Feature Selection using Multiobjective Optimization for Aspect based Sentiment Analysis

Md Shad Akhtar[1], Sarah Kohail[2], Amit Kumar[1], Asif Ekbal[1] and Chris Biemann[2]

[1] IIT Patna, India; {shad.pcs15,amit.mtmc14,asif})@iitp.ac.in
[2] Universität Hamburg, Germany; {kohail, biemann}@informatik.uni-hamburg.de

Abstract. In this paper, we propose a system for aspect-based sentiment analysis (ABSA) by incorporating the concepts of multi-objective optimization (MOO), distributional thesaurus (DT) and unsupervised lexical induction. The task can be thought of as a sequence of processes such as aspect term extraction, opinion target expression identification and sentiment classification. We use MOO for selecting the most relevant features, and demonstrate that classification with the resulting feature set can improve classification accuracy on many datasets. As base learning algorithms we make use of Support Vector Machines (SVM) for sentiment classification and Conditional Random Fields (CRF) for aspect term and opinion target expression extraction tasks. Distributional thesaurus and unsupervised DT prove to be effective with enhanced performance. Experiments on benchmark setups of SemEval-2014 and SemEval-2016 shared tasks show that we achieve the state of the art on aspect-based sentiment analysis for several languages.

Keywords: Sentiment Analysis, Aspect based Sentiment Analysis, MOO

1 Introduction

The number of internet users in social media platforms has increased exponentially during the last few years and so does the amount of user written reviews about a product or service. Feedback available through reviews is useful to manufactures for upgrading or enhancing the quality of their products as well as for the users to take informed decisions. However, due to a large number of user reviews, it is quite in-feasible to scroll through all the reviews.

Sentiment analysis is the area of study that target to identify the sentiments (positive, negative or neutral) of the users based on the opinions and emotions expressed in the reviews written either for a particular product or service or any of its aspects (or feature/attribute). Classification of sentiment at document or sentence level may not always satisfy the need of user's requirement. They might need more precise information *i.e.* sentiment related to a particular aspect (or feature) of any product or service. Aspect-level analysis [15] is concerned with finding sentiment on fine-grained levels: aspects are features of the product or service that has been discussed in any user review. A common benchmark set up for ABSA was introduced in SemEval 2014 shared task 4 [24], and then subsequently extended in SemEval 2016 shared task 5 [23].

SemEval-2014 ABSA: The two tasks of SemEval-2014 of our interest in the work are (a) *aspect term extraction* and (b) *sentiment with respect to aspect terms*. The first task deals with the identification of all the aspect terms present in the review, while the second task predicts sentiment polarity for aspects. For example, in the review below there are two aspect terms (or opinion target expression, OTE), *Pasta* and *waiter*. Sentiment towards these two aspect terms are contrasting in nature. For the first aspect term (*Pasta*) it has positive sentiment while the second aspect term (*waiter*) conveys negative sentiment.

*"**Pasta** was good but the **waiter** was rude."*

SemEval-2016 ABSA: In 2016, the task was modified and extended to three subtasks i.e. (a) *aspect category detection*, (b) *opinion target expression (OTE) identification* and (c) *sentiment towards aspect category and OTE tuple*. Aspect category can be seen as the generalization of the aspect terms. The goal of aspect category detection task is to find the pre-defined set of *entity#attribute* pairs towards which the opinion is expressed in a review. The second task i.e. OTE is same as aspect term extraction of SemEval-2014. The sentiment classification task (third task) tries to predict the polarity for each aspect category and OTE tuple. As an example, in the following review category (i.e. *entity#attribute pair*) is *FOOD#QUALITY* while the OTE is *Food*. The opinion towards the <*entity#attribute, OTE*> tuple is negative.

*"**Food** was okay, nothing great."*

In our current work, we target to solve both the tasks of SemEval-2014 and the last two tasks (i.e. OTE and sentiment classification) of SemEval-2016 for ABSA.

Literature shows that most of the existing works in sentiment analysis focused primarily on document [27] and sentence [12] level. Some of the earlier approaches for aspect term extraction are based on frequently used noun and noun phrases [25, 4, 11]. In [11], the authors proposed the method which identifies frequently present noun phrases from the text based on association rule mining. This type of approach works well when frequently occurring terms are strongly co-related with certain types (e.g. noun), but many times fail when frequency of terms, which used as the aspects are very low. Supervised machine learning techniques [31, 21] are being widely used with the emergence of various labeled datasets. Some other techniques for extracting aspect terms include manually specified subset of the Wikipedia category [9] hierarchy, semantically motivated technique [27] and unsupervised clustering technique [25]. Phrase dependency tree [29] is also helpful in aspect term extraction. Recently, a detailed survey on the aspect based sentiment analysis has been presented in [26].

