Fantastic Beasts and Where to Find Them

Out of the Box Sentiment Classification

Fuat Can Beylunioğlu

July 25, 2019

Fuat Can Beylunioğlu

(Siz) (bizim)

Çekoslovakya-lı-laş-tır-abil-di-k-ler-imiz-den misin-iz?

- Turkish is an agglutinating language having a complex morphology.
- Due to its complex structure rule based approaches couldn´t achieve good accuracy and no good baseline has been proposed.
- No big corpus available to train contextualized word representations or multi-task learning

Table: # of online comments per company and platform

Corpus	Source	
Technology	Twitter	196079
	Facebook	22767
	Web	55960
Other	Twitter	18080
	Facebook + Web	21435

Table: Distribution of sentiment categories w.r.t. company

	Positive	Neutral	Negative	Total
Technology	14.87%	58.36%	26.77%	274806
Insurance	7.17%	52.79%	40.05%	18721
Car	3.35%	93.29%	3.36%	11943
Rental 1	9.58%	78.29%	12.13%	3340
Rental 2	10.17%	89.22%	0.62%	5508

- Developing a Twitter sentiment classifier based on the company's own tweet corpus.
- Developing a social media sentiment classifier (Twitter, Facebook and other web) using company's own corpus.
- Developing a general social media sentiment classifier

1. Word tokens
    - "Laubalilikten" "hoşlanmam" "damat"
2. Stem & Suffixes
    - "Laubali", "lik", "ten", " ", "hoşlan", "ma", "m", " ", "damat"
3. Uni+Bigrams - r"\S{1,2} | \s"
    - "La", "ub", "al", "il", "ik", "te", "n", " ", "ho", "şl", "an", "ma", "m", " ", "da", "ma", "t"

- Inside the box classification
    - (Tech Tweet -> Tech Tweet)
- Out of platform classification
    - (Tech Tweet -> Tech Other)
- Out of topic classification
    - (Tech Tweet -> All Other)

along with using

- 100%
- 50%
- 25%
- and 12.5% of the training set

Accuracy by Task

Task	Model (All CNN)	12.50%	25.00%	50.00%	100.00%
Inside the Box	word	94.08%	94.92%	96.44%	99.63%
	suffix + stem	94.36%	**94.95%**	**96.56%**	99.66%
	char	**93.84%**	94.77%	96.31%	**99.72%**
Out of Platform	word	85.46%	85.30%	85.56%	86.39%
	suffix + stem	**86.04%**	85.24%	86.22%	**87.21%**
	char	84.70%	**85.99%**	**86.46%**	85.10%
Out of Platform and Topic	word	80.88%	80.65%	78.76%	79.94%
	suffix + stem	**86.52%**	**83.62%**	**84.74%**	**84.07%**
	char	82.13%	82.98%	82.49%	82.23%
Out of Topic	word	83.31%	85.12%	85.33%	80.78%
	suffix + stem	**86.62%**	**85.60%**	**86.41%**	**85.41%**
	char	82.19%	79.69%	77.20%	82.35%