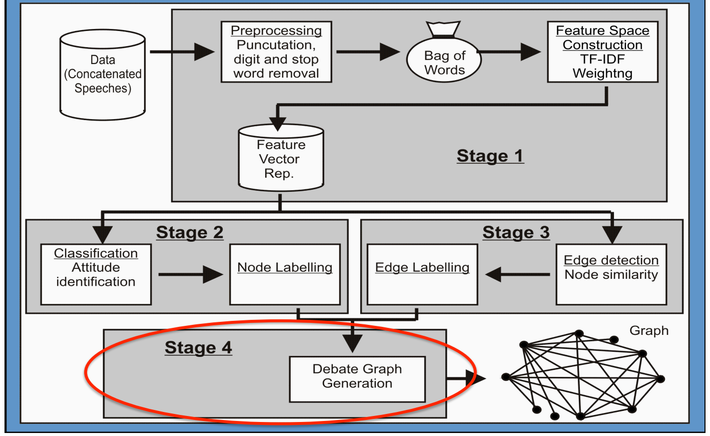
Node Labelling

- Sentiment (text) classification applied to a training set to determine each speaker’s “attitude" (positive or negative).
- Training set (UKHCS2) included the known vote associated with each concatenated speech.
- To reduce the size of the ``search space” χ^2 feature selection was used to identify the top k words that served as the best discriminators.
- Nodes labelled according to detected attitude.

Edge Identification and Labelling

- Edges between node pairs established when the speeches associated with two nodes (speakers) are deemed to be similar.
- We used the cosine similarity measure.
- Similarities between all node pairs determined using an “affinity” matrix.
- An edge is deemed to exist if a similarity value is greater than the average pair-wise similarity.
- Edges are labelled using the terms “support" and “oppose". Support if the linked nodes have the same attitude, oppose otherwise.

Framework



Data (Concatenated Speeches)

Preprocessing Puncutation, digit and stop word removal

Bag of Words

Feature Space Construction TF-IDF Weightng

Feature Vector Rep.

