Aspect-Based Opinion Mining from Product Reviews Using Conditional Random Fields

Amani K. Samha, Yuefeng Li, Jinglan Zhang
Science and Engineering Faculty, Queensland University of technology
Brisbane 4000, Queensland, Australia
asamha@gmail.com, {y2.li, Jinglan.zhang} @qut.edu.au

Abstract

Product reviews are the foremost source of information for customers and manufacturers to help them make appropriate purchasing and production decisions. Natural language data is typically very sparse; the most common words are those that do not carry a lot of semantic content, and occurrences of any particular content-bearing word are rare, while co-occurrences of these words are rarer. Mining product aspects, along with corresponding opinions, is essential for Aspect-Based Opinion Mining (ABOM) as a result of the e-commerce revolution. Therefore, the need for automatic mining of reviews has reached a peak. In this work, we deal with ABOM as sequence labelling problem and propose a supervised extraction method to identify product aspects and corresponding opinions. We use Conditional Random Fields (CRFs) to solve the extraction problem and propose a feature function to enhance accuracy. The proposed method is evaluated using two different datasets. We also evaluate the effectiveness of feature function and the optimisation through multiple experiments.

Keywords: Opinion Mining, Customer reviews, Product reviews, Conditional random fields, Feature Function.

1 Introduction

The growth of world-wide web platforms such as social media, forums, blogs and product reviews has led people to post their opinions and benefit from others' past experiences. User-generated reviews have become an exciting reference in most fields, such as business, education and e-commerce, as they contains opinionated information about services and products (Moghaddam, Jamali and Ester 2011). Analysing such information enhances the decision-making process when selling, buying and providing services. In the business world, for example, reviews help to improve the way that services or products are offered and eliminate customer dissatisfaction. Obtaining such information will guarantee that feedback is delivered to the manufacturer or service provider. For potential customers, it creates awareness from others' past experiences and thus enhances the decision-making process. The ability to post reviews is provided by many e-commerce websites, such as Amazon, Yahoo Shopping and eBay, among others, and allows customers to post their opinions freely. While this seems straightforward, the process becomes complicated when there are large numbers of reviews. Therefore, the enormous number of online opinionated customer reviews creates the need for systems to gather important information, analyse it, and extract useful knowledge to ensure that end-users can benefit with minimal effort. Opinion mining is classified into three branches: the document level, which aims to provide an overall opinion; the sentence level, which produces opinions based on the sentence; and the feature level, which examines each feature in the review. This is known as ABOM (Himmat and Salim 2014; Liu 2012; Liu and Zhang 2012).

ABOM, which is the base case study of this work, involves several tasks. First, it aims to efficiently identify and extract product entities, which include the actual product, its components, functionality, attributes and the aspects of the product (Ding, Liu and Zhang 2009). The next task is to find the corresponding opinions for each entity extracted from relevant reviews. Opinions are also known as 'sentiments', which are the adjectives that are given by users to describe the product. A number of researchers have attempted to solve the opinion mining problem using different approaches via supervised, unsupervised and semi-supervised learning. These include rule-based methods (Guo et al. 2009; Hu and Liu 2004a, 2004b; Liu, Hu and Cheng 2005; Moghaddam and Ester 2010), statistical methods (Guo et al. 2009; Wang, Lu and Zhai 2010; Choi and Cardie 2010; Titov and McDonald 2008) and lexicon approaches(Zhao and Li 2009; Noy 2004; Zhang et al. 2011; Taboada et al. 2011; Wogenstein et al. 2013).

In this paper, we study the problem of ABOM as a sequence labelling problem, and propose a computational technique to model ABOM of product reviews. Recent research has shown that the sequence labelling approaches based on conditional relations enhance the accuracy and performance of unstructured prediction problems. There are some proposed models for sequence labelling tasks, such as CRF (Lafferty, McCallum and Pereira 2001), Hidden Markov Models (HMM) (Eddy 1996) and Max-Margin Markov Networks (Roller 2004), among others. These models have shown enormous improvement and considerable success in certain practical tasks, such as natural language processing, pattern recognition and information extraction. We employ a supervised learning approach using the CRF model to identify and extract aspects, as well as extract and map

opinions as a sequence labelling problem. CRF is a class of statistical modelling methods often applied in pattern recognition and machine learning that is used for structured prediction. This is particularly important for opinion mining of product reviews. We propose techniques for selecting the best features for the proposed CRF model and optimising its accuracy.

The goal of our work includes identifying product entities and mapping them to the corresponding opinions along with their orientation as a subjective ABOM pattern, which is represented as the form of a single word or multi-word expressions. CRF is used to encode known relationships between reviewers' opinions and construct consistent interpretations of the reviews. With this approach, CRF predicts the sequence of labels for a given input sequence. Here, the reviews were considered as input sequences and POS tags and opinion tags were used as output labels. The Center for SprogTeknologi (CST) online tagger was used for performing POS tagging and opinion tagging was done manually. To tune and evaluate the CRF model, we trained and tested the model with an annotated dataset, obtained from Hu and Liu (2004a) and Marcińczuk and Janicki (2012). The essentials tasks of POS tagging and opinion tagging are described below.

Unlike other systems that consider a single feature of the entity, in which previous work considers nouns and/or noun phrases to be product aspects, our method attempts to find the best combination of features that makes the word eligible to be a product aspect. These features include, but are not limited to tokenisation, part of speech tagging, chunking, word distance features and position features. Our experimental results confirm the effectiveness and accuracy of the proposed solution. The major contributions of this paper include:

1. Hand annotation of the dataset.
2. Proposing statistical frameworks to automatically find the ABOM pattern by considering linguistics to expand the list of words are that are likely to be product aspects, then mapping the relationships to corresponding opinions, without considering domain knowledge and based only on strict matches.
3. Extracting all possible aspects and opinions and improving the accuracy of aspect and opinion extractions by proposing a technique to select the best feature functions considering three inputs to the CRF model (Labels | Words, POS tagging, Chunking) = $(T| W, P , C)$.
4. Identifying and mapping the relationships and boundaries between product aspects and opinions by combining basic linguistic features and n-grams, where all the comparison were made based on strict matches only.

