Aspect Based Sentiment Analysis using Sentiment Flow with Local and Non-local Neighbor Information

Shubham Pateria
Samsung R&D Institute
Bangalore, India
s.pateria@samsung.com

Abstract

Aspect-level analysis of sentiments contained in a review text is important to reveal a detailed picture of consumer opinions. While a plethora of methods have been traditionally employed for this task, majority focus has been on analyzing only aspect-centered local information. However, incorporating information from non-local neighbor aspects may capture richer context and enhance sentiment prediction. This may especially be helpful to resolve poor prediction due to ambiguities in review text. The context around an aspect can be incorporated using semantic relations within text and inter-label dependencies in the output. On the output side, this becomes a structured prediction task. However, non-local label correlations are computationally heavy and intractable to infer for structured prediction models like Conditional Random Fields (CRF). Moreover, some prior intuition is required to incorporate non-local context. Thus, inspired by previous research on multi-stage prediction[1], we propose a two-level model for aspect-based analysis. The proposed model uses predicted probability estimates from first level to incorporate neighbor information in the second level. The model is evaluated on data taken from SemEval Workshops and Bing Liu's review collection. It shows comparatively better performance against few existing methods. Overall, we get prediction accuracy in a range of 83-88% and almost 3-4 point increment against baseline (first level only) scores.

1 Introduction

The voice of consumer is growing stronger. With numerous platforms now available for providing reviews, consumers find it easy to share their opinions and sentiments about a product, service or other subjects. Thus, it becomes essential to analyze such reviews in order to identify consumers preferences and grievances. Sentiment analysis for consumer reviews (or general text) is a prominent research area. Such analysis can be done on various levels - global (collection of text), sentence-level (where sentiment is assigned to one full sentence) or aspect-based. In Aspect Based Sentiment Analysis (ABSA), the problem of interest is to estimate sentiment associated with a specific aspect within a review. An example is given below,

Example 1. The **movie** had a $brilliant_+$ story. The **location** was $awesome_+$ and I must highly $praise_+$ the **camera-work**. However, I have to differ regarding the **acting**. It is hard to comprehend XYZ's **style** in such a role. Her on-screen **presence** is not the usual; I have to say, her **act** left me in a very bad_- mood. The rest of the **cast** was ok_o...$average_o$ at best.

Here, the text in bold marks aspect and italic text marks sentiment-indicators. Also, (+,-,o) underscore notations indicate positive, negative and neutral sentiments, respectively. Usually, there can be one or several aspects within a single sentence. The sentiment associated with any aspect can mostly be inferred by checking the terms associated with it (e.g., *awesome* - **location**). However, this may not be very

[1] (Krishnan and Manning, 2006; Hoefel and Elkan, 2008)

beneficial if the statement is ambiguous. For e.g., *'I have to differ regarding the acting'* does not clearly indicate any sentiment on its own. Such ambiguities may be found frequently in reviews. Due to varied styles of different review writers, use of uncommon (even obscure) terms or phrases, terms or phrases conveying conflicting sentiments, or even due to limited data, a prediction system may be expected to mis-classify sentiments. One way to identify such ambiguities is by using prediction confidence scores (discussed in Section 4.2).

While addressing ambiguities, it may be assumed that there generally is some inherent *flow* in the sentiments. A review may have elements of discourse, such that, discourse-markers[2] like 'and', 'also', 'but' etc. can be used to identify sentiment flow or transition. Such terms may or may not be explicitly used, but presence of flow can be assumed. This idea of flow is not new. Recently, *Sentiment Flows* have been studied by Wachsmuth et al. (2015). They incorporated flow information while predicting global sentiments and also identified some frequent types of flows (Wachsmuth et al., 2014). Analysis using sentiment flow requires neighbor information. Here, a neighbor can be local or non-local. In Example 1, initial sentiment flow is positive. Then, the flow is broken by *However* and the sentences that follow are ambiguous. Towards the end, *very bad mood* sets a negative polarity towards **act**. For multi-class classification, it is difficult to predict sentiment associated with **acting** just from information of its local neighbors (**style** and **camera-work**). It is important to incorporate distant aspect **act**'s sentiment to approximately predict the negative shift in mood. This task can become more complex with more involved semantics (such as in reviews by expert critics).

The neighbor-dependencies can be holistically modeled by also considering correlations among polarity labels, thus making ABSA a structured prediction task. Previously, modified version of Condition Random Field (CRF) classifier has been proposed to predict local sentiments (Mao and Lebanon, 2006). However, while CRFs perform well for local dependencies, they may not be very suitable in standard form for ABSA after we consider non-local neighbors as well because inference over long-range would be expensive (Kazama and Torisawa, 2007; Krishnan and Manning, 2006). Moreover, we believe it would also be beneficial to incorporate textual terms surrounding local and non-local neighbors as input features. This would be difficult without some pre-intuition about non-local neighbor sentiments, lest the the input representation itself becomes complex.

To address these issues, we propose a two-level model for ABSA. The proposed model first performs classification using a baseline set of features. Based on this, the probability estimates are obtained which give indication about ambiguities, as well as, preliminary information about neighbor sentiments. Another classifier on top of this uses the local and non-local neighbor information (first-stage probabilities as well as textual terms) for prediction. In this paper, we discuss a preliminary work on this model. The rest of the paper is organized as follows. The structure used for internal representation of reviews is discussed in Section 3. The classification model using SVM classifier in both stages is discussed in Section 4. Further, in order to test a linear-chain CRF at second stage of our model, an independent experiment is performed using available CRF software. The CRF experiment uses different setup from SVM+SVM model and thus it is not meant for comparison with SVMs, but for independent evaluation. This is briefly discussed in sub-section 4.5 . An evaluation of the models is discussed in Section 5.