The performance of any classifier is influenced by the feature set used to represent the train and test dataset. Feature selection [16, 17] is the technique of automatically selecting a subset of relevant features for a classifier. By removing the irrelevant and redundant set of features, we can construct a classifier with reduced computational costs. In the present work we pose the problem of feature selection in machine learning model with respect to optimization framework. In particular we use evolutionary optimization based algorithms for finding the most optimized features set. Some of the prior works that made use of evolutionary optimization techniques can be found in [7, 8], in which the authors focused on named entity recognition task in multiple languages.

One of the novel contributions of the proposed technique is to study the effectiveness of unsupervised pre-processing steps to the target task. The major problems in applying machine learning algorithms for solving different problems is the non-availability of large annotated corpus, which results in issues with vocabulary coverage.

We explore possibilities arising from the use of unsupervised part-of-speech (PoS) induction [1] and lexical expansion [19]. Unsupervised PoS induction [1] is a technique that induces lexical-syntactic categories through the statistical analysis of large, raw text corpora. As compared to linguistically motivated PoS-tags, the categories are usually more fine-grained, i.e. the linguistic class of nouns is split into several induced classes that also carry semantic information, such as days of the week, professions, mass nouns etc. As shown in [2], these induced categories as features results in improved accuracies for a variety of NLP tasks. Since the induction of PoS is entirely language independent and sensitive to domain vocabulary as shown in [1], we expect further improvements when combining these features with MOO.

While unsupervised PoS tagging primarily targets syntactic categories, we utilize lexical expansion [19] to explore semantic characteristics. Lexical expansion is also an unsupervised technique that needs a large corpus for the induction, and is based on the computation of a distributional thesaurus (DT) [14].

2 Method

In this section at first, we briefly introduce multi-objective optimization (MOO), and then discuss the approach of feature selection for the tasks. Finally, we discuss the feature sets which we use for the different tasks.

2.1 Brief overview of MOO

In our daily life we have to face the situations where we have to deal with more than one objectives. MOO focuses on optimization of more than one objectives simultaneously in contrast to single objective optimization where algorithm focuses on only one objective. Many decision problems have been solved using the concept of MOO. Mathematically, a general MOO [5] can be defined as follows :

$$\begin{aligned} \text{Minimize/Maximize } & f_m(x), m = 1, ..., M \\ \text{subject to } & g_j(x) \geq 0, j = 1, ..., J; \\ & h_k(x) = 0, k = 1, ..., K; \\ & x_i^L \leq x_i \leq x_i^U, i = 1, ..., N \end{aligned}$$

The solution of above general equation is x, a vector $(x_1, x_2...., x_N)^T$ of size N, where $x_i; i = 1, ..., N$ represents the decision variables. The above specified optimization problem has J number of inequality constraints and K number of equality constraints. $g_i(x)$ and $h_k(x)$ represent the constraint functions.

2.2 Non-dominated sorting genetic algorithm-II (NSGA-II)

We use the non-dominated sorting genetic algorithm (NSGA-II) [5] as the optimization technique to develop our models. This algorithm uses the search capability of GA and tries to minimize the fitness functions (*i.e.* objective functions). It starts with creating the parent population P_0 of size N. Each of the candidates in the population is called chromosome. For each of the chromosome, the fitness function is computed. By applying binary tournament selection, crossover and mutation operators on parent population P_0, a child population Q_0 of size N is created. In t^{th} iteration, a combined population of parent and child population R_t = P_t + Q_t is formed. All the candidates of R_t are sorted according to non-dominated sorting algorithm thus producing non-dominated sets F_1, F_2, .. F_n with decreasing fitness values. We then select N chromosomes from these sets. In case of selecting a subset of chromosomes of equal fitness value we apply crowding distance operator as in NSGA-II [5] to break the deadlock. Crowding distance prefers chromosomes that lies in the less dense regions. Above sequence of steps runs until either the specified number of generation or the stopping criteria is met.

We use NSGA-II due its following positive points: (i) Low computation cost $O(MN^2)$, where no. of objective functions = M and Population size = N; (ii) High elitism property, where in each iteration best individuals from the parent and child populations are preserved; (iii) Diversity in population without having to specify any parameter manually; and (iv) Easy to use.

2.3 Problem formulation

Performance of any classifier highly depends on the feature set what we use. Generally, we use heuristics based approach to select a subset that optimized the fitness function. This process is very time consuming. On the other hand, an automated method of feature selection might be able to identify the most relevant features.

