Adjective Intensity and Sentiment Analysis

Raksha Sharma, Mohit Gupta, Astha Agarwal, Pushpak Bhattacharyya
Dept. of Computer Science and Engineering
IIT Bombay, Mumbai, India
{raksha,mohitgupta,astha11,pb}@cse.iitb.ac.in

Abstract

For fine-grained sentiment analysis, we need to go beyond zero-one polarity and find a way to compare adjectives that share a common semantic property. In this paper, we present a semi-supervised approach to assign intensity levels to adjectives, *viz. high, medium* and *low*, where adjectives are compared when they belong to the same semantic category. For example, in the semantic category of EXPERTISE, *expert, experienced* and *familiar* are respectively of level *high, medium* and *low*. We obtain an overall accuracy of 77% for intensity assignment. We show the significance of considering intensity information of adjectives in predicting star-rating of reviews. Our intensity based prediction system results in an accuracy of 59% for a 5-star rated movie review corpus.

1 Introduction

Sentence intensity becomes crucial when we need to compare sentences having the same polarity orientation. In such scenarios, we can use intensity of words to judge the intensity of a sentence. Words that bear the same semantic property can be used interchangeably to upgrade or downgrade the intensity of the expression. For example, *good* and *outstanding* both are positive words from the QUALITY category, but the latter can be used to intensify positive expression in a sentence.

There are several manually or automatically created lexical resources (Liu, 2010; Wilson et al., 2005b; Wilson et al., 2005a; Taboada and Grieve, 2004) that assign a fixed positive (+1) or negative (−1) polarity to words, making no distinction among them in terms of their intensity. This paper presents a semi-supervised approach to assign intensity levels to adjectives, *viz. high, medium* and *low*, which share the same semantic property. We have used the semantic frames of FrameNet-1.5 (Baker et al., 1998) to obtain these semantic categories. Our approach is based on the idea that the most intense word has higher contextual similarity with *high* intensity words than with *medium* or *low* intensity words. We use the intensity annotated movie review corpus to obtain the most intense word for a semantic category. Then, cosine similarity between word vectors of the most intense word and other words of the category is used to assign intensity levels to those words. Our approach with the used resources is shown in figure 1.

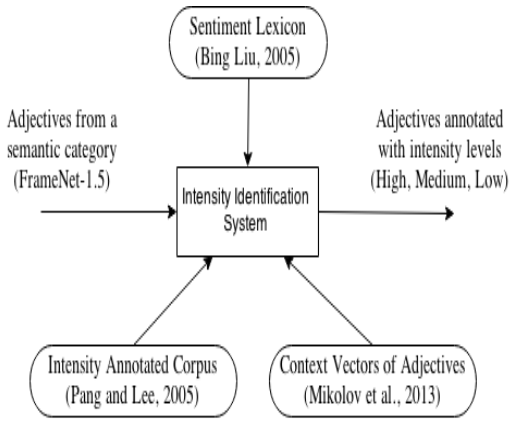

Figure 1: Intensity Analysis System

Our Contribution: Corpus based approaches suffer from the data sparsity problem. Our approach tackles this problem by using word vectors for intensity assignment (Section 2.3). It also provides a better overall accuracy (77%) than current state of the art when compared with gold-standard intensity levels (Section 6.2). In addition to this, we show that accuracy of the star rating prediction task improves when we incorporate our intensity levels as features in addition to standard features such as unigrams (Section 6.3).

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2520–2526,
Lisbon, Portugal, 17-21 September 2015.

2 Idea Used for Deriving Adjectival Scale

In this paper, we dealt with 52 semantic (polar) categories of the FrameNet data and derived the polarity-intensity ordering among adjectives for each category. Examples of these semantic categories with a few words that belong to the category are as follows.

- INTELLIGENCE: *Brainy, brainless, intelligent, smart, dim etc*.
- CANDIDNESS: *Honest, dishonest, trustworthy, reliable, gullible etc*.
- EMOTION: *Sad, upset, appalled, tormented, gleeful, happy, pleased etc*.

