Semi-supervised Sentiment Analysis for Under-resourced Languages with a Sentiment Lexicon

Peng Liu, Cristina Marco, Jon Atle Gulla

{peng.liu, cristina.marco, jon.atle.gulla}@ntnu.no
Web Intelligence and Semantics Lab

NTNU
Norwegian University of Science and Technology

Background

- A great amount of user opinions are stored online.
- Opinion Mining and Sentiment Analysis have significant and valuable influence on government decision, market advertising and recommender system.

NTNU
Norwegian University of Science and Technology

Background

- Two sentiment analysis methods:
  - ➢ Using a big training corpus to train a supervised learning algorithm.
  - ➢ Making use of a sentiment lexicon in order to perform sentiment analysis on any type of text, such as rule-based method.
- A common challenge of both approaches is the lack of sufficiently big and representative training corpora and sentiment lexicons.

NTNU
Norwegian University of Science and Technology

Background

- The aim of this paper is two-fold:
  - Firstly, we want to present the results of using semi-supervised machine learning on an available training corpus.
  - Secondly, we seek to determine the impact of using a general sentiment lexicon for semi-supervised learning.

NTNU
Norwegian University of
Science and Technology

Related Work

- Most approaches use classification algorithms to determine the polarity of a text, such as Support Vector Machines (SVM), Bayesian Networks, and decision trees, among others.
  - ➢ [Habernal et al. 2015], [Lin et al. 2012] and [Singh and Hussain 2014].
- Lexicon-based approaches
  - ➢ There are a number of lexical resources for this research field, such as SentiWordNet, WordNet-Affect, SentiSense, Opinion Lexicon, Subjectivity Lexicon and MPQA Opinion Corpus, etc.
  - ➢ [Ortega et al. 2013], [Bhaskar et al. 2015], [Chikersal et al. 2015].

NTNU
Norwegian University of Science and Technology

Sentiment Classification

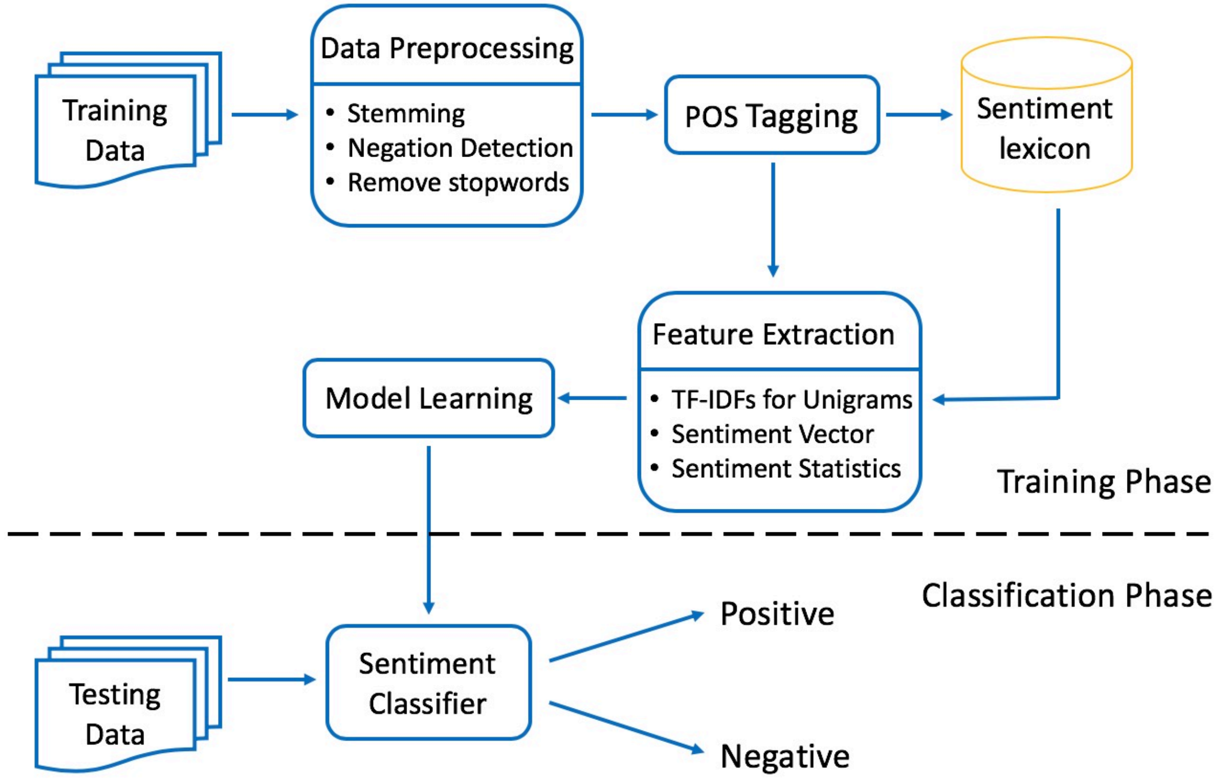

Figure 1: Framework of the proposed sentiment analysis approach.

