Feature Extraction Methods for

Semantic Orientation Based Approaches to Sentiment Analysis

Basant Agarwal
Department of Comp. Engg.
Malaviya National Institute of Technology, Jaipur
thebasant@gmail.com

Namita Mittal
Department of Comp. Engg.
Malaviya National Institute of Technology, Jaipur
nmittal@mnit.ac.in

Vijay Kumar Sharma
Department of Comp. Engg.
Malaviya National Institute of Technology, Jaipur
sharmavijaykumar55@gmail.com

Abstract

Sentiment analysis means to extract opinion expressed in natural language text about a specific topic. Sentiment analysis research has increased tremendously in recent time due to its huge business applications and importance. In this paper, various sentiment-rich features are extracted using part-of-speech (POS) patterns and dependency rules. Next, more composite, hybrid and split features are created from the basic POS rule based features and dependency features. Further, semantic orientations of all these features are computed using both supervised and unsupervised methods. Finally, semantic orientations of all the features are aggregated to determine the polarity of the document. All the experiments are performed on standard movie review and book review datasets. Experimental results show the effectiveness of the proposed split features that performs better than other features.

1 Introduction

With the tremendously increasing number of Web applications, people are using blogs, discussion boards, reviews, social networking websites to express their opinion about a specific topic (Cambria et al. 2013). The explosion of online contents has increased the demand of analysing online contents, in order to know what people think about a specific topic. Companies may use this information for improving their products and customers may find this information useful in taking purchasing decisions.

To recognize the polarity of a given text, polarity words like “good”, “bad” “excellent” etc. are key indicators for creating a machine learning model for sentiment classification. Other lexicon based approaches aggregate the semantic orientation values of these polarity words to determine the overall sentiment of a document. Sometimes, however, these individual polarity words are incapable of incorporating actual sentiment of the text. Individual words can have different polarity for different domains. For example, “unpredictable” word may have a negative polarity in auto-mobile review, with phrase “unpredictable steering”, but it could have positive polarity for movie review with the phrase “unpredictable story” (Turney 2002). Contextual and syntactic information is important for sentiment analysis. Two-word features extracted by POS patterns and dependency relations can incorporate important information for sentiment analysis. However, effectiveness of these two-word features is limited due to limited coverage (Cambria et al. 2014). In this paper, new feature extraction methods are investigated that increases the performance of sentiment analysis model by increasing the coverage. New composite, hybrid and split features are experimented in addition to unigrams, POS patterns and dependency features for semantic orientation based methods for sentiment analysis. Supervised and unsupervised settings are used to determine the semantic orientation of the features extracted.

This paper is organized as follows. Section 2 presents the related work. Various sentiment-rich feature extraction methods are discussed in Section 3. Section 4 presents the method for calculation of Semantic Orientation of the features. Section 5 discusses the method for aggregating semantic orientation values of the features. Further, Experimental setup and results

Proceedings of ICON-2013: 10th International Conference on Natural Language Processing, pp: 225-230, 2013
Noida, India, 2013

are discussed in Section 6. Finally, Section 7 presents the conclusion.

2 Related Work

Sentiment analysis research can be broadly categorized into machine learning based approaches (Agarwal et al. 2013a), semantic orientation based approaches (Agarwal et al. 2013b) and knowledge based approaches (Cambria et al. 2011). This paper focuses on the semantic orientation based approaches for sentiment analysis. Initial work for identifying the semantic orientation of words is done by Hatzivassiloglou et al. (1997). They developed a supervised learning method for calculating the semantic orientation of adjectives. Turney (2002) proposed an unsupervised method for detecting the polarity of a movie review document. Initially, they extracted two-word phrases using fixed POS based patterns, then semantic orientation of those phrases are computed using Point-wise Mutual Information (PMI) method. Finally, overall polarity of the document is recognized by aggregating the semantic orientation of all the phrases. Fei et al. (2004) constructed phrase patterns with adjectives, adverbs, prepositions, conjunctions, noun, and verbs. Further, semantic orientations of these phrases are computed using unsupervised method. Dependency tree of a sentence produces syntactic relation among words in the sentence. Several researchers have investigated the importance of these syntactic relations for sentiment analysis (Nakagawa et al. 2010). Thet et al. (2007) generated dependency tree of a sentence and split the sentence into clauses. Further, contextual sentiment score of each clause is determined for further detection of sentiment of the document.

