One word embedding to rule them all? Political text classification using CNNs

Dr. Kakia Chatsiou
achats@essex.ac.uk

ESRC Business and Local Government Data Research Centre
School of Computer Science and Electronic Engineering
University of Essex

Research Questions

- Does the choice of word embedding for sentence topic classification with CNNs improve accuracy, as applied to political text
  - Train (English Manifestos) → Test (English Manifestos)
  - Comparison between more context free (Word2Vec, Glove, ELMo) with more context sensitive approaches (BERT)
- For political text, could we port trained CNN sentence topic classifier across different discourse styles such as press briefings?
  - Train (English Manifestos) → Test (COVID19 Press Briefings)
  - ditto

Background

- Topic classification of domain-specific types of political text
  - Supervised
  - Unsupervised
    - LDA – Hofmann 1999; Blei 2003
    - Structural Topic Modelling – Lindstedt 2019; Jacobs & Tchotschel 2019
    - Neural Networks – Zirn et al, 2016; Glavas, Nanni & Ponzetto, 2017; Bilbao-Jayo and Almeida 2018a
- Domain transfer of political manifestos classification to other political texts
  - Nanni et al 2016; Zirn et al 2016; Bilbao-Jayo & Almeida 2018b

NN Architecture

1. Word vectors of training English manifestos dataset sentences created using different word embeddings:
   - Word2Vec (M1); GloVe (M2); ELMo (M3); BERT (M4)
   - 300 vector size; 60 x d for convolution space
2. Vectors are fed to CNN
   - Convolution operations (100 filters; 3 different filter sizes [2xd,3xd,4xd]; dimensionality reduction using *1-max-pooling*, concatenated)
3. A dropout rate of 0.5 is applied (Srivastava et al 2014) [regularisation, prevent overfitting]
4. Softmax computes probability distribution over the labels
   - Sentence-level topic classification, using information local within sentence
5. Optimization using Adam optimiser with original parameters (Kingma & Ba 2017)

4 experiments:

- M1: CNN with Word2Vec
- M2: CNN with GloVe
- M3: CNN with ELMo
- M4: CNN with BERT

Datasets

Manifestos Project Corpus (English Subset) – Volkens et al, 2020b

- Expert manually annotated political manifestos of political parties' manifestos
- 7 policy domains and 56 subdomains
- English subset of 115 manifestos ~approx. 86K annotated sentences

Domain 1 (External Relations)	6.5%
Domain 2 (Freedom and Democracy)	4.42%
Domain 3 (Political System)	10.64%
Domain 4 (Economy)	25.45%
Domain 5 (Welfare and Economy of Life)	31.77%
Domain 6 (Fabric of Society)	11.20%
Domain 7 (Social groups)	9.99%

Table 1: Domain Codes' distribution in the English subset of the Manifestos Corpus used for training the CNN classifier.

COVID19 Press Briefings corpus – Chatsiou, 2020

- English subset of 717 press briefings ~approx. 62K sentences in total
- Annotated: 20 press briefings; 1740 sentences
- 7 policy domains of Manifestos

Domain 1 (External Relations)	0.74%
Domain 2 (Freedom and Democracy)	0.47%
Domain 3 (Political System)	11.58%
Domain 4 (Economy)	33.99%
Domain 5 (Welfare and Economy of Life)	34.62%
Domain 6 (Fabric of Society)	15.02%
Domain 7 (Social groups)	3.58%

Table 3: Manifest Project Domain Codes' distribution in the manually annotated subset of the COVID-19 corpus.

Evaluation

- M1: CNN with Word2Vec
- M2: CNN with GloVe
- M3: CNN with ELMo
- M4: CNN with BERT

Does the choice of word embedding for topic classification with CNNs improve accuracy, as applied to political text?

For political text, could we port trained CNN topic classifier across different discourse styles such as press briefings?

Experiment	Accuracy	F1
M1	65.79%	61.11
M2	68.15%	64.93
M3	72.84%	68.42
M4	87.52%	74.68

Table 2: Domain results of all models using political manifestos

Experiment	Accuracy	F1
M1	50.65%	48.62
M2	54.18%	48.82
M3	60.74%	57.07
M4	68.65%	64.58

Table 4: Domain results of all models using COVID-19 Press briefings corpus

Conclusion & Future work

- We built a sentence-level political discourse classifier
- Used existing human expert annotated corpora of English political manifestos (Manifestos Project)
- Tested accuracy & performance of a CNNs classifier using different word embeddings
- Showed that sentence-level CNN + BERT outperforms others
- Applied same pretrained models to a different type of political discourse type (COVID-19 press briefings) – no additional training

Next steps:

- Conduct similar experiments with
  - (56) subdomain manifestos classes
  - Other NNs (LSTMs)
  - More languages in manifestos corpus

THANK YOU – Questions?