Stage 1

Stage 2

Classification Attitude identification

Node Labelling

Stage 3

Edge Labelling

Edge detection Node similarity

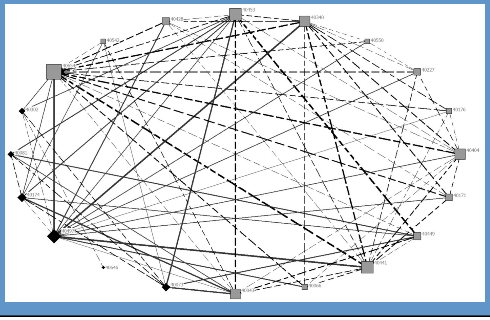
Stage 4

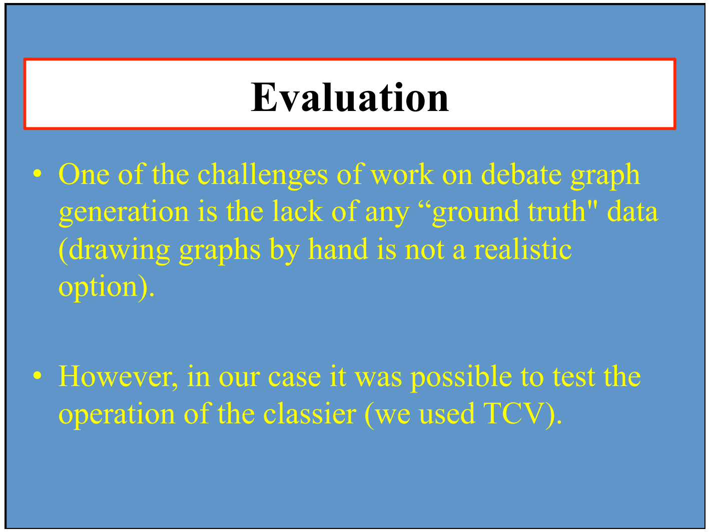
Debate Graph Generation

Graph

Debate Graph Generation

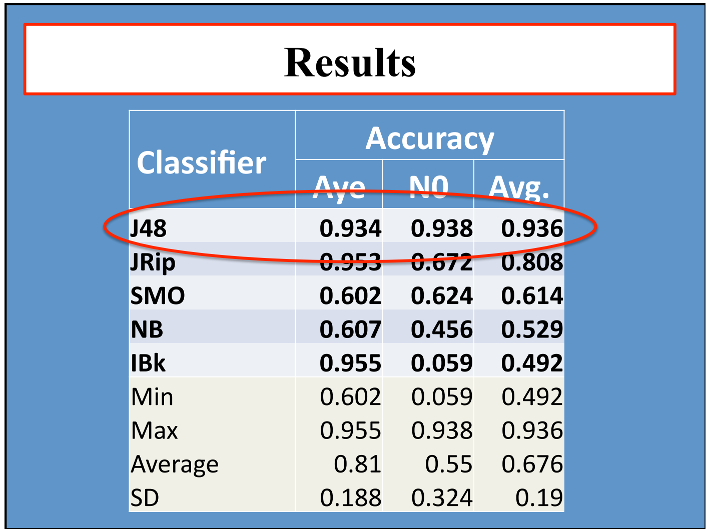
- Graph generation is conducted using the outputs from Stages 2 and 3, and is fairly straight forward (we used *NetDraw*).





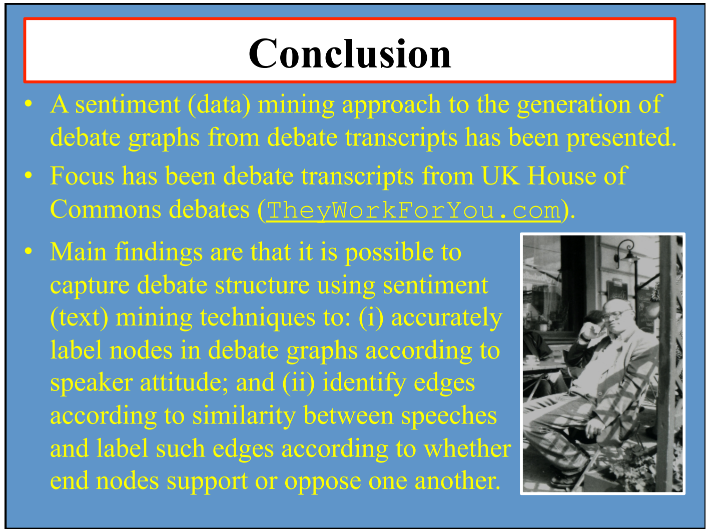
Evaluation

- One of the challenges of work on debate graph generation is the lack of any “ground truth" data (drawing graphs by hand is not a realistic option).
- However, in our case it was possible to test the operation of the classier (we used TCV).



Results

Classifier	Accuracy		
	Aye	NO	Avg.
J48	**0.934**	**0.938**	**0.936**
JRip	**0.953**	**0.672**	**0.808**
SMO	**0.602**	**0.624**	**0.614**
NB	**0.607**	**0.456**	**0.529**
IBk	**0.955**	**0.059**	**0.492**
Min	0.602	0.059	0.492
Max	0.955	0.938	0.936
Average	0.81	0.55	0.676
SD	0.188	0.324	0.19



Conclusion

- A sentiment (data) mining approach to the generation of debate graphs from debate transcripts has been presented.
- Focus has been debate transcripts from UK House of Commons debates (TheyWorkForYou.com).
- Main findings are that it is possible to capture debate structure using sentiment (text) mining techniques to: (i) accurately label nodes in debate graphs according to speaker attitude; and (ii) identify edges according to similarity between speeches and label such edges according to whether end nodes support or oppose one another.