The use of natural language processing techniques for identifying different forms of politics-related content in large cross-platform datasets

Mykola Makhortykh
(IKMB, University of Bern)

Together with Ernesto de León, Aleksandra Urman, Clara Christner, Maryna Sydorova, Silke Adam, Michaela Maier, and Teresa Gil-Lopez

Introduction: NLP and politics-related content

- The combination of the recent advancements in NLP/ML and growing volume of digital data enables new possibilities for platform data analysis
- Classic supervised machine learning approaches (CML) and deep learning (DL): from SVM and decision trees to CNNs and transformer models
- Specifically, it allows addressing complex tasks regarding nuanced forms of content detection (e.g. politics-related or populism-related content)
- These tasks are important for understanding how individuals engage with online content and related phenomena of selective/ incidental exposure, AI-driven content personalization, consumption of news / disinformation

Introduction: Why cross-platform NLP is relevant?

- The growing volume of digital content is supplemented by our increasing capacities to capture such content across platforms
- Web-tracking (Adam et al., 2024) or browser data donations (Stubenvoll & Binder, 2024) trace individuals' cross-platform behavior resulting in large datasets (e.g. 2.4 million pages from 80k domains; Adam et al., 2024)
- To process such datasets, we need cross-platform methods both for data preprocessing (e.g. HTML parsing) and analysis (e.g. politics detection)
- However, the practical implementation is challenging due to different HTML architectures and different content types that result in high noise

Introduction: So, what we do?

- For the large SNF-DFG-funded project led by Prof.Dr. Adam (Bern) and Prof.Dr. Maier (Kaiserslautern), we developed and compared NLP approaches for cross-platform detection of politics-related information
- The project uses web-tracking to investigate online behavior of Swiss and German citizens and the consumption of content dealing with different forms of politics and populist radical right (PRR) ideas
- Two waves of data collection in 2020 with extreme diversity of domains (with many domains being blacklisted)
- Under these conditions, we wanted to know if cross-platform NLP approaches are feasible and, if yes, then which ones we better use

Methodology: Politics-related content detection

- Politics-related: content mentioning political actors in CH, DE, and around the world| societal issues (e.g., economy or climate change)
- Level of classification: document level
- Training data [TD]: 4,023 articles from Swiss/German news websites (e.g. Blick and Bild); journalistic tags used as labels for political / non-political
- Approaches compared:
  - Dictionaries (log-likelihood based on TD | Comparative Political Agendas + custom enrichment | combination)
  - CSML: five models [Bernoulli naive Bayes (BNB), multinomial naive Bayes (MNB), logistic regression (LR), passive aggressive (PA)]
  - DL: CNN | LSTM | BERT

Methodology: PRR-related content detection

- PRR-related: content containing elements of populism, nativism, or authoritarianism [ideally together, but here things get complicated]
- Level of classification: sentence-to-document level
- Training data (TD): 27,430 manually annotated sentences coming from a large sample of tracking data engaged with Swiss/German users
- Three groups of approaches + ensemble models:
  - Dictionaries (log-likelihood based on TD | Gründl (2020) populism dictionary | combination)
  - CSML: five models [Bernoulli naive Bayes (BNB), multinomial naive Bayes (MNB), logistic regression (LR), passive aggressive (PA)]
  - DL: CNN | LSTM | BERT

Methodology: Preprocessing

- Preprocessing is used to decrease data complexity / improve classification performance (Grimmer & Stewart, 2013); of particular importance for the cross-platform data due to the high volume of noise
- For cross-platform data, the first step of preprocessing concerns the extraction of text from HTML; hence, we compared a selection of parsers
- Besides, we tested six modes of text preprocessing: 1) no preprocessing; 2) stopword removal (NLTK list of German stopwords); 3) stemming (Cistem stemmer); 4) stemming + stopword removal; 5) lemmatization (Scapy lemmatizer for German); 6) lemmatization + stopword removal

