Adaptation and Use of Subjectivity Lexicons for Domain Dependent Sentiment classification

Rahim Dehkharghani,, Berrin Yanikoglu, Dilek Tapucu, and Yücel Saygın

Faculty of Engineering and Natural Sciences,

Sabanci University

Istanbul, Turkey

Sabanci University

Sentiment Analysis Research Group

Content

- Introduction
- Suggested approach
  - Features Based on SentiWordNet
  - Features Based on Subjective Words
- Classification
- Results
- Discussion

10/12/2012

Sentiment Analysis of Reviews

- **Definition:**
  - Sentiment Analysis is the task of extracting the attitude (positivity, objectivity or negativity) of a text (in natural language).
- **Motivation:**
  - Producers and companies like to know the ideas of their customers about their services and products.
  - Manual extraction of the sentiment of a text is time consuming, so making this task automatic can save a lot of time.
- **Problem:**
  - Extracting the sentiment (positivity or negativity) of a text review in two domains (Hotel and Movie).

Approaches to Sentiment Analysis

- Lexicon-based approach
  - Semantic orientation of words in a review are obtained from a domain-independent polarity lexicon such as SentiWordNet.
  - Features (average polarity, purity,…) are extracted from these word polarities.
- Supervised methods
  - Machine Learning techniques are used to establish a domain-specific model from a large corpus of labelled reviews.
  - Although these methods are typically more successful, collecting a large training data is often a problem.
- Often the review is seen as a bag-of-words.

In this Work

- Combining Domain-independent and Domain-specific resources and using Machine Learning methods

- Proposing new features based on seed word sets

Used Resources

- Domain-independent Resource (SentiWordNet)
  - We used the positivity and negativity values of each word in this resource.

- Domain-specific Resource (SubjWords)
  - Extracted from seed word list of Liu and Hu [Liu and Hu, 2005], based on their occurrence in a specific domain (Hotel and Movie).

Bing Liu, Minqing Hu and Junsheng Cheng. "Opinion Observer: Analyzing and Comparing Opinions on the Web" To appear in *Proceedings of the 14th international World Wide Web conference (WWW-2005)*, May 10-14, 2005, in Chiba, Japan

Feature List

Using SentiWordNet	**F1: Average polarity of all words** **F2: Average polarity of negative words** **F3: Average polarity of positive words** **F4: Average polarity of last 3 sentences** **F5: Average polarity of first 3 sentences**
Using SubjWords	F6: Cumulative frequency of positive words F7: Cumulative frequency of negative words F8: Proportion of positive to negative words F9: Weighted probability of positive words F10: Weighted probability of negative words

SentiWordNet

- A WordNet based polarity lexicon
- It associates words with positivity, negativity and objectivity values

$$\text{PosScore}(w) + \text{NegScore}(w) + \text{ObjScore}(w) = 1$$

- Word Polarity

$$\text{Pol}(w) = \text{PosScore}(w) - \text{NegScore}(w)$$

- Review Polarity

$$AP(R) = \frac{1}{|R|} \sum_{w_i \in R} \text{Pol}(w_i)$$

10/12/2012

Features based on SentiWordNet

- F1: Average polarity of all words
- F2: Average polarity of negative words
- F3: Average polarity of positive words
- F4: Average polarity of last three sentences
- F5: Average polarity of first three sentences

SubjWords

- **InitialSeedWords** : Seed Word List used in [Liu and Hu, 2005]
- Includes 2005 Positive and 4783 negative words.
- **SubjWords** : a subset of InitialSeedWords based on the occurrences of those words in 500 pos. and 500 neg. reviews of two domains.
  - **Hotel Domain**: 671 positive and 1393 negative words
  - **Movie Domain**: 1093 positive and 1977 negative words

10/12/2012

Features based on SubjWords

- F6: Cumulative frequency of positive words
- F7: Cumulative frequency of negative words
- F8: Proportion of positive to negative words
- F9: Weighted probability of positive words
- F10: Weighted probability of negative words

Features based on SubjWords

F6: Cumulative frequency of positive words

$$F_6(r) = \sum_{t_i \in PS}^{n} tf(t_i, r)$$

F7: Cumulative frequency of negative words

$$F_7(r) = \sum_{t_i \in NS}^{n} tf(t_i, r)$$

F8: Proportion of positive to negative words

$$F_8(r) = \frac{p+1}{n+1}$$

Features based on SubjWords

F9: Weighted probability of positive words

$$F_9(r) = p * (1 - P_+(p))$$

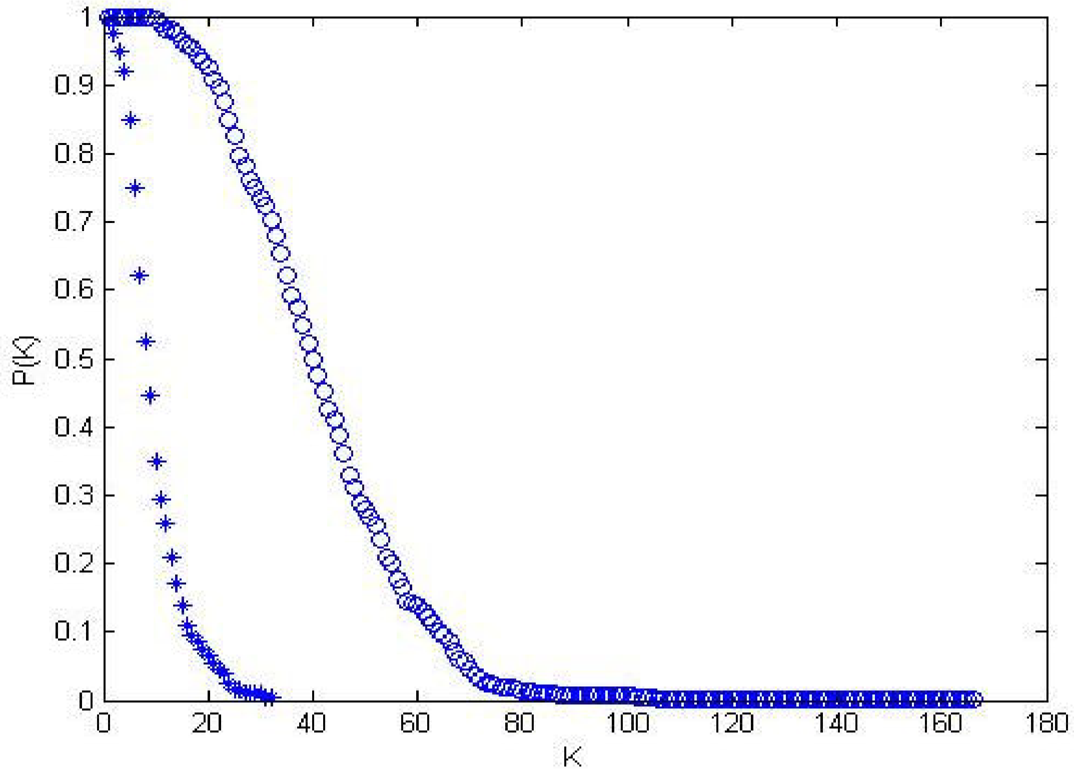
F10: Weighted probability of negative words

$$F_{10}(r) = n * (1 - P_-(n))$$

Features based on SubjWords

Plot of $P_+(p)$ as a function of p ('*' represents the hotel and 'o' represents the movie domain).

Domai n	Hotel	Movie
Avg. # of words per review	157	734



10/12/2012

Suggested Approach

- Preprocessing the text
  - Tokenizing
  - POS tagging

- Extracting the feature vector for each review
  - Features using domain independent resource (SentiWordNet)
  - Features using domain-specific resource (SubjWords)

- Classifying the reviews

Datasets

- Hotel Domain
  - TripAdvisor corpus
  - 250000 customer-supplied reviews of 1850 hotels
  - randomly selected 6000 reviews half positive, half negative

- Movie Domain [B. Pang and L. Lee, 2004]
  - all reviews in this domain including 1000 positive and 1000 negative reviews

- The TripAdvisor website. http://www.tripadvisor.com (2011), [TripAdvisor LLC].
- B. Pang and L. Lee, “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts,” in Proceedings of the ACL, 2004, pp. 271–278.

Example

- Fantastic Experience. We booked a room on Valentines day and the experience was fantastic. The lady at the front desk was very helpful and rather friendly. When we arrived there where fresh pastries and brewed coffee waiting at the front desk. The room was very spacious, the cable was good. The shower had enough pressure and the temperature of the water never changed once it was set. We took a walk up Lombard street and ended up in the Italian district surrounded by great restaurants and a live night life. The parking was a little tight so if you have a big car leave it at home. Besides that minor detail we felt safe, comfortable and will return when staying in San Francisco in the near future.

- **Features vector:**

$$(\underbrace{-0.346 , 0.231 , -0.005 , 0 , 0.19}_{\text{SentiWordNet-based features}} , \underbrace{11 , 0 , 0.295 , 0 , 12}_{\text{SeedWord-based features}} , 1)$$

Classifiers

- ➢ SVM classifier
- ➢ Multilayer Perceptron (Neural Networks)
- ➢Logistic Classifier

- ➢Used Tool: Weka 3.6
- ➢ Classification metod: 5 Fold Classification

Classifier	SVM	Neural Networks	Logistic Classifier

Parameters	Nu = 0
	loss =0.1
	epsilon= 0.001
	cost to 1.0
		Learning rate =0.3
		hidden layers= 1
		validation threshold=20
			No parameter set

10/12/2012

Results

Domain	Feature Subset	SVM	Neural Networks	Logistic Classifier
Hotel	Basic:F1-F5	81:58	81:24	81:47
	Pos/Neg. Ratio: F8	83:37	82:78	82:21
	Weight. Pol.: F9,F10	84:45	83:08	82:99
	Cumul. TF.: F6, F7	83:56	84:15	83:07
	F1-F5 + F8	**86.36**	**86.80**	**86.10**
	F6-F7 + F8	**84:52**	**84:51**	**83:43**
	F9-F10 + F8	**85.07**	**83.48**	82.48
	F6-F7 + F8-F10	**84:50**	**84:39**	**83:02**
	All: F1-F5 + F6-F10	**87.10**	**87.08**	**87.51**
Movie	Basic:F1-F5	62:60	62:00	64.2
	Pos/Neg. Ratio: F8	67:95	67:50	68:30
	Weight. Pol.: F9,F10	69:25	65:85	65:75
	Cumul. TF.: F6, F7	70:65	70:25	71:05
	F1-F5 + F8	**69.10**	**67.50**	**70.45**
	F6-F7 + F8	67:20	**71.25**	**72.25**
	F9-F10 + F8	**70.30**	**70.15**	**70.80**
	F6-F7 + F8-F10	68:80	**70.95**	**72.75**
	All: F1-F5 + F6-F10	**68.45**	**71.65**	**72.85**

Discussion on Results

➢The best feature group in isolation is based on cumulative term frequencies (F6 and F7).

➢ The accuracy of domain-specific features, F6-F10 is better than the accuracy of domain-independent ones, F1-F5.

➢The most useful addition is the positive to negative word ratio (F8) which is mostly positive.

➢ In both domains the best results are obtained using all features, except for one experimental setup (the accuracy of the SVM in the movie domain is highest using only F6 and F7, which may be due to suboptimal parameter optimization in SVMs).

Conclusion and Future Work

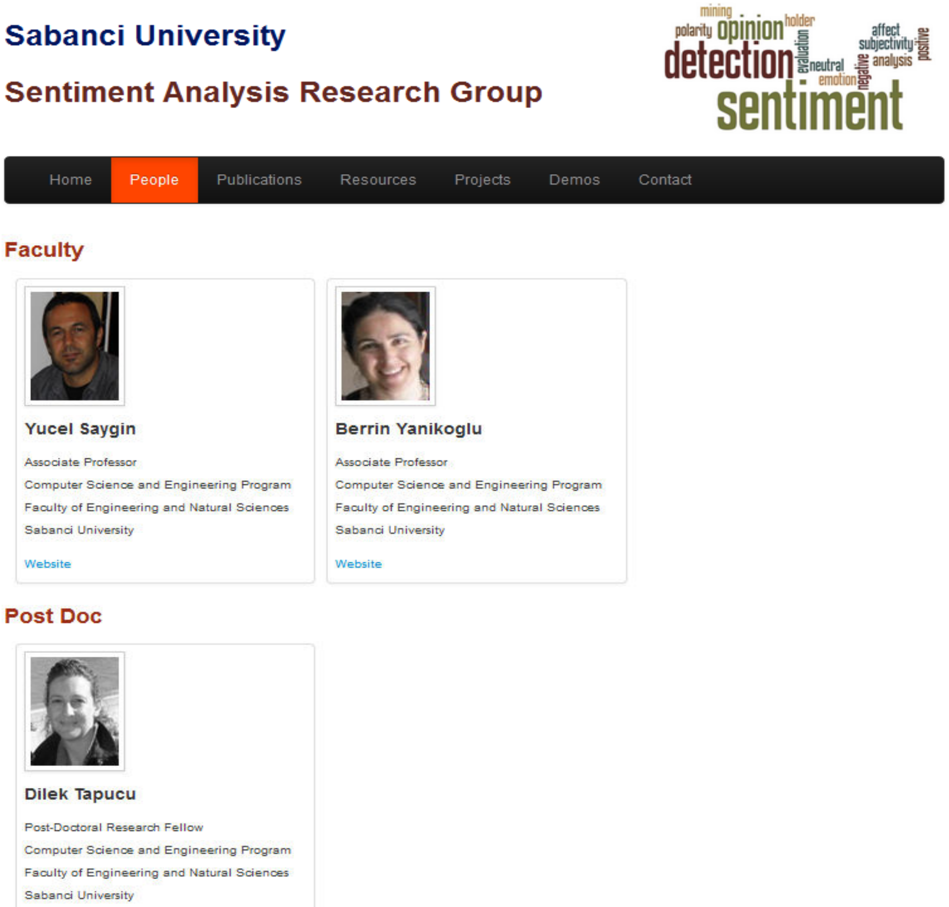
- **We proposed**
- A hybrid approach for sentiment analysis in two domains
- Using two different resources: Domain-specific and Domain-independent
- A few new features based on seed word sets
- **Future Work**
- We will extend this work by adapting it to Turkish and use it in a bigger project named SARE

10/12/2012

Thanks for your attention!

For any questions, contact:

http://sentilab.sabanciuniv.edu



Sabanci University

Sentiment Analysis Research Group

Faculty

Yucel Saygin

Associate Professor

Computer Science and Engineering Program

Faculty of Engineering and Natural Sciences

Sabanci University

Website

Berrin Yanikoglu

Associate Professor

Computer Science and Engineering Program

Faculty of Engineering and Natural Sciences

Sabanci University

Website

Post Doc

Dilek Tapucu

Post-Doctoral Research Fellow

Computer Science and Engineering Program

Faculty of Engineering and Natural Sciences

Sabanci University

Email :

- rdehkharghani@sabanciuniv.edu