2 Related Works in ABSA

Three major steps in ABSA are aspect-term extraction, category detection and polarity estimation. There have been significant amount of work in these areas. Major work related to ABSA has appeared in SemEval Workshops[3]. Some notable contributions in these workshops discuss good practical methods for aspect and category extraction (Brun et al., 2016; Khalil et al., 2016; Toh et al., 2016; Saias, 2015). A lot of work has been done on aspect extraction, however, we would like to focus our discussion on sentiment prediction. The basic form of sentiment analysis at sentence-level or aspect-level uses local context of an aspect for input feature representation. Some notable works include that by Nakagawa et al. (2010) who use dependency-tree structures to model local word interactions; Choi and Cardie (2008)

[2] https://en.wikipedia.org/wiki/Discourse_marker

[3] http://alt.qcri.org/semeval2016/task5/, http://alt.qcri.org/semeval2015/task12/

apply inference rules for polarity reversals. Moreover, deep learning methods have also been explored for aspect-level (Wang et al., 2015) and sentence-level (Socher et al., 2013) analysis by exploiting vector representations of aspect-related terms. Discourse information has been very much favored to expand the context around sentiment targets. Discourse-based analysis has been profoundly covered in some previous works (Somasundaran et al., 2009a; Somasundaran et al., 2008; Somasundaran et al., 2009b; Mukherjee et al., 2012). These works cover different types of discourse relations, in detail, for sentiment analysis. Polanyi and Zaenen (2006) discuss valence-shift over sentences due to discourse markers. It is also natural to consider discourse for sentiment flow, as will be discussed later in this paper. Discourse has been used to model neighbor relations as well. Pang and Lee (2004) have explored the consistency of sentiment between neighbors. Also, Zhou et al. (2011) use the sentiment consistency or contrast as constraint on polarity assigned to neighbors. Lazaridou et al. (2013) encoded discourse relations into their supervised classifier's input features. Similar techniques are used in our model, however, for both local and non-local context. Apart from relational structures within text, it is also beneficial to model correlation among polarity labels. An important work in this direction is by Mao and Lebnon (2006) who introduced a modified CRF model to predict ordinal polarity labels. Wachsmuth et al. (2015) follow this work and discuss sentiment flow adaptability across domains. They also discuss ideas to constraint the label sequence lengths. The grouping method discussed later in this paper is inspired by their work. However, these works are still constrained to correlations between adjacent labels only. Some of the notable and relevant work exploring long distance (non-local) information are: work by Somasundaram et al. (2008) on opinion target relations and its application as constraint on predicted sentiments; work by Zhang et al. (2013) on using discourse relations as constraints for Markov Logic Network; and of special interest is the work by Yang and Cardie (2014) who incorporate discourse and coreference constraints into Posterior Regularization (PR) of a CRF. The works discussed above use some form of discourse or coreference relations for feature embedding or inference constraints. However, while we use discourse markers to separate aspect context, we do not restrict neighbor feature embedding based on discourse only. Instead, we propose a two-level model with a base level prediction from which a probability distribution over sentiment labels can be obtained. This serves as an intermediate intuition about sentiments. Using this, at second level, we sample important non-local neighbor units to embed non-local context into input features. Thus, such sampling is not necessarily restricted to presence of coreference or discourse. However, following the work of Yang and Cardie (2014), it would be interesting to explore the use of CRF at second level with PR constraints involving base level probabilities.

3 Review Representation

In this section, we discuss the structure of a review in the form of a graph of aspect-centered nodes.

3.1 Review Structure

A review can be modeled as a non-directed graph of connected aspects and sentiment nodes. The graph is depicted in Figure 1. For simplicity, the sentiment values are not shown as separate nodes but included within the aspect nodes. We define following attributes of the review graph:

- Aspect-Units (or Units): A Unit U is a node in review graph, and the basic entity which bundles the parameters associated with an aspect. Formation of units is discussed in Section 3.2

  U : (*aspect-terms*; *related-terms*; *sentiment information*)

- Groups: A group G is a cluster of continuous units with similar sentiment. In Example 1, under the assumption of sentiment flow, "The movie had a $brilliant_+$ story", "The location was $awesome_+$" and "I must highly $praise_+$ the camera-work" can be grouped under positive sentiment category. Similar grouping applies to other sentences or phrases as well. Consider the polarity sequence in Example 1, PP**P**-XXX-**NO**O, where P is positive label, N is negative, O is neutral and X ambiguous. After the grouping mentioned above, we consider that labels marked in bold can represent the polarity of a group of same sentiments. Henceforth, these will be referred to as *terminal labels*. A group G contains information about the cluster of units as,

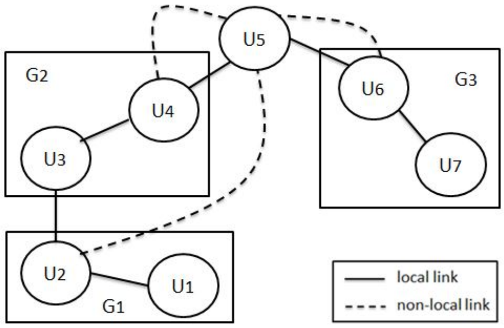


Figure 1: Review Structure around unit U_5.

