Active Learning for Classifying Political Tweets

Erik Tjong Kim Sang*, Marc Esteve del Valle, Herbert Kruitbosch and Marcel Broersma

netherlands eScience center

by SURF & NWO

Problem description

Goal: study the behavior of politicians on social media

Means: tens of thousands of manually annotated tweets

Problem: new tweets need annotation, which requires a lot of work

Examples: tweet intention

CampaignTrail: Tomorrow we'll make a difference: at 16:00 @mariannethieme @PartijvdDieren is a guest in #zeelandkiest http://t.co/D0SoxDkZ @omroepzeeland

Critique: You don't need to be an economist to understand that we need to get rid of mortgage interest tax deduction. #EenVandaag

News: Roemer will not give the right a majority – VK Dossier Elections of 2012 – VK http://t.co/Zg7YZgWn #SP

Methods

We use Facebook's fastText for classifying new tweets based on annotations of old tweets

The baseline performance is not very high (50%, humans reach 71%), so we look for additional methods

We use language modeling and active learning to improve base performances

Active learning

Active learning is a method for selecting the most promising candidate training data for annotation

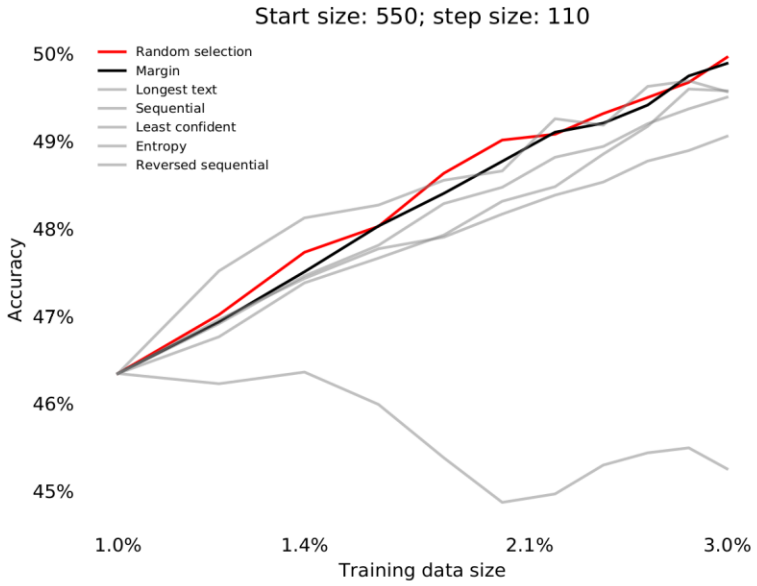
In 2001, Banko & Brill showed that active learning can reduce the required training data with more than 99%

We compare four different methods with three baselines, among which random data selection

Experiment design

1. We started training with 1% of our data
2. Next active learning was used to select the best 0.1% of the remaining data
3. Another extra 0.1% was selected at random
4. Step 2 and 3 were repeated ten times
5. Performances were measured after each step

Results



Train size	Accuracy	Method
80.0%	55.6±0.3%	All training data
3.0%	50.0±0.9%	Random selection
3.0%	49.9±0.9%	Margin
3.0%	49.6±0.7%	Longest text
3.0%	49.6±0.9%	Sequential
3.0%	49.5±1.0%	Least confident
3.0%	49.1±0.8%	Entropy
3.0%	45.3±1.3%	Revrsd sequential
1.0%	46.3±0.8%	Baseline

Why did this not work?

- Perhaps we had too few extra data: 79 times our initial set while Banko and Brill had 999 times more
- Maybe our initial performance is not good enough take advantage of active learning (Dasgupta 2011)
- Or did our language model decrease the benefits of the training data selection methods?

Concluding remarks

We applied active learning for improving the task of labeling political tweets

For our data set, we did not find an improvement of active learning over random training data selection

We provided a few possible reasons for this result

THE END