NTNU
Norwegian University of Science and Technology

Sentiment Classification

- Negation detection

Norwegian *bokmål*	Norwegian *nynorsk*	English
ikke	*ikkje*	'not'
ei	*ei*	'not'
nei	*nei*	'no'
aldri	*aldri*	'never'
neppe	*neppe*	'hardly'
ingen, inga, intet	*ingen, inga, inkje*	'none, any'

Table 1: Negation words in the Norwegian language.

- POS tagger for Norwegian bøkmål [1]

[1] Cristina Marco, Peng Liu, and Jon Atle Gulla. Cross-lingual sentiment analysis for under-resourced languages using machine translation and sentence embeddings. "Under review".

Sentiment Classification

- Input features for machine learning classifier.

Features	Description	Type
TF-IDF	TF-IDFs for Unigrams	Discrete
Sentiment Vector	Sentiment score from the sentiment lexicon according to part-of-speech	Discrete
Statistical Features	1) The minimum/maximum sentiment score of the input document. 2) The number of negative/positive words of the input document. 3) The sum of negative/positive score in the input document. 4) If the sum of negative score is higher than the positive score.	Discrete

Table 2: Sentiment features used in this paper.

Sentiment Classification

- Machine Learning Algorithms
  - Gaussian Naive Bayes (NB)
  - Logistic Regression (LR)
  - Support Vector Machine (SVM)
  - Neural Networks (NN)

Experiments

- Two external resources
  - Training corpus -- Norwegian Review Corpus (NoReC) [2]
  - Norwegian Sentiment lexicon [1] -- This lexicon contains 33,224 synsets and 35,035 wordsenses.

Datasets	Full Review Corpus	Simplified Review Corpus
#Reviews	31,671	15,713
#Pos. reviews	23,477	13,156
#Neg. reviews	8,194	2,557
Imbalance ratios	2.87	5.15

Table 3: Some statistics of the datasets.

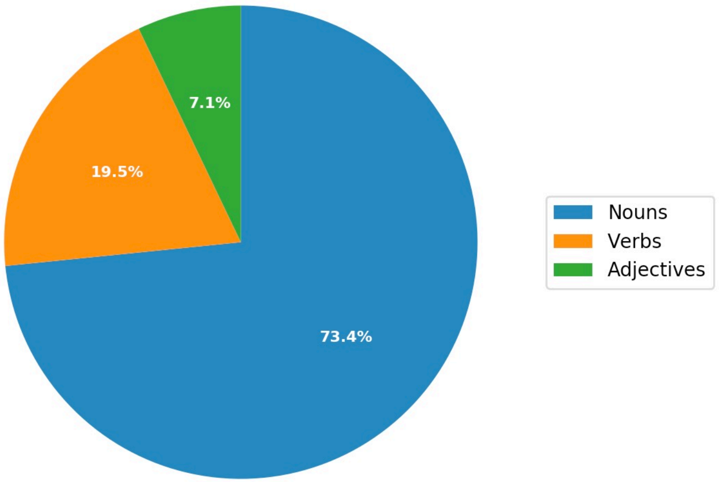

Figure 2: The distribution of synsets per morphological category in Norwegian sentiment lexicon.

[2] https://github.com/ltgoslo/norec.

NTNU
Norwegian University of Science and Technology

Experiments

- Sentiment classification results

Datasets	Full Review Corpus	Simplified Review Corpus
NB	0.7439	0.8428
LR	0.8333	0.9257
SVM	**0.8372***	**0.9296***
NN	0.8159	0.9251

Table 3: The AUC score of sentiment classification results.

NTNU
Norwegian University of
Science and Technology

Experiments

- Effect of different features

Features	NB	LR	SVM	NN
TF-IDF	0.7346	0.8232	0.8310	0.7982
SV	0.6757	0.7365	0.7363	0.6734
SS	0.5906	0.6207	0.6223	0.6184
TF-IDF + SV	**0.7440***	0.8298	0.8348	0.8027
TF-IDF + SS	0.7356	0.8269	0.8340	0.7810
SV + SS	0.6752	0.7423	0.7428	0.6693
TF-IDF + SV + SS	0.7439	**0.8333***	**0.8372***	**0.8159***

Table 4: The AUC score on full review corpus with different features.

Table 5: The AUC score on simplified review corpus with different features.

Features	NB	LR	SVM	NN
TF-IDF	0.8399	0.9176	0.9251	0.9143
SV	0.7673	0.8145	0.8147	0.7856
SS	0.6698	0.7182	0.7177	0.7237
TF-IDF + SV	**0.8438***	0.9198	0.9247	0.9176
TF-IDF + SS	0.8398	0.9229	**0.9305***	0.9093
SV + SS	0.7691	0.8299	0.7292	0.7904
TF-IDF + SV + SS	0.8428	**0.9257***	0.9296	**0.9251***

Conclusions

- To our knowledge, this is the first paper that explores semi-supervised sentiment analysis using a sentiment lexicon for Norwegian.
- The use of features obtained from the general sentiment lexicon improves the results significantly.

NTNU
Norwegian University of Science and Technology

Thank you!

NTNU
Norwegian University of
Science and Technology