Type	Markers
Additive	and, or, also, therefore, furthermore, consequently, thus, as a result, hence, subsequently, eventually, in addition, additionally, moreover, as well as
Contrast	though, although, however, but, despite, yet, still, nonetheless, nevertheless, in spite

Figure 2: Common discourse markers with shallow categorization.

G (*polarity*; *count*; *distance*)

Thus, polarity sequence in Example 1 can be reduced to this form as G(P;3;1)-G(X;3;0)-G(N;1;1)-G(O;2;2). Here, distance is taken from ambiguous group.

- Links: A Link *L* is a connection between two units or groups or a unit and a group. A link may or may not contain information about the sentiment flow. Three types of links are used:
  L{+} : Additive links. Terms like 'and', 'Moreover', 'Also' etc. make L{+} links.
  L{c} : Contrast links. Terms like 'but', 'however' etc. make L{c} links.
  L{x} : Undefined links, where a clear connection between two units may not exist. However, sentiment flow (or transition) can still exist. For instance, in Example 1, there no clear link between "**act** left me in a very bad_- mood" and "The rest of the **cast** was ok_o" This would be considered as L{x}.

3.2 Aspect Unit Formation

In order to form a unit, the terms in a review related to an aspect are extracted using parse relations. We use Stanford Parser (Marneffe et al., 2006) for this purpose. Consider following sentence : *The rice was of fine quality but very salty. Also, dessert was a bit expensive. However, service was brilliant.* A selective approximate parsing is shown in Figure 3. Here. *but* is an internal connector. *Also* and *However* are terminal connectors. Using the dependency relations, the units can be formed as follows: U_1 : {The rice was of fine quality but very salty}, U_2 : {dessert was a bit expensive}, U_3 : {service was brilliant}, with connections as: U_1–**(Also)**–U_2–**(However)**–U_3. *Also* and *However* become part of the Links.

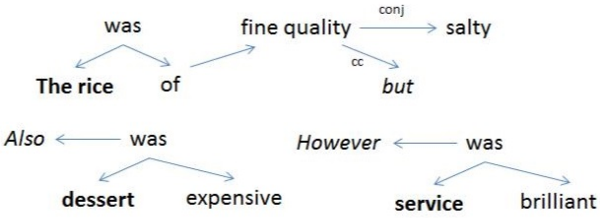


Figure 3: Approx. parsing.

4 Classification Model

The overall sentiment prediction process is divided into 2 stages: Base Prediction and Level-2 Prediction. Consider a feature set described for each aspect unit as $\phi(U_i)$ = (ϕ_1, ϕ_2...ϕ_m), where U_i is the i-th aspect unit in a sequence and any ϕ_j is a feature. Thus, input consists of a set ϕ = { $\phi(U_1)$, $\phi(U_2)$,...,$\phi(U_N)$

}. The aim of first stage or Level-1 prediction is to obtain an intermediate set of sentiment labels $\hat{P}$ = { $\hat{P}(U_1)$, $\hat{P}(U_2)$...$\hat{P}(U_N)$ } along with probability estimates for each. These predictions are used to form groups and sample neighbor information to be added to each input vector $\phi(U_i)$ for final prediction. We discuss our model with-respect-to 3 class classification (positive, negative and neutral) below.

4.1 Baseline Features

These are aspect-centered features formed using text surrounding a given aspect term(s).

Sentiment Scores : The scores of sentiment-indicator terms are aggregated into feature sets,
$\phi_{1k}(U_i) : (score_{positive}, score_{negative}, score_{neutral})$

Here, *k* indicates *kth* type of score-set. We use five score-set types (3 from lexicon corpus + 1 keyword-based + 1 neutral terms) as discussed below:

Sentiment Lexicons from external corpus: Bing Liu's lexicons (Bing Liu, 2012), MPQA subjectivity clues (Wiebe et al., 2005) and SentiWordnet (Stefano et al., 2010) lexicon corpus are used to obtain scores. Bing Liu's and MPQA corpus provide binary scores (positive: 1, negative: 0). These are used as binary features. SentiWordnet provides a range of scores for positive and negative categories.

Category Keywords: Apart from lexicons available in the external corpus, there may be terms which convey sentiments relative to categories. For e.g., *the acting was cheap* conveys negative sentiment while *the price was cheap* is positive despite the same term *cheap*. Such keywords are extracted by dividing the review data into category-specific documents and obtaining TF-IDF scores to identify frequent keywords and corresponding sentiment types. Frequency thresholds of min:0.3 & max:0.8 are set, based on best performance in our experiment.

Neutral terms: Several terms which occur in neutral sentences are not scored in external corpus. These are extracted by identifying frequent words or bi-grams in a collection of neutral sentences. The most frequent ones used in this paper are: 'average', 'normal', 'simple', 'okay', 'ok', 'not great', 'nothing great', 'mediocre', 'not good', 'decent', 'as expected', 'reasonable', 'moderate', 'typical', 'alright', 'fair'.

The score assigned to each sentiment-indicator is also subject to negation. In case of negation, binary scores are simply reversed. For SentiWordnet scores, negation is made in proportion to the scores as: pos = pos + $\frac{(neg-pos)}{2}$ and neg = neg + $\frac{(pos-neg)}{2}$. Here, pos and neg are positive and negative scores, respectively. A significant work on negation problem has been done by Zhu et al. (2014).Moreover, a unit may contain multiple sentiment terms. Thus, the scores are aggregated and normalized. However, as discussed in Section 3.2, the terms within an aspect-unit may be connected by discourse markers. The simplest strategy that can be used is to weigh the sentiment-indicators' scores according to their position in a unit and their relation with discourse markers (Mukherjee et al., 2012).