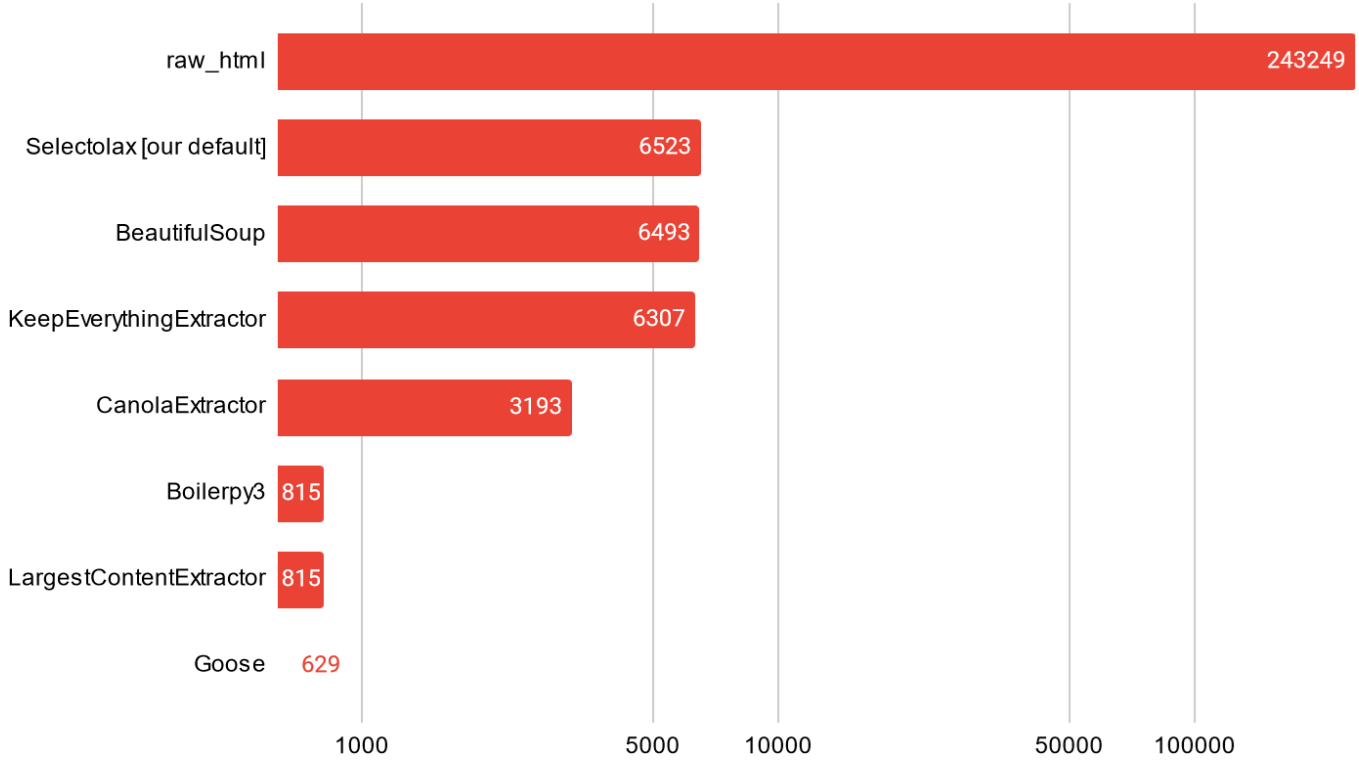
Methodology: Test datasets

- Politics-related content detection:
  - test-train split: 805 journalistic articles
  - low noise set: 594 documents from a few large platforms
  - high noise set: 262 documents from tracking data
- PRR classification:
  - test-train split: 4,782 manually annotated sentences
  - low noise set: 300 sentences from hyperpartisan / journalistic websites manually annotated
  - high noise set: 192 documents from tracking data (hyperpartisan / non-hyperpartisam websites manually annotated)