The rest of the paper is organized as follows: *section 2* explains the related work; *section 3* stated the problem of ABOM, *section 4* describes the design, train and test of the CRF model. *Section 5* is the feature function of the CRF model. *Section 6* is the experiment and error analysis. Finally, *section 7* includes discussion and future work.

2 Related work

As mentioned above, and according to Pang and Lee (2008), the opinion mining task can be classified as follows, based on the extraction task: word/phrase level, sentence level (Wiebe et al. 2004; Moghaddam and Ester 2011; Hu and Liu 2004a) or document level (Turney 2002; Pang, Lee and Vaithyanathan 2002; Yu et al. 2008; Lim et al. 2010; Liu, Hu and Cheng 2005). Many studies on opinion mining have been conducted at the document level, which aims to find the orientation of the review rather the precise likes and dislikes reported. Turney (2002) used point-wise mutual information to calculate the average semantic orientation of the extracted phrases to determine the polarity of the whole document. (Hatzivassiloglou 2000) proposed a statistical supervised method that works by combining dynamic adjectives, semantic oriented adjectives and gradable adjectives as a simple subjective classifier. (Pang and Lee 2002) studied the effectiveness of sentiment classification using machine learning techniques with movie review data. As a result of the general orientation of the whole review, the mining process missed the detail of what likes and dislikes the review contained. To address this problem, more research was conducted at the sentence and phrase levels. The concept of mining aspects and corresponding opinions was first addressed by Hu and Liu (2004) using information extraction techniques and based on aspect frequency. These approaches were useful when associating aspect extraction with the fact that aspects are most commonly nouns. However, such models highlight the limitations of not extracting infrequent aspects and also by the fact that some extracted nouns are not aspects. Proprdue and Et (2005) improved the Hu and Liu's system by developing a system to remove frequent nouns that are not aspects, such that it achieved high precision but low recall; however, this failed to solve the problem of infrequent aspect extraction.

In general, ABOM (Samha, Li and Zhang 2014) comes under phrase-level opinion mining, and aims to produce a detailed sentiment analysis at the aspect level. Vivekanandan and Aravindan (2014) categorized the ABOM approaches into three groups: first, the frequency-based approaches, which are based on frequent aspects of products. These assume that the frequent aspects are more important than non-frequent aspects (Hu and Liu 2004a; Baccianella, Esuli and Sebastiani 2009; Zhuang et al. 2006). Second, the relational-based approaches map relations between aspects and opinions and assume that the closest are more likely to be accurate (Zhuang et al. 2006; Hu and Liu 2004a, 2004b). Finally, the model-based approaches aim to overcome the limitations of the other approaches. Some of the commonly used supervised learning techniques are HMM (Abbasi Moghaddam 2013) and CRF (Qi and Chen 2010; Huang et al. 2012; Jakob and Gurevych 2010; Xu et al. 2010). In this paper we have used CRF and attempted to overcome some of the limitations of other models.

Jin, Ho and Srihari (2009a, 2009b) have considered opinion mining as a sequence labelling problem built under HMM (lexicon-based) using linguistic features. HMM models assume that each feature is generated independently and ignore the underlying relationships

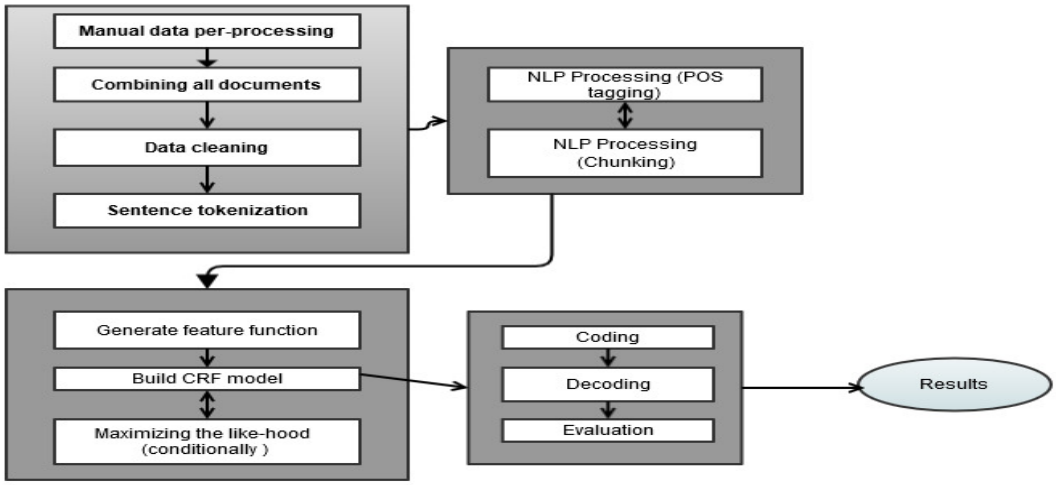


Figure 1: Architecture of the proposed model

between the actual words and labels, as well as the overlapping features (Qi and Chen 2010). CRF overcomes these limitations because it is a discriminative model that models the overlapping dependent features (Peng and McCallum 2006). Choi et al. (2005) view sentiment analysis as a hybrid task information extraction problem that combines CRF as a sequence tagging task and AutoSlog (Riloff 1996) to learn the extraction patterns. Even though their system employs extraction learning with CRF, it showed a recall of 54% with exact match. CRF has been implemented in different languages, linguistic features (Xu et al. 2010) where the strict match performance was around 50%. Here, we have developed a CRF model that can address this problem and extract frequent and infrequent product aspects, along with their corresponding orientations.

3 Aspect-Based Opinion Mining

3.1 Problem Statement

Let $D = \{d_1, d_2, \ldots, d_n\}$ be a set of opinionated documents, where each d_i consists of a set of reviews $R = \{r_1, r_2, \ldots, r_n\}$. Let $S = \{s_1, s_2, \ldots, s_n\}$ be a set of sentences, where each s_i consists of words $W = \{w_1, w_2, \ldots, w_n\}$, the corresponding part of speech tags s $P = \{p_1, p_2, \ldots, p_n\}$, and the corresponding chunking phrases $C = \{c_1, c_2, \ldots, c_n\}$.