Bi-grams formed using terms in a unit. Bi-grams around negation terms are taken separately.
$\phi_{21}(U_i)$: *(bi-grams around negation)*, $\phi_{22}(U_i)$: *(other bi-grams)*,

A binary feature for **aspect-category** type can be used (if category has been extracted). This feature has minor effect (Table 3).
$\phi_3(U_i) : (category\ type)$

Local Context window: sentiment score features from previous and next aspect units.
$\phi_4(U_i)$: $\{\phi_{11}...\phi_{1n}\}(U_{i-1})$ and $\{\phi_{11}...\phi_{1n}\}(U_{i+1})$ (n=5 in this case)

4.2 Level-1 or Base Prediction Model

Base prediction is performed using feature set $\{\phi_{11}...\phi_{1n}, \phi_{21}, \phi_{22}, \phi_3, \phi_4\}$ (n=5 in this case). The distribution of features is non-linear as well as high-dimensional and Support Vector Machine (SVM) classifier with Radial Basis Function (RBF)[4] kernel is well-suited for this task due to its high-dimensional mapping and good margin. We use classifier from scikit-learn SVM (SVC) package[5]. A set of primary prediction labels $\hat{P}$ = { $\hat{P}(U_1)$, $\hat{P}(U_2)$...$\hat{P}(U_N)$ } (N = no. of aspects in a single review) is obtained from the base model using confidence scores over the *c* classes (c=3 in this case).

Confidence Scores or Probabilities

The scikit-learn SVM package provides two methods to obtain such scores[6]. First is the *predict_proba* function which provides probability distribution over different classes based on multi-class variant for Platt Scaling (Wu et al., 2004). Second is the *decision_function* which indicates the distance of input points from the hyperplane (or decision boundary). The prediction method (*predict*) of SVM uses *decision_function*. Platt Scaling based estimation may cause disagreement between outcome of *predict* function and the obtained $\arg\max$ (*predict_proba*). However, in the experiments, we found that $\arg\max$ (*predict_proba*) always corresponds to true class label when prediction is strong (high probability assigned to one class). When the classifier fails, the probability values across different classes have small separation (section 4.3). Thus, we stick to Platt Scaling (*predict_proba*) and obtain following:

proba(U_i) : $\{l_1, l_2, ..., l_c\}$, gives probability distribution (summing to 1) over *c* classes, such that,

$\hat{P}(U_i)$ = $\arg\max$ (*proba*(U_i))

4.3 Ambiguity Criteria

As discussed in Section 4.2, *proba* values are used as confidence scores for base predictions. The ambiguous units are identified using *proba* by detecting low difference between any two probability values (low confidence). Following criteria is used for detection,

```
∀ ($l_q$, $l_r$) ∈ proba($U_i$), where q ≠ r,
    if |$l_q$ - $l_r$| ≤ T1 and ($l_q$ > T2 or $l_r$ > T2) then
        $U_i$ is ambiguous
```

In our experiment with 3-class classification, we set *T1* = 0.20 and *T2* = 0.33 based on observations made on available data.[7]

4.4 Level-2 Prediction Model

Having obtained *proba* values, the final requirement is to predict polarity label set *P* = {*P*(U_1), *P*(U_2)...*P*(U_N)}. The process of Level-2 model training and prediction are discussed below.

Training

Firstly, in order to incorporate local and non-local neighbor information, the neighbor units are grouped as described in Figure 4. Assignment of ambiguous unit's sentiment to a group is avoided here. This ensures that neighbor information consists of high confidence values during final prediction stage. After grouping, feature set $\mathbf{F}(U_i)$ is produced for a unit U_i as follows:

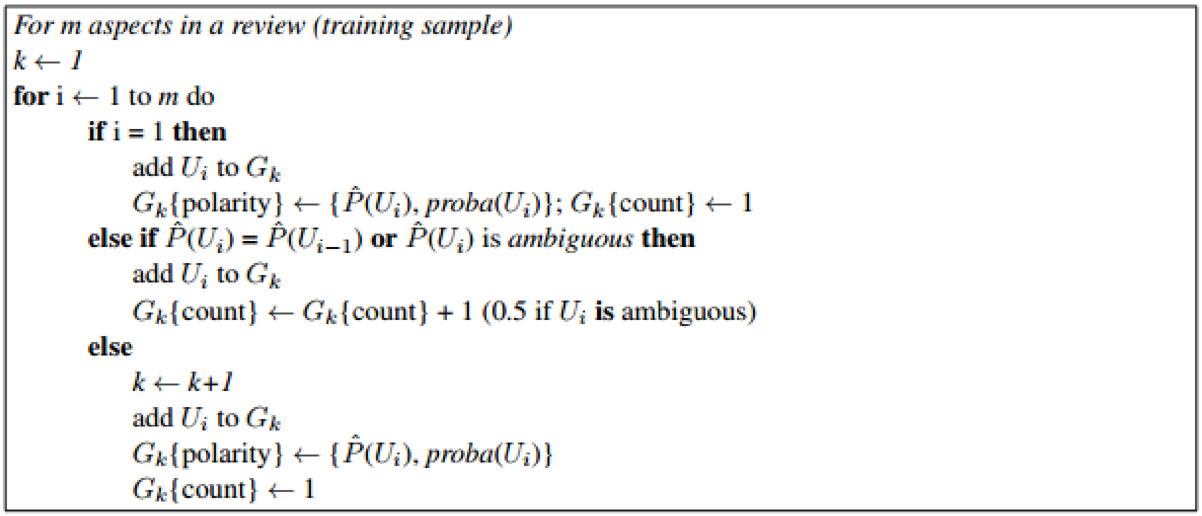
f_1 : list of *G*(polarity) values for max. 3 groups before unit, f_2 : list of *G*(polarity) values for max. 3 groups after unit. If unit U_i lies within *kth* group G_k, then the group is temporarily divided into