Our algorithm to assign intensity levels to adjectives is based on the following ideas:

2.1 What Does Intensity Annotated Corpus Tell About The Intensity of Words?

Rill et al. (2012) showed that an intensity annotated polar corpus can be used to derive the intensity of the adjectives. A high intensity word will occur more frequently in high intensity reviews. For example, the word *excellent* is found 118 times, while *average* is found only 16 times in 5-star rated movie reviews (Section 3). Based on this distribution, we use a weighted mean formula to find intensity of the words from the corpus. We call it **Weighted Normalized Polarity Intensity (WNPI)** formula. For a 5-star intensity rating corpus, the $WNPI$ formula is as follows:

$$WNPI(word) = \frac{\sum_{i=1}^{5} i * C_i}{5 * \sum_{i=1}^{5} C_i} \quad (1)$$

where C_i is the count of the $word$ in i-star reviews.

2.2 Need Significant Occurrence of A Word

The $WNPI$ formula gives a corpus based result, hence can give biased scores for words which occur less frequently in the corpus. For example, in our movie review data-set, the word *substandard* occurs only 3 times in the corpus, and these occurrences happen to be in 1-star and 2-star reviews only. Hence, the $WNPI$ formula assigns a higher score to *substandard*. To avoid such a bias, we integrate $WNPI$ formula with Chi-Square test. Sharma and Bhattacharyya (2013) used Chi-Square test to find significant polar words in a domain. We use the same categorical Chi-Square test in our work.

2.3 How to Get Intensity Clue for All Words?

A combination of $WNPI$ formula and Chi-Square test cannot assign intensity scores to adjectives, which are not present in the corpus. To overcome this data sparsity problem, we restrict the use of $WNPI$ formula to identify the most intense word in each category. We explore precomputed context vectors of words, presented by Mikolov et al. (2011) (Section 3), to assign intensity levels to remaining words of the semantic category:

Case-1 Words which have less number of senses: These words will have a limited set of context words. Hence, their context vectors will also be based on these limited words. Example: *excellent, extraordinary, amazing, superb, great etc.*

Case-2 Words which have many senses: These words will have a large set of context words. Hence their context vectors will be based on a set of large number of words. Example: *good, fair, fine, average etc.*

Inferences:

1. Two words expressing similar meaning, and satisfying case-1 will have similar context. Hence, their word vectors will exhibit high cosine similarity. Whereas a word satisfying case-2 will be less similar to a word satisfying case-1.

2. The classical *semantic bleaching theory*[1] states that a word which has less number of senses (possibly one) tends to have higher intensity in comparison to a word having more senses. Considering *semantic bleaching* phenomenon as a base, we deduce that words which satisfy case-1 tend to be *high* intensity words while words satisfying case-2 are *low* intensity words.

Hence, we conclude that *high* intensity words (case-1) have higher *cosine similarity* with each other than with *low* or *medium* intensity words (case-2). Therefore, cosine similarity with a *high* intensity word can be used to obtain intensity ordering for remaining words of the category.

3 Data and Resources

This section gives an overview of the corpus and lexical resources used in our approach.

Semantic Categories: We worked with frames of FrameNet-1.5 (Baker et al., 1998). A frame

[1]The *semantic bleaching phenomenon* in words was reported in US edition of *New York Times*: `http://www.nytimes.com/2010/07/18/magazine/18onlanguage-anniversary.html?\_r=0`

Rating	Definition	Size
0	Totally painful, unbearable picture	179
1	Poor Show (dont waste your money)	1057
2	Average Movie	888
3	Excellent show, look for it	1977
4	A must see film	905

Table 1: Review ratings with their definitions and number of reviews.

represents a semantic property and contains words bearing the property. We explored the FrameNet data manually and found 52 frames (semantic categories) with polar semantic properties.

Intensity Annotated Corpus: To identify a *high* intensity word for a semantic category, we use a movie review corpus[2] (Pang and Lee, 2005) of 5006 files. Each review is rated on a scale of 0 to 4, where 0 indicates *an unbearable movie* and 4 represents *a must see film*. Table 1 describes the meanings of the rating scores with the count of reviews in each rating. We can infer that increase in rating corresponds to increase in positive intensity and decrease in negative intensity.

Sentiment Lexicon: To identify the polarity orientation of words, we use a list of positive (2006) and negative (4783) words[3] (Liu, 2010). We manually assign polarities to universally polar words like *enduring, creditable* and *nonsensical*, which are missing in this lexicon, using other standard lexicons. We found a total of 218 such missing words.

Context Vectors: We use the precomputed context vectors of words generated using Recurrent Neural Network Language Model (RNNLM) (Mikolov et al., 2013). The RNN is trained with $320M$ words from the broadcast news data.

