

UMCC_DLSI_Prob: A Probabilistic Automata for Aspect Based Sentiment Analysis

Yenier Castañeda

Armando Collazo

Elvis Crego

Jorge L. Garcia

DI, University of Matanzas

Matanzas, Cuba

{yenier.castaneda,
armando.collazo}@umcc.cu,
elvis.crego@mtz.cu,
jorge.garcia@infonet.umcc.cu

Yoan Gutiérrez

David Tomás

Andrés Montoyo

Rafael Muñoz

DLSI, University of Alicante

Alicante, Spain

{ygutierrez, dtomas, montoyo,
rafael}@dlsi.ua.es

Abstract

This work introduces a new approach for aspect based sentiment analysis task. Its main purpose is to automatically assign the correct polarity for the aspect term in a phrase. It is a probabilistic automata where each state consists of all the nouns, adjectives, verbs and adverbs found in an annotated corpora. Each one of them contains the number of occurrences in the annotated corpora for the four required polarities (i.e. positive, negative, neutral and conflict). Also, the transitions between states have been taken into account. These values were used to assign the predicted polarity when a pattern was found in a sentence; if a pattern cannot be applied, the probabilities of the polarities between states were computed in order to predict the right polarity. The system achieved results around 66% and 57% of recall for the restaurant and laptop domain respectively.

1 Introduction

Sentiment analysis is increasingly viewed as a vital task from both an academic and a commercial standpoint. Textual information has become one of the most important sources of data to extract useful and heterogeneous knowledge. *“Texts can provide factual information, such as: descriptions, lists of characteristics, or even instructions to opinion-based information, which would include reviews, emotions, or feelings. These facts have motivated dealing with the identification and extraction of opinions and*

sentiments in texts that require special attention.” (Gutiérrez, et al., 2014). Sentiment Analysis or “Subjectivity Analysis” in (Liu, 2010) is defined as the computational treatment of opinions, sentiments and emotions expressed in a text. In order to automatically treat the subjectivity, we need lexical resources that allow the detection and evaluation of the affective/ subjective charges in texts, its polarity and intensity.

Regarding research carried out for linguistic patterns identification and its polarity in texts, it is worth mentioning works on: adjectives (Hatzivassiloglou and McKeown, 1997) (Wiebe, 2000); adjectives and verbs (Turney, 2002) (Wilson, et al., 2005) (Takamura, et al., 2007); and also verbs and names (Esuli and Sebastiani, 2006). WordNet (Fellbaum, 1998) has also been used for the collection of opinion adjectives and verbs (Kim and Hovy, 2005) to determine the semantic orientation of the terms depending on their notes (Esuli and Sebastiani, 2005), for the adjective extraction (Andreevskaja and Bergler, 2006) or opinion mining (Esuli and Sebastiani, 2007).

Inspired on Hidden Markov models (Baum and Petrie, 1966) and following the idea that words combinations are finite in an evaluation text, we decided to create a finite automata in graph form to represent all these relations extracted from a training corpus. For the creation of this automata we utilised different resources, such as WordNet and OpinionFinder Subjectivity Lexicon. Also, different extracted patterns based on (Cazabón, 1973) were applied.

This paper is structured as follows: In section 1.1 is described the task 4 of SemEval2014 (Pontiki, et al., 2014) where this system was presented. Section 2 presents the description of the automata and how it was built. The polarity assignation method using the trained automata is

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Page numbers and proceedings footer are added by the organisers. Licence details: <http://creativecommons.org/licenses/by/4.0/>

described in section 3. Finally, in section 4 and 5 are shown the results and conclusions, respectively.

1.1 Task Description

The SemEval2014 task 4 (Pontiki, et al., 2014) was divided into four subtasks: 4.1 Aspect term extraction; 4.2 Aspect term polarity; 4.3 Aspect category detection; and 4.4 Aspect category polarity.

This paper is focused on subtask 4.2 which is described as follows:

Given one or more Aspect Terms within a sentence, it is necessary to determine whether the polarity of each Aspect Term is positive, negative, neutral or conflict (i.e., both positive and negative). For example:

```
"I loved their fajitas" →  
"fajitas": positive  
"I hated their fajitas, but  
their salads were great" →  
"fajitas": negative,  
"salads": positive  
"The fajitas are their first  
plate" → "fajitas": neutral  
"The fajitas were great to  
taste, but not to see" →  
"fajitas": conflict.
```

Each participant was permitted to submit two kinds of runs for this task:

Constrained: Using only the provided training data and other resources, such as lexicons.