3.2 Problem definition

Given a sequence of words, $W = \{w_1, w_2, \ldots, w_n\}$, with the corresponding part of speech tags $P = \{p_1, p_2, \ldots, p_n\}$ and the corresponding chunking phrases for each word $C = \{c_1, c_2, \ldots, c_n\}$. The ABOM task can be defined as a sequence labelling problem. We employ CRF to find the most likely sequence of labels $T = \{t_1, t_2, \ldots, t_n\}$

3.3 The big picture

Figure 1 illustrated the architecture of the whole model and Figure 2 shows an overview picture of the whole model. We started with labelling of the dataset using the tags listed in Table 2. However, data first needs to be pre-processed and cleaned. So all abnormal characters are removed from the text using regular expressions. Then we used the OpenNLP (Baldridge 2005) to detect and split sentences.

After manually labelling the cleaned data, we prepare the dataset for build a CRF model. We use OpenNLP to do the POS tagging and chunking for all words to satisfy the input equation (T|W, POS, Chunking). After that, we train the CRF model with the feature function. Finally, we tested the CRF model and generate results.

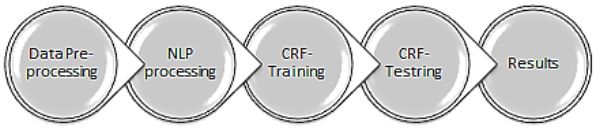


Figure 2: System big picture

3.4 Data Set preparation

3.4.1 Entity Definition

The focus is to define and extract product entities and corresponding opinions then label the training dataset using tags. According to Banitaan et al. (2010) and Glance, Hurst and Tomokiyo (2004) there are different categories of entities (Table 1). However, the broad overview categorises them into four entity groups that represent different types of words in the review text. These four categories are components, functions, features and opinions. As an example, (Table 1) includes an example of entity categories related to the word *'camera'* (Glance, Hurst and Tomokiyo 2004). Some entities may not fit in any categories. Therefore, we can form a fifth category, called *'other'*, and leave it open for any suggested categories that not belong to any of these four entity category.

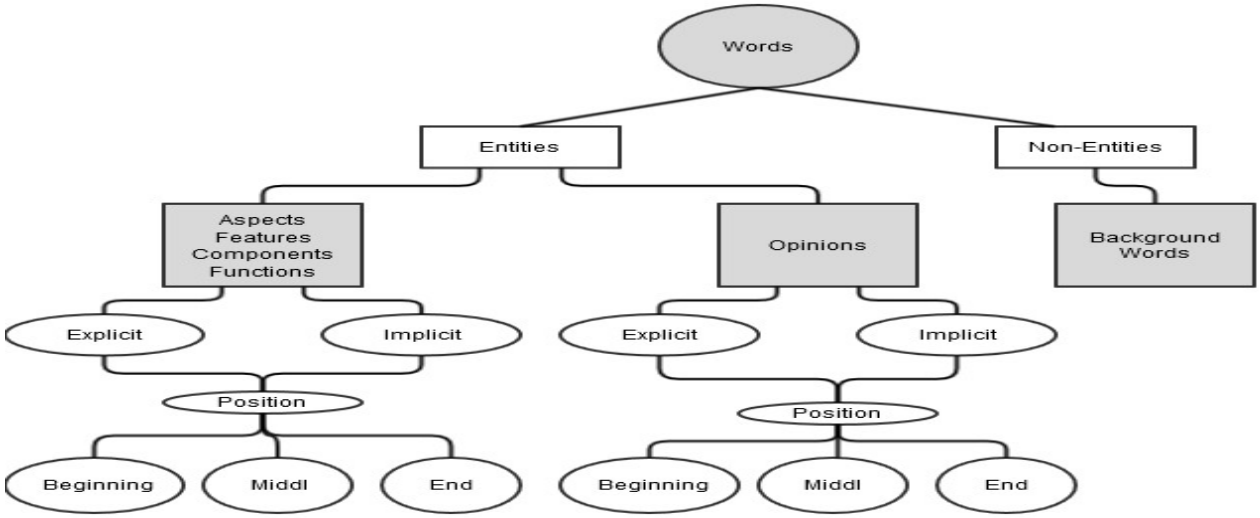


Figure 3: Dataset tagging process

Entity	Description
Components	Physical objects of a camera, including the camera itself, the LCD, viewfinder and battery
Functions	Capabilities provided by a camera, including movie playback, zoom and autofocus
Features	Properties of components or functions, such as colour, speed, size, weight, and clarity
Opinions	Ideas and thoughts expressed by reviewers on the product, its features, components or functions
Other	Other possible entities defined by the domain

Table 1: Entity categories

3.4.2 Pre-processing

Pre-processing is a necessary step, since the dataset is raw and must be prepared for training and then for testing. At this stage, all abnormal characters and HTML tags, such as *<b>, [ ], "",* are removed. Next, all sentences are combined into one single document, and then sentences were detected using OpneNLP tools (Baldridge 2005).

3.4.3 Dataset tagging process

According to the entity definition, this experiment defined five types of tags, where the tags are based on entities, defined in Figure 3, divided into two main categories. The first category is *'Features'*, and includes the product itself, its components, functions, features, attributes and the aspects of the product. Each category is based on its meaning, both explicit and implicit.

Then we used the most positional and represented labels following the Beginning-Middle-End (BME) labelling schema: *B-Target*, identifying the beginning of feature/opinion target; *M-Target*, identifying the middle position of the word, where it may have more than one middle tag. Finally is the *E-Target*, which represents the end position of the word in the sentence.