3 Feature Engineering

Sentiment orientation based approaches for sentiment analysis works in three phases. First of all sentiment-rich features are extracted. Further, semantic orientations of these sentiment-rich features are computed and finally overall semantic orientation of the document is determined by aggregating the semantic orientations of all the features in the document. Various types of sentiment-rich features are extracted in this paper as discussed in the subsequent subsections.

3.1 Unigrams

Semantic orientation based approaches relies on the sentiment-rich words like adjectives, adverbs (Turney, 2002). Such words are used generally to express sentiments in text (Hatzivassiloglou et al. 1997). For example, “*this_DT was_VB a_DT great_JJ movie_NN*”, here word “*great*” is an adjective and shows positive sentiment. Other words like “*this”, “was”, “a”, “movie”* are not conveying any sentiment in the text. All the unique words in the corpus are considered as features if they conform to a specific POS tag i.e. adjective, Adverb, Noun, and Verb.

3.2 POS Pattern based feature

Phrases are very useful for extraction of syntactic, contextual information which is very important for sentiment analysis.

S.no.	First Word	Second Word
1	JJ	NN/NNS
2	RB/RBR/RBS	JJ
3	JJ	JJ
4	NN/NNS	JJ
5	RB/RBR/ RBS	VB/VBD/VBG
6	VB/VBG/VBD	NN/NNS
7	VB/VBG/VBD	JJ/JJR/JJS
8	JJ	VB/VBD/VBG
9	RB/RBR/RBS	RB/RBR/RBS

Table 1.POS patterns

For example, attaching an adverb like “very” with a polarity adjective “good” will increase the intensity of the word “good”. This information may be useful for sentiment classification. In addition, phrases are capable of capturing contextual information like “not good”, “unpredictable story”, “amazing movie” etc. Therefore, two-word phrases are extracted that conform to the predefined pattern. These POS pattern are given in Table 1.

3.3 Dependency Features

A deeper linguistic analysis of syntactic relations may be important for sentiment analysis. Several researchers have used syntactic patterns for sentiment analysis. Dependency tree of a sentence produces syntactic relation information from the text. Wiebe et al. (2005) investigated that syntactic patterns are very effective for subjective detection which is a prior step to sentiment classification. Table 2 presents the dependency relations which are used to extract the sentiment-rich dependency features from the text.

S.No.	Relation	Meaning	Example
1	Acomp	adjectival complement	(look, good)
2	Advmod	adverbial complement	(cool, pretty)
3	Amod	adjectival modifier	(performance, poor)
4	Dobj	direct object	(appreciated, actor)
5	Neg	negation modifier	(happy, not)
6	Nsubj	nominal subject	(good, actors)
7	Rcmod	relative clause modifier	(film, exhilarate)
8	Xcomp	open clause complement	(bored, watching)
9	Cop	Copula	(beautiful, is)
10	Ccomp	clausal complement	(happens, bored)

Table 2. Dependency relations

3.4 Composite Features

New composite features are created by combining POS based and dependency relation based features as discussed in previous subsection. Phrases extracted using POS based fixed patterns are not enough in extracting all the sentiment-rich phrases. Also, POS based phrases can incorporate contextual information but are not efficient in extracting syntactic information unlike from dependency features which is also important for sentiment analysis. For example, "*This movie is very impressive and effective*." POS tagged sentence is as follows "*This_DT Movie_NN is_VBZ very_RB impressive_JJ and_CC effective_JJ*". Phrase "*very impressive*" would be extracted from this sentence using POS pattern based feature extraction method. However, there are more sentiment-rich phrases which may be useful for the sentiment analysis that can be extracted using dependency features. Phrases extracted using dependency relations are as follows. *nsubj(impressive, movie*), *nsubj(effective, movie)*, *cop(impressive, is)*, *advmod(impressive, very)*, *advmod(effective, very)*. By combining POS based and dependency based phrases would incorporate contextual and syntactic information from the text. Therefore, composite features take advantage of both types of features for sentiment analysis.