For a given set of features F and classification quality measures, we aim to determine the feature subset f of F *i.e.* $f \subseteq F$, which optimize all of the fitness functions. In our case we use precision, recall, F1-measure, accuracy and number of features as the objective functions. For each of the tasks we build two frameworks as follows:

- **Framework 1**: In this framework we optimize two objective functions i.e. *number of features* (minimize) and *F-measure* (maximize) for aspect term extraction or OTE. For sentiment classification objective function are *number of features* and *accuracy*.
- **Framework 2**: Here for aspect term and OTE extraction tasks, we use two objective functions: *precision* and *recall*. For sentiment classification, we optimize *accuracy of each class* as the fitness function. We maximize all of the fitness values.

2.4 Problem encoding

If total number of features are N, then the length of chromosome will be N. All bits are randomly initialized to 0 or 1 and each represents one features. If the i^{th} bit of a chromosome is 1 then the i^{th} feature participates in constructing the framework otherwise not. Each chromosomes in the population (P) are initialized in the same way.

2.5 Fitness computation

To determine the fitness value of a chromosome, following procedures are executed. Let F is the number of 1's present in the chromosome. We develop the model by training CRF for aspect and opinion term expression extraction and SVM for sentiment classification on selected features i.e. F. The motive is to optimize the values of objective functions using the search capability of NSGA-II. To evaluate the objective functions we perform 5-fold cross validation.

2.6 Features

We identify and implement a variety of features for aspect term and opinion target expression (OTE) extraction tasks. The set of features that we use in this work along with their descriptions are presented in Table 2. For sentiment classification, Table 1 lists set of features we use for the datasets of SemEval-2014 shared task. For SemEval-2016 shared task we use unigram, bigram and expansion scores based on the prior works reported in [13] as features for English. For Dutch language we use unigram, bigram and expansion score as features.

Feature	**Description**
	Sentiment Classification
Aspect term and context:	We convert the actual forms of aspect terms in lower case character and use it as a feature along with the actual aspect terms. The polarity orientation of aspect term heavily depends on local context words where it appears. We include succeeding five and preceding five terms of aspect term to provide contextual information.
Lexicon:	Sentiment lexicons are the useful resources, which provide important information for predicting the sentiment. For computing lexicon sentiment score we consider the preceding five and following five tokens of the aspect term. We use following set of lexicons. – **MPQA:** We take the help of MPQA subjectivity lexicon [28] which contains a list of words denoting the negative, positive and neutral sentiments. – **Bing Liu lexicon:** For each token in training and test set we define the values in the following way: -1 for negative; 1 for positive and 2 for those do not appear in Bing Liu lexicons [6]. Then, we define two features: 1. We calculate sum of sentiment score of all the words that appear in context of target aspect term and use as a feature. 2. We also compute the sum of the sentiment scores of only those words which have *direct dependency relation* with the target aspect term. – **SentiWordNet lexicon:** This is one of the most widely used lexicons for sentiment analysis. We compute sentiment score of all words that appear in surrounding context (previous-5 and next-5) of the target aspect term. – **Other lexicons:** Apart from above mentioned lexicons we also use AFINN [22], NRC Hashtag, Sentiment 140 [30] and NRC Emotion [20] lexicons for calculating the score and use them as features.
Domain-Specific Words:	We hand-made a list of words from general intuition and web that describes domain-specific information. For e.g. *yummy*, *over cooked* etc. are some of the sentiment bearing words for restaurant domain. We assign score 1, -1 and 2 to each positive, negative and words that are missing in the list respectively. We compute the feature value based on local context [-5..5] of the aspect term.

Table 1. Feature set for sentiment classification for SemEval-2014 dataset.

Feature	Description
Aspect Term Extraction	
Word, local context & PoS	Word and local context play a significant role in determining the aspect term and opinion target expression. We use the current token, its lower case form and local context [-5..5] as features. Part-of-speech (PoS) information is useful to capture the syntactic property, thus we use PoS information of current and context tokens [-2..2] as features.
Head word and PoS	Generally, aspect terms belong to the category of noun phrase. The head word of the noun phrase along with its PoS tags are used as the feature.
Prefix and suffix	Suffix and prefix of fixed length character sequences are trimmed from each token and used as the features for our model. Here, we use prefix and suffix of current and context tokens [-1,0,1] as the features.
Frequent aspect term	We generate a list of frequently (more than 4) occurring OTEs from the training set. We define a binary feature for the presence or absence of extracted OTEs.
Dependency Relation	Here we define two features: i. when the current token is governor via relations nsubj, amod or dep and ii. when the current token is dependent via relations nsubj, dobj or dep. For example in the review below food is the dependent on *lousy* via relation nsubj. However, *food* is a governor of *the* via relation det. Therefore, for the token *food*, we use the relation 'nsubj' as the first feature and *null* for the second. *The **food** was lousy.*
Character n-grams	Character n-gram is a contiguous sequence of n character extracted from a given token. We extract character bigram, trigram and four-gram of the current token and use them as features in our model.
Orthographic feature	Many of the aspect terms start with the capital letter. We define a feature which checks whether the current token starts with the capitalized letter or not.
Semantic Orientation (SO) Score:	*SO* score [10] is the measurement of negative or positive sentiment expressed in a phrase. *SO* score of each token is computed with Point wise mutual information (PMI) as follows: $$SO(w) = PMI(w, prev) - PMI(w, nrev)$$ Here PMI is measurement of association of token w with respect to negative $nrev$ or positive $prev$ reviews.
DT features	Distributional thesaurus (DT) gives the lexicon expansion of the token based on similar context. In [3], the authors have used it for lexical expansion of text by virtually expanding every content word in the text with the list of most similar words from the DT. It is very helpful in unseen texts. We obtain top 5 DTs of current token and top 3 DTs of context tokens [-2,-1,1,2] as the features.
Expansion score	OTEs and aspect terms have opinions around them. Opinions are regularly lexicalized with words found in sentiment lexicons. We compute sentiment score based on induced lexicons as computed in [13] by considering the window size of 10 (preceding 5 and following 5 of the current token). We use expansion score of context tokens [-2..2] as the features in our model.
Unsupervised PoS tag	In [1], the authors implement a system which takes a reasonable amount of tokenized and unlabeled text without PoS information mentioned as input and induces number of word clusters. We use the tag of context tokens [-2..2] as the features in our system.
We additionally extract the following set of features only for English language.	
Chunk information	A text can have multi-word aspect term or OTE. To identify the boundaries of these multi-words, we use chunk information of context tokens[-1,0,1] as features.
Lemma	Lemmatization trims the inflectional forms and derivationally related forms of a token to a common base form. We use lemma of the current token as the feature.
WordNet	Tokens from the same lexical category are grouped into synsets in WordNet [18]. We extract top four noun synsets of each token and use as feature. This feature can be helpful in clustering tokens with identical sense, thus assist the system in finding the unseen aspect terms more accurately. For example, senses *lunch* and *dinner* are related to the sense *meal* in the WordNet hierarchy. So for the scenario where aspect term *lunch* appears in the test set but was unseen in the train set, this feature will guide the system to identify it as an aspect term more accurately.
Named entity information	We extract named entity information of the current token with Stanford CoreNLP tool, and use the NER sequence labels as features.

Table 2. Feature set for aspect term and OTE extraction

3 Experiments and Analysis

3.1 Datasets

For experiments we use the benchmark datasets of SemEval-2014 [24] and SemEval-2016 [23] shared tasks on ABSA. SemEval-2014 datasets belongs to English only cov-

ering laptop and restaurant domains while SemEval-2016 datasets contains reviews from different languages i.e. English, Spanish, Dutch and French.

3.2 Results and Analysis

We perform various experiments with different feature combinations for all the languages and domains. We divide our feature sets in two category: Standard features (Std.Fea) and Non-standard features (i.e. DT based features). Standard features correspond to the features as mentioned above *except* DT, Unsupervised POS tag and expansion score in case of aspect term and OTE extraction. For sentiment classification standard features denote the features mentioned above *except* the expansion scores. Results of aspect term and OTE extraction are reported in Table 3a. This shows how DT based features help in improving the performance. Feature selection based on MOO aids in achieving better performance. This shows how effectively MOO selects the most relevant sets of features for each domain and language. The system attains better performance with a smaller set of features compared to the scenario where the exhaustive feature set is used. Results for sentiment classification on SemEval 2014 and SemEval 2016 datasets are reported in Table 3b and 3c, respectively. We perform several experiments and observe that by adding the expansion score, we do not get much increment in English, but for Dutch this feature is very effective. It is also observed that after applying MOO, performance of the system improves significantly. With a much smaller number of features we obtain the increments of 4.76% and 3.66% in restaurant (38 vs. 20) and laptop (38 vs. 12) domain, respectively. For SemEval-2016 data, we perform sentiment classification for English and Dutch. Since the number of features for this task is too little, we do not apply MOO.

3.3 Comparisons

As explained in previous section, we perform experiments on SemEval-2014 and SemEval-2016 datasets. We compare the performance of our proposed systems with those submitted in the shared tasks. Results are shown in Table 4. In non-English languages we are at first position among all the teams. In these languages for aspect terms extraction, we are 3.18%, 4.87% and 13.24% ahead in Spanish, French and Dutch respectively compared to the other top teams who participated in the challenge. The performance that we achieve for sentiment classification is also satisfactory.

3.4 Feature selection: Analysis

Performance of any classification problem fully depends on the features used for solving the problem. Here we show the set of features selected for each of the languages English, French, Spanish and Dutch. Results are reported in Table 5, Table 6 and Table 7 for aspect term extraction, OTE and sentiment classification respectively.

4 Error Analysis

In subsequent subsections, we present analysis of error encountered by the proposed method. Due to space constraints we only show analysis for SemEval-2016 datasets.

Datasets	Parameters	Feature sets			After MOO	
		Std.Fea	Std.Fea + DT	Std.Fea + DT + UnPos	M1	M2
Restaurant (SemEval-2014)	No.of Fea	68	83	88	43	-
	Precision	82.29	83.09	83.30	-	83.78
	Recall	75.39	76.27	76.10	-	76.98
	F1-mea	78.69	79.54	79.53	79.92	-
Laptop (SemEval-2014)	No.of Fea	69	84	89	24	-
	Precision	82.79	84.26	83.36	-	83.76
	Recall	62.53	62.23	62.84	-	64.67
	F1-mea	71.25	71.59	71.66	70.56	-
English (SemEval-2016)	No.of Fea	63	78	83	24	-
	Precision	74.40	75.60	75.54	-	75.84
	Recall	60.78	61.76	62.58	-	62.09
	F1-mea	66.90	67.98	68.45	68.21	-
French (SemEval-2016)	No.of Fea	47	62	67	11	-
	Precision	70.64	71.81	70.85	-	70.91
	Recall	67.38	68.61	68.46	-	66.76
	F1-mea	68.97	70.18	69.64	67.85	-
Spanish (SemEval-2016)	No.ofFea	47	62	67	26	-
	Precision	71.84	76.23	76.85	-	74.28
	Recall	66.19	62.55	63.81	-	69.28
	F1-mea	68.90	68.72	69.73	70.33	-
Dutch (SemEval-2016)	No.ofFea	47	62	67	15	-
	Precision	64.88	64.22	66.10	-	66.57
	Recall	61.93	61.12	62.73	-	62.46
	F1-mea	63.37	62.63	64.37	65.01	-

(a)

Datasets	Parameters	Feature sets		After MOO	
		Std.Fea	Std.Fea + Exp.score	M1	M2
Restaurant (SemEval-2014)	No.ofFea	37	38	20	-
	Positive	88.59	88.73	-	89.56
	Negative	54.59	54.08	-	60.20
	Neutral	34.69	32.65	-	39.79
	Conflict	14.28	14.28	-	0.00
	Overall	72.48	72.13	76.89	-
Laptop (SemEval-2014)	No.ofFea	37	38	12	-
	Positive	76.83	76.53	-	78.29
	Negative	64.84	64.84	-	70.31
	Neutral	32.54	31.95	-	40.82
	Conflict	06.25	06.25	-	12.50
	Overall	61.31	61.01	64.67	-

(b)

	English		Dutch	
	Std. Fea	Std.Fea + Exp.Score	Uni + Bi-gram	Uni + Bi + Exp.Score
Positive	88.70	88.54	82.92	83.73
Negative	76.47	76.96	68.72	70.61
Neutral	11.36	11.36	03.03	03.03
Overall	81.83	81.84	73.73	74.87

(c)

Table 3. Results: (a) aspect term and OTE extraction. (b) sentiment classification SemEval-2014, (c) sentiment classification SemEval-2016. M1 and M2 corresponds to framework 1 and framework 2 of Section 2.3. Std.Fea refers to feature set of section 2.6 except DT, Unsupervised PoS and expansion score.

Datasets	**Task**	**Rank**	**Diff. from top team**
Restaurant (SemEval-2014)	Aspect Ext.	3/28	3.78
Laptop (SemEval-2014)	Aspect Ext.	3/27	1.56
English (SemEval-2016)	OTE	3/19	3.89
Spanish (SemEval-2016)	OTE	1/4	-
French (SemEval-2016)	OTE	1/2	-
Dutch (SemEval-2016)	OTE	1/2	-
Restaurant (SemEval-2014)	Senti.	5/31	4.06
Laptop (SemEval-2014)	Senti.	6/31	5.04
English (SemEval-2016)	Senti.	8/27	6.28
Dutch (SemEval-2016)	Senti.	4/4	2.94

Table 4. Comparisons with other teams

4.1 OTE

1. Long OTEs (more than two tokens) are often not correctly identified by our system. For instance, the 4-token aspect term in the below example is completely ignored by the proposed method.

Domains	Framework	Features																														
		Word&Context	Char n-grams (2,3,4)	Chunk	Dependency	Dependency (Head)	Dependency (Modifier)	Ortho (digit)	DT (-2..2)	Freq. Aspect List1	Freq. Aspect List2	HeadWord	HeadWord PoS	Lemma	LowerCase	NER (-1..1)	PMI	Rounded PMI (-1..1)	PoS (-2..2)	1-Char Prefix(-1..1)	2-Char Prefix(-1..1)	3-Char Prefix(-1..1)	Expansion score (-2..2)	RoundExpansion score (-2..2)	SentiWordNet (-2..2)	Round SentiWord (-2..2)	Stopword	1-Char suffix (-1..1)	2-Char suffix (-1..1)	3-Char suffix (-1..1)	Noun synset	Unsupervised PoS (-2..2)
Restuarant	M1	-3,-1..3	2,3	0,1	-	✓	✓	-	-2..0	-	✓	-	-	-	-	0	-	-1	-2..0	1	-	-	✓	-	0,2	-2,2	✓	✓	-1,0	✓	✓	-1,2
	M2	-3,-1..1,3	4	-1,1	✓	✓	✓	-	-2,0,2	✓	✓	-	✓	-	✓	-	✓	0	-2,0	-1,1	-1	1	✓	0,2	-2..0,2	0,2	✓	✓	-1	-1,0	✓	-1..2
Laptop	M1	-1,0,3	2,4	✓	-	-	✓	✓	0	-	✓	-	✓	-	-	-1	0,1	1	-2,1	-	-	-1	2	1	-	-1	-	-	-	-	-	-
	M2	2	✓	✓	✓	✓	✓	✓	-2,0	-	✓	-	✓	-	-	-1,0	0,1	-1,1	-2..1	0	-	0	-1,1	0,2	-1,2	-	-	0	-	-1,0	✓	2

Table 5. Features selected for aspect term extraction: SemEval-2014. M1 and M2 corresponds to framework 1 and framework 2 of Section 2.3. Here ✓represents that feature has been selected. Number denotes the context of current token. Example:- (-3..1) represents that context token from previous 3 upto next 1 have been selected.

Language	Framework	Features																														
		Word and Context	Char n-gram(2,3,4-gram)	Chunk information(-1,0,1)	Dependency (Head)	Dependency (Modifier)	Distri. Thesaurus(-2,-1,0,1,2)	Frq. Aspect List1	Freq. Aspect List2	Head Word	Head Word POS	Lemma	Lower Case	NER	Ortho	POS (-2,-1,0,1,2)	1-Char Prefix(-1,0,1)	2-Char Prefix(-1,0,1)	3-Char Prefix(-1,0,1)	4-Char Prefix	Expansion score(-2,-1,0,1,2)	Round. Exp.Score(-2,-1,0,1,2)	SentiWord(-2,-1,0,1,2)	Round. SentiWord(-2,-1,0,1,2)	1-Char Suffix(-1,0,1)	2-Char Suffix(-1,0,1)	3-Char Suffix(-1,0,1)	4-Char Suffix	Noun Synset	Unsupervised POS(-2,-1,0,1,2)
English	M1	0,3,4	2..4		✓		-2..1						✓		✓	-2,0,2	0,1		-1			-1,1,2		2	-1						
	M2	-4..-2,1,2,4,5	4	0,-1	✓		-2..0,2	✓				✓		✓	✓	-1..2	0,1	✓	0,1	✓	-1,1,-2	-1,1	1	-2	✓	0	✓		✓	-2,0,2
Dutch	M1	-1,1		-			-1,0					-		-		-2,-1			-1,0	✓	-2	-2	-	-		-1	1		-	-1,0
	M2	-4,-2,-1,1		-	✓		-2..0			✓	✓	-	✓	-	✓	-2,0,1		0,1	1	✓	-2,-1,1	-2	-	-	0,1	✓	1		-	-1,0
Spanish	M1	-5,-1,0	2,3	-		✓	-1,0	✓	✓			-		-	✓	-2..0	0	1	-1	✓	-1	-2,1	-	-	0	1	-1		-	-1,0
	M2	-4,-3,5	✓	-			-1		✓			-	✓	-	✓	-2,-1	0,1	-1	-1,0	✓	-1,0	-2,-1,1	-	-	0,1	1	0		-	-1..1
French	M1	-3,-1		-			-2,0					-		-		-2,-1	1	0,1				2	-	-		-1			-	
	M2	-5,-3..0,2	2,3	-		✓	-1,0		✓	✓	✓	-		-		-2,-1,1	✓	0	-1,0		-2,1,2	-2..0	-	-	-1	-1		✓	-	-1

Table 6. Features selected for OTE: SemEval 2016.

*... with delectable **creamed Washington russet potatoes** and crisp ...*

2. Many times our system identifies the two OTEs which is associated with 'and' as single OTE. For example, ***server*** and ***food*** are the two aspect terms in the following review but our system predict ***server and food*** as an aspect term due to the presence of 'and' in between the two.

*... our wonderful **server** and **food** made the experience a very positive one ...*

3. The proposed method faced difficulties in identifying opinion target which contains special characters (e.g. **'**, **-**). For example

*The **pizza's** are light and scrumptious .*

Features	Restaurant		Laptop	
	Model 1	Model 2	Model 1	Model 2
Word and Context	-4..3	-4,-2..1	-4,-2..3,5	-5,-2..4
Bi-gram	2		2	-2,2
Bing Lexicon	✓	✓	✓	✓
Bing Direct Lexicon	✓	✓	✓	✓
SentiWord	✓	✓		✓
PMI	✓	✓		✓
MPQA	✓	✓	✓	✓
Domain Specific Word	✓	✓	-	-
Sentiment-140 lexicon (Unigram,Bigram)	✓	Unigram		
NRC Hashtag lexicon (Unigram,Bigram)	✓	Bigram		
AFINN lexicon				
NRC Emotion	✓		✓	
Expansion Score		✓		

Table 7. Feature selection of sentiment classification - 2014

4. Some instances of the OTE that form the last token in the text are not classified by our system. For e.g. opinion target term ***pepperoni*** appears at the end of the review, which is not identified by the proposed system.

   *...big thick **pepperoni**.*

4.2 Sentiment Classification

1. Presence of negative term like *n't, never, but* etc. always change the polarity orientation in the text but our system fails to capture such orientation on few instances specially with smaller sentences. For e.g. *"Not good !"*. It should be classified as negative but our system classifies it as positive.
2. In many cases our system is biased to a particular review. It means if a review has more than one OTE, our system assigns same class to all of them, even it is different.

   *... the **fish** is unquestionably fresh , **rolls** tend to be inexplicably bland.*

   Here for *fish* and *rolls*, we have positive and negative review respectively. But, our system classifies positive in both cases.

5 Conclusion

In this paper, we reported on experiments for improving the quality of aspect-based sentiment analysis (ABSA) and its subtasks. We have discussed two contributions in detail: 1) the use of features from unsupervised lexical acquisition and 2) the use of multi-objective optimization (MOO) for feature selection. Experimental results on three subtasks of ABSA for six languages and several domains show consistent improvements in all experiments for unsupervised lexical acquisition features. MOO was able to improve the classification / extraction accuracy in some cases, but always resulted in a much more compact model.

The strength of our approach is the combination of unsupervised features and feature selection: Since unsupervised features do not require language-specific processing,

they apply to all natural languages where sufficient raw corpora are available. However, unsupervised acquisition necessarily retains a certain amount of noise. The MOO feature selection mechanism counterbalances this by only retaining features (of any kind), which contribute to a good performance. This, we are able to afford the experimentation with more features and feature combinations since this setup allows us to over-generate features and have them selected automatically.

In future work, we would like to further automatize the process by extending our approach to several base classifiers, making their selection and their ensemble also subject to automatic optimization techniques.

References

1. Biemann, C.: Unsupervised Part-of-Speech Tagging in the Large. Research on Language and Computation 7(2-4), 101–135 (2009)
2. Biemann, C., Giuliano, C., Gliozzo, A.: Unsupervised Part-of-Speech Tagging Supporting Supervised Methods. In: Proceedings of RANLP. vol. 7, pp. 8–15. Borovets, Bulgaria (2007)
3. Biemann, C., Riedl, M.: From Global to Local Similarities: A Graph-Based Contextualization Method using Distributional Thesauri. In: Proceedings of the 8th Workshop on TextGraphs in conjunction with EMNLP. pp. 39–43. Seattle, USA (2013)
4. Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G.A., Reynar, J.: Building a Sentiment Summarizer for Local Service Reviews. In: WWW Workshop on NLP in the Information Explosion Era. vol. 14, pp. 339–348. Beijing, China (2008)
5. Deb, K.: Multi-objective Optimization using Evolutionary Algorithms, vol. 16. John Wiley & Sons (2001)
6. Ding, X., Liu, B., Yu, P.S.: A Holistic Lexicon-based Approach to Opinion Mining. In: Proceedings of the 2008 International Conference on Web Search and Data Mining. pp. 231–240. New York, USA (2008)
7. Ekbal, A., Saha, S.: Multiobjective Optimization for Classifier Ensemble and Feature Selection: An Application to Named Entity Recognition. International Journal on Document Analysis and Recognition (IJDAR) 15(2), 143–166 (2012)
8. Ekbal, A., Saha, S.: Simulated Annealing Based Classifier Ensemble Techniques: Application to Part of Speech Tagging. Journal Information Fusion 14(3), 288–300 (2013)
9. Fahrni, A., Klenner, M.: Old Wine or Warm Beer: Target-Specific Sentiment Analysis of Adjectives. In: Proceedings of the Symposium on Affective Language in Human and Machine, AISB. pp. 60–63. Aberdeen, Scotland
10. Hatzivassiloglou, V., McKeown, K.R.: Predicting the Semantic Orientation of Adjectives. In: Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics. pp. 174–181. Madrid, Spain (1997)
11. Hu, M., Liu, B.: Mining and Summarizing Customer Reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 168–177. Seattle, USA (2004)
12. Kim, S.M., Hovy, E.: Determining the Sentiment of Opinions. In: Proceedings of the 20th International Conference on Computational Linguistics. pp. 1367–1373. Geneva (2004)
13. Kumar, A., Kohail, S., Kumar, A., Ekbal, A., Biemann, C.: IIT-TUDA at SemEval-2016 Task 5: Beyond Sentiment Lexicon: Combining Domain Dependency and Distributional Semantics Features for Aspect Based Sentiment Analysis. In: 10th International Workshop on Semantics Evaluation (SemEval-2016), ACL. pp. 311–317. San Diego, USA (2016)
14. Lin, D.: Automatic Retrieval and Clustering of Similar Words. In: Proceedings of the 17th Inter. Conference on Computational Linguistics-Volume 2. pp. 768–774. Stroudsburg (1998)

15. Liu, B.: Sentiment Analysis and Opinion Mining, vol. 5. Morgan & Claypool Publishers (2012)
16. Liu, H., Motoda, H.: Feature Selection for Knowledge Discovery and Data Mining, vol. 454. Springer Science & Business Media, Norwell, USA (2012)
17. Liu, H., Yu, L.: Toward Integrating Feature Selection Algorithms for Classification and Clustering. IEEE Transactions on Knowledge and Data Engineering 17(4), 491–502 (2005)
18. Miller, G.A.: Wordnet: A Lexical Database for English. Communications of the ACM 38(11), 39–41 (1995)
19. Miller, T., Biemann, C., Zesch, T., Gurevych, I.: Using Distributional Similarity for Lexical Expansion in Knowledge-based Word Sense Disambiguation. In: Proceedings of COLING. pp. 1781–1796. Mumbai, India (2012)
20. Mohammad, S.M., Turney, P.D.: Crowdsourcing a Word–Emotion Association Lexicon. Computational Intelligence 29(3), 436–465 (2013)
21. Mukherjee, A., Liu, B.: Aspect Extraction through Semi-supervised Modeling. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. pp. 339–348. Jeju Island, Korea (2012)
22. Nielsen, F.Å.: A new ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs. In: Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages. pp. 93–98. Heraklion, Greece (2011)
23. Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., Clercq, O.D., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., JimÃl'nez-Zafra, S.M., EryiÄ§it, G.: Semeval-2016 Task 5: Aspect Based Sentiment Analysis. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016). pp. 19–30. San Diego, USA (2016)
24. Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., Manandhar, S.: Semeval-2014 Task 4: Aspect Based Sentiment Analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). pp. 27–35. Dublin (2014)
25. Popescu, A.M., Etzioni, O.: Extracting Product Features and Opinions from Reviews. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. pp. 339–346. Stroudsburg, USA (2005)
26. Schouten, K., Frasincar, F.: Survey on aspect-level sentiment analysis. IEEE Transactions on Knowledge and Data Engineering 28(3), 813–830 (March 2016)
27. Turney, P.D.: 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. pp. 417–424. Philadelphia, USA (2002)
28. Wiebe, J., Mihalcea, R.: Word Sense and Subjectivity. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. pp. 1065–1072. Sydney, Australia (2006)
29. Wu, Y., Zhang, Q., Huang, X., Wu, L.: Phrase Dependency Parsing for Opinion Mining. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3. pp. 1533–1541. Singapore (2009)
30. Zhu, X., Kiritchenko, S., Mohammad, S.M.: NRC-Canada-2014: Recent Improvements in the Sentiment Analysis of Tweets. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). pp. 443–447. Dublin, Ireland (2014)
31. Zhuang, L., Jing, F., Zhu, X.Y.: Movie Review Mining and Summarization. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management. pp. 43–50. Virginia, USA (2006)