Automated Analysis of Online Behaviour on Social Media

Erik Tjong Kim Sang
Online Behaviour: project description

• We have 55,000 labeled political tweets

• We have 250,000 unlabeled tweets

• Train a machine learner to accurately classify the unlabeled tweets

• Target performance: 67% accuracy (= human performance)
Labels: intention of political tweets

- **CampaignTrail**: Tomorrow is about something at 16:00 @mariannethieme @PartijvdDieren is a guest in #zeelandkist http://t.co/D0SoxDkZ @omroepzeeland
- **Critique**: You don't need to be an economist to understand how to deal with mortgage deduction. #EenVandaag
- **News**: Roemer will not give the right a majority – VK Dossier Elections of 2012 – VK http://t.co/Zg7YZgWn #SP
We used a standard tool for text classification: fastText

fastText classifies texts based on their words and characters

It can learn from labeled and unlabeled text

Available at: https://github.com/facebookresearch/fastText
Project results

- Improved the best score for this task from 50% to 55%
- Successful application of learning from unlabeled data
- Gained new knowledge related to word vectors and active learning