3.5 Hybrid Features

New hybrid feature set is created by considering composite features and unigram features. In hybrid features, sentiment information from both unigrams and two-word features are taken into consideration for detection of overall semantic orientation of the document (Bakliwal et al., 2011). Contribution of unigrams and two-word features is determined empirically in computation of overall semantic orientation; it shows that two-word features are more important as compared to unigrams for sentiment analysis as it contains more sentiment information. Semantic score of a document is computed using Eq. (1)

$$SO(doc.)=(3/4)*two-word\_features+(1/4)*unigram \qquad ....(1)$$

3.6 Split Features

Main problem with two-word features (phrase) is the coverage. Two-word features matches very infrequently within other documents, sometimes none of the features of testing document is previously seen in training document, that document is difficult to classify in the positive or negative polarity document. Therefore, split features are investigated to increase the coverage of two-word features. In this method, to derive the semantic orientation of the overall two-word feature, it is split into two unigrams and semantic orientation of each word is aggregated on various combination methods as given in Table 3.

Combinations	I	II	I	II	I	II	I	II
	+	+	-	-	+	-	-	+
1	Avg		Avg		Avg		Avg	
2	Avg		Avg		Avg		Max	
3	Avg		Avg		Max		Max	
4	Avg		Max		Max		Max	
5	Max		Max		Max		Max	
6	Max		Max		Max		Avg	
7	Max		Max		Avg		Avg	
8	Max		Avg		Avg		Avg	
9	Avg		Avg		Max		Avg	
10	Avg		Max		Avg		Avg	
11	Avg		Max		Max		Avg	
12	Max		Max		Avg		Max	
13	Max		Avg		Max		Max	
14	Max		Avg		Avg		Max	

Table 3 Possible combination for splitting two-word features

The process to extract split features is as follows. Initially, phrases are extracted using POS based patterns and dependency relation, then, if semantic orientation of extracted phrase is not available already (that two-word feature doesn't occur in training document), in that case, that feature is divided into two unigrams, further, semantic score of each word would be combined by various combination of average and maximum functions as given in Table 3. For example combination 1 case I in Table 3, if 1st word of the two-word feature is having positive semantic orientation and 2nd word is having positive semantic orientation, than average semantic orientation of both the unigrams is taken as semantic orientation of that two-word feature.

4 Semantic Orientation

After extraction of various sentiment-rich features, semantic orientation of each feature is computed by two methods. (1) Supervised Method (2) Unsupervised Method

4.1 Supervised Method

Computation of semantic orientation of the feature is based on the assumption that if a feature is occurring frequently and predominantly in one class (positive or negative), then that feature would have high polarity. If a feature has high positive polarity value that indicates that feature has occurred mostly in positive documents. Point-wise Mutual Information (PMI) is generally used to calculate the strength of association between a feature and positive or negative documents. It is defined as follows (Kaji, 2007).

$$PMI\ (c,pos) = \ log_2 \frac{P(c,pos)}{P(c)P(pos)} \qquad \text{....(2)}$$

$$PMI\ (c,neg) = \ log_2 \frac{P(c,neg)}{P(c)P(neg)} \qquad \text{....(3)}$$

Here, *P(c,pos)* is probability of a feature that it occurs in positive documents i.e. frequency of the positive documents in which feature occurs divided by total number of positive documents. *P(c,neg)* is the probability that a feature occurs in negative document i.e. frequency of negative documents in which feature occurred divided by total number of negative documents. Polarity value of the feature is determined by their PMI value difference (Turney, 2002). Semantic orientation of a feature (p) is computed using Eq. (5).

$$SO(p) = \ PMI(c,pos) - PMI(c,neg) \qquad \dots (4)$$

$$SO(p) = \ log_2 \frac{P(c,pos)/P(pos)}{P(c,neg)/P(neg)} \qquad \dots (5)$$

4.2 Unsupervised Method

Drawback with supervised method is the requirement of large labelled training dataset, since large labelled training dataset is very difficult to obtain for every domain. Hence, unsupervised methods are important to investigate. Therefore, unsupervised method is investigated to compute the semantic orientation of the feature.