Tag	Labels	Examples
Background words	(B)	I(B) bought(B) this(B)
Explicit aspect or feature	(Feature_B) (Feature_M) (Feature_E)	to(Feature_B) use(Feature_E)
Implicit aspect or feature	(Feature_B_Imp) (Feature_M_Imp) (Feature_E_Imp)	affordable (Feature_B_Imp)
Positive and negative explicit opinions	(Opinion_B_P/N_Exp) (Opinion_M_P/N_Exp) (Opinion_E_P/N_Exp)	Inexpensive (Opinion_E_P_Exp)
Positive and negative implicit opinions	(Opinion_B_P/N_Imp) (Opinion_M_P/N_Imp) (Opinion_E_P/N_Imp)	real(Opinion_B_P_Imp) buy(Opinion_E_P_Imp)

Table 2: Tags

4 CRF model Design, Train and Test

Product features are mostly nouns or noun phrases; whereas opinions are adjectives or adjectival phrases that are most likely appear closer to the nouns. Natural language is usually a sequence of words that form sentences as a meaningful sequence based on grammatical rules. Therefore, the sequence is a sentence and a word is a primary element of it. There are enormous elements that we can assign to each individual word, such as parts of speech, chunking and more. Therefore, the problem of ABOM can be formulated as a sequence-labelling task. The solution to the sequence-labelling problem is based on natural language processing techniques, where we aim to assign a single label to each element in a sequence. First-order CRF (Lafferty, McCallum and Pereira 2001; McDonald and Pereira 2005; Sutton and McCallum 2006) considers the dependencies between at most three adjacent labels.

CRF was proposed by (Lafferty, McCallum and Pereira 2001). It is a probabilistic method for extracting and labelling sequential data that encode dependencies between different entities of a sequence, and typically outperforms other supervised learning algorithms, such as

support vector machine learning. It has demonstrated high performance in information extraction, particularly in entity recognition (Klinger and Friedrich 2009). CRFs are resolved according to undirected graphical models over sets of random variables. It is formally defines as follows: Let $G = (V, E)$, a considering undirected graph, let$Y = (Y_v)$, $v \in V$ where each node $\in V$ is corresponding to each of the random variables that $\in Y$, and (X, Y) is a CRF illustrated, in (Figure 4). X is a set of variables *'input'* over the observation sequence to be labelled and Y is a set of random variables *'output'* over the corresponding labelling to be predicted. In this paper, the CRF model works as an extraction model that computes the probability of $Y = (T)$, which represents the probability of the sequence of hidden labels to the sequence of input $X = (W, P, C)$, which represents the observed labels, that aims to find the most probable label sequence $Y's$, given an observation sequence in the problem of sequence label modelling. Therefore, we are looking to represent a distribution over a large number of random variables using only local functions requiring only a small number of variables.

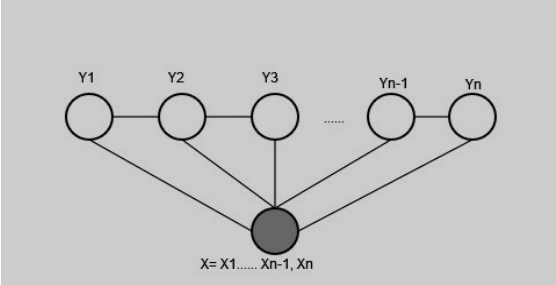


Figure 4: Linear CRF graphical structure

CRF defined as:

$$P(Y \mid X) = \frac{1}{Z(X)} \prod_{j=1}^{n} \psi_j (X, Y)$$

With a normalisation factor:
$Z(x) = \sum y \prod i\epsilon N \phi (y_i, x_i)$ and a feature suction is defined as: f_k as $\phi_i (y_i, x_i) = \exp(\sum_k \lambda_k f_k(y_i, x_i))$. In prediction, we output the most probable label that maximize the likehood of $\hat{Y} = \arg_y MAX P(Y|X)$.

The main task of this paper, is viewing ABOM as a sequential tagging problem which use a set of statistical and natural language features to train the liner-chine CRF. The relation between aspects and opinions are mapped by understanding the syntactic based on observations.

5 Feature Selection

Feature selection is a very important step in any information extraction system. Thus, modelling the perfect subset of features is significant to increase the performance of the ABOM model. In this paper, the extracted features are used to map the relationships between observation labels and hidden labels. Since the product aspects/features are mostly likely to be nouns or/and noun phrases and opinions are most likely to be adjectives or/and adjectival phrases, selecting features is based on natural language processing techniques and a probabilistic language model. These features are divided into two categories:

5.1 Basic features

Basic features are linguistic features. These features are extracted using the OpenNlp toolkit (Baldridge 2005) as follows:

- Token feature f1: This represents the string of the current token in which every word of the text is a token w_i. Tokenisation worked well in Zhang and Liu (2014) and Jakob and Gurevych (2010). In this paper, the token is the value of the actual word of the sentence. It values each token in the sentence by the natural word and the position of the word in the sentence indexed by relative position to the word.
- Part-of-speech tagging feature $f2$ and chunking features f3: are two syntactic features (Marcińczuk and Janicki 2012) that examine the phrase level in depth, considering the token and its surrounding words. f2 is used to classify each $w_n \in W$ into one of a set of tags, such as verbs, nouns or adjectives, while $f3$ is used to classify each $w_n \in W$ to the applicable chunk based on phrases. f2 and f3 are used to map the relationship between product aspects and opinions. Here, we used Part of Speech Tagger from the Open NLP toolkit (Baldridge 2005).
- Chunking feature f3 : text chunking is used to recognise the relatively simple syntactic structure of sentences. POS tagging shows the product aspects at a word level only; however, some product aspects are noun phrases, which are more likely to be nearest to the opinion words (Tjong Kim Sang and Buchholz 2000). For chunking we used the Chunker tools from the OpneNLP toolkit (Baldridge 2005) that was trained on conll2000 (Tjong Kim Sang and Buchholz 2000) shared task data.
- Sentence segmentation feature f4: this feature is used to segment each review into sentences. This feature helps to find the boundaries of the opinionated sentences.

5.2 Advanced features

Advanced features are the basic features mixed with certain statistical features to form rules, as follows:

N-grams features $f5$: since POS tagging and chunking map the synaptic structure of the sentence in a simple way, n-gram was added as a feature, as it performs well in sentiment classification (Pak and Paroubek 2010; Dave, Lawrence and Pennock 2003; Pang, Lee and Vaithyanathan 2002). From this point, we experimented with the best settings usage of unigrams, bigrams, and trigrams and combined them with the basic features, as shown in (Table3) .

- Context features $f6$ considers the token feature$f1$ to obtain contextual information, where the tokens near the target token may indicate its type and to which category it

Item sequences	Attributes	Description
1	w[t-2], w[t-1], w[t], w[t+1], w[t+2],	(5 features of trigram words)
2	w[t-1]|w[t], w[t]|w[t+1],	(2 features of bigram words)
3	pos[t-2], pos[t-1], pos [t], pos [t+1], pos [t+2],	(5 features of trigram POS tagging)
4	pos[t-2]|pos[t-1], pos[t-1]|pos[t], pos[t]|pos[t+1], pos[t+1]|pos[t+2],	(4 features of POS tagging relations (2-order))
5	pos[t-2]|pos[t-1]|pos[t], pos[t-1]|pos[t]|pos[t+1], pos[t]|pos[t+1]|pos[t+2]	(3 features of trigram POS tagging relations (3-order))
6	chunk[t-2], chunk [t-1], chunk [t], chunk [t+1], chunk [t+2],	(5 features of trigram chunk tags)
7	chunk [t-2]|chunk [t-1], chunk [t-1]|chunk [t], chunk [t]|chunk [t+1], chunk [t+1]|chunk [t+2],	(4 features of chunk tagging relations (2-order))
8	chunk [t-2]|chunk [t-1]|chunk [t], chunk[t-1]| chunk [t]|chunk [t+1], chunk [t]|chunk [t+1]| chunk [t+2]	(3 features of trigram chunk tagging relations (3-order))

Table 3: CRF advanced features

belongs. This works using f2 and f3 features as added features to the neighbouring words of different n-grams, where we study the surrounding words in combination with other features, such as n-grams, POS and chunking. Therefore, we formed rules based on observations using f5, as shown in Table 3.

- Position of the word featuref_7: we used tags applicable for the word's position in the sentence, for instance, *_B* 'beginning of sentence', *_M* 'middle of the sentence' and *_E* is 'end of sentence'.

The combination of both feature sets increased the accuracy of the CRF model. Some definitions are necessary to clarify the reading of the features:

- W is the word features: include word at position t-2, t-1, t, t+1, t+2: trigram of words.
- w[t-1]|w[t]: associations between words features: represents the concurrency of bigram of words.
- Corresponding with POS: part of speech tag and chunk: Chunker Tag.

6 Experimental Framework

CRFsuite (Okazaki 2007), a fast implementation of CRF (Lafferty, McCallum and Pereira 2001), was used to train our model. In the training phase, CRFsuite predicted some wrong labels; for instance, the product aspects that we were interested in might be a single word or consist of multi-word strings; however, we needed some scripts to help with the dataset. Therefore, we wrote few Python scripts that aim to align the CRFsuite output tags with the original input labelled file. We then evaluated the work by calculating precision, recall and F-score measures on the actual word, post-tagging and chunking recognition rather than individual words.

Additionally, we divided the actual and predicted aspects into four categories: correct (self-explanatory), missed (actual chunks not identified by the model), wrong label (word sequences that were correctly extracted but wrongly classified), and false positives (self-explanatory) to obtain a more detailed picture.

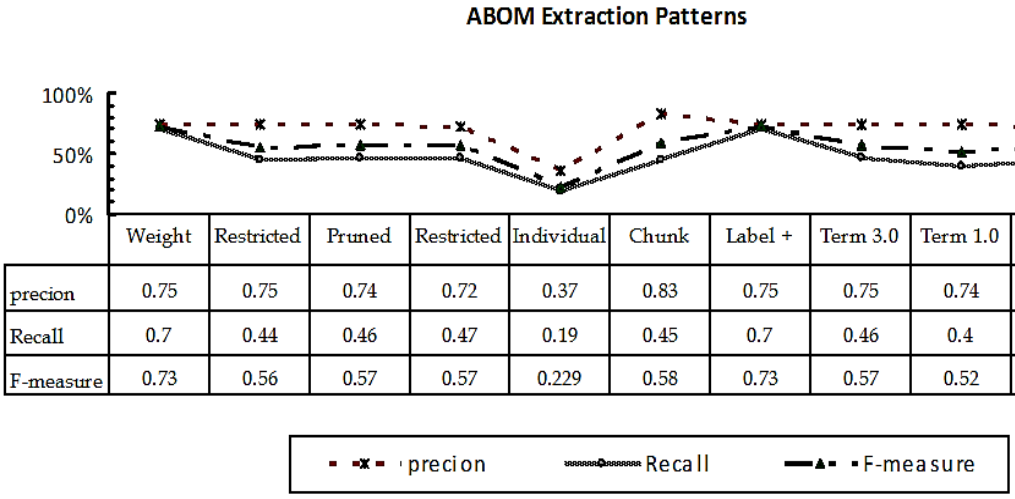
Performance	Individual label assignment	Chunk recognition	Label + POS+ chunk
Precision	0.37	0.83	0.75
Recall	0.19	0.45	0.7
F-measure	0.229	0.58	0.73
Correctly identified chunks	-	0.45	0.50
Missed chunks	-	0.53	0.49
Incorrectly labelled chunks	-	0.01	0.017
False positives	-	0.08	0.212

Table 4: Model label performance

We examined the model's performance by measuring the accuracy at every level of the experiment. We began by measuring the performance of the model using the labels alone, which we consider as the baseline, as shown in (Table 4), where it shows poor performance in general. Due to the limited matched examples in the dataset, rare tags did not occur often enough to generalise from them, especially for context-dependent features. However, the word itself is a suitable predictor of the label. On the other hand, if the model is too heavily trained on words, then it will not be able to make good predictions for words that it has never seen a common occurrence when dealing with natural language data.

We then added the chunking tags to the labels, which improved precision and recall. The performance was improved by using the actual word, POS tags and chunking. From this point, the actual word, the label, the POS tag and the chunking tag were used in model experiments along with several different feature sets.

