Identification and Analysis of Aspect in Text-based terms

Sakshi Mehta[1], Manish Kumar[2], Gitika Sharma[3]

[1]Assistant Professor, Chandigarh University, India, mehtasakshi1191@gmail.com

[2]Assistant Professor, Chandigarh University, India, mksidhu255@gmail.com

[3]Assistant Professor, Chandigarh University, India, gitikasharma41@gmail.com

Abstract **Semantic Analysis or Opinion Mining is considered as a great idea in the research area due to the rapid growth in number of online documents on the Web. This vital information is generally present in the form of text. It's now been a trend to mark a review in the form of an acknowledgement for the availed services online. This practice leads us to the concept of Aggregate Numeric Rating. This rating or score adds up company's reputation from the collected reviews across the web. In this paper, we have focused on two major concepts: Aspect Identification (to identify words and clauses referring to a review subject) and Aspect-Based Sentiment Analysis (to find out the hidden idea or meaning of each sentimental notation). In order to identify aspects of the targeted entities, we have deployed two models: Conditional Random Fields and Association Mining Algorithm. On the other hand, for SA, we would use a method based on VADER which will extract sentiment notations sentence-by-sentence.**

Keywords* — *Sentiment Analysis, Numeric Rating, Aspect Identification.

I. Introduction

With more and more users getting active in the field of sentiment reviewing, text-based reviews have become a major throwback in making decisions. This collected data can be either a product review or a suggestion or normally a type of feedback for the targeted entity. This acknowledgment can be either positive or negative. These reviews are accompanied with a numeric value called as rating, which is further aggregated as overall or average rating (score) for a particular subject. Thus, it is now desirable to deploy a system that can rate the important aspects of a subject separately so that crucial information is in the hands of those who have some kind of preferences. The need of the hour is to have a system that can use the recorded reviews and such a method is known as Semantic Analysis or Opinion Mining. The useful source of sentiment analysis for such a model is a Text-based review which is useful in building models that can determine sentiment polarity. The procedure starts with noticing the sentiment of various attributes of the product. High or low rating of the product depends on the polarity of the product either positive or negative. Finding the correct attribute or aspect is done through Aspect Identification that searches the words and phrases referring to a particular aspect of the product. After identifying these aspects, Sentiment Analysis moves further to find out the polarity of each aspect. During this process, we come across sentiment lexicons which are helpful in classifying adjectives on the basis of their sentiment polarity.

A. Natural language processing

It's an area of computer science combined with artificial intelligence relating computer and human languages by driving meaningful information from the human-generated text. NLP works in collaboration with Machine Translation or Learning that translates sentences from one platform (language) to another. Earlier approach involved translating sentences word-by-word and suffered from syntax and semantics issues. Today modern NLP research identifies problems like speech-to-text conversion, text-based question answering, automatic spell-check and much more. In the advertisement area, NLP is widely used to extract customized interests of the user which prove beneficial to the companies.

We have gathered some data related to the use of NLTK (Natural Language ToolKit) **[1]** and Stanford's CoreNLP toolkit [2] on the basis of literature survey. These packages are Python-based and provides a large set of functions and datasets useful for natural language processing

II. Datasets and Text Features

A. Datasets

To move through any of the steps involved in Aspect-Identification, it is very important for the data to contain information about the categorization of words. Information about identifying which words are aspect-terms, or are a part of which aspect categories, and whether each instance of a term is referred to positive or negative range. It limits the effectiveness of methods which are based on features

DOI : 10.18231/2454-9150.2018.0428

like domain and require different data sets for these methods to prove effective. All the aspect terms and categories are related to sentiment polarity from the defined set {"positive", "negative", "neutral"}.

Data is available from two domains: medicine and hospital reviews. The 2014 datasets are stored as sentences, aspect terms are provided for sentences in the datasets of both domains, and aspect categories are provided for sentences in the dataset of the hospital domain. For each aspect term, character offsets are mentioned as "from" (beginning) and "to" (end). The 2015 datasets are stored as reviews in two different domains: laptop reviews and restaurant reviews. Each review is provided as a list of sentences in order and each sentence is associated with zero or more aspect categories. The 2016 dataset is provided in two different formats. One is identical to the 2015 dataset format. The other is a review-based format that stores sentences and aspect categories separately. Each review consists of a list of sentences and a separate list of the aspect categories within the review.

The format summarization can be as follows: 2014 datasets tracks particular aspect terms and their related polarities, also aspect categories for the Restaurant dataset that are not explicitly linked to aspect terms. 2015 datasets are more specific as they identify specific entity-attribute combinations that form aspect categories, as well as target aspect terms for the Restaurant dataset that explicitly link aspect terms to aspect categories. 2016 datasets identify specific entity-attribute combinations that form aspect categories that are found within a review as a whole, rather than individual sentences. Hence, they become more general.

B. Text features: Token-level features

In this step, each sentence is broken into tokens that hold words and punctuation marks using Penn Treebank tokenizer within NLTK [3]. The original token text is stored in combination with its lowercase version. Porter Stemmer [4] has been used to store the stem of a word after removing all prefixes and suffixes

C. Text features: Sentence-Level Features

Using the sentence-level context, index of each token is stored with 0 being the foremost token of the targeted sentence. To tag POS (Part-of-Speech) for each token, POS tagger using Penn Treebank tagset [3] is used. Every token holds the prior and successor tokens in the sentence. In the case where there are no such texts, a default value is thus stored.

III. Aspect Identification

A. Problem Description

Aspect term extraction is defined to be a process which specifies or denotes those words or phrases that resemble some different aspect of the targeted subject in the given text. We usually come across two types of aspect forms like explicit and implicit. In this paper, we would be focusing only on the implicit ones. We would be requiring new sets of training data for each new domain under examination. The state comes when each such set would require thousands of sentences then this task becomes quite infeasible to go through.

The next challenge which we would be facing in aspect identification is to have a balance between accuracy and robustness. This means more and more accurate models would require more detailed training data accompanied with their respective polarities. Thus, as per conclusion we would be focusing on both supervised and unsupervised training strategies.

B. Conditional Random Fields: Sequential Labelling

This methodology is best suited for issues like POS tagging, shallow parsing and entity recognition [5]. We would discuss about CRF model and Hidden Markov Model, both for sequence labelling. These models are generalized single label models that have their description in Bayes Classifier and Maximum Entropy models. Using the above methodology, many authors have been able to define feature functions allowing a stream of output features related to each word in the sentence.

Using the HMM Model[6], we have two streams: X as hidden states and Y as output. Related to our problem description, we define each output x(i) as a stream of features. We would be using various features related to a word rather than just focusing on the word only. After the application of this model, it is seen that count of possible label combinations come out to be very large but can be removed during training.

This step needs a set of training data defined as {x(i) , y(i) } where i tends from 1 to N and N denotes number of documents taken. Every sentence has n(i) tokens. For any sentence i, y(i)= {y1(i), y2(i), ..., yn(i) } and x(i)={ x1(i), x2(i),.....,xn(i) }; where x(i) denotes sequence of IOB2 Labels and y(i) denotes stream of feature vectors. In combination to this, one more technique is employed known as Regularization [7] that aims at smoothening the parameters by a miss (penalty) for overfitting. In this paper, we would also be using CRFSuite, which is a software implementation of CRF and allows many optimization algorithms for our problem description[8].

If we start our evaluation based on distinct aspect terms then we would require a comparison between already defined aspect terms and the actual distinct aspect terms. Next if we consider evaluation based on instances of each aspect term then it would lead to overconfidence in models and we are left with identifying only the most common terms (though with accuracy) but the accuracy level slows down for the rest of the terms. Thus, we would be implementing both evaluation criteria by using 70% of data for training and the left 30% for testing.

While implementing CRF-Suite, we would be implementing two optimization algorithms namely L-BFGS and stochastic gradient descent. Table 3.1 shows the results when we apply CRF on distinct aspect terms. Table 3.2 shows the results when we apply CRF on instances of aspect terms. L-BFGS, a quasi-Newton method, is helpful in solving problems where we have a large number of parameters. Stochastic gradient descent (SGD) [9] works out in the cases of random data points.

TABLE 3.1 Applying CRF on distinct aspect terms.

Algorithm	Dataset	Precision	Recall	F-measure
L-BFGS	Medicine	0.7331	0.5636	0.6061
SGD	Medicine	0.6501	0.5123	0.5673
AP	Medicine	0.6395	0.4987	0.4846
PA	Medicine	0.5644	0.5234	0.5930
AROW	Medicine	0.4321	0.5712	0.4980
L-BFGS	Hospital	0.6543	0.3856	0.5051
SGD	Hospital	0.4653	0.3191	0.4106
AP	Hospital	0.5675	0.2875	0.3953
PA	Hospital	0.5335	0.3943	0.4961
AROW	Hospital	0.4345	0.3745	0.4565

TABLE 3.2 Applying CRF on instances of aspect terms

Algorithm	Dataset	Precision	Recall	F-measure
L-BFGS	Medicine	0.8175	0.7854	0.8002
SGD	Medicine	0.7978	0.6974	0.7562
AP	Medicine	0.8053	0.7027	0.7173
PA	Medicine	0.8193	0.7884	0.7937
AROW	Medicine	0.7061	0.7291	0.6993
L-BFGS	Hospital	0.8325	0.7035	0.7042
SGD	Hospital	0.6829	0.6557	0.5917
AP	Hospital	0.7969	0.4749	0.5573
PA	Hospital	0.8042	0.6498	0.7334
AROW	Hospital	0.7867	0.6443	0.6515

C. Association Mining Method

This is a rule-based method that constructs a given list of itemsets, having noun and noun phrases in every sentence, then filters them to extract the aspect terms. The idea behind its functionality is a fact that reviewers often specify similar words while mentioning aspects terms and thus, such words tend more to become aspect terms [10].

The method starts by generating initial itemsets [11]. These are nothing but a list of noun and noun phrases for each sentence. Another point to consider here is that we might have pairs or triplets of the above selected noun and noun phrases, so such terms are also regarded as candidate terms or itemsets. To avoid large number of terms, we reduce the set of such terms by defining only “frequent’ itemsets as per defined level “m”. The remaining terms are thus ignored.

In order to reduce these terms, two pruning methods are adopted. The first one implements an adjustable frequency parameter named as “p-support” which is useful in counting a candidate term provided it is not a subset of another candidate term within a given same sentence. It further defines a minimum p-support that acts as a threshold. If for a candidate, p-support is less than p and is also a subset of some other term then clearly, we would be negating such terms.

The other pruning method works on frequent itemsets. In this pairs and triplets are considered as candidate terms. For any term within a sentence, this method defines the maximum distance between any two adjacent words in the given term. This distance means the number of tokens by which they are apart in the sentence. A threshold “w” is also defined and if value exceeds w then the term is regarded as “non-compact” within the sentence. If this practice is frequently followed then the whole term is ignored.

While evaluating, we come across two important issues namely, the case of one-word aspect terms that are nouns and are easily traceable using POS tagger. The other case is multi-word aspect terms where noun phrases are to be identified. This way or method of identifying noun phrases is called Shallow Parsing [12]. Thus, we attempted to use NLTK's "Regexp" (regular expression) feature that works on a pre-defined search pattern [13] to extract specific patterns in the text. In addition to this, we also examined default named-entity chunker within NLTK[14].

IV. Aspect Based Sentiment Analysis

A. Problem Description

In this paper, another focusing area is the sentiment of every aspect term per sentence. There might be the cases where only aspect categories are known to us rather than actual terms. In this context, we try to estimate sentiment of each occurrence of an aspect term in the given text with accuracy. The case where we have multiple aspect terms in single sentence then the word related to one aspect term might be wrongly associated to another term. Thus, we implement a method to focus on identifying sentiment of instances of aspect terms. Secondly, our main aim is to use aspect categories than aspect terms due to its multiple benefits. That is, we will have lesser number of such categories that will lead us to have limited data to have accurate rating.

B. VADER Based Method

VADER meaning Valence-Aware Dictionary for sentiment reasoning is a rule-based model designed to perform sentiment analysis per sentence. It was trained to work on online media text including movies and product reviews. It is capable of performing classification and performing unsupervised testing even on newly-added data and domain-related data.

DOI : 10.18231/2454-9150.2018.0428

It works on sentiment lexicon thus making it suitable for analyzing reviews online. It also works well on both newly-traced data and data across domains. It involves valence scores that hold information about sentiment intensity and polarity. This score ranges from -4 to +4.