4 Gold Standard Data Preparation

We asked five annotators to assign words to different intensity levels: *high, medium*, and *low*. Annotators were given positive and negative words of each category separately. The level chosen by a majority of annotators is selected as the gold standard intensity level for the word. To compute agreement among five annotators, we used *fleiss' kappa*, and obtained a score of **0.61**.

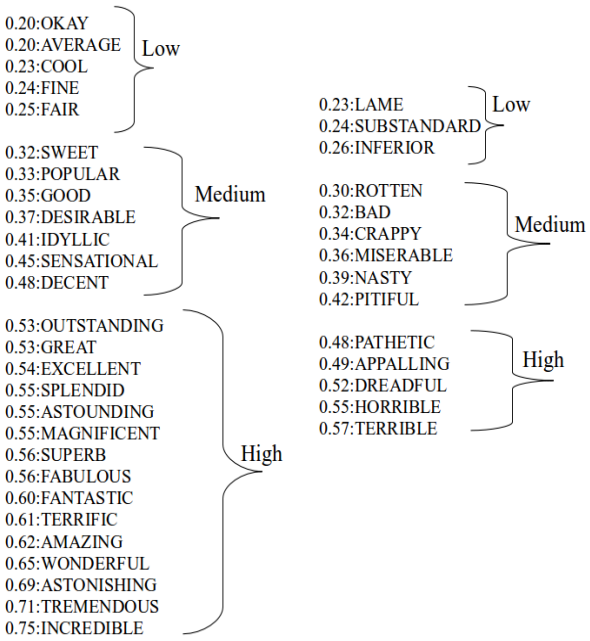

Figure 2: Intensity scale for QUALITY category, where *extraordinary* was found as *Pos-pivot* and *awful* as *Neg-pivot*.

5 Identification of Intensity of Adjectives

In this section, we give a step-by-step description of our approach.

Step 1: Find Intensity of Words

We calculate polarity-intensity of each word of a semantic category using $WNPI$ formula (eq. 1). Based on the polarity orientation of a word, the $WNPI$ formula uses intensity interpretation of star-rating as shown in table 2. The variable i of the $WNPI$ formula refers to these star ratings (intensity levels). The polarity orientation of an observed word is obtained using Bing Liu's lexicon.

Star-Rating / Word-Orientation	0	1	2	3	4
Positive	1	2	3	4	5
Negative	5	4	3	2	1

Table 2: Interpretation of star rating as intensity scores of reviews for positive and negative words.

Step 2: Find Pivot Using Chi-Square Test

The word which gives the highest Chi-Square

[2]Written and rated by four authorized movie critics. Available at: `http://www.cs.cornell.edu/people/pabo/movie-review-data/`

[3]Available at: `http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html\#datasets`

score with the highest intensity score as per $WNPI$ is set as *pivot* (*Pos-pivot* and *Neg-pivot*). The Chi-Square test helps us to exclude the biased words, which are getting high intensity scores by the $WNPI$ formula, just by chance (Section 2.2).

Step 3: Obtain Similarity Scores with Pivot Further, we compute the *cosine similarity* between the context vectors of the pivot and the other words of the category. We use *Pos-pivot*, if the observed word is positive and *Neg-pivot*, if the observed word is negative.

Step 4: Assign Intensity Level to Words Finally, we arrange similarity scores obtained above in decreasing order, and place 2 break points in the sequence where consecutive similarity scores differ the most. We set these breakpoints as the thresholds for intensity levels.

Figure 2 shows the intensity scale obtained by our approach for the QUALITY category, where *extraordinary* was found as *Pos-pivot* and *awful* as *Neg-pivot*.

6 Experiments And Results

To evaluate the performance of our approach, we consider three measures: accuracy with the gold-standard data, comparison with state of the art and accuracy for the *star rating prediction task*.

6.1 Evaluation Using Gold Standard Data

We compute accuracy as the fraction of adjectives for which the predicted intensity level is the same as the gold standard level. We obtained an overall accuracy of 77% across 52 polar categories, containing a total of 697 adjectives.

6.2 Comparison with State of The Art

Ruppenhofer et al. (2014) showed that a corpus based method called MeanStar approach performs the best for intensity ordering task among existing approaches (De Melo and Bansal, 2013; Kim and de Marneffe, 2013; Fahrni and Klenner, 2008; Dragut et al., 2010) for polar semantic categories. Figure 3 shows the comparison between MeanStar and our approach for four semantic categories[4]. For first three categories, our approach performs better than MeanStar and for EXPERTISE we obtain the same level of accuracy. MeanStar approach gives an overall accuracy of 73% across 52 polar categories, which is significantly lesser than the accuracy obtained with our approach. MeanStar approach does not assign intensity score to words missing from the corpus. While, 88 out of 122 missing words are assigned correct intensity levels by our approach.