CRFsuite allows the possibility of providing scaling values for each feature; this value is multiplied by the learned weight when predicting the value of a label, making it possible to adjust the importance of the feature to some degree. We modified the feature extraction script to allow scaling values to weight the value of the word-based features or part-of-speech based features more



	Weight	Restricted	Pruned	Restricted	Individual	Chunk	Label +	Term 3.0	Term 1.0	Term 3.0	Term 2.0
precion	0.75	0.75	0.74	0.72	0.37	0.83	0.75	0.75	0.74	0.73	0.71
Recall	0.7	0.44	0.46	0.47	0.19	0.45	0.7	0.46	0.4	0.45	0.44
F-measure	0.73	0.56	0.57	0.57	0.229	0.58	0.73	0.57	0.52	0.56	0.55

heavily. (Table 5) contains the results of preliminary tests on an 80/20 split of the manually tagged data using several combinations of scale values.

Weights	**Term 3.0 POS 1.0**	**Term 1.0 POS 3.0**	**Term 3.0 POS 2.0**	**Term 2.0 POS 3.0**
Individual tags				
Precision	0.36	0.35	0.35	0.36
Recall	0.20	0.19	0.20	0.20
F-score	0.24	0.23	0.24	0.24
Extraction task				
Precision	0.75	0.74	0.73	0.71
Recall	0.46	0.40	0.45	0.44
F-score	0.57	0.52	0.56	0.55
Correct	0.46	0.40	0.45	0.44
Missed	0.52	0.57	0.52	0.53
Label error	0.00	0.01	0.01	0.02
False positive	0.20	0.20	0.21	0.21

Table 5: 80/20 data with different scales

Performance	**Restricted tags**	**Pruned sentences**	**Restricted tags, pruned sentences**
Individual tags			
Precision	0.44	0.36	0.43
Recall	0.26	0.21	0.27
F-Score	0.30	0.24	0.32
Extraction task			
Precision	0.75	0.74	0.72
Recall	0.44	0.46	0.47
F-Score	0.56	0.57	0.57
Correct	0.44	0.46	0.47
Missed	0.54	0.51	0.51
Label error	0.00	0.02	0.01
False positive	0.19	0.20	0.23

Table 6: Extracted tags results

Because there were so few examples of some of the labels, we tried consolidating some of them. Specifically, we combined the *_M* and *_E* tags into one group *(_M)*. The *_B* tag marks the beginning of an aspect, but we can detect the end when we reach a *B* tag, or a subsequent *_B* tag. This yielded a more marked improvement in performance on individual tag performance, but had little effect on performance in the extraction task. Since the number of background tags is so much larger than the number of aspect tags, we also removed sentences that did not contain opinions and trained the model on the more limited dataset (see Table 6).

10 -fold cross validation of the dataset		
Precision	0.75	0.75
Recall	0.49	0.51
F-score	0.59	0.61
Correctly identified chunks	0.49	0.51
Missed chunks	0.48	0.46
Incorrectly labelled chunks	0.01	0.01
False positives	0.21	0.20

Table 7 10-cross validation

Combining the methods yielded the following results on the 80/20 split (scaling value for term features = 3.0, reduced tag set, pruned sentences).

Tags	**B**	**Feature**	**Opinion**
Weight	**-0.25**	**0.25**	**0.25**
P	0.91	0.72	0.74
R	0.96	0.66	0.71
F	0.94	0.7	0.73
Tags	**B**	**Feature**	**Opinion**
Weight	**0**	**0.5**	**0.5**
P	0.91	0.73	0.75
R	0.97	0.67	0.7
F	0.94	0.7	0.73

Table 8: Extraction task after weighting

The item accuracy for training the CRF of 10-fold cross validation ranged between 83% and 87%.

At this stage, we noticed low recall in extracting aspects and opinion; therefore we tried to balance the labels by giving less weight to the dominant tag and high weight to the features and opinions, where the item accuracy was 2,674/2,812 (0.95%), and the instance accuracy was 111/184 (0.60%).

6.1 Experiment set up and results

In this experiment, we used two datasets of product reviews, in order to train and test the system with different data. One was collected by Qi and Chen (2010) from Yahoo Shopping of different cameras. The other

dataset was collected by Hu and Liu (2004a, 2004b) from Amazon for nine different products. Five random cameras were chosen from both datasets, which gave 1,025 full reviews, consisting of 2,500 opinionated sentences. Then dataset was then tagged using the tag sets described in (Table 2).

For each review, each sentence was hand-labelled, which accumulated of 35,877 terms, with the distribution of labels illustrated in (Figure 3). Words belonging to any product aspects and opinions had *_B, _M*, or *_E*, infixes according to whether they are the first word in a phrase representing the aspect, a word in the middle of the phrase, or the last word in the phrase (some of these tags were combined as described in the methods section). Any word that did not belong to these categories received a background tag, *B*. The distribution of labels is shown in (Table 9).

Most terms were unambiguously associated with a particular label; the average overlap in the sets of terms in all pairs of labels was *3%* and *82%* of the terms received a unique label.

We extracted the following features for each word: the word itself; the part-of-speech of the word; nearby words and part-of-speech tags in a window of configurable size, indexed by relative position to the word. The n-grams of words and parts-of-speech of the length of the window size containing the word; and a *'beginning of sentence'* or *'end of sentence'* tag where applicable. Part-of-speech tagging was done with the default Maxent tagger of the nltk library (trained on the Penn treebank corpus).

6.2 Error Analysis

The error analysis indicated some detected mistakes that we face during experiment. Most of the errors were due to the nature of the data, since it does not following a constant sentence structure, in which case the proposed CRF model would not detect the pattern easily.

Label	No of labels
B	28,541
Feature_B	2,526
Feature_M	359
Feature_E	881
Feature_B_Imp	192
Feature_M_Imp	42
Feature_E_Imp	33
Opinion_B_N_Exp	439
Opinion_M_N_Exp	163
Opinion_E_N_Exp	248
Opinion_B_P_Exp	1,549
Opinion_M_P_Exp	179
Opinion_E_P_Exp	525
Opinion_B_N_Imp	25
Opinion_M_N_Imp	15
Opinion_E_N_Imp	18
Opinion_B_P_Imp	55
Opinion_M_P_Imp	45
Opinion_E_P_Imp	42

Table 9: Distribution of labels

Unsurprisingly, precision was high. Many of the correctly identified aspects occurred many times in the training set (for example, *'camera'* was extracted as a feature in 29 of 34 appearances and *'great'* in 18 of 20). Almost 50 aspects were correctly extracted despite occurring only once.