[4] http://research.cs.tamu.edu/prism/lectures/pr/pr_l19.pdf

[5] http://scikitlearn.org/stable/modules/generated/sklearn.svm.SVC.html

[6] http://scikit-learn.org/stable/modules/svm.html#scores-probabilities

[7] Since, three polarity classes are used, a homogeneous distribution will allot probability close to 0.33 to each. If either l_q or l_r has value greater than 0.33 (*T2*), assuming third value to be 0.33, then a l_q value of 0.53 will make l_r 0.13. In this case, l_q can be chosen as non-ambiguous and the difference between l_q and third value would be 0.20 (0.53-0.33), set as *T1*.



```
For m aspects in a review (training sample)
k ← 1
for i ← 1 to m do
    if i = 1 then
        add $U_i$ to $G_k$
        $G_k$\{polarity\} ← \{$\hat{P}(U_i)$, $proba(U_i)$\}; $G_k$\{count\} ← 1
    else if $\hat{P}(U_i) = \hat{P}(U_{i-1})$ or $\hat{P}(U_i)$ is ambiguous then
        add $U_i$ to $G_k$
        $G_k$\{count\} ← $G_k$\{count\} + 1 (0.5 if $U_i$ is ambiguous)
    else
        k ← k+1
        add $U_i$ to $G_k$
        $G_k$\{polarity\} ← \{$\hat{P}(U_i)$, $proba(U_i)$\}
        $G_k$\{count\} ← 1
```

Figure 4: Group formation using level-1 predictions.

$\{G_{k1}, U_i, G_{k2}\}$

f_3 : list of G(count) values for max. 3 groups before unit, f_4 : list of G(count) values for max. 3 groups after unit, f_5 : list of distances of max. 3 groups before unit, f_6 : list of distances of max. 3 groups after unit. The orders for these are maintained as per f_1 and f_2.

f_7 : link (type) between U_i and immediately previous group, f_8 : link (type) between U_i and immediately next group,

f_9 : Local feature set $\{\phi_{11}...\phi_{1n}, \phi_{21}, \phi_{22}, \phi_3, \phi_4\}(U_i)$, with ϕ_4 modified as

$\phi_4 = \{\phi_{11}...\phi_{1n}, \phi_{21}, \phi_{22}, \phi_3\}(U_{Terminal})$,

embedding the features of *terminal units* of max. 3 groups before and max. 3 after.

f_{10} : $\alpha(U_i)$, where

$$\alpha(U_i) = \begin{cases} 0 & , if\ U_i\ is\ ambiguous \\ argmax\ (proba(U_i))\ +\ 1 & , otherwise \end{cases}$$

These features are used to train a SVM classifier.

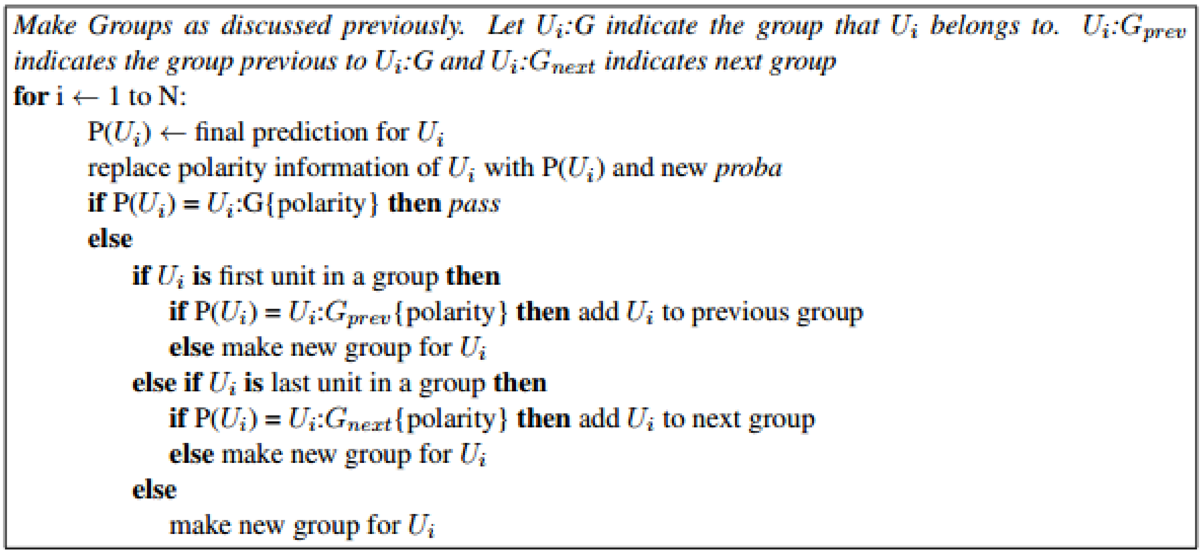
Prediction

The prediction on new data (or evaluation data) is made in a sequential manner (one-by-one). The $\hat{P}$ and *proba* are already available at this stage.

The final output is the required polarity set $P = \{ P(U_1), P(U_2)...P(U_N)\}$.

