A Language-Based Approach to Fake News Detection Through Interpretable Features and BRNN

Yu Qiao[1], Daniel Wiechmann[2] and Elma Kerz[1],
[1] RWTH Aachen University
[2] University of Amsterdam

3rd International Workshop on Rumours and Deception in Social Media (RDSM) co-located with the 28th International Conferences on Computational Linguistics (COLING 2020)

December 13, 2020

The phenomenon of fake news and its impact

- 'Disinformation' and 'fake news':
  - Various types of false or inaccurate information typically relating to emerging and time-sensitive events (for reviews, see e.g. Shu et al. 2017)
- Rapid and extensive spread can have significant negative impact on individuals and society:
  1. Propagandists persuade individuals to accept biased or false beliefs
  2. Disrupt balance of authenticity of the news ecosystem and change the way how real/authentic news are interpreted and responded to
  3. Increases political polarization, decreases trust in public institutions, and undermines democracy

Challenges associated with fake news detection

- Automated fake news detection still in early ages
- What characteristics make automated fake news detection uniquely challenging?
  a. Content is diverse in terms of topics, styles and media platforms
  b. Draws on diverse linguistic styles to distort truth
  c. Typically relates to newly emerging, time-critical events

Automated Fake News Detection

- Three approaches (for reviews and overviews, see below)
  1. Knowledge-based approaches
  2. Context-based (propagation-based) approaches
  3. **Language-based** (including style-based and stance-based) approaches

References

1. K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective. ACM SIGKDD explorations newsletter, 19(1):22–36, 2017.
2. R. Oshikawa, J. Qian, and W. Y. Wang. A survey on natural language processing for fake news detection. arXiv preprint arXiv:1811.00770, 2020.
3. M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff and Benno Stein. A Stylometric Inquiry into Hyperpartisan and Fake News. arXiv preprint arXiv:1702.05638, 2017
4. X. Zhou and R. Zafarani. Fake news: A survey of research, detection methods, and opportunities. arXiv preprint arXiv:1812.00315, 2020.
5. X. Zhang and A. A. Ghorbani. An overview of online fake news: Characterization, detection, and discussion. Information Processing Management, 57(2):102025, 2020.

Advantages of language-based approaches

1. Enable near real-time feedback (proactive rather than retroactive), i.e. they are not restricted to being applied only *a posteriori* and
2. they are scalable (see, Potthast et al., 2017)

Types of language-based approaches

1. Readability/style-based features (Potthast et al., 2017)
2. discourse/rhetorical features (Rubin et al., 2015; Shu et al., 2017)
3. Word embedding techniques (Rashkin et al., 2017; Wang, 2017, Ahmed et al., 2018; Kula et al., 2020; Goldani et al., 2020)

Accuracy and Interpretability

1. Word embeddings have proven to be particularly successful
2. However, latent features are not human interpretable
3. We thus need both accurate but also understandable models (Rudin, 2019; Loyola-Gonzalez, 2019)

Contribution and Approach

Contribution

- We respond to recent calls for more explainable (white-box) approaches to fake news detection
- Features of language use informed by contemporary theories of human language learning and processing

