SENTIMENT ANALYSIS IN TURKISH FOR MOVIE REVIEWS

Caner Mercan
Doğan Altunbay
Elif Eser
M. Burak Şenol

Instructor: Prof. Dr. Fazlı Can

OUTLINE

- Introduction
- Motivation
- Related Works
- Methodology
- Evaluation
- Conclusion

INTRODUCTION

With the World Wide Web, the quantity of textual information available is immense. As of March 2011 there were around 300 million websites. [1]

Facts and opinions…

In its essence, opinion mining or sentiment analysis deals with opinion.

[1] March 2011 Web Server Survey, NetCraft, news.netcraft.com

INTRODUCTION

- The main problem in sentiment analysis is to define a formal representation of unstructured textual data and classify it as positive, negative or neutral.

In this project, we are dealing with Turkish movie reviews and extract the reviewers' general opinion on movies. Our project is able to automatically analyze reviews and classify them based on the **5-star** scale.

MOTIVATION

- With the World Wide Web, it's been easier to reach opinions of movie reviewers.

- Most of movie review web sites also allow their movie critics to vote movies.

- Thus, we are able to evaluate our system considering the votes that the critics give as actual votes.

RELATED WORKS

Positive vs. Negative; Binary Outcomes:

- Turney et al. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the ACL*, 2002
- Dave et al. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of WWW*, 2003

Multiple Outcomes:

- Bo Pang and Lillian Lee, Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*, 2005.

METHODOLOGY

Feature Extraction

- *Removing Stop Words*
- *Roots and Polarity*
- *Parts of Speech*
- *N-Gram of Words*
- *Output of Processed List*

Classification

- *PRank*
- *SVM*

FEATURE EXTRACTION

In this part, we make a given critics document ready to process in classification.

- *Removing Stop Words*
  - The program takes a document and turns it to a word list belonging to a sentence.
  - Turkish stop word list that we utilize is taken from The Natural Language Processing Group, Fatih University

FEATURE EXTRACTION

- *Roots and Polarity*
  - With Zemberek, we are able to analyze each word in Turkish morphologically.

Git-me-yecek-im

[root:git type:FIIL]

Structure:

FIIL_KOK

FIIL_OLUMSUZLUK_ME

FIIL_DONUSUM_ECEK

ISIM_SAHIPLIK_BEN_IM

FEATURE EXTRACTION

- *Parts of Speech*
  - Use of the adjectives for the unigrams.
    - Proven to have more meaningful & informative information compared to other structures.
  - Use of all word structures for the n-grams.
    - Need to protect the general structure of the sentence to extract meaningful n-grams.

FEATURE EXTRACTION

- *Parts of Speech*
  - Negativity of a sentence
    - Defined as the number of combined negative words/affixes. If the total negativity is preserved, then other words in the sentence is also made negative by adding '_' prefix.
    - Olağanüstü, şahane, fevkalade bir film olmuş!
      - Olaganustu sahane fevkalade
    - Bu film hakkında berbat, rezalet, iğrenç diyenler hiçbir şey bilmiyorlar!
      - _berbat _igrenc

FEATURE EXTRACTION

- *N-Gram of Words*
  - We use n-gram representations of all processed words
    - Some words are more meaningful together. For example "güzel" has a positive meaning but "en güzel" has a more powerful positive meaning.

N-gram	Frequency
film ol	49
ilk film	36
ol film	31
_film _ol	27
en iyi	26
film izle	25
cok sev	23
tur film	21
film festival	20
_cok _sev	16
cok iyi	12
altin portakal	10

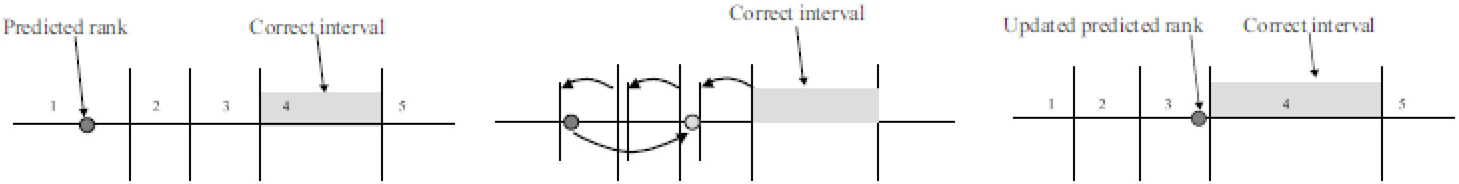
FEATURE EXTRACTION

- *Output of Processed List*
  - Build of feature vectors with
    - Only unigrams -> adjectives
    - N-grams (unigram, bigram …)
  - Feature matrices with varying threshold values.
    - Adjectives and n-grams having only more than k-threshold value occurences in the training data are selected
      - *k = 5,10,15…*

  - *Culmination of Feature Extraction Techniques* → *Feature Matrix*

CLASSIFICATION

- *PRank*
  - Perceptron Ranking algorithm as presented in [1]
  - We will use a subset of the documents as a train set.
    - The alogirthm keeps a set of weights associated with each feature and a set of boundaries for ranks.



SVM

- We use the extracted features of the training set to train an SVM model and upon obtaining the test set; we classify it based on its processed features with our previously trained model.

[1] K. Crammer, Y. Singer, Pranking with Ranking, School of Computer Science and Engineering, The Hebrew University

CLASSIFICATION

For both of the classification techniques, we are able to evaluate its performance to the already present other studies. The performance measure of our study is **F-Measure** in order to drop the bias under good majority class estimates but bad minority class results.

CONCLUSION

Sentiment analysis is a hard task.

In our project, we dealt with movie reviews from feature extraction process to classification.

Moreover, we pointed out to the fact of the lack of studies in Turkish reviews, hence applied methodologies that are deemed useful and practical in their foreign language counterparts.