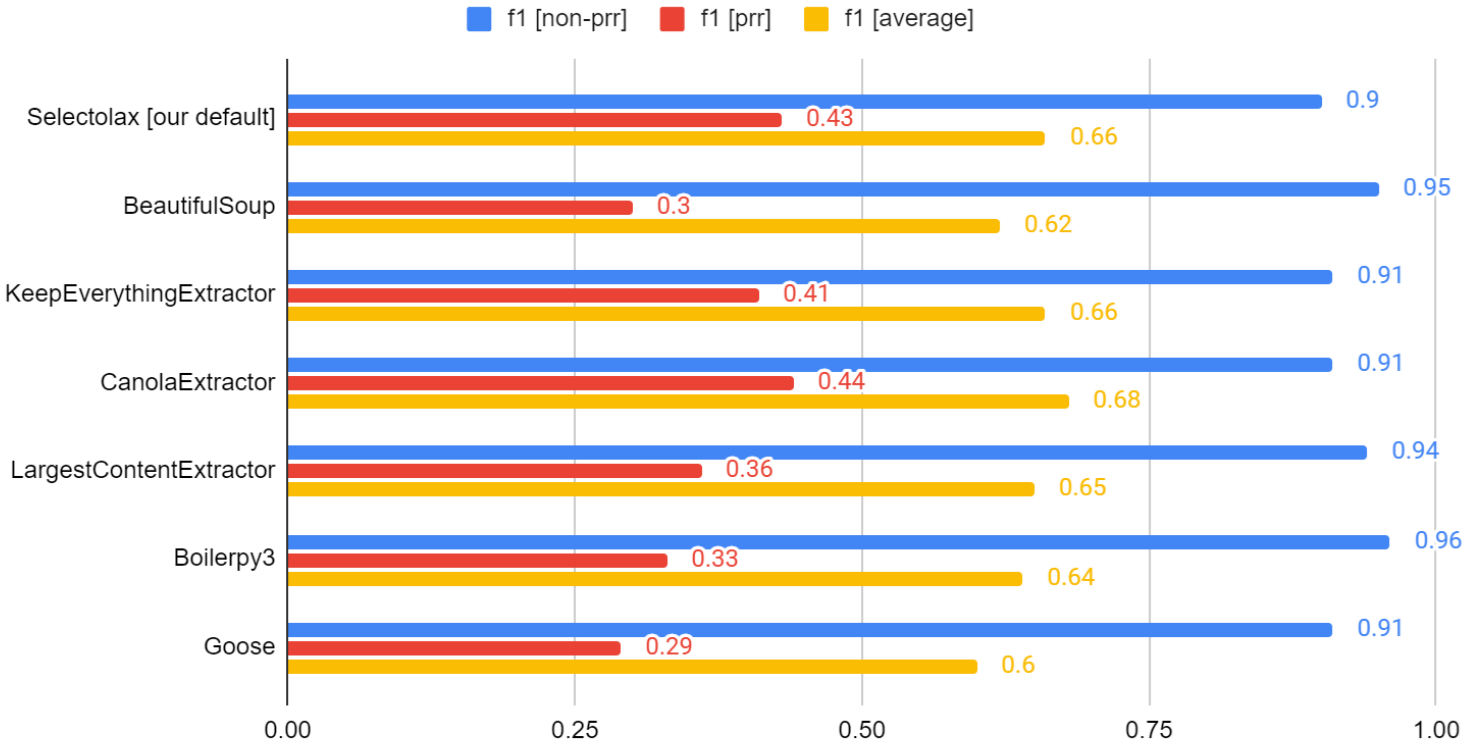
Methodology: Important limitations

- Not a comprehensive comparison of models due to feasibility reasons; some popular CSML and DL models are excluded (e.g. random forests)
- Limited fine-tuning and no cross-validation due to the focus on comparing low-cost implementations (fixed test sets instead)
- No comparison with AI-based labelling due to the project being implemented before the recent advancements in the field
- Difficulties with capturing PRR as a complex construct both for training data and for detection; eventually, we shifted the strategy towards combining individual PRR models [BERT-based]

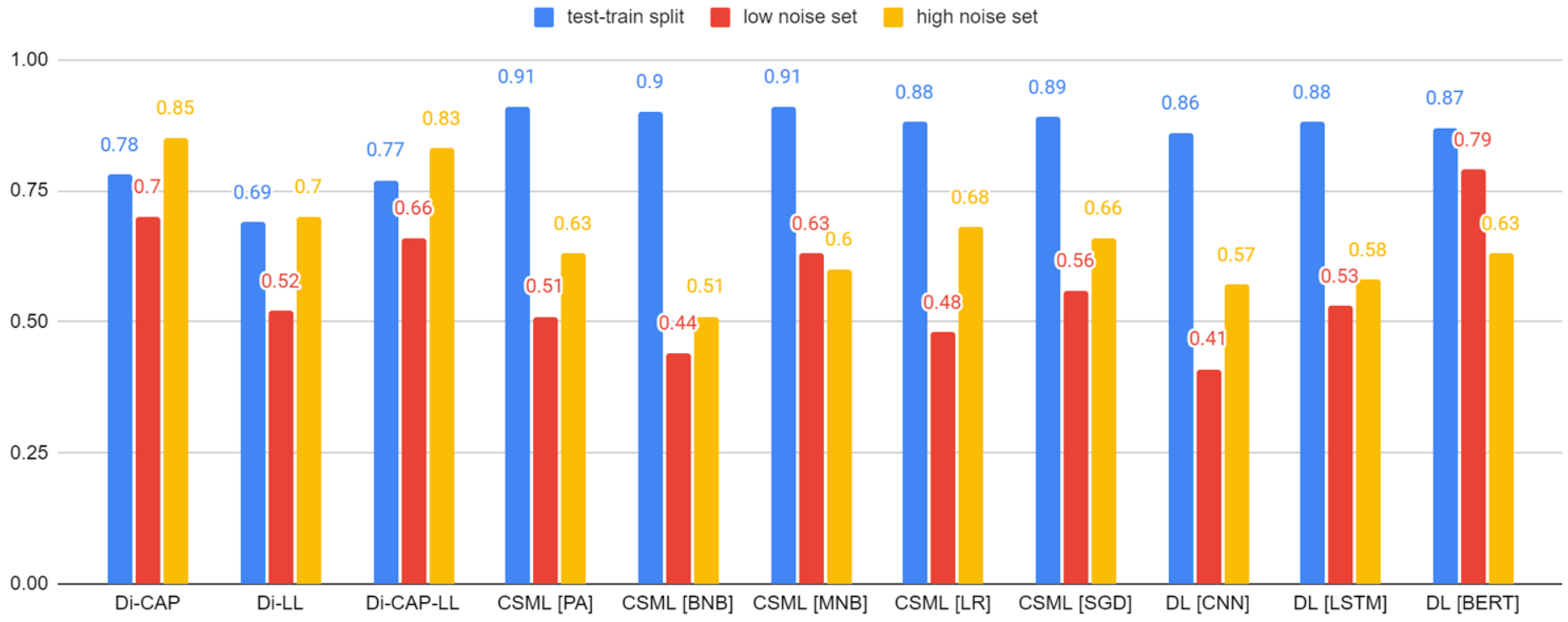
Parsing: N of characters for #948465 per parser



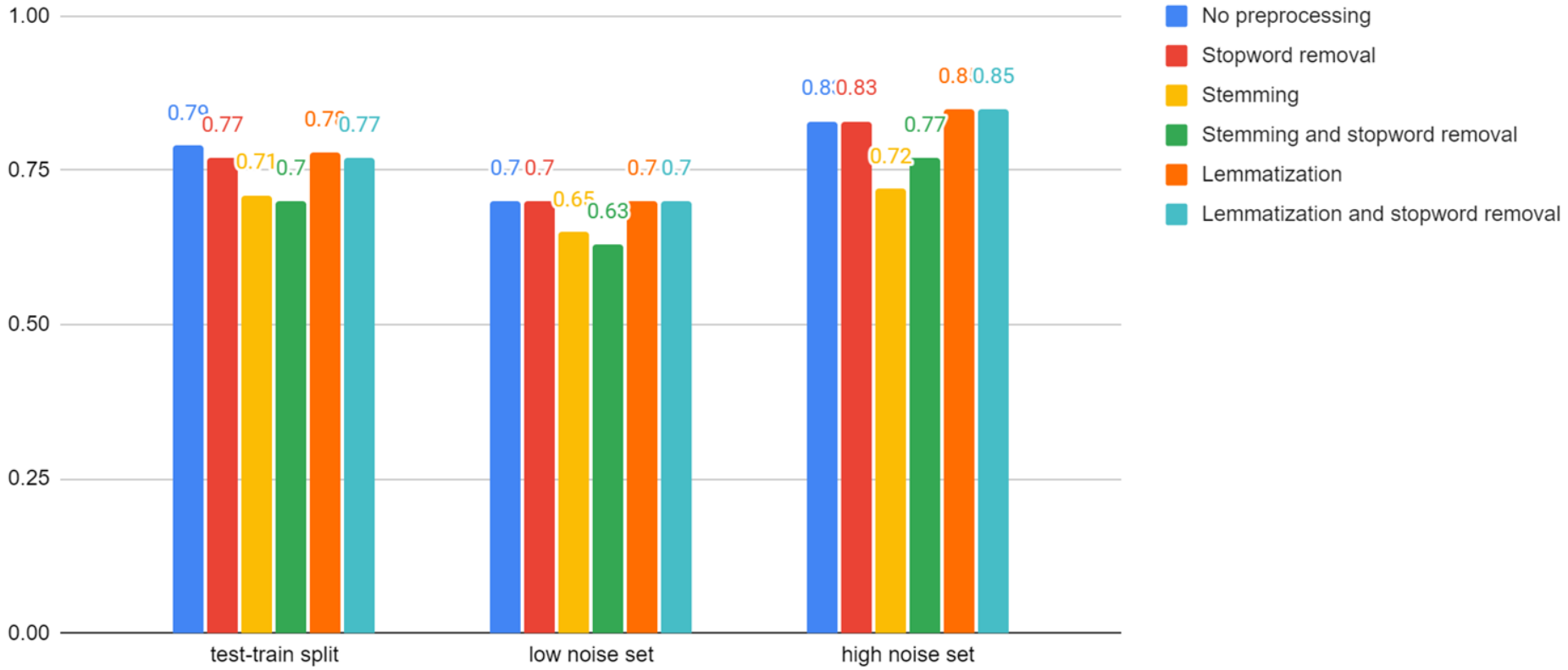
Parsing: Impact on text classification [CSML for PRR]



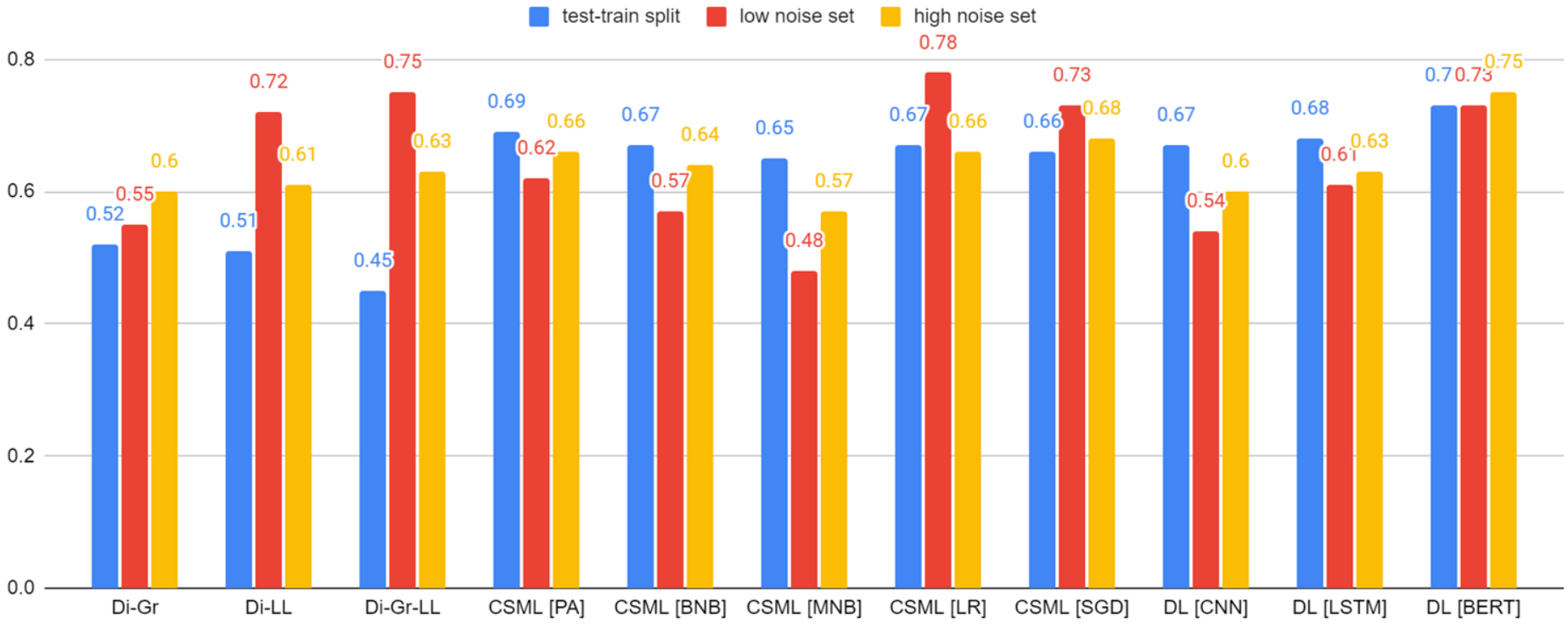
Detection: Politics [lem] [F1 average]



Preprocessing: Politics [Di-CAP] [F1 average]



Detection: PRR [individual models] [lem] [F1 average]



Conclusions: What these findings tell us

- Cross-platform detection of complex forms of textual content is becoming possible due to development of NLP but it still remains a complicated task
- In some cases, simpler approaches (e.g. dictionaries) can outperform complex DL models, albeit it depends on what we want to detect and how
- Generally, dictionaries can be a solid alternative for issue-specific task, whereas DL shines for complex argumentation-related tasks
- Importance of validating NLP techniques against diverse test sets and considering pre-processing: especially for HTML-to-text transformation
- Documentation and models are openly available [https://osf.io/e8xtb/] with the paper on politics-related detection released as a preprint []