Unconstrained: Using additional data for training. Teams were asked to report what resources they used for each submitted run. The training dataset, provided by the organiser of the Task 4 challenge, consists of two domain-specific datasets which contain over 6,500 sentences with fine-grained aspect-level human-authored annotations. These domains are:

Restaurant reviews: This dataset consists of over 3000 English sentences from the restaurant reviews of (Ganu, et al., 2009) that were adapted to the task.

Laptop reviews: This dataset consists of over 3000 English sentences extracted from customer reviews of laptops.

2 The automata

The automata was represented as a graph $G = (S, T)$ whose vertexes constitute the group of finite states $S = [s_1, s_2, s_3, \dots, s_n]$ while the

edges represent the transitions $T = [t_1, t_2, t_3, \dots, t_n]$ of going from one state to another.

Our finite automata involves the following features:

1. Group of finite states: all the verbs, nouns, adverbs, adjectives that were extracted from the training dataset (see Section 2.1) using Freeling 3.1 language analyser (Atserias, et al., 2006), or Aspect Terms (that may be formed by several words). In every state the automata stores the occurrences W_i^p , where p is one of the following polarity classes: positive, negative, neutral, conflict or undefined, i being the index of the current state in the graph.
2. Finite alphabet: a sentences set which contains one or more Aspect Terms to which should be assigned a polarity.
3. Initial state: first word of the sentence.
4. Transition state ($Ts_{i,j}$ and $Ts_{j,i}$): each transition between two states contains $W_{i,j}^p$ and $W_{j,i}^p$, where p is positive, negative, neutral or conflict, i is the current state, and j is the next state.
5. End state: last word of the sentence.

If we could not determine the polarity classification for a state or transition, then we set it as *undefined polarity*.

2.1 Training the automata

In order to create the automata the training dataset provided for the SemEval2014 tasks 4¹ was used.

In the automata, each word of a sentence forms a state which is connected to the following word. This connection forms a transition between the two words. This method is repeated until the last word of the sentence is reached. If the word already exists in the automata, both its state and all the transitions (from and to that word) are adjusted, increasing in one the W_i^p , $W_{i,j}^p$ and $W_{j,i}^p$ of the polarity value initially assigned in the corpus.

The transitions from words to Aspect Terms with their respective polarities allow to go through those words with undefined polarities to the target Aspect Terms. This event is done for finding the most probably polarity according to the training discoveries. Same thing happens with transitions from an Aspect Term to a word, but in this case from the polarity of the Aspect Term to *undefined polarity*.

¹<http://alt.qcri.org/semEval2014/task4/>

On the other hand, if the word is not an Aspect Term its state do not change at all, since the dataset only annotates the Aspect Terms, so we do not know the polarity of those words that are not an Aspect Term.

To solve this issue we decided to make use of other resources to enhance the automata, so that the probability for finding a polarity for a word in the automata increases with the expansion of the dictionary. We used the Opinion Finder Subjectivity Lexicon (OFSL) (Wilson, et al., 2005) to adjust the state and transitions of the words in the automata. To address the adjustment, for every word of OFSL (according to the classification of the sentiment polarity) that exists in the graph represented by automata, the respective value of polarity of W_i^p , $W_{i,j}^p$ and $W_{j,i}^p$ is increased in one. We also used WordNet 3.0 to obtain the synonyms and antonyms of the words in the automata to form new states and transitions. Synonyms were given the same polarity as the related word, whereas antonyms took the opposite polarity. The subjectivity clues extracted by the patterns detected in the training dataset were used as well (See section 3.2).

In Table 1 we show the terminology used for the patterns.

Symbol	Description
[]	Optional word
/!	Subjectivity clue
/l	Compare by lemma
AT	Aspect Term

Table 1: Pattern symbols

Examples:

```
[DT] AT [PRP] [RB] be/l [VBG/!]
RB/! [JJ/!] [RB/!]

[RB/!] [DT] JJS/! [DT] [NN] AT [VB]
[NN/!]

[DT] JJ/! NN PRP VBD VB [DT] AT
AT be/l [DT/!] JJ/! [PRP/!] [RB/!]
```

Note the use of the POS tags such as DT, NN, VBD, and others were taken from the result of the pos-tagging process performed by Freeling 3.1. Using this tool the incoming texts were split into parts (sentences) for the following processes.

For instance, in the sentence “This MacBook Pro is excellent” the subjectivity clue for the Aspect Term **MacBook Pro** is **excellent**; so its states and transitions get adjusted the same way as the Aspect Term. Figure 1 describes this example, where p_i is W_i^p , p_{ij} is $W_{i,j}^p$ and p_{ji} is $W_{j,i}^p$ means the occurrence for positive polarity (negative, neutral and conflict polarities were omitted by lack of space). Both states and transitions are represented.

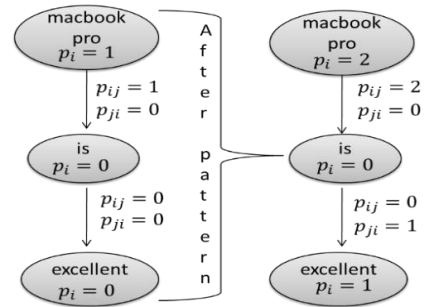


Figure 1: Adjusting states and transitions after pattern analysis.

3 Polarity Assignment

Before predicting the polarity of the Aspect Terms, each sentence is divided by its connectors (conjunctions, prepositions and adverbs, extracted using Freeling), forming the corresponding phrases. For instance, the sentence “Where Gabriela personally greets you **and** recommends you what to eat” is divided into the phrase “Where Gabriela personally greets you” and the phrase “recommends you what to eat” by connector **and**.

3.1 Selection criteria

If only one polarity is found then that is the polarity for the Aspect Term. On the other hand, if more than one polarity is found, the polarity for the Aspect Term is the most repeated one.

Note that if both positive and negative are the most repeated polarities we set conflict as the polarity for the Aspect Term.

If no polarities are found at all, we assignee neutral to the Aspect Term.

3.2 Assigning polarity using patterns

We detected different patterns which allowed us to extract those words that influence on the Aspect Term polarity in the phrase (See section 2.1).

For each phrase subjectivity clue i , we calculate the most probable polarity $Pp_i = \max_p(W_i^p)$, if i has a state in the automata. After that, we apply our selection criteria described in section 3.1.

If no polarities are found at all, we process the phrase in the next steps.

3.3 Assigning polarity using the automata

For each Aspect Term in the phrase we get the sentence it belongs to and we calculate $Pt_{i,j} = \frac{W_{i,j}^p}{\sum W_{i,j}^p}$ in that sentence, where $Pt_{i,j}$ is the most probable polarity of $Ts_{i,j}$ (j being the Aspect Term), if such a transition existed. If no polarity

is found, then we calculate $Pt_{j,i} = \frac{w_{ji}^p}{\sum w_{ji}^p}$ again if such a transition existed.

In case of applying aforementioned processes without finding out a concrete polarity for the target Aspect Terms, we perform other steps to try to find one or more polarities for the Aspect Term.

First, we verify whether the Aspect Term is part of a phrase which was matched to a pattern but no polarity was found as explained in section 3.2, if so we get the subjectivity clues of the phrase and for each subjectivity clue we calculate $TS_{i,j}$, where i is the Aspect Term index and j corresponds to the subjectivity clue index. If no polarity is found, then we calculate $TS_{j,i}$.

If no polarities are found after this step, we proceed to do the same as above, but this time for each word in the sentence.

Lastly, if no polarities are found, Pp_i is obtained for each word i in the sentence if i has a state in the automata.

Training				Evaluation		
Test	Patterns	WordNet	OFSL	Pattern/Automata only	Automata only	Accuracy (%)
1	X	X	X	X		58.0
2	X			X		57.9
3	X		X	X		57.9
4	X	X	X		X	54.0

Table 2: Evaluation over restaurant domain

With test 4 it is evident that it is better to use two methods combined than only one of them, since the patterns indicate the words that assign polarity to the Aspect Term, making the automata more precise with this information at the time of assigning the correct polarity. Otherwise, if a pattern is not encountered we need to analyse the words that are closer to the Aspect Term determining the polarity according to the context. In addition, not always is assigned a polarity to it in case of the pattern found in the context is empty. Table 3 shows the results of our system in comparison with the best of the challenge SemEval2014 subtask 4.2.

Test	Constrained	Unconstrained
Rest	66.5	66.8
Laptop	56.1	57.0
BRR	80.9	77.6
BRL	70.4	66.6

Table 3: Test subtask 4.2 (BRR: Best Ranked for Restaurant; BRL Best Ranked for Laptop)

The system behaved the same as the training stage on the competition although the accuracy increased.

After performing these steps we apply our selection criteria to assign the polarity to the Aspect Term in question. As can be seen, our proposal is focused on the application of an exhaustive exploration of the automata in order to classify Aspect Terms with the target polarities.

4 Results and Discussion

In order to evaluate the accuracy of the system several tests were run. Table 2 shows some of the tests using SemEval2014 task4 Baseline for the Restaurant reviews. We did the same evaluation for Laptop reviews and the results obtained were very similar to those shown in Table 2 for Restaurant. We used `semeval_base.py`² script to split the dataset into a train and a test part using an 80:20 ratio. Despite tests 1, 2 and 3 results do not vary much, it is evident that using the three training resources yields our best accuracy.

5 Conclusions and future works

This work introduces a new approach for aspect based sentiment analysis. For that, a probabilistic automata was created where the states are formed by the nouns, adjectives, verbs and adverbs found in the annotated corpora, based on their occurrence. The transitions between states are also taken into account. A set of patterns were defined in order to extract the words that influence on an Aspect Term, also known as subjectivity clues, and then we predicted their polarity using the automata's probabilities. A system was developed following this approach to participate on SemEval2014 competition, obtaining an accuracy of 66% for restaurant reviews and 57% for laptop reviews.

As future works we plan to deal with the fact that this automata only involves states represented by the words lack extracted from the training data. So, the previously unseen aspect terms which do not correspond to any state in the automata, are not recognised in many cases as far as the polarity is concerned. To address this issue we plan to

² <http://alt.qcri.org/semeval2014/task4/data/semeval14-absa-base-eval-valid.zip>

expand the aspect term dictionary using Wikipedia definitions. On the other hand, we plan to use a disambiguation method to select the exact WordNet synset and then to reduce the polysemy of the automata's words. Finally, to smooth the probabilities it would be interesting to study different balances in order to get new improvements for the system.

Acknowledgments

This research work has been partially funded by the University of Alicante, Generalitat Valenciana, Spanish Government and the European Commission through the projects, "Tratamiento inteligente de la información para la ayuda a la toma de decisiones" (GRE12-44), ATTOS (TIN2012-38536-C03-03), LEGOLANG (TIN2012-31224), SAM (FP7-611312), FIRST (FP7-287607) and ACOMP/2013/067.

References

- Alina Andreevskaia and Sabine Bergler, 2006. *Mining WordNet for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses*. Trento, Italia, s.n.
- Jordi Atserias et al., 2006. *FreeLing 1.3: Syntactic and semantic services in an opensource NLP library*. Genoa, Italy, s.n.
- Leonard Baum and Ted Petrie, 1966. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics*, pp. 1554--1563.
- María Cazabón, 1973. *Patterns of English*. s.l.:Editorial Pueblo y Educación.
- Andrea Esuli and Fabrizio Sebastiani, 2005. Determining the semantic orientation of terms through gloss classification. *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pp. 617-624.
- Andrea Esuli and Fabrizio Sebastiani, 2007. *PageRanking WordNet Synsets: An Application to Opinion Mining*. Prague, Czeck Republic, s.n., pp. 424-431.
- Andrea Esuli and Fabrizio Sebastiani, 2006. *SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining*. Genova, IT, s.n., pp. 417-422.
- Christiane Fellbaum, 1998. *WordNet. An Electronic Lexical Database*. University of Cambridge: s.n.
- Gayatree Ganu, Noemie Elhadad and Amélie Marian, 2009. *Beyond the stars: Improving rating predictions using review text content*. Rhode Island, s.n.
- Yoan Gutiérrez, Andy González, Roger Pérez, José I. Abreu, Antonio Fernández Orquín, Alejandro Mosquera, Andrés Montoyo, Rafael Muñoz and Franc Camara, 2014. UMCC_DLSI-(SA): Using a ranking algorithm and informal features to solve Sentiment Analysis in Twitter. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pp. 443--449.
- Vasileios Hatzivassiloglou and Kathleen McKeown, 1997. *Predicting the Semantic Orientation of Adjectives*. Madrid, Spain, s.n., pp. 174-181.
- Soo-Min Kim and Eduard Hovy, 2005. *Automatic Detection of Opinion Bearing Words and Sentences*. Jeju Island, Republic of Korea, s.n.
- Bing Liu, 2010. Sentiment Analysis and Subjectivity. In: *Handbook of Natural Language Processing*. Boca Raton: s.n., pp. 627-666.
- Maria Pontiki, Dimitrios Galanis, John Pavlopoulos, Haris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar, 2014. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland.
- Hiroya Takamura, Takashi Inui and Manabu Okumura, 2007. *Extracting Semantic Orientations of Phrases from Dictionary*. s.l., s.n., pp. 292-299.
- Peter Turney, 2002. *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews*. Philadelphia, Pennsylvania, s.n., pp. 417-424.
- Janyce Wiebe, 2000. *Learning Subjective Adjectives from Corpora*. Austin, Texas, s.n.
- Theresa Wilson, Janyce Wiebe, and Paul Hoffmann, 2005. *Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis*. Vancouver, Canada., s.n.