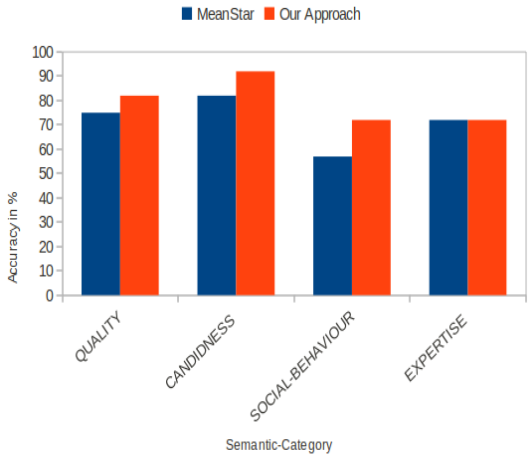

Figure 3: Accuracy obtained with MeanStar and our approach

6.3 Evaluation Using Star Rating Prediction

There have been several successful attempts at sentiment polarity detection in the past (Turney, 2002; Pang et al., 2002; Pang and Lee, 2004; Mohammad et al., 2013; Svetlana Kiritchenko and Mohammad, 2014). However, prediction of star ratings still considered as a challenging task (Qu et al., 2010; Gupta et al., 2010; Boteanu and Chernova, 2013). We implemented three systems to evaluate the significance of intensity annotated adjectives in star rating prediction task.

System 1: A rule based system based on the concept that *negatively high intense* words will occur more frequently in the *low star reviews* and *positively high intense* words will occur more frequently in the *high star reviews*. This system uses the following function I to assign intensity score to a review r:

$$I(r) = \frac{\sum_{i=1}^{3} i * C_i^P - \sum_{i=1}^{3} i * C_i^N}{3 * (\sum_{i=1}^{3} C_i^P + \sum_{i=1}^{3} C_i^N)} \quad (2)$$

where C_i^P and C_i^N respectively represent sum of the term-frequencies of positive and negative adjectives with intensity i.

Eq. 2 gives us an intensity score between -1 and $+1$ for each review. We need four breakpoints on these intensity scores to map intensity scores

[4]We have used the same semantic categories and intensity annotated movie review corpus in our work as used by Ruppenhofer et al. (2014).

into 5-star ratings. We learn these breakpoints by maximizing accuracy for the training data[5] over all possible breakpoints.

System 2: In this system, we consider intensity of each adjective as $+1$ or -1 as per its polarity, and then uses eq. 2 to find review intensity.

System 3: This is an SVM based system which uses four different types of features: (a) unigrams, (b) unigrams with the modification that if adjective belongs to our intensity annotated adjective list, then feature value is intensity of the adjective, (c) and (d) use the scores coming from eq. 2 as an additional feature over those in (a) and (b) respectively.

System	Accuracy(%)	MSE	MAE
1	42.28	0.94	0.69
2	27.33	1.12	0.86
3(a)	55.81	0.63	0.50
3(b)	57.21	0.56	0.47
3(c)	58.71	0.57	0.46
3(d)	**59.21**	**0.54**	**0.45**

Table 3: Comparison of rating prediction systems, where MSE is the Mean Squared Error and MAE is the Mean Absolute Error

Table 3 shows the results obtained with the above systems. System 3(d) achieves the maximum accuracy depicting that inclusion of intensity information with the standard features improves the star rating prediction significantly.

7 Related Work

Sentiment analysis on adjectives has been extensively explored in NLP literature. However, most of the works addressed the problem of finding polarity orientation of adjectives (Hatzivassiloglou and McKeown, 1997; Wiebe, 2000; Fahrni and Klenner, 2008; Dragut et al., 2010). The first work in the direction of adjectival scale was done by Hatzivassiloglou and McKeown (1993). They exploited linguistic knowledge available in the corpora to compute similarity between adjectives. However, their approach did not consider polarity orientation of adjectives, they provided ordering among non-polar adjectives like, *cold, lukewarm, warm, hot*.

[5]We use 80% of the star-rated movie review corpus as training data and 20% as test data. The results reported in table 3 are based on the 20% test data.