Most of the items that were missed occurred only once or twice. The highest single number of misses was five out of 34 instances of *'camera'*. The next highest were all four occurrences of *'sensor'* (feature), and three of five occurrences of *'photos'* (feature).

These four chunks were correctly extracted, but assigned the wrong label, such as *user (Opinion_P_Exp / Feature_Imp)* and *quality (Opinion_P_Exp / Feature)*. There seems to be a trend of mis-characterising the polarity of opinions, and perhaps mistaking opinions for some features. The false positives are the most interesting errors, these are some examples: *product (None / Feature_Imp) (1), rechargeable battery (None / Feature) (1), LCD (None / Feature) (1*). Most, if not all, are entirely reasonable extracts. Some of the extractions, especially the features, are clearly mis-tagged in the original dataset: *'large view screen'*, *'rechargeable battery'*, *'wide angle coverage'*, *'viewfinder'*, etc. Camera models, such as *'Kodak camera'* and *'Canon XS'* were also correctly identified.

7 Discussion and Conclusion

In this paper, we have analysed the ABOM problem. We propose a CRF-based method to extract all possible aspects and corresponding opinions in reviews and integrate basic linguistic features with statistical features and combined features. As a result, the model achieves high performance.

We were able to achieve high performance when applying CRFs to opinion mining by the selected feature functions. However, when attempting to improve the performance, this seemed to be determined by the limitations of the dataset rather than the defects of the technique. We had 2,500 sentences, and only 60% of them expressed explicit opinions and features.

We considered using a bootstrapping process to augment our data; however, the performance was not as we expected. We wrote a bootstrapping script that used votes from several models to output sentences where all models agreed, but the danger is that agreement might not be a good indication of correctness in this case. This script would nonetheless be useful in easing the process of manually annotating data, as it would be easier to correct tags than to assign them from scratch. Incorrectly, tagged data is also a problem, particularly when there is a limited opinionated dataset that are manually tagged. The impact of a mistake is much greater since it less likely to be overshadowed by correct instances when there are not many of the latter. Nonetheless, it is clear from the comparison with the baseline use of word frequency that the ability of CRF to exploit context results is definitely helpful. Further work might include adding features based on semantics, as well as improving the quality of the training data by adding more opinionated data. In future work, domain knowledge will be added to the identification process and then integrated with the use of current features to enable more effective features.

8 Acknowledgment

The authors would like to thank Mahnoosh Kholghi for her opinions and recommendations.

9 References

Abbasi Moghaddam, S. (2013): *Aspect-based opinion mining in online reviews*. Applied Sciences, School of Computing Science.

Baccianella, S., Esuli, A. and Sebastiani, F. (2009): Multi-facet rating of product reviews. *Advances in Information Retrieval*, 461–472, Springer.

Baldridge, J. (2005): The opennlp project. http://opennlp.apache.org/index.html. Accessed 2 February 2012.

Banitaan, S., Salem, S., Jin, W. and Aljarah, I. (2010): A formal study of classification techniques on entity discovery and their application to opinion mining. edited, 29–36, ACM.

Choi, Y. and Cardie, C. (2010): Hierarchical sequential learning for extracting opinions and their attributes. *Proc. ACL 2010 Conference Short Papers*, 269–274, ACL.

Choi, Y., Cardie, C., Rilof, E. and Patwardhan, S. (2005): Identifying sources of opinions with conditional random fields and extraction patterns. *Proc. Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Vancouver, Canada, ACL, doi: 10.3115/1220575.1220620.

Dave, K., Lawrence, S. and Pennock, D.M. (2003): Mining the peanut gallery: Opinion extraction and semantic classification of product reviews, 519–528, ACM.

Ding, X., Liu, B. and Zhang, L. (2009): Entity discovery and assignment for opinion mining applications. *Proc. 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Paris, France, ACM, doi: 10.1145/1557019.1557141.

Eddy, S.R. (1996): Hidden Markov models. *Current Opinion in Structural Biology* **6**(3):361–365.

Glance, N., Hurst, M. and Tomokiyo, T. (2004): Blogpulse: Automated trend discovery for weblogs. *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*.

Guo, H., Zhu, H., Guo, Z., Zhang, X.X. and Su, Z. (2009): Product feature categorization with multilevel latent semantic association. *Proc. 18th ACM Conference on Information and Knowledge Management*, 1087–1096, ACM.

Himmat, M. and Salim, N. (2014): Survey on product review sentiment classification and analysis challenges. *Proc. First International Conference on Advanced Data and Information Engineering (DaEng-2013),* 213–222. Herawan, T., Mat Deris, M. and Abawajy, J. (eds). Springer, Singapore.

Hu, M. and Liu, B. (2004a): Mining and summarizing customer reviews. *Proc. 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 168–177, ACM.

Hu, M. and Liu, B. (2004b): Mining opinion features in customer reviews. *AAAI*, 755–760.

Huang, S., Liu, X., Peng, X. and Niu, Z. (2012): Fine-grained product features extraction and categorization in reviews opinion mining. *IEEE 12th International Conference on Data Mining Workshops,* 680–686, IEEE.

Jakob, N. and Gurevych, I. (2010): Extracting opinion targets in a single- and cross-domain setting with conditional random fields. *Proc. 2010 Conference on Empirical Methods in Natural Language Processing*, Cambridge, MA, USA, ACL.

Jin, W., Ho, H.H. and Srihari, R,K. (2009a): A novel lexicalized HMM-based learning framework for web opinion mining. *Proc. 26th Annual International Conference on Machine Learning*, 465–472, Citeseer.

Jin, W., Ho, H.H. and Srihari, R.K. (2009b): OpinionMiner: A novel machine learning system for web opinion mining and extraction. *Proc. 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Paris, France, 1195–1204, ACM.