Approach

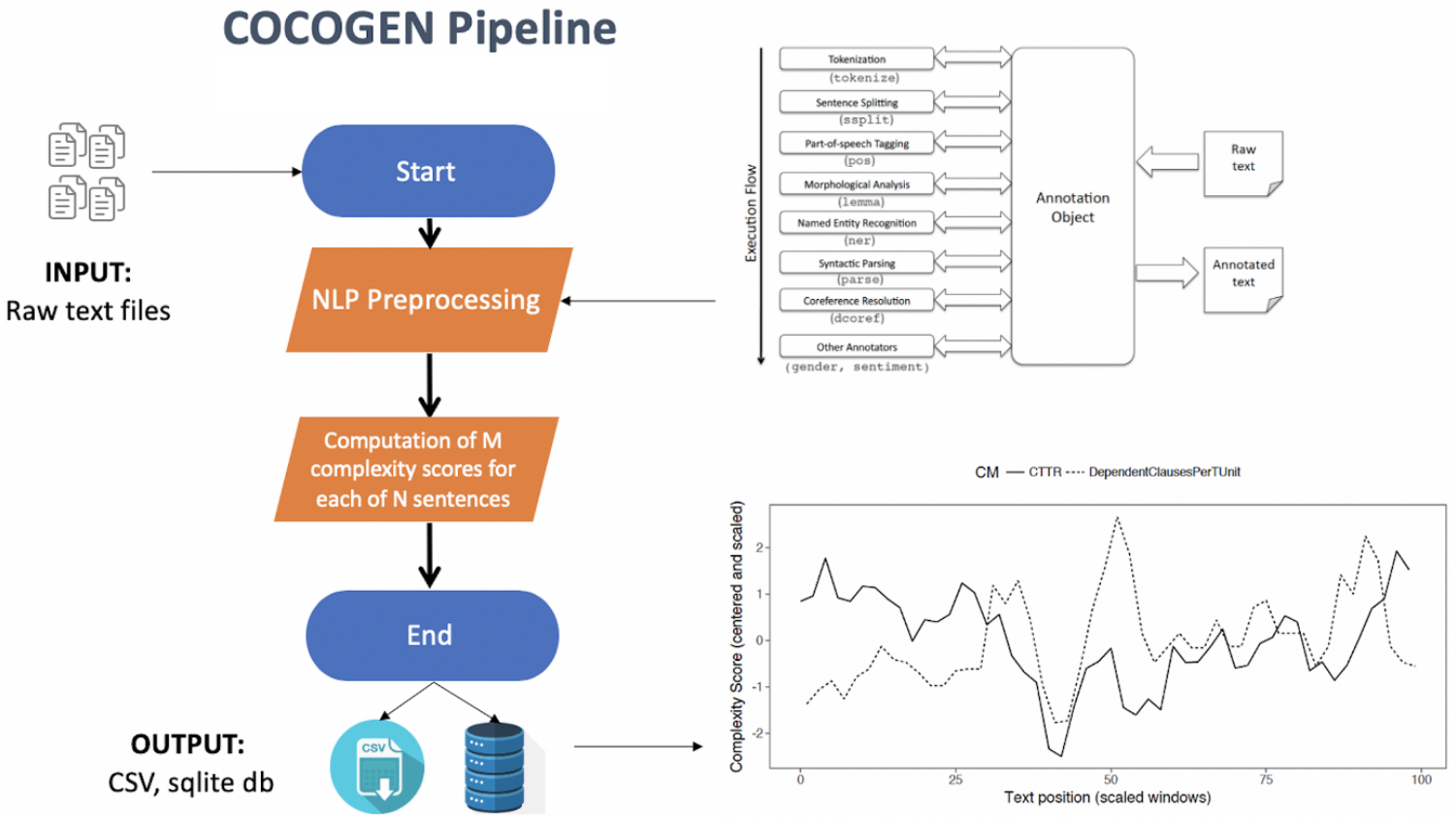
- Perform series of experiments on two benchmark datasets: bi-directional recurrent neural network classification models trained on interpretable features
- Employ high-resolution language analysis afforded by CoCoGen, our computational tool that calculates within-text distributions of feature scores
- Approach achieves similar results as best performing black box models
- Report on ablation experiments geared towards assessing feature importance

Datasets

Datasets were selected based on their complementary attributes in terms of text types and the granularity of the veracity labels

1. ISOT (Ahmed et al., 2018)
   - 'Entire article' dataset comprising 20k+ real and fake news texts (average text length: 400 words)
   - Binary veracity labels (real vs. fake):
     - Real (truthful) news articles crawled from Reuters.com
     - Fake news articles were collected from unreliable websites that were flagged by politifact.com and Wikipedia
2. LIAR (Wang, 2017):
   - 'Claims dataset' comprising 12k+ real-world short statements (average text length: 17 words) sampled from various contexts (news releases, TV or radio interviews, campaign speeches)
   - Six-way veracity labels:
     - Each statement was labeled by an editor from politifact.com on a six-level ordinal scale of truthfulness (pants-fire, false, barely-true, half-true, mostly true & true)

CoCoGen - Automated Text Analysis based on Within-text Distributions of feature scores



