Explicit Sentiment Analysis with Language Patterns about Uncertainty

Our goal with this project is divided into two parts. First of all we extract a dataset about semantic uncertainty from the web archive data [2]. This dataset is then compared to Sentiment140 [1] to clarify how well our dataset conforms with this sentiment classification dataset. In the second part a sentiment classifier based on DistilBERT and finetuned on SST-2 [3] is trained on our dataset using transfer learning. This model is benchmarked on the previous mentioned dataset to see if our dataset is suited to improve sentiment classification.

	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
Invest.	whether	247	investigate	177	whether	73	investigate	221	whether	26	whether	52
	investigate	222	examine	160	investigate	44	examine	183	if	3	if	20
	examine	183	whether	96	test	25	whether	169	remain to be seen	2	whether or not	7
	study	102	study	88	examine	23	study	101	question	1	assess	3
	determine	90	determine	67	determine	20	determine	87	determine	1	evaluate	3
Cond.	if	418	if	14	if	85	if	99	if	65	if	254
	would	238	would	6	would	46	would	52	would	50	would	136
	will	80	until	2	will	20	will	20	will	21	will	39
	until	40	could	1	should	11	should	11	until	16	until	15
	could	30	unless	1	could	9	could	10	could	9	unless	14

Figure 1: The most frequent semantic cues in the English corpora [4]

We will use specific language patterns about uncertainty to extract samples from web archive data. An overview of the patterns can be found here [4, p. 43]. To classify these samples into positive/negative sentiments we will mainly use GPT-3. Since we cannot fully trust results based on GPT-3 we will verify some of the labels manually. We plan to analyse our dataset based on what topics the internet is most uncertain about and how those topics changed over time using circular packing charts. After the labeling process we want to compare strong positive/negative topics of our data with those of the dataset mentioned above.

In the next step we train our model. We prevent train-test leakage with the known train-validation-test split method. Furthermore we exclude all Twitter domain names from the web archive because Sentiment140 is based on statements from Twitter. The ratios of this split depends on the amount of samples we actually get in the end. We then compare our model and the baseline model with regards to the performance on Sentiment140 dataset.

This leads us to the following two research questions:

- Does finetuning on uncertain statements improve sentiment classification?
- What are topics the internet is most uncertain about and have those topics changed over time?

The project will be split in the following three work packages:

- Dataset extraction from the web archive data
- Labeling, analyzing, visualizing the dataset and comparing it to Sentiment140
- Train a model and evaluate it on said dataset

Deliverables of the project are:

- A Dataset consisting of only uncertain statements (including datasheet)
- Visualizations and analysis of said dataset
- A DistilBERT model trained on said dataset (including model card)

---

[1] Go, Alec ; Bhayani, Richa ; Huang, Lei: Twitter sentiment classification using distant supervision. In: *Processing* 150 (2009), 01

[2] Kiesel, Johannes ; Kneist, Florian ; Alshomary, Milad ; Stein, Benno ; Hagen, Matthias ; Potthast, Martin: Reproducible Web Corpora. In: *Journal of Data and Information Quality* 10 (2018), dec, Nr. 4, S. 1–25

[3] Sanh, Victor ; Debut, Lysandre ; Chaumond, Julien ; Wolf, Thomas: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In: *CoRR* abs/1910.01108 (2019). – URL `http://arxiv.org/abs/1910.01108`

[4] Vincze, Veronika: Uncertainty detection in natural language texts. In: *PhD, University of Szeged* 141 (2014)