4.5 Experiment with CRF

In this paper, we have focused on method to incorporate non-local context information into input representation of aspect units. However, such method makes i.i.d assumption for output labels. Under sentiment flow property, there must be correlations between polarity labels as well, both adjacent and long-distance. Devising an efficient structured prediction model using long-distance dependencies is beyond scope of current work and is kept for future. Instead, we experiment with simple linear-chain CRF to get a glimpse into its performance on review text. CRFSUITE (Okazaki, 2007) is used to build a CRF classifier in python. This software provides an internal implementation of linear-chain (first-order Markov) CRF (Sutton and McCallum, 2010). This classifier is used at Level-2 of our model instead of SVM. However, SVM is preferred as base classifier (Level-1) due to its maximum-margin advantage (Hoefel and Elkan, 2008). For CRF (Level-2), features f_1 to f_{10} are used. However, prediction is made over full sequence of output labels and features $\mathbf{F}(U_1)$ to $\mathbf{F}(U_N)$ are fed together. Thus, the grouping is



```
Make Groups as discussed previously. Let $U_i$:G indicate the group that $U_i$ belongs to. $U_i$:$G_{prev}$
indicates the group previous to $U_i$:G and $U_i$:$G_{next}$ indicates next group
for i ← 1 to N:
        P($U_i$) ← final prediction for $U_i$
        replace polarity information of $U_i$ with P($U_i$) and new proba
        if P($U_i$) = $U_i$:G{polarity} then pass
        else
            if $U_i$ is first unit in a group then
                if P($U_i$) = $U_i$:$G_{prev}${polarity} then add $U_i$ to previous group
                else make new group for $U_i$
            else if $U_i$ is last unit in a group then
                if P($U_i$) = $U_i$:$G_{next}${polarity} then add $U_i$ to next group
                else make new group for $U_i$
            else
                make new group for $U_i$
```

Figure 5: Final Prediction.

done only once unlike that described in Figure 5. For CRFUITE settings, LBFGS algorithm is used, with 'max_iterations' equal to 1000.

5 Evaluation

5.1 Experiment Setup

The data for experiment is obtained from SemEval Workshop (2016, 2015) data-sets for ABSA (Pontiki et al., 2016). The data is provided for Restaurant domain in English language and contains labels for aspect-terms, category and polarity. Additionally, data is also obtained from Bing Liu's Consumer Review collection (5 + 9 product data)[8]. This data is for Consumer Electronics (CE) domain, in English language, and has ordinal labels (-3, -2, -1, +1, +2, +3). For 3-class classification, (-1, +1) values are mapped to *neutral*, (+2, +3) to *positive* and remaining to negative class. The data divisions are given in Table 1. Bing Liu's data[9] is divided into 70:30 ratio for training and evaluation. Also, data for number of transitions is given in Table 2. While all reviews are used for experiments, reviews with more than 3 transitions are of special interest to study non-local dependencies. The SemEval - 2016 training data is a mix of SemEval - 2015 training & evaluation data. So, 2015 data is used only for comparison. The data-sets are not balanced; for e.g., in 2016 data, the proportions of pos:neg:neutral instances are 1:1/2:1/15, approximately. Thus, before training, the class weights are balanced in the SVM predictor. Aspect-categories are not provided in Bing Liu's data. Thus, the category related baseline features are dropped for this data.

Two types of training and evaluation (or test) setup are used. In *setup1*, only base model is used. The base model is trained on entire training set after 10-fold cross-validation. Then, predictions made on the test set. In *setup2*, 10-fold cross validation is performed on the training set with base model. However, this time the *proba* values for each validation partition are saved. Finally, the *proba* values for all validation partitions of training data are available, so the Level-2 model is trained on the entire training set. The combined model is then used for predictions on test set. Similar process has been used for multi-stage prediction previously (Krishnan and Manning, 2006)

The system is built using scikit-learn and NLTK (Bird et al., 2009) packages in Python 2.7. Parameters of SVM are set using Grid Search. For our experiments, C=100 and gamma in the range of 0.001 to 0.005 work well (gamma = 0.001 is chosen). RBF kernel is used and decision_function type is one-vs-rest. Before feeding into the model, the data is cleared of stopwords (using NLTK stopword list) and special

[8] https://www.cs.uic.edu/ liub/FBS/sentimentanalysis.html

[9] ipod and powershot files excluded due to low context information

Data	#Aspects	pos	neg	neut
SemEval 2016				
Restaurant - training	2500	1650	750	100
Restaurant - evaluation	859	613	206	40
SemEval 2015				
Restaurant - training	1654	1198	404	53
Restaurant - evaluation	845	457	347	45
Bing Liu's data				
All (5 + 9 products)	3933	2130	1036	767

Table 1: Approximate divisions for review data.

#Terminal Labels	#Reviews
SemEval 2016 data	
> 3	53 (training)
	21 (evaluation)
= 3	71 (training)
	19 (evaluation)
< 3	242 (training)
	53 (evaluation)
SemEval 2015 data	
> 3	29 (training)
	25 (evaluation)
= 3	49 (training)
	20 (evaluation)
< 3	183 (training)
	53 (evaluation)
Bing Liu's data	
> 3	148
= 3	120
< 3	321

Table 2: No. of reviews according to terminal labels (i.e. no. of transitions in polarity).

characters. The data is also lemmatized and all terms converted to lowercase. Also, in order to reduce the size of feature set, only 2000 best bi-grams are selected using Chi-square function[10].