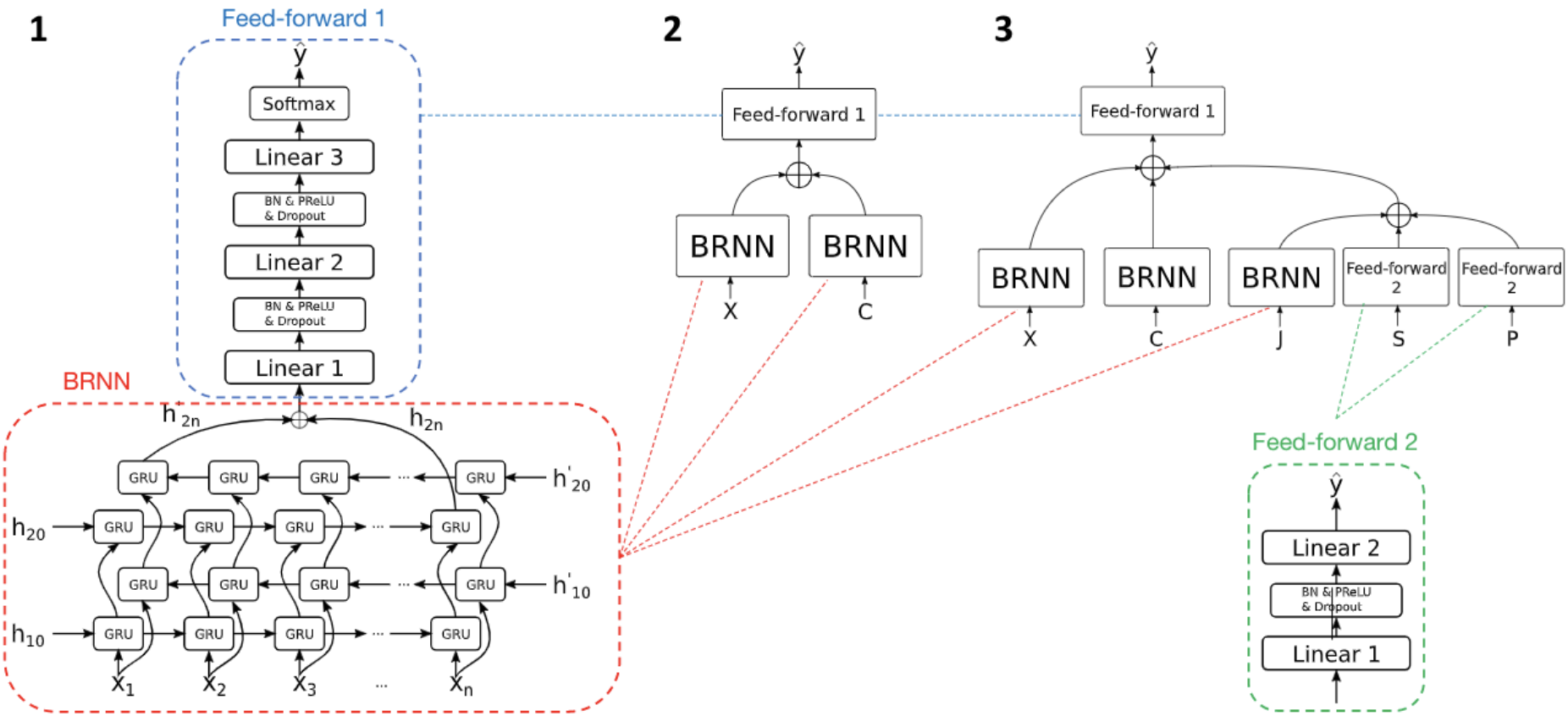
(see Ströbel et al., 2018; Kerz et al., 2020a; Kerz et al., 2020b, for details and recent applications of the tool)

Feature sets

Feature group	Size	Subtypes	Example/Description
Syntactic complexity	18	Length of production unit Subordination Coordination Particular structures	e.g. mean length of clause e.g. clauses per sentences e.g. coordinate phrases per clause e.g. complex nominals per clause
Lexical richness	12	Lexical density Lexical diversity Lexical sophistication	e.g. ratio contents words / all words e.g. type token ratio e.g. words on General Service List
Register-based n-gram frequency	25	Spoken ($n \in [1, 5]$) Fiction ($n \in [1, 5]$) Magazine ($n \in [1, 5]$) News ($n \in [1, 5]$) Academic ($n \in [1, 5]$)	measures of frequencies of n-grams of order 1-5 from five language registers
Information theory	3	$\text{Kolmogorov}_{\text{Deflate}}$ $\text{Kolmogorov}_{\text{Deflate Syntactic}}$ $\text{Kolmogorov}_{\text{Deflate Morphological}}$	measures use Deflate algorithm and relate size of compressed file to size of original file
LIWC-style	60	2300 words from > 70 classes	classes include e.g. function, grammar perceptual, cognitive and biological processes, personal concerns, affect, social, basic drives, ...
Word-Prevalence	36	crowdsourcing-based corpus-based	measures capture information on word frequency, contextual diversity and semantic distinctiveness differentiated across language variety (US, UK) and gender (male, female)

Classification Models

- **Classification Model:** 2-layer bidirectional RNN with GRU cells followed by a 3-layer FFNN with PReLU as activation function.
- **Meta-data Encoders For Liar Dataset:**
  - contextual meta-data and job titles: word embedding + BRNN
  - party affiliation and speaker: one-hot encoding + 2-layer FFNN



CoCoGen - Automated Text Analysis based on Within-text Distributions of feature scores

Algorithm 1: Feature ablation algorithm

Input: N training instances with feature group set $F = \{f_1, \ldots, f_D\}$
Input: m feature groups to remove at each step
Result: *list* containing the feature group importance rank order
begin
  $t \leftarrow 0$
  $list \leftarrow []$
  while $|F| > 0$ **do**
    Train a classifier with $|F|$ input feature groups;
    Compute $S_{i,t}, i \in F$;
    Find $f_{i_1}, \ldots, f_{i_m}$, where $S_{i_1,t}, \ldots, S_{i_m,t}$ are m largest among all $S_{i,t}(i \in F)$ in descending order;
    $list \leftarrow list.append([f_{i_1}, \ldots, f_{i_m}])$;
    $F \leftarrow F - \{f_{i_1}, \ldots, f_{i_m}\}$;
    $t \leftarrow t + 1$;
  return *list*

Experiments

- **Dataset Splitting** train/dev/test with 80/10/10
- **Loss function:**
  - Cross Entropy Loss: $\mathcal{L}(\hat{Y}, Y) = -\sum_{i=1}^{C} y_i \log(\hat{y}_i)$
  - For classifier using ordinal information, binary cross entropy was used: $\mathcal{L}(\hat{Y}, Y) = -\frac{1}{C}\sum_{i=1}^{C}(y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i))$
  - where $\hat{Y} = (\hat{y}_1, \hat{y}_2, , \hat{y}_C)$ is the prediction of classifier and $Y = (y_1, y_2, \ldots, y_C)$ is the target. C is the number of categories.
- **Baseline:** structurally equivalent BRNN classifiers based on sentence embeddings from Sentence-BERT (SBERT)

Results - Classification experiments

Dataset	Model	Accuracy	Precison	Recall
ISOT	LSVM unigram 50k[1]	0.920	–	–
	LSTM-glove[2]	**0.998**	–	–
	CAPSULE-glove[3]	**0.998**	–	–
	BRNN SBERT	0.997	0.997	0.997
	BRNN CoCoGen	**0.993**	0.993	0.993
LIAR	Bi-LSTM 300-dim word2vec[4] embeddings (Google News)	0.233	–	–
	CNN 300-dim word2vec[4] embeddings (Google News) + context + speaker profile	**0.274**	–	–
	CAPSULE-glove + Party[3]	0.240	–	–
	CAPSULE-glove + State[3]	0.243	–	–
	CAPSULE-glove + Job[3]	0.251	–	–
	BRNN SBERT (ordered)	0.270	0.296	0.249
	BRNN CoCoGen (ordered)	0.237	0.217	0.207
	BRNN CoCoGen (ordered) + context	0.253	0.281	0.238
	BRNN CoCoGen (ordered) + context + speaker profile	**0.272**	0.304	0.258

[1]=Ahmed et al., 2018; [2]=Kula et al., 2020; [3]=Goldani et al., 2020;
[4]= Wang, 2017

Results - Feature ablation experiments

Dataset	Feature Group	Accuracy base model	Accuracy after drop
ISOT	LIWC	0.993	0.942
	Syntactic	0.991	0.964
	Lexical	0.988	0.915
	N-grams	0.989	0.763
	Info theory	0.979	0.822
	Word-prevalence	0.933	0.482
LIAR	Lexical	0.255	0.217
	LIWC	0.252	0.204
	Syntactic	0.232	0.193
	N-grams	0.224	0.188
	Word-prevalence	0.210	0.190
	Info theory	0.209	0.193

Conclusion and Outlook

- We demonstrated that neural network classification models that are trained on interpretable features motivated by current theories of language processing and learning can compete with SOTA black box models using word embeddings.
- Future work:
  1. extending the approach presented here to more benchmark datasets (including COVID-19)
  2. extending it to fake news detection in German

EXTRA: Category Encoding

- **ISOT Dataset:** Binary Classification, fake news or not.
  - One-hot encoding.
- **Liar Dataset:** six-way classification
  - One-hot encoding.
  - Ordinal information:
    pants-fire<false<barely-true<half-true<mostly-true<true
  - category k is encoded by $(y_1, y_2, \ldots, y_{C-1})$, where $y_i = 1$ for $i < k$ and otherwise 0
  - category 0: $(0, 0, 0, \ldots, 0)$
    category 1: $(1, 0, 0, \ldots, 0)$
    category 2: $(1, 1, 0, \ldots, 0)$
    $\ldots$
    category C: $(1, 1, 1, \ldots, 1)$