The task of ordering adjectives according to their polarity-intensity has recently received much attention due to the vital role of intensity analysis in several real world tasks. Kim et al. (2013) interpreted the continuous space word representation by demonstrating that vector off-set can be used to derive scalar relationship amongst adjectives. Their approach provided relationship among all the adjectives independent of their semantic property. De Melo and Bansal (2013) used a pattern based approach to identify intensity relation among adjectives, but their approach had a severe coverage problem. They also did not consider the semantic property of adjectives, assuming one single intensity-scale for all adjectives.

Ruppenhofer et al. (2014) provided ordering among polar adjectives that bear the same semantic property. Their approach was completely corpus dependent, it was not able to derive intensity of those adjectives which were not found in the corpus. We have used the same star-rated movie review corpus in our work as used by Ruppenhofer et al. (2014) and found 122 polar adjectives which are absent from the corpus. Our system is able to identify intensity levels for these missing adjectives. Moreover, we obtained an improvement of 4% in overall accuracy.

8 Conclusion

In this paper, we have proposed an approach that assigns intensity levels to domain independent adjectives, *viz. high, medium* and *low*. The important feature of our approach is that it is not fully corpus dependent, hence is able to assign intensity to adjectives that are absent in the corpus. We have reported that the overall results are better than the recently reported corpus based approach and fairly close to human agreement on this challenging task.

The use of adjectives with their intensity information can enrich existing sentiment analysis systems. We have shown the significance of considering intensity information of adjectives in predicting the intensity of movie reviews.

9 Acknowledgment

We heartily thank English linguists Rajita Shukla and Jaya Saraswati from CFILT Lab, IIT Bombay for giving their valuable contribution in gold standard data creation.

References

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics.

Adrian Boteanu and Sonia Chernova. 2013. Unsupervised rating prediction based on local and global semantic models. In *2013 AAAI Fall Symposium Series*.

Gerard De Melo and Mohit Bansal. 2013. Good, great, excellent: Global inference of semantic intensities. *Transactions of the Association for Computational Linguistics*, 1.

Eduard C Dragut, Clement Yu, Prasad Sistla, and Weiyi Meng. 2010. Construction of a sentimental word dictionary. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM.

Angela Fahrni and Manfred Klenner. 2008. Old wine or warm beer: Target-specific sentiment analysis of adjectives. In *Proc. of the Symposium on Affective Language in Human and Machine, AISB*.

Narendra Gupta, Giuseppe Di Fabbrizio, and Patrick Haffner. 2010. Capturing the stars: predicting ratings for service and product reviews. In *Proceedings of the NAACL HLT 2010 Workshop on Semantic Search*. Association for Computational Linguistics.

Vasileios Hatzivassiloglou and Kathleen R McKeown. 1993. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics.

Vasileios Hatzivassiloglou and Kathleen R McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics*. Association for Computational Linguistics.

Joo-Kyung Kim and Marie-Catherine de Marneffe. 2013. Deriving adjectival scales from continuous space word representations. In *EMNLP*.

Bing Liu. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2.

Tomas Mikolov, Anoop Deoras, Daniel Povey, Lukas Burget, and Jan Cernocky. 2011. Strategies for training large scale neural network language models. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises*.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics.

Lizhen Qu, Georgiana Ifrim, and Gerhard Weikum. 2010. The bag-of-opinions method for review rating prediction from sparse text patterns. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics.

Sven Rill, Jörg Scheidt, Johannes Drescher, Oliver Schütz, Dirk Reinel, and Florian Wogenstein. 2012. A generic approach to generate opinion lists of phrases for opinion mining applications. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*. ACM.

Josef Ruppenhofer, Michael Wiegand, and Jasper Brandes. 2014. Comparing methods for deriving intensity scores for adjectives. *EACL 2014*, 117.

Raksha Sharma and Pushpak Bhattacharyya. 2013. Detecting domain dedicated polar words. *IJCNP 2013*, pages 661–666.

Xiaodan Zhu Svetlana Kiritchenko and Saif M. Mohammad. 2014. Sentiment analysis of short informal texts. 50.

Maite Taboada and Jack Grieve. 2004. Analyzing appraisal automatically. In *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text (AAAI Technical Re# port SS# 04# 07), Stanford University, CA, pp. 158q161. AAAI Press*.

Peter D Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics.

Janyce Wiebe. 2000. Learning subjective adjectives from corpora. In *AAAI/IAAI*.

Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005a. Opinionfinder: A system for subjectivity analysis. In *Proceedings of hlt/emnlp on interactive demonstrations*. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005b. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics.