A Vector Space Model for Subjectivity Classification in Urdu aided by Co-Training

Smruthi Mukund
CEDAR
University at Buffalo
smukund@buffalo.edu

Rohini K. Srihari
CEDAR
University at Buffalo
rohini@cedar.buffalo.edu

Abstract

The goal of this work is to produce a classifier that can distinguish subjective sentences from objective sentences for the Urdu language. The amount of labeled data required for training automatic classifiers can be highly imbalanced especially in the multilingual paradigm as generating annotations is an expensive task. In this work, we propose a co-training approach for subjectivity analysis in the Urdu language that augments the positive set (subjective set) and generates a negative set (objective set) devoid of all samples close to the positive ones. Using the data set thus generated for training, we conduct experiments based on SVM and VSM algorithms, and show that our modified VSM based approach works remarkably well as a sentence level subjectivity classifier.

1 Introduction

Subjectivity tagging involves distinguishing sentences that express opinions from sentences that present factual information (Banfield 1982; Wiebe, 1994). A wide variety of affective nuances can be used while delivering a message pertaining to an event. Although the factual content remains the same, lexical selections and grammatical choices can considerably influence the affective nature of the text. Recognizing sentences that exhibit affective behavior will require, at the least, recognizing the structure of the sentence and the emotion bearing words.

To date, much of the research in this area is focused on English. A variety of reliable resources that facilitate effective sentiment analysis and opinion mining, such as polarity lexicons (Senti-WordNet [1]) and contextual valence shifters (Kennedy and Inkpen, 2005) are available for English. The MPQA corpus of 10,000 sentences (Wiebe *et al.,* 2005) provides detailed annotations for sources of opinions, targets, speech events and fragments that indicate attitudes for the English newswire data. The IMDB corpus contains 10,000 sentences categorized as subjective and objective in the movie review domain. Clearly, English is well supported with resources. There are other widely spoken resource poor languages that are not as privileged. When we consider social media, limiting our analysis to a language like English, however universal, will lead to loss of information. With the advent of virtual keyboards and extended Unicode support, the internet is rapidly getting flooded by users who use their native language in textual communication. There is a pressing need to perform non-topical text analysis in the multilingual paradigm.

Subjectivity analysis is a precursor to numerous applications performing non-topical text analysis like sentiment analysis, emotion detection, and opinion extraction (Liu *et al.*, 2005; Ku *et al.,* 2006; Titov and McDonald, 2008). Creating the state-of-the-art subjectivity classifier using machine learning techniques require access to large amounts of annotated data. For less commonly taught languages like

[1] http://swn.isti.cnr.it/download_1.0/

Coling 2010: Poster Volume, pages 860–868,
Beijing, August 2010

Urdu, Hindi, Bengali, Spanish and Romanian, the resources required to automate subjectivity analysis are either very sparse or unavailable. Generating annotated corpus for subjectivity detection is laborious and time consuming.

However, several innovative techniques have been proposed by researchers in the past to generate annotated data and lexical resources for subjectivity analysis in resource poor languages. Mihalcea *et al.,* (2007) and Banea *et al.,* (2008) used machine translation technique to leverage English resources for analysis in Romanian and Spanish languages. Wan (2009) proposed a co-training technique that leveraged an available English corpus for Chinese sentiment classification. Wan (2008) focused on improving Chinese sentiment analysis by using both Chinese and English lexicons.

Unfortunately, not much work has been done in the area of subjectivity analysis for the Urdu language. This language lacks annotated resources required to generate even the basic NLP tools (POS tagger, NE tagger etc.) needed for text analysis. In order to facilitate subjectivity analysis in Urdu language, we annotated a small set of Urdu newswire articles for emotions (§2). The sentence level annotations provided in this dataset follow the annotation guidelines proposed by Wiebe *et al.,* (2003). Although tremendous effort was put into generating this corpus, the data set is not very comprehensive and contains only about 500 sentences marked subjective. This is definitely insufficient to train a suitable subjectivity classifier.

1.1 Issue with unbalanced data set

A subjectivity classifier is a binary classifier. A traditional binary classifier is trained using universal representative sets for positive and negative categories. But in subjectivity analysis, especially for languages like Urdu that have no annotated data, generating universal representative sets is extremely difficult and almost an impossible task. Assimilating the negative set is especially a delicate task as the set should be carefully pruned of all the positive samples. Also, detecting subjectivity in a sentence is highly personalized. Annotators are sometimes prejudiced while marking samples. This bias, however small, produces errors with some true positive samples being unintentionally missed and categorized as negative. Traditionally, research in machine learning has assumed the class distribution in the training data to be reasonably balanced. However, when the training data is highly imbalanced, i.e., the number of positive examples is very small, the performance of text classification algorithms such as linear support vector machine (SVM) (Brank and Grobelnik, 2003), naïve Bayes and decision trees (Kubat and Matwin, 1997) are adversely affected.

In order to achieve a balanced training set, Japkowicz (2000) duplicates positive examples (oversampling) and discards negative ones (downsizing). Kubat and Matwin (1997) discard all samples that are close to the positive set to avoid misclassification. Chan and Stalfo (1998) have trained several classifiers on different balanced data subsets, each constructed to include all positive training samples and a set of negative samples of comparable size. The predictions are combined through stacking.

For the task of subjectivity analysis, especially in the multilingual paradigm where the data set is highly unbalanced, using one of the techniques proposed above will yield benefit. To the best of our knowledge, co-training technique has not been applied before for the subjectivity detection task, in particular, for the Urdu language.

1.2 Contribution

Our first contribution is inspired by the work of Luo *et al.,* (2008). We propose a similar co-training technique that helps to create a likely negative set (objective sentences) and a filtered positive set (subjective sentences) simultaneously from the unlabeled set. We use two learning models trained using the linear SVM algorithm iteratively. In every iteration of co-training, the likely positive samples are filtered. The iterative process terminates when no more positive samples are found. The final negative set is the likely negative set, considered as the universal representative set for the non-subjective category. The likely positive sample set is appended to the already existing positive set (annotated set). The SVM models are trained using part of speech, unigrams and emotion bearing words, as features.

The second contribution of this work includes training a state-of-the-art Vector Space Model

(VSM) for Urdu newswire data using the data sets generated by the co-training method. Experiments that use the SVM classifier are also performed. The results show that the performance of the proposed VSM based approach helps to achieve state-of-the-art sentence level subjectivity classifier. The F-Measure of the VSM subjectivity classifier is 82.72% with 78.7% F-measure for the subjective class and 86.7% F-Measure for the objective class.

2 Data Set

The data set used to generate a subjectivity classifier for Urdu newswire articles is obtained from BBC Urdu[2]. The annotating efforts are directed towards achieving the final goal- *emotion detection* in Urdu newswire data and the annotation guidelines are based on the MPQA standards set for English.

The repository of articles provided by BBC is huge and needs to be filtered intelligently. Two levels of filters are applied. – *date* and *keyword search*. The *date* filter is applied to retrieve articles of three years, starting year 2003. The *keyword* based filter consists of a set of seed words that are commonly used to express emotions in Urdu -*ghussa (~anger), pyar (~love)* etc. Clearly, this list will not cover all possible linguistic expressions that express emotion and opinion. But it is definitely a representative of a wide range of phenomena that naturally occurs in text expressing emotions.

The data retrieved is parsed using an in-house HTML parser to produce clean data. To date, we have 500 articles, consisting of 700 sentences annotated for emotions. There are nearly 6000 sentences that do not contain any emotions making it highly unbalanced. This data set is divided into testing and training sets with 30% and 70% of the data respectively. Co-training is performed only on the 70% training set that consists of 470 subjective sentences and about 4000 objective sentences. The purpose of co-training here is to remove samples that are close to subjective from the objective set and create a likely negative set. The samples removed are the likely positive set. This set of 4000 objective sentences can be considered as the un-annotated set.

[2] http://www.bbc.co.uk/urdu/

3 Co-Training

Identifying sentences that express emotions in Urdu newswire data is not trivial. Subjective sentences do not always contain individual expressions that indicate subjectivity. Analysis is highly dependent on the contextual information. Wiebe *et al.,* (2001) reported that nearly 44% of sentences in the MPQA corpus (English newswire data) are subjective. In newswire data, though most facts are reported objectively, there are cases when the tone of the sentence is very intense indicating the existence of emotion. Consider Example 1.

Example 1:
Political news headline

بھارت کا پاکستان کے ساتھ **جامع مذاکرات** سے انکار ،**بھارتی لیکچر سننے کے خواہاں نہیں**

[bhart ka pakstan kE sath jame mZakrat sE ankar, bharty lykcr snnE kE KwahaN nhyN]

*[India refuses to have a **dialog** with Pakistan, Indians **are not willing to listen to the lecture**]*

Common Urdu

انڈیا نے پاکستان سے **بات چیت** کارنے سے انکار کر دیا ہے

*[India refuses to **talk** to Pakistan]*

Clearly, the news headline is extremely intense and strongly expresses the opinion of India on Pakistan. However, the statement in common Urdu is not as affective.

Example 2:

انصاری نے کہا، میری رائے میں عامر سہیل ایک بد دماغ اور ضدی شخص ہیں

[anSary nE kha "myry ray^E myN eamr shyl ayk bd dmaG awr Zdy XKS hyN"]

[Ansari said, "according to me Aamir Sohail is one crazy and stubborn man"]

Statements in quotes that express emotions are subjective as shown in example 2.

Consider example 3. Here, identifying the words that indicate subjectivity is not straight forward. The phrase, *"found it very difficult to hide his smile"* is indicative of the emotion experienced by *"Habib Miya"*.

Example 3:

رقم کی اس وصولی پر یہ حبیب میاں کے لئے **بہت مشکل** تھا کہ وہ اپنی مسکراہٹ چھپا سکیں

[rqm ky as wSwly pr yh Hbyb myaN kE ly^E bht mXkl t\|ha kh wh apny mskrahT c\|hpa skyN]

[At this event of money collection, Habib Miyan found it very difficult to hide his smile.]

There are also several false positives that make subjective detection hard task. Example 4 is an objective sentence despite the usage of word *"pyar" ~ love*, an emotion bearing word.

Example 4:

انضمام کا نیا پیار کا نام انزی پڑا ہے

[n|Zmam ka nya pyar ka nam anzy pRa hE]

[The new nickname for Inzaman is Inzi]

Expressive elements in Urdu sentences were marked with an inter-annotator agreement of 0.8 kappa score. Though high, there still exists a bias that can influence classification especially when the number of sentences in the positive set is relatively less. In order to obtain a reliable positive and negative set for training a learning algorithm, we adopt a semi-supervised learning technique of co-training. *Co-training* (Blum and Mitchell, 1998) is similar to self-training in that it increases the amount of labeled data by automatically annotating unlabeled data. The intuition here is that if the conditional independence assumption holds, then on an average each selected document will be as informative as a random document, and the learning will progress. Co-training differs from self-training as it uses multiple learners to do the annotation. Each learner offers its own perspective that when combined gives more information. This technique is especially effective when the feature space of a particular type of problem can be divided into distinct groups and each group contains sufficient information to perform the annotation. In other words, co-training algorithm involves training two different learning algorithms on two different feature spaces. The learning of one becomes conditionally independent of the other and the prediction made by each classifier is used on the unlabeled data set to augment the training data of the other.

A traditional co-training classifier is trained and later applied on the same unlabeled data set. Theoretically such classifiers are not likely to assign confident labels. In this work, the proposed co-training method differs from the traditional co-training method in that the two classifiers are based not on two different feature spaces but on two different training data sets with the same feature space.

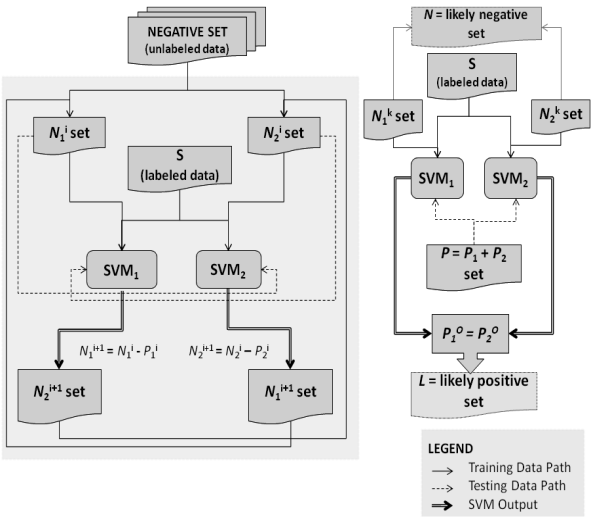

Figure 1: Co-Training model

Figure 1 explains the overall working of the model. The negative set (which can also be the unlabeled set) is split into two equal parts N_1 and N_2. S represents the positive annotated set. Two linear SVM classifiers are trained iteratively to purify the negative data set. SVM_1 is trained using $S+N_1^i$ and SVM_2 is trained using $S+N_2^i$ data sets. In every iteration *i*, N_1^i data set is evaluated using SVM_2 model and N_2^i data set is evaluated using SVM_1 model. The samples that are classified as positive in a given iteration *i* are binned into sets P_1^i and P_2^i respectively. These samples are removed from N_1^i and N_2^i data sets to create new N_1^{i+1} and N_2^{i+1} sets that are used for training in the next iteration *i+1*. The iterations continue until no positive samples are marked by both SVM_1 and SVM_2 models. The final set of likely negatives is $\mathbf{N} = N_1^k + N_2^k$ sets, where N_1^k and N_2^k are sets created in the last *k* iteration of the algorithm. In order to obtain the likely positive set, the final $P_1 = \{P_1^1 + P_1^2 + \ldots. + P_1^k\}$ and $P_2 = \{P_2^1 + P_2^2 + \ldots. + P_2^k\}$ sets are combined and tested using the SVMs modeled in the last *k* iteration of the co-training algorithm. Similar to the traditional co-training method the samples that are marked positive by both classifiers ($P_1^o = P_2^o$) are considered to be the likely positive set **L**.

Several features are used to train the SVM learning models used for co-training. The best performance is obtained when word unigrams, parts of speech and likely emotion words are used as features.

This technique of co-training provides us with a relatively huge set of likely positive samples

(close to 400 sentences). Sentences in this set were examined by the annotators and nearly 60% of the sentences were subjective or near subjective in nature (Example 5 and 6).

Labels	R %	P %	IF %	AF %
Unigram				**52.63**
1	**18.64**	74.57	**29.83**	
-1	95.4	62.35	75.44	
Unigram+Bigram				50.25
1	**14.40**	85	**24.63**	
-1	98.19	61.82	75.87	

Table 1: Performance of the model using un-balanced data set[3]

Labels	R %	P %	IF %	AF %
Annotated positive + likely positive + likely negative				**62.95**
1	**39**	70	**50.09**	
-1	87.28	67.34	79.9	
Annotated positive + likely negative				55.42
1	30	61.2	**40.26**	
-1	86.1	64.23	73.57	

Table 2 – Performance of the model after co-training method

Table 1 shows the performance of the SVM model using the unbalanced data set for training. Table 2 shows the performance of the same model using data generated after co-training.

Example 5:

پوتن نے کہا کہ لوگ دوسروں کی آنکھ میں تنکا دیکھ لیتے ہیں لیکن اپنی آنکھ میں پڑا شہتیر انہیں نظر نہیں آتا ۔

[pwtn nE kha kh lwg dwsrwN ky Ank|h myN tnka dyk|h lytE hyN lykn apny Ank|h myN pRa Xhtyr anhyN n|zr nhyN Ata .]

[Potan said people who see dust in others eyes never realize that it is their eyes that are filled with dirt.]

The above example is a metaphor indicating extreme anger.

Example 6:

عطاء الرحمن شیخ کا کہنا ہے کہ بارہ اگست کو انہیں ان کے بیٹوں کے سامنے مکمل طور پر برہنہ کر کے پریڈ کرائی گئی

[e|ta& alrHmn XyK ka khna hE kh barh agst kw anhyN an kE byTwN kE samnE mkml |twr pr brhnh kr kE pryD kray^y gy^y]

[etlaalrahman said that on 12th Aug they made him parade naked in front of his children.]

Example 6 indicates extreme sad emotion. Such examples were found in the likely positive set.

4 Features

Features that are commonly used to train a subjectivity classifier for English are word unigrams, emotion keywords, part of speech information and noun patterns (Pang *et al.,* 2002). Due to difference in syntactic structure, vocabulary and style, features that work for English may not work for Urdu. Also, Urdu is handicapped by the lack of resources required to perform basic NLP analysis. However, it is worth exploring the English feature set as subjectivity is more a semantic phenomenon. Efforts to generate likely emotion word lexicons and subjectivity patterns for the Urdu language are underway. The sections that follow summarize the experimented features.

4.1 Word Unigrams

Unigram word features are very informative. Three different approaches are tried for selecting the unigrams. The first method involves selecting only those words that occur more than twice in the dataset. This eliminates proper nouns (low frequency named entities do not generally contribute towards subjectivity detection) and spelling errors (Pang *et al.,* 2002). In the second method, only words that are adjectives and verbs along with the surrounding case markers are accounted for as features. This has the advantage of drastically reducing the feature set. The third method involves including the nouns as well to the feature set. A simple list of stop words (common Urdu words – pronouns such as *'us', 'is', 'aap', 'un'*, salutations like *'shabba khair', 'aadab'* and honorifics along with punctuations and special symbols) are eliminated. The features are represented as Boolean features for the SVM model. The value is 1 if the feature word appears in the sentence to be classified and 0 otherwise. The best performance is obtained for the first method that considers all words with frequency greater than 2. This conforms to what is shown by Pang *et al.,* (2002) for classification of English movie reviews.

4.2 Part of Speech (POS) Information

The work done by Mukund and Srihari (2009) provides suitable POS and NE tagger for Urdu.

[3] Convention used across tables - Label 1: subjective sentences Label -1: objective sentences R: Recall P: Precision IF: Individual F-Measure AF: Average F-Measure.

This POS tagger is used to generate parts of speech tags on the acquired data set (§3). The POS tags associated with adjectives, verbs, common nouns and auxiliary words are considered and used as Boolean features for the SVM model. The proper noun words are normalized to one common word *"nnp"* and are assigned the common noun tag. For the English language, when building a subjectivity classifier for review classification, the use of POS information did not benefit the system (Kennedy and Inkpen, 2006). However, for Urdu, the performance of the co-training model with POS information showed 1.2% improvement (table 3).

4.3 Likely Emotion Lexicon

In order to facilitate simple keyword based detection of subjectivity, access to a lexicon consisting of likely emotion words is needed. Unfortunately, no such lexicon is available off the shelf for Urdu. In this work, an Urdu specific emotion list is generated that contains translations from the English emotion list released by SemEval (2007) '*WordNet affect Emotion List*'. Words for each emotion category - sadness (sad), fear, joy (happy), surprise, anger and disgust are obtained for Urdu by using an Urdu-English dictionary. The list is pruned manually and corrected to remove errors. Simple keyword lookup on the Urdu annotated corpus has an emotion detection rate of 29.27%. This shows that although the contribution of the emotion lexicon for subjectivity classification is not significant, it contains information which when used along with other features aid subjectivity detection.

4.4 Patterns

Extracting syntactic patterns contribute towards the affective orientation of a sentence (Riloff *et al.,* 2003). The Apriori algorithm (Agarwal and Srikant, 1994) for learning association rules is used here to mine the syntactic word patterns commonly used in the positive and negative data set. The length of the candidate item set k = 4. Starting from a small set of seed words (likely emotion words) and the associated POS tags, POS sequential patterns like "adverb verb verbtransitive sentencemarker", "noun noun casemarker verbtransitive", etc., that are most commonly found in subjectivity set are extracted. 23 patterns that strongly indicate subjectivity were found by this method and included as features to train the SVM learning algorithm.

4.5 Confidence Words

The confidence word list positively aids the VSM classifier (§5). The words in the likely emotion list are not the only ones that contribute towards the emotion orientation of a sentence and also, not all of these words contribute effectively. There are several stop words (eliminated while accounting for unigrams) (esp. case markers) that contribute significantly for categorization. In order to identify all the keywords that actually contribute to subjectivity categorization, a technique proposed by Soucy and Mineau (2004) is used.

The confidence weight of a given word w, based on the number of documents it is associated with under each category, is measured using the Wilson Proportion Estimate (Wilson, 1927). In order to compute the confidence of w for a specific category, the number of positive and negative documents associated with w has to be determined. A document is positive if it belongs to that category and negative otherwise. Thus, two kinds of word confidence metrics are computed, $C_{POS:w}$ and $C_{NEG:w}$ as given below.

$$C_{POS:w} = \frac{\left(\hat{p}_{POS:w} + \frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2} \sqrt{[\hat{p}_{POS:w}(1-\hat{p}_{POS:w}) + z^2_{\alpha/2}/4n]/n} \right)}{(1 + z^2_{\alpha/2}/n)} \quad \text{..... (Eq. 1)}$$

$$C_{NEG:w} = \frac{\left(\hat{p}_{NEG:w} + \frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2} \sqrt{[\hat{p}_{NEG:w}(1-\hat{p}_{NEG:w}) + z^2_{\alpha/2}/4n]/n} \right)}{(1 + z^2_{\alpha/2}/n)} \quad \text{..... (Eq. 2)}$$

where n is the total number of positive and negative documents, $\hat{p}_{POS:w}$ is the ratio of the number of positive documents which contain w to the total number of documents, and $\hat{p}_{NEG:w}$ is the ratio of the number of negative documents which contain w to the total number of documents. The normal distribution is used when $n > 30$.

Note that equations 1 and 2 give a range of values for $C_{POS:w}$ and $C_{NEG:w}$. If the lower bound of $C_{POS:w}$ is greater than the upper bound of $C_{NEG:w}$, we say that w is likely to be a word in that category. Now, we compute the strength of a word S_w in a particular category as

$$S_w = \begin{cases} \log(2 \cdot mPRF) & ; \text{if } lb(C_{POS:w}) > ub(C_{NEG:w}) \\ 0 & ; \text{otherwise} \end{cases}$$

..... **(Eq. 3)**

where *mPRF* is given by

$$mPRF = \frac{lb(C_{POS:w})}{lb(C_{POS:w}) + ub(C_{NEG:w})}$$

..... **(Eq. 4)**

and lb(…) and ub(…) are the lower and upper bounds of their arguments, respectively.

Equations 1 through 4 generated a very good set of keywords that are used as category word features in the SVM learning model. For VSM, the strength value is used as a boost factor along with the *tf-idf* weight when calculating the similarity score (table 3).

5 Final Subjectivity Classifier

Wiebe *et al.,* (2005) and Pang *et al.,* (2002) have shown that an SVM based approach works well for subjectivity classification. Riloff *et al.,* (2003) have conducted experiments that use Bag-Of-Words (BoW) as features to generate a Naïve Bayes subjectivity classifier for the MPQA corpus in English. This method has an accuracy of 73.3%. Su and Markert (2008) use BoW features termed as lexical features on the IMDB corpus to generate an accuracy of 60.5%. Das and Bandyopadhyay (2009) use a CRF based approach to generate a subjectivity classifier for Bengali data with a precision of 72.16% for news and 74.6% for blogs domain. The same approach has a precision of 76.08% and 79.9% on the two domains respectively. Impressive results for emotion detection are obtained by Danisman and Alpkocak, (2007) who use a VSM based approach. They show that their approach works much better than a traditional SVM based approach commonly used for emotion detection.

In this work, we conduct subjectivity classification experiments using two different learning algorithms – linear SVM and VSM. The best performance is obtained using the VSM model as shown in table 4. All experiments are conducted on the data set obtained after applying the co-training technique.

5.1 VSM algorithm

The final subjectivity classifier is based on the VSM approach. Inspired by the work done in "*Feeler*" (Danisman and Alpkocak, 2007), a similar technique is used to train the final subjectivity classifier for Urdu. The algorithm is explained in table 3. The similarity metric is modified to include the confidence score for each word (pt.5). In VSM, documents and queries are represented as vectors, and the cosine angle between them indicates the similarity.

1.	$d_i = <w_{1i}, w_{2i}, ... w_{ni}>$ where w_{ki} is the weight of the k^{th} term in document i , d_i is the document vector. w_{ki} is computed using *tf-idf* weighting scheme.
2.	$M_j=\{d_1,d_2,\ldots,d_c\}$ where M_j is each class (subjective and objective)
3.	Model vector for an arbitrary class E_j is created by taking the mean of d_j vectors $$E_j = \frac{1}{M_j}\sum_{d_i \in M_j}^{|M_j|} d_i$$ where $\|M_j\|$ represents number of documents in M_j.
4.	The whole system is represented with a set of model vectors, $D=\{E_1,E_2,...,E_s\}$ where s represents the number of distinct classes to be recognized.
5.	The normalized similarity between a given query text Q, and a class, E_j, is defined as follows: $$sim(Q,E_j) = \sum_{k=1}^{n}(w_{kq} + conf) * E_{kj}$$ *conf* is the confidence factor applied for lexical terms found in the word list.
6.	classification result is, $VSM(Q) = \arg\max(sim(Q,E_j))$

Table 3: VSM Algorithm for subjectivity Classification

Labels	R %	P %	IF %	AF %
Before Co-Training (all data)				**62.95**
1	65.85	70.85	**67.4**	
-1	85.58	83.33	84.44	
After Co-Training (pruned data)				**86.73**
1	**72.88**	85.57	**78.72**	
-1	91.29	82.60	86.73	

Table 4: VSM approach, using all training data and using pruned training data (L+N+true)

The confidence metric (strength) for each term is calculated using the Wilson proportion estimate (§4.4) and added to the term score as the boost factor. Q is the test set. Model vectors are obtained using the data set that consists of true set (annotated positive samples), likely positive set L and likely negative set N. Sets L and N are obtained from the co-training method. The results are shown in table 4.

The power of SVM cannot be ignored. Pang *et al.,* (2002) use SVM to generate a subjectivity (polarity) classifier for English. Our second set of experiments is conducted to measure the performance of a linear SVM classifier for subjectivity analysis on the Urdu newswire data. The data set used for training is the pruned data set

obtained after applying the co-training technique. The features used and the performance of the model with each feature is documented in table 6.

Labels	R %	P %	IF %	AF %
Unigrams + POS				64.2
1	40.67	71.1	**51.75**	
-1	88.29	67.74	76.67	
Unigrams + POS + Patterns				65.68
1	43.22	72.34	**54.11**	
-1	88.29	68.69	77.26	
Unigrams + POS + Patterns + emotion words				**67.31**
1	48.31	70.81	**57.43**	
-1	85.88	70.09	77.19	

Table 6: SVM classifier on Urdu newswire data

In order to provide a better understanding of the power of the VSM technique, we applied this model on the IMDB data set. The training data consists of 4000 positive (subjective) and 4000 negative (objective) samples. Since the data set is already balanced, we skip the co-training method. Our aim here is to test the working of VSM classifier. The test set consists of 1000 positive and 1000 negative samples. The classification result on this data set is shown in table 5. The results are comparable to the state-of-the-art performance of English subjectivity classifier that uses SVM (Wiebe *et al.,* 2005).

Labels	R %	P %	IF %	AF %
Balanced training				**78.01**
1	64	90.57	**75**	
-1	93.18	71.68	81.03	

Table 5: VSM approach on IMDB data set

6 Analysis of results

In this work, experiments were conducted using two different classification approaches; 1. VSM based 2. SVM based. Results in table 4 indicate that the VSM technique when combined with the modified boost factor (confidence measure) can be a very powerful technique for sentence level classification tasks. When model vectors were constructed using the entire training set (highly unbalanced), the performance was at 62% F-Measure with the subjectivity detection rate of 70.85%. Post co-training, using the modified model vectors obtained from the pruned data set generated better scores. The increase in the recall of negative class and the increase in the overall F-Measure can be attributed to (i) increase in the positive samples (~likely positive set), and (ii) cleaner negative set (no near positive samples).

The results in table 6 for the SVM classifier also indicate the benefits of co-training. The subjectivity classification performance show positive improvement. Although the performance of the SVM model is not as good as the VSM model, addition of each feature shows an improvement in the subjectivity recognition rate. This performance indicates that the feature sets explored definitely contain positive information necessary for accurate detection.

The poor performance of SVM (over VSM) can be attributed to 1. lack of balanced data for training a traditional SVM model and, 2. small number of positive samples. In VSM the problem of unbalanced data set in a way is overcome by using the confidence score at the time of calculating similarity. If these factors are compensated, the performance of the SVM model will significantly improve.

7 Conclusion

This research provides interesting insights in modeling a subjectivity classifier for Urdu newswire data. We show that despite Urdu being a resource poor language, techniques like co-training and statistical techniques based on *tf-idf* and word unigrams coupled with confidence measures help model the state-of-the-art subjectivity classifier. We demonstrate the power of the co-training technique in generating likely negative and positive sets. The number of near subjective samples in the likely positive set suggests that this method can be used as an adaptive learning technique to enable the annotators produce more samples. For a task like emotion detection, that requires fine grained analysis, sentences need to be analyzed at the semantic level and this goes beyond simple keyword based approach. Our efforts are now concentrated in this direction.

References

Agrawal R, Srikant R. 1994. Fast Algorithms for Mining Association Rules. *In Proc. Of the Intl. Conf on Very Large databases.* Santiago, Chile. Sept. Pp. 478-499.

Banea, C., Mihalcea, R., Wiebe, J., and Hassan, S. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of EMNLP-2008.*

Banfield, A. 1982. *Unspeakable Sentences*. Routledge and Kegan Paul, Boston.

Blum, A. and Mitchell, T. 1998. Combining labeled and unlabeled data with co-training. *Proceedings of the eleventh annual conference on Computational learning theory, ACM. p. 100.*

Brank, J., Grobelnik, M., Milic-Frayling, N., and Mladenic, D. 2003. Training text classifiers with SVM on very few positive examples. *Technical Report* MSR-TR-2003-34, Microsoft Corp.

Chan, Philip K. and Stolfo J. Salvatore. 1998. Toward Scalable Learning with Non-Uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection. *Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining* (KDD-98), August 27–31, 1998, New York City, New York, USA, pp. 164–168. AAAI Press.

Danisman, T., and Alpkocak, A. 2008. Feeler: Emotion Classification of Text Using Vector Space Model. *AISB 2008 Convention Communication, Interaction and Social Intelligence, p. 53.*

Das, A., and Bandyopadhyay, S. 2009. Subjectivity Detection in English and Bengali: A CRF-based Approach. *Seventh International Conference on Natural Language Processing* (ICON 2009), December. Hyderabad, India.

Japkowicz Nathalie. 2000. Learning from Imbalanced Data Sets: A Comparison of Various Strategies. *In Nathalie Japkowicz (ed.), Learning from Imbalanced Data Sets*: Papers from the AAAI Workshop (Austin, Texas, Monday, July 31, 2000), AAAI Press, Technical Report WS-00-05, pp. 10–15.

Kennedy, A, & Inkpen, D. 2005. *Sentiment classification of movie and product reviews using contextual valence shifters*. In Workshop on the analysis of informal and formal information exchange during negotiations (FINEXIN 2005)

Ku, L. W., Liang, Y. T., and Chen, H. H. 2006. Opinion extraction, summarization and tracking in news and blog corpora. In *Proceedings of AAAI-2006*.

Kubat, Miroslav and Matwin Stan. 1997. Addressing the curse of imbalanced training sets: one-sided selection. *Proc. 14th ICML*, Nashville, Tennessee, USA, July 8–12, 1997, pp. 179–186.

Liu, B., Hu, M., and Cheng, J. 2005. Opinion observer: Analyzing and comparing opinions on the web. *In Proceedings of WWW-2005*.

Luo, N., Yuan, F., and Zuo, W. 2008. Using CoTraining and Semantic Feature Extraction for Positive and Unlabeled Text Classification. *International Seminar on Future Information Technology and Management Engineering*.

Mihalcea, R., Banea, C., and Wiebe, J. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of ACL-2007*.

Mukund, S., and Srihari, R.K., 2009. NE Tagging for Urdu based on Bootstrap POS Learning. *Third International Workshop on Cross Lingual Information Access*: Addressing the Information Need of Multilingual Societies (CLIAWS3), NAACL - 2009, Boulder, CO.

Pang, B., Lee, L., and Vaithyanathan, S. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on EMNLP*, pages 79–86.

Riloff, E., Wiebe, J., and Wilson, T. 2003. Learning subjective nouns using extraction pattern bootstrapping. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003* - Volume 4, Edmonton, Canada: Association for Computational Linguistics, pp. 25-32.

Soucy, P., and Mineau, G. W. 2005. Beyond tfidf weighting for text categorization in the vector space model. *International Joint Conference on Artificial Intelligence*, Citeseer, p. 1130.

Su, F., and Markert. K. 2008. From words to senses: a case study of subjectivity recognition. *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, ACL*, pp. 825-832.

Titov, I., and McDonald, R. 2008. A joint model of text and aspect ratings for sentiment summarization. *In Proceedings of ACL-08:HLT*.

Wan, X. 2008. Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. In *Proceedings of EMNLP-2008*.

Wan, X. 2009. Co-Training for Cross-Lingual Sentiment Classification. *In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Association for Computational Linguistics, pp. 235-243.

Wiebe, J. 1994. *Tracking point of view in narrative*. Computational Linguistics, 20(2):233-287.

Weibe, J., Bruce, R., and O'Hara, T. 1999. Development and use of a gold standard data set for subjectivity classifications. *In Proc. 37th Annual Meeting of the Assoc. for Computational Linguistics (ACL-99).*

Wiebe, J., and Riloff, E. 2005. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. *Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics*.

Wiebe, J., Wilson, T., and Cardie, C. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, volume 39, issue 2-3, pp. 165-210.

Wilson, B. Edward. 1927. *Probable Inference, the Law of Succession, and Statistical Inference*. Journal of the American Statistical Association, Vol. 22, No. 158 (Jun., 1927), pp. 209-212.