5.2 Results and Discussion

The measure used for performance evaluation is the prediction Accuracy[11]. The results of evaluation are provided in Table 3. The results for *setup1* are listed under 'Base Model' and that for *setup2* are under 'Base + Level-2'. For base model, it can be seen that bi-grams and aggregated sentiment scores are the most significant features. It is to be noted that discourse-based aggregation shows good improvement in accuracy. This is expected because discourse can inherently help in incorporating consistency, contrast or negation of sentiment over a sentence, which is difficult to achieve by simple aggregation rules. We believe that with more detailed use of discourse relations, the performance can be further improved. On top of the base model, the combined 'Base + Level-2' model shows higher accuracy scores. Thus, embedding non-local neighbor features does provide richer context information, thereby also resolving sentiments associated with ambiguous sentences or phrases.

The performance of CRF (Level-2) in our experiment is below SVM (Level-2). This may seem counter-intuitive since a CRF should be able to model inter-label dependencies well. However, we wish to emphasize that the experiment with CRF is not aimed at comparison against SVM, but to check the performance of available CRF tool on review data. Firstly, the crfsuite library used for this experiment uses a linear-chain CRF model. If our intuition about non-local dependency holds, then linear-chain CRF should not be sufficient for a performance much superior than SVM. Secondly, the setup of SVM (level-2) is different from the input and output setup for crfsuite. For crfsuite, the full sequence of aspect unit features is fed as input, and full sequence of output labels is predicted at once. On the other hand, in the SVM (level-2) model we form new groups (or modify existing group information) as new labels are predicted. Thus, the context information provided as input changes as prediction proceeds. This is to include as much context information as possible during prediction. The difference in setup does not necessarily mean that CRF should perform below SVM, but it makes a definite comparison unfeasible. Devising a structured prediction method suited for non-local dependencies among polarity labels is kept as a future work. Such work would be more suitable to perform comparison between structured prediction and discrete prediction (i.e., with inter-label independence assumption).

The comparison of our two-stage model against few previous proposals is given in Table 4. Our

[10] http://scikit-learn.org/stable/modules/feature_selection.html#univariate-feature-selection

[11] http://scikit-learn.org/stable/modules/model_evaluation.html#accuracy-score

Method	SemEval-16	Bing Liu's
Base Model		
Base (SVM)		
Only n-grams	77.43	78.45
Only sentiment scores - external + neutral (simple aggregation)	79.60	79.20
Only sentiment scores - external + neutral (discourse-based aggregation)	81.75	81.05
Only sentiment scores - all	82.25	81.05*
Sentiment scores + n-grams	83.36	81.48
All features	83.44	81.48*
Base (SVM) + Level-2 (SVM)		
All features	**87.30**	**83.90**
Base (SVM) + Level-2 (CRF)	86.01	81.80

Table 3: Approximate accuracy scores. (× *category-specific features dropped for Bing Liu's data.)

Method	Accuracy
SemEval 2015 data	
SENTIUE (Saias, 2015)	78.70
ECNU (Zhang and Lan, 2015)	78.11
Base (SVM) + Level-2 (SVM)	**82.80**
SemEval 2016 data	
XRCE (Brun et al., 2016)	**88.12**
IIT-TUDA (Kumar et al., 2016)	86.73
Base (SVM) + Level-2 (SVM)	87.30

Table 4: Comparison against top-2 systems in previous SemEval workshops.

model performs relatively better for SemEval-2015 data. For SemEval-2016 data, the top-scoring system 'XRCE' shows better result. XRCE uses a feedback mechanism which provides information about feature relevance and cross-validation errors to the Feature Design step. We do not explore or replicate XRCE's design in detail. However, we believe that their feedback mechanism leads to more robust features and results in better accuracy.

5.3 Conclusion and Future Work

Concepts like *Sentiment Flow* and *Discourse relations* are important to address semantics of text for sentiment analysis. Such approach basically concerns with: (1) Incorporating local as well as non-local neighbor information as features and (2) structured prediction of a sequence of polarity labels with some constraint on correlation between non-adjacent labels. This problem needs to be approached in holistic manner. However, under non-locality assumption, the existing methods for structured prediction may become too complex. Moreover, relations like discourse and coreference can also be used to embed non-local context as features. However, it is important to extend this concept towards more data-driven approach. Thus, in this paper, we propose a multi-level model where a probability distribution obtained from first level can be used to incorporate non-local neighbor features, in addition to discourse, for further level of prediction. We show that multi-level model with non-local information can achieve some improvement in aspect-based prediction. The model evaluated on different data-sets performs in the 83-88% accuracy range. Nonetheless, it is important to explore more robust methods in theory and practice. The methods proposed by Kazama and Torisawa (2007), and Collins (2002) provide good basis for this. Also, it would be interesting to explore CRF with Posterior Regularization constraints taken from first level predictions, building upon work by Yang and Cardie (2014). Moreover, prediction can be improved by aggregating sentiment at sentence or phrase-level using discourse markers. The technique used in this paper is rudimentary. In Rhetorical Structure Theory, much intricate discourse relations have been proposed and there have been interesting works in this area exploring richer discourse concepts. Thus, in further work, we would expand the discourse relations used in sentiment aggregation as well as for linking aspect-units.

References

Henning Wachsmuth, Johannes Kiesel, and Benno Stein. 2015. Sentiment Flow–A General Model of Web Review Argumentation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 601–611.

Henning Wachsmuth, Martin Trenkmann, Benno Stein, and Gregor Engels. 2014. Modeling Review Argumentation for Robust Sentiment Analysis. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, 553–564.

Yi Mao and Guy Lebanon. 2006. Isotonic conditional random fields and local sentiment flow. *Advances in neural information processing systems*, 961–968.