	SEED WORDS
Positive	fantastic, satisfying, mood, superb, rare, terrific, memorable, realistic, natural, excellent, brilliant, hilarious, incredible, effectives, powerful, amazing, wonderful, strong, surprisingly
Negative	waste, boring, worst, stupid, mess, awful, ridiculous, lame, unfunny, tedious, ludicrous, terrible, bore, blame, guilty, laughable, dull, dumb, poor, painful, embarrass, insult, lousy, fake

Table 4. Selected seed words for movie reviews

Semantic orientation of words and phrases are computed using the manually created positive and negative seed word list. List of positive and negative seed words are given for movie review dataset in Table 4. Basic intuition behind this method is almost same as supervised method that if a feature occurs frequently with positive seed words and also does not occur frequently with negative seed words then that feature would have high positive polarity value. Semantic orientation of a feature is computed using Eq (6).

$$SO(C) = log_2 \frac{p(c, pos\_seed\_word)}{p(c, neg\_seed\_word)} \qquad .....(6)$$

Here, p(c, pos_seed_word) is the probability of feature occurring with positive seed words and *p(c, neg_seed_word)* is the probability of feature occurring with negative seed words in unlabelled document corpus.

5 Semantic Orientation Aggregation

After computation of Semantic Orientation (SO) of all the features of the training documents, a lexicon of various features with their SO values is developed. Further, for the testing document, initially features are extracted and then SO values of these features are retrieved from the developed lexicon. Finally, summing up the semantic orientations of all the features from the document would give the overall SO of the document. If the overall SO is positive, the document is labeled as positive-polarity document else it is labeled as negative-polarity document.

6 Experimental Setup and Discussion

To evaluate the effectiveness of the proposed methods for sentiment analysis, two publically available standard datasets are used. First dataset is Movie review dataset also known as Cornell's Dataset (Pang & Lee, 2004). Another dataset is book review dataset, provided by amazon product reviews (Blitzer et al. 2007). Both the datasets contains 2000 movie reviews consisting of 1000 positive and 1000 negative reviews.

For all the experiments, initially training dataset is created by randomly selecting 700 positive and 700 negative documents. Then, Remaining 300 positive and 300 negative documents are used for the testing of the proposed approach. Accuracy is used as a performance evaluation measure which is calculated by dividing total testing documents to the correctly classified document.

6.1 Results and discussions

Accuracies for various features using supervised and unsupervised methods for movie review and book review datasets are reported in Table 5. First of all, POS based unigrams features are experimented with supervised and unsupervised methods. Accuracy with unigram features is considered to be baseline accuracy. In all the experiments, supervised methods intuitively produce better results as compared to their respective unsupervised method. For example, unigram features gives an accuracy of 75.33% and 72.33% respectively for supervised and unsupervised methods using movie review dataset. However, unsupervised method may be used with effective features for such a domain in which labeled dataset is a problem.

Further, POS based patterns are considered as features for sentiment analysis with intuition that these features carry more sentiment information as compared to simple unigrams. These features improve the performance over unigrams. For example, performance increases up to 81.50% (+8.19%) with supervised method for movie review domain. Further, dependency feature are explored as these features can extract more syntactic information as compared to the simple POS pattern based features. These features produces better results as compared to unigrams and POS pattern based features due to more coverage and incorporation of syntactic information. For example, dependency features give an accuracy of 82.50% (+9.51%) using supervised method for movie review dataset. Further, performances of composite features are investigated, which gives an accuracy of 88.83% (+17.9%) for movie review dataset. These features improve the performance of sentiment analysis for both supervised and unsupervised methods. It is due to the fact that these features provide more coverage of the features unlike previous experimented features.

Features		Accuracy (In %)			
		Movie review		Book review	
		Unsusupervised (In %)	Supervised	Unsusupervised	Supervised
Unigrams		72.33	75.33	76.67	77.17
POS pattern based features		71.6	81.50	74.33	79.00
Dependency features		75.50	82.50	78.17	84.33
Composite features		79.67	88.83	82.17	88.33
Hybrid features		80.17	90.17	83.50	89.67
Splitting features	POS patterns	76.83	85.83	78.83	86.17
	Composite features	81.50	91.67	84.17	90.17

Table 5 Accuracies (In %) for various features with movie review and book review datasets

Next, hybrid features are experimented for sentiment classification; it incorporates the information of both unigrams and two word features. For example, these features give an accuracy of 90.17% for movie review dataset with supervised learning method.