Some axioms for VADER to determine valence score are as follows:

1) An increase in the score can be seen when we have punctuations like exclamation points.
2) The cases where we have full-word capitalization also show a higher magnitude of score.
3) A new concept "Degree Modifiers", a set of adverbs, also affects the magnitude of the valence score.
4) A shift in sentiment polarity is seen with the presence of "but" conjunction.
5) Lastly, if we come across a trigram before the lexical feature then it is helpful in determining whether negation corresponds to an opposite polarity.

After all this, VADER returns four scores, one each for positive, negative, neutral and compound score that explicitly refer to the intensity of polarity within a sentence

C. Evaluation

On evaluating each occurrence of an aspect term with accuracy, focus is kept on predicted and true label values.

TABLE 3.1 Results on aspect terms

Accuracy		0.6055		
Label	Precision	Recall	F-Measure	Domain
Positive	0.7910	0.6842	0.6954	Medicine
Negative	0.5034	0.3905	0.4026	Medicine
Neutral	0.5119	0.7013	0.6294	Medicine

The main measurements used for evaluation are basically accuracy, precision, recall and F-measure with all labeled as "positive", "negative" and "neutral". Based on the above evaluation, if we are provided with a dataset having quantitative review scores plus aspect categories then we would have more accurate ratios of positive to negative sentiments.

V. Conclusion

Through this paper, we tried to find out some basic features of an aspect-based review system or model. We tried to concentrate on the idea of developing an annotated dataset for training models in regard to aspect identification and aspect-based semantic analysis. Talking about aspect identification, we employed two algorithms: Sequential Learning Model known as Conditional Random Field and Association Mining Algorithm, for supervised and unsupervised training and testing respectively. CRF showed that when the parameters are trained using L-BFGS, we get an effective classifier model for noticing aspect terms. While the evaluation results of Association Mining Algorithm proposed a futuristic idea of exploring noun phrases accurately. On the other hand, we also described a rule-based sentiment analysis model, VADER, to track the sentiment of aspect terms and categories.

A very important area that comes through this paper is Aspect Aggregation meaning to track those aspect terms which are alike or belong to an overreaching aspect category. This criterion was achieved by using pre-defined categories allowing supervised training for the clustering issues. Both review-level and sentence-level data tend to be not that easy to present a good range of domains but when talking of smaller number, this proves quite feasible.

Lastly, for future explorations, unsupervised clustering may also be used in which clusters would be identified by their very frequent aspect terms.

References

[1] S. Bird, "NLTK: the natural language toolkit. In Proceedings of the COLING/ACL on Interactive presentation sessions", *Association for Computational Linguistics*, pages 69–72, 2006.

[2] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit", *ACL System Demonstrations*, pages 55–60, 2014.

[3] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of English: The Penn Treebank", *Computational Linguistics*, 19(2):313–330, 1993.

[4] M. F. Porter, "An algorithm for suffix stripping. Program", 14(3):130–137, 1980.

[5] F. Sha and F, "Pereira. Shallow parsing with conditional random fields", *Pro- ceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Volume 1, pages 134–141. Association for Computational Linguistics, 2003.

[6] I. Pavlopoulos. Aspect based sentiment analysis. Athens University of Economics and Business, 2014.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incom- plete data via the EM algorithm", *Journal of the Royal Statistical Society. Series B (methodological)*, pages 1–38, 1977.

[8] M. Hu and B. Liu. Mining and summarizing customer reviews. *In Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177. ACM, 2004.

[9] S. Vishwanathan, N. N. Schraudolph, M. W. Schmidt, and K. P. Murphy, "Ac- celerated training of conditional random fields with stochastic gradient methods", *Proceedings of the 23rd International Conference on Machine Learning*, pages 969–976. ACM, 2006.

[10] M. Hu and B. Liu, "Mining and summarizing customer reviews", *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177. ACM, 2004.

[11] S. P. Abney, "Parsing by chunks", *Principle-based parsing*, pages 257–278. Springer, 1991.

[12] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text", *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[13] S. Vishwanathan, N. N. Schraudolph, M. W. Schmidt, and K. P. Murphy. Accelerated training of conditional random fields with stochastic gradient methods. *In Proceedings of the 23rd International Conference on Machine Learning*, pages 969–976. ACM, 2006.

[14] S. Bird. NLTK: the natural language toolkit. *In Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics, 2006.

DOI : 10.18231/2454-9150.2018.0428