Jun'ichi Kazama and Kentaro Torisawa. 2007. A new perceptron algorithm for sequence labeling with non-local features. *Proc. of EMNLP-CoNLL*, Volume 7, 315–324.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. *LREC*.

Swapna Somasundaran, Galileo Namata, Janyce Wiebe, and Lise Getoor. 2009. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, 170–179. Association for Computational Linguistics.

Swapna Somasundaran, Galileo Namata, Lise Getoor, and Janyce Wiebe. 2009. Opinion graphs for polarity and discourse classification. *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, 66–74. Association for Computational Linguistics.

Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. 2008. Discourse level opinion relations: An annotation study. *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, 129–137. Association for Computational Linguistics.

Subhabrata Mukherjee and Pushpak Bhattacharyya. 2012. Sentiment Analysis in Twitter with Lightweight Discourse Analysis. *Proceedings of COLING*, 1847–1864. Association for Computational Linguistics.

Livia Polanyi and Annie Zaenen. 2006. Contextual valence shifters. *Computing attitude and affect in text: Theory and applications*, 1–10.

Bing Liu. 2012. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologiese.*

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *LREC*, 165—210.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *Proceedings of the 9th international conference on Language Resources and Evaluation (LREC).*

Xiaodan Zhu, Hongyu Guo, Saif Mohammad, and Svetlana Kiritchenko. 2014. An Empirical Study on the Effect of Negation Words on Sentiment. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 304–313.

Ting-Fan Wu, Chih-Jen Lin, and Ruby C. Weng. 2004. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, Volume 5, 975–1005.

Maria Pontiki, Dimitris Galanis, and others. 2016. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 19–30. Association for Computational Linguistics.

Michael Collins. 2002. Ranking algorithms for named-entity extraction: Boosting and the voted perceptron. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 489–496. Association for Computational Linguistics.

Caroline Brun, Julien Perez, and Claude Raux. 2016. XRCE at SemEval-2016 Task 5: Feedbacked Ensemble Modeling on Syntactico-Semantic Knowledge for Aspect Based Sentiment Analysis. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 277–281. Association for Computational Linguistics.

Ayush Kumar, Sarah Kohail, Amit Kumar, Asif Ekbal, and Chris Biemann. 2016. IIT-TUDA at SemEval-2016 Task 5: Beyond Sentiment Lexicon: Combining Domain Dependency and Distributional Semantics Features for Aspect Based Sentiment Analysis. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 1129–1135. Association for Computational Linguistics.

Talaat Khalil and Samhaa R. El-Beltagy. 2016. NileTMRG at SemEval-2016 Task 5: Deep Convolutional Neural Networks for Aspect Category and Sentiment Extraction. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 271–276. Association for Computational Linguistics.

Zhiqiang Toh and Jian Su. 2016. NLANGP at SemEval-2016 Task 5: Improving Aspect Based Sentiment Analysis using Neural Network Features. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 282–288. Association for Computational Linguistics.

Bo Wang and Min Liu. 2015. Deep Learning for Aspect-Based Sentiment Analysis. https://cs224d.stanford.edu/reports/WangBo.pdf.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, Volume 1631.

José Saias. 2015. Sentiue: Target and Aspect based Sentiment Analysis in SemEval-2015 Task 12. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 767–771. Association for Computational Linguistics.

Zhihua Zhang and Man Lan. 2015. ECNU: Extracting Effective Features from Multiple Sequential Sentences for Target-dependent Sentiment Analysis in Reviews. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 767–771. Association for Computational Linguistics.

Arjun Mukherjee and Bing Liu. 2012. Aspect extraction through semi-supervised modeling. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, 339–348. Association for Computational Linguistics.

Vijay Krishnan and Christopher D Manning. 2006. An effective two-stage model for exploiting non-local dependencies in named entity recognition. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 1121–1128. Association for Computational Linguistics.

Guilherme Hoefel and Charles Elkan. 2008. Learning a two-stage SVM/CRF sequence classifier. *Proceedings of the 17th ACM conference on Information and knowledge management*, 271–278. Association for Computing Machinery.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of Conditional Random Fields (CRFs). http://www.chokkan.org/software/crfsuite.

Charles Sutton and Andrew McCallum. 2010. An introduction to conditional random fields. arXiv preprint arXiv:1011.4088. homepages.inf.ed.ac.uk/csutton/publications/crftutv2.pdf.

Steven Bird, Edward Loper, and Ewan Klein. 2009. Natural Language Processing with Python. O'Reilly Media Inc.

Tetsuji Nakagawa, Kentaro Inui, and Sadao Kurohashi. 2010. Dependency tree-based sentiment classification using CRFs with hidden variables. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 786-794. Association for Computational Linguistics.

Y. Choi and C. Cardie. 2008. Learning with compositional semantics as structural inference for subsentential sentiment analysis. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 793-801. Association for Computational Linguistics.

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

Lanjun Zhou, Binyang Li, Wei Gao, Zhongyu Wei, and Kam-Fai Wong. 2011. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 162–171. Association for Computational Linguistics.

Angeliki Lazaridou, Ivan Titov, and Caroline Sporleder. 2013. A bayesian model for joint unsupervised induction of sentiment, aspect and discourse representations. *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics*, 1630-1639. Association for Computational Linguistics.

Qi Zhang, Jin Qian, Huan Chen, Jihua Kang, and Xuanjing Huang. 2013. Discourse level explanatory relation extraction from product reviews using first-order logic.

Bishan Yang and Claire Cardie. 2014. Context-aware Learning for Sentence-level Sentiment Analysis with Posterior Regularization. *ACL(1)*, 325-335. Association for Computational Linguistics.