Combinations	Correctly classified positive docs	Correctly classified negative docs	Correctly classified	Accuracy (In %)
1	278	207	485	81.83
2	278	217	495	83.50
3	274	218	492	83.00
4	**270**	**233**	**503**	**85.83**
5	278	205	483	81.50
6	279	199	478	80.67
7	280	194	474	80.00
8	283	181	464	78.33
9	275	210	485	80.83
10	274	222	496	82.67
11	270	228	498	83.00
12	280	203	483	81.50
13	281	195	476	80.33
14	284	187	471	79.50

Table 6 Accuracy for all the possible combination for splitting features with POS pattern based features using movie review dataset

In further experiments, split features are investigated for sentiment analysis. In these features all the possible methods of splitting two-word features are empirically experimented and due to limitation of size, results for POS pattern based features with movie review dataset are reported in Table 6. It is observed from the experiments that fourth combination produces best results. Therefore, finally, results for splitting all the features with this combination are reported in Table 5. Split features produce the best results with composite features among all the features. For example, it produces the accuracy of 91.67% (+21.69%) for with supervised setting using movie review dataset. The main possible reason is the increased coverage and incorporation of syntactic and contextual information.

7 Conclusion

Sentiment analysis model depends on the efficient feature extraction methods for better classification results. In this paper, various sentiment-rich features are extracted like unigrams, POS based pattern based features and dependency relations based features. POS pattern and dependency relation based features are important for extracting contextual and syntactic information which is very useful for sentiment analysis. However, effectiveness of these two-word features is limited due to coverage; this paper proposes methods to improve the performance of sentiment analysis by increasing the coverage. Further, for determination of semantic orientation of the features both supervised and unsupervised methods are investigated. Experimental results show that proposed split features performs better than other features for sentiment analysis due to increased coverage. Supervised methods performs better than unsupervised methods, however, with new proposed split features by increasing the coverage unsupervised methods can also give well performance and may be very useful for the domains in which labeled training dataset is a problem. In future, more methods for feature extraction may be explored that can incorporate semantic information from the text.

References

Agarwal B., Mittal N.,(2013a) "Optimal Feature Selection for Sentiment Analysis", In CICLing 2013. Vol-7817, pages-13-24.

Agarwal B., Sharma VK, Mittal N.,(2013b) "Sentiment Classification of Review Documents using Phrases Patterns", In the International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp.1577-1580.

Cambria E., Song Y., Wang H., Hussain A. (2011). "Isanette: A common and common sense knowledge base for opinion mining", Proceedings of IEEE ICDM, pp. 315-322.

Cambria E., Schuller B., Xia Y., Havasi C. (2013). "New Avenues in Opinion Mining and Sentiment Analysis", IEEE Intelligent Systems 28(2), pp 15-21.

Cambria E., White, B. (2014). "Jumping NLP Curves: A Review of Natural Language Processing Research", IEEE Computational Intelligence Magazine 9(2)

Pang B., Lee L. 2004. "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts" In ACL, pp. 271–278.

Wiebe J., Riloff. E. 2005."Creating subjective and objective sentence classifiers from unannotated texts". In CICLing 2005, pp 486-497.

Nakagawa T., Inui K., Kurohashi S. 2010. "Dependency Treebased Sentiment Classification using CRFs with Hidden Variables", In Human Language Technologies: Annual Conference of the North American Chapter of the ACL, pp:786–794.

Poria S., Gelbukh A., Hussain A., Howard N., Das D., Bandyopadhyay S., (2013) "Enhanced SenticNet with Affective Labels for Concept-based Opinion mining." In IEEE Intelligent Systems, vol 28, no 2, pp 31-38.

Hatzivassiloglou V. McKeown K. R., "Predicting the seman-tic orientation of adjectives", ACL, pp. 174-181. 1997

Blitzer J., Dredze M., Pereira F. 2007. "Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification", Proc. Assoc. Computational Linguistics., pp 440-447.

Thet TT, Na JC., Khoo CSG.,Shakthikumar S. 2009. "Sentiment Analysis of Movie Reviews on Discussion Boards using a Linguistic Approach", In the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion, pp. 81-84.

Fei Z., Liu J., Wu G. 2004. "Sentiment classification using phrase pattern", In Proceedings of the Fourth Inter-national Conference on Computer and Information Technology (CIT'04), pp. 1147-1152 .

Turney PD. 2002. "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews", ACL, pp.417-424.

Kaji N., Kitsuregawa M. 2007. "Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents" ,ACL., PP. 1075-1083.

Bakliwal A., Arora P., Patil A., Verma V. 2011. "Towards enhanced opinion classification using NLP techniques", IJCNLP, pp. 101-107.