UNIVERSITÄT
LEIPZIG

Project Presentation

Explicit Sentiment Analysis with Language Patterns about Uncertainty

04th July 2022

Jan Albrecht, Paul Brassel, Pascal Singer

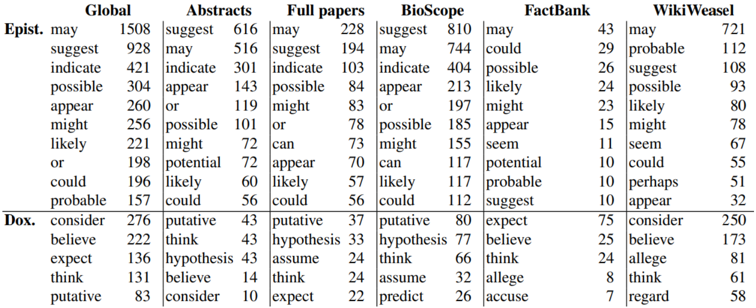
Goals

1. Extract a dataset about semantic uncertainty from the web archive data.
   - Use specific language patterns about uncertainty
   - Classify samples into positive/negative sentiments
   - Compare dataset to Sentiment140[1]
2. Train a sentiment classifier based on DistilBERT on our dataset using transfer learning.
   - Baseline classifier is finetuned on SST-2[2]
   - Benchmark our classifier on Sentiment140

[1] https://huggingface.co/datasets/sentiment140

[2] https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english

Language Patterns



	Global		Abstracts		Full papers		BioScope		FactBank		WikiWeasel	
Epist.	may	1508	suggest	616	may	228	suggest	810	may	43	may	721
	suggest	928	may	516	suggest	194	may	744	could	29	probable	112
	indicate	421	indicate	301	indicate	103	indicate	404	possible	26	suggest	108
	possible	304	appear	143	possible	84	appear	213	likely	24	possible	93
	appear	260	or	119	might	83	or	197	might	23	likely	80
	might	256	possible	101	or	78	possible	185	appear	15	might	78
	likely	221	might	72	can	73	might	155	seem	11	seem	67
	or	198	potential	72	appear	70	can	117	potential	10	could	55
	could	196	likely	60	likely	57	likely	117	probable	10	perhaps	51
	probable	157	could	56	could	56	could	112	suggest	10	appear	32
Dox.	consider	276	putative	43	putative	37	putative	80	expect	75	consider	250
	believe	222	think	43	hypothesis	33	hypothesis	77	believe	25	believe	173
	expect	136	hypothesis	43	assume	24	think	66	think	24	allege	81
	think	131	believe	14	think	24	assume	32	allege	8	think	61
	putative	83	consider	10	expect	22	predict	26	accuse	7	regard	58

Figure 1: The most frequent cues in the English corpora.[3]

[3] **http://doktori.bibl.u-szeged.hu/id/eprint/2291/1/Vincze_Veronika_tezis.pdf**, p. 43

Dataset

- Extract data from web archive
- Annotate with Twitter-roBERTA-base-sentiment model[4]
- Topic extraction for web archive dataset
- Compare topic frequency and sentiment with Sentiment140 (Twitter)
- Questions: Which topics is the internet most uncertain about? Have those changed over the years?

[4] **https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest**

Model

- Train on extracted web archive dataset
- Leave out Twitter URLs of Web Archive to prevent Train-Test-Leakage
- Can transfer learning on exclusively uncertain language samples improve sentiment detection?