Klinger, R. and Friedrich, C.M. (2009): Feature subset selection in conditional random fields for named entity recognition. *RANLP*, 185–191.

Lafferty, J., McCallum, A. and Pereira, F.C.N. (2001): Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Lim, E.P., Nguyen, V.A., Jindal, N., Liu, B. and Lauw, H.W. (2010): Detecting product review spammers using rating behaviors. *Proc. 19th ACM International Conference on Information and Knowledge Management*, 939–948, ACM.

Liu, B. (2012): Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies* **5**(1):1–167.

Liu, B. and Zhang, L. (2012): A survey of opinion mining and sentiment analysis. *Mining Text Data*, 415–463.

Liu, B., Hu, M. and Cheng, J. (2005): Opinion observer: Analyzing and comparing opinions on the web. *Proc. 14th International Conference on the World Wide Web*, 342–351, ACM.

Marcińczuk, M. and Janicki, M. (2012): Optimizing CRF-based model for proper name recognition in Polish texts. *Computational Linguistics and Intelligent Text Processing*, 258–269: Springer.

McDonald, R. and Pereira, F. (2005): Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinformatics* **6**(Suppl 1): S6.

Moghaddam, S. and Ester, M. (2010): Opinion digger: An unsupervised opinion miner from unstructured product reviews. *Proc.19th ACM International Conference on Information and Knowledge Management*, 1825–1828, ACM.

Moghaddam, S. and Ester, M. (2011): AQA: Aspect-based opinion question answering. *IEEE 11th International Conference on Data Mining Workshops*, 89–96, IEEE.

Moghaddam, S., Jamali, M. and Ester, M. (2011): Review recommendation: Personalized prediction of the quality of online reviews. *Proc 20th ACM International Conference on Information and Knowledge Management*, 2249–2252, ACM.

Noy, N.F. (2004): Semantic integration: A survey of ontology-based approaches. *SIGMOD Record* **33**(4):65–70. doi: 10.1145/1041410.1041421.

Okazaki, N. (2007): crfsuite. http://www.chokkan.org/software/crfsuite.

Pak, A. and Paroubek, P. (2010): Twitter as a corpus for sentiment analysis and opinion mining. *LREC*, 1320–1326.

Pang, B. and Lee, L. (2008): Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* **2**(1–2):1–135.

Pang, B., Lee, L. and Vaithyanathan, S. (2002): Thumbs up? Sentiment classification using machine learning techniques.' *Proc. ACL-02 Conference on Empirical Methods in Natural Language Processing, Volume 10*, 79–86, ACL.

Peng, F. and McCallum, A. (2006): Information extraction from research papers using conditional random fields. *Information Processing & Management* **42**(4):963–979.

Qi, L. and Chen, L. (2010): A linear-chain CRF-based learning approach for web opinion mining. *Web Information Systems Engineering–WISE 2010*, 128–141, Springer.

Riloff, E. (1996): Automatically generating extraction patterns from untagged text. *Proc. National Conference on Artificial Intelligence*, 1044–1049.

Roller, B., Taskar, C. and Guestrin, D. (2004): Max-margin Markov networks. *Advances in Neural Information Processing Systems* **16**: 25.

Samha, A.K., Li, Y. and Zhang, J. 2014. Aspect-based opinion extraction from customer reviews. *arXiv preprint arXiv:1404.1982*.

Sutton, C. and McCallum, A. (2006): An introduction to conditional random fields for relational learning, vol. 2: *Introduction to statistical relational learning*. MIT Press.

Taboada, M., Brooke J., Tofiloski, M., Voll, K. and Stede, M. (2011): Lexicon-based methods for sentiment analysis. *Computational Linguistics* **37**(2): 267–307.

Titov, I. and McDonald, R. (2008): Modeling online reviews with multi-grain topic models. *Proc. International Conference on the World Wide Web*, 111–120, ACM.

Tjong, K.S., Erik, F. and Buchholz, S. (2000): Introduction to the CoNLL-2000 shared task: Chunking. *Proc. 2nd Workshop on Learning Language in Logic and 4th Conference on Computational Natural Language Learning,* Volume 7, 127–132, ACL.

Turney, P.D. (2002): Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, 417–424, ACL

Vivekanandan, K. and Aravindan, J.S. (2014): Aspect-based opinion mining: A survey. *International Journal of Computer Applications* **106**:

Wang, H., Lu, Y. and Zhai, C. (2010): Latent aspect rating analysis on review text data: A rating regression approach. *Proc. 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 783–792, ACM.

Wiebe, J., Wilson, T., Bruce, R., Bell, M. and Martin, M. (2004): Learning subjective language. *Computational Linguistics* **30**(3):277–308.

Wogenstein, F., Drescher, J., Reinel, D., Rill, S. and Scheidt, J. (2013): Evaluation of an algorithm for aspect-based opinion mining using a lexicon-based approach. *Proc. 2nd International Workshop on Issues of Sentiment Discovery and Opinion Mining*, 5, ACM.

Xu, B., Zhao, T.J., Zheng, D.Q. and Wang, S.Y. (2010): Product features mining based on conditional random fields model. *International Conference on Machine Learning and Cybernetics (ICMLC)*, 3353–3357: IEEE.

Yu, L., Ma, J., Tsuchiya, S. and Ren, F. (2008): Opinion mining: A study on semantic orientation analysis for online documents. *7th World Congress on Intelligent Control and Automation,* 4548–4552, IEEE.

Zhang, L. and Liu, B. (2014): Aspect and entity extraction for opinion mining, *Data Mining and Knowledge Discovery for Big Data,* 1–40. Chu, W.W. (ed), Springer, Berlin Heidelberg.

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M. and Liu, B. (2011): Combining lexicon-based and learning-based methods for Twitter sentiment analysis. *HP Laboratories Technical Report HPL-2011,* 89.

Zhao, L. and Li, C. (2009): Ontology-based opinion mining for movie reviews. *Knowledge Science, Engineering and Management*, 204–214.

Zhuang, L., Jing, F., Zhu, X.Y. and Zhang, L. (2006): Movie review mining and summarization. *Proc. 15th ACM International Conference on Information and Knowledge Management*, 43–50.