Dictionary-Based Sentiment Analysis applied to specific domain using a Web Mining approach

Laura Cruz
Universidad Nacional de San Agustín, Perú
lcruzq@unsa.edu.pe

José Ochoa
Universidad Católica San Pablo, Perú
jeochoa@ucsp.edu.pe

Mathieu Roche
TETIS
Cirad, Cnrs
AgroParisTech, Irstea, France
mathieu.roche@cirad.fr

Pascal Poncelet
LIRMM, Cnrs
Université Montpellier, France
pascal.poncelet@lirmm.fr

Abstract

In recent years, the Web and social media are growing exponentially. We are provided with documents which have opinions expressed about several topics. This constitute a rich source for Natural Language Processing tasks, in particular, Sentiment Analysis. In this work, we aim at constructing a sentiment dictionary based on words obtained from web pages related to a specific domain. To do so, we correlate candidate opinion words, seed words and domain using $AcroDef_{MI3}$ and TrueSkill methods. This dictionary-based approach is compared to the SentiWordNet lexical resource. Experimental results show suitability of our approach for multiple domains and infrequent opinion words.

1 Introduction

In recent years, the Web and social media are growing exponentially, this constitute a rich source for Sentiment Analysis tasks. Companies are increasingly using the content in these media to make better decisions (Marrese-Taylor et al., 2013). Social networking sites are being used for expressing thoughts and opinions about products by users (Amine et al., 2014). In this context, Sentiment Analysis involves the process of identifying the polarity of opinionated texts. These opinionated texts are highly unstructured in nature and thus involves the application of Natural Language Processing techniques (Varghese and Jayasree, 2013). As a rule, documents have opinionated texts about several topics. Words used to express opinions about some topics can be specific and highly correlated to a particular domain (Duthil et al., 2011). Likewise, while we may find that *The chair is black*, such an adjective would be unusual in a movies domain. To tackle these issues both machine learning and dictionary-based approaches have been proposed in the literature. A machine learning method that applies text-categorization techniques has been proposed by (Pang and Lee, 2004). In such method, graphs, minimum cut formulation, context and domain have been considered to extract subjective portions of documents.

On the other hand, dictionary based approaches are unsupervised in nature. In general, these methods assume that positive (negative) adjectives appear more frequently near a positive (negative) seed word (Harb et al., 2008). An unsupervised learning algorithm for classifying reviews (thumbs up or thumbs down) has been adopted by (Turney, 2002; Wang and Araki, 2007). A review classification is given by the average semantic orientation of their phrases which contain either adjectives or adverbs. A phrase semantic orientation is computed using the mutual information between the given phrase and the word *excellent* minus the mutual information between the given phrase and the word *poor*. Therefore, a phrase has a positive semantic orientation when it has good associations and a negative semantic orientation when it has bad associations, as shown by equation 1.

$$SO(phrase) = \log \frac{hits(phrase \text{ NEAR excellent}) \cdot hits(\text{poor})}{hits(phrase \text{ NEAR poor}) \cdot hits(\text{excellent})} \quad (1)$$

In this work, words used to express opinions

are learned. To do so, positive and negative seed words (e.g. good, excellent, bad) are used to extract adjectives near seed words. To correlate candidate words, seed words and domain, $AcroDef_{MI3}$ and TrueSkill methods are proposed. Experimental results show suitability of our proposal. Several domains (e.g movies, agricultural) were used to compare our approach to SentiWordNet.

The paper is organized as follows. The Methodology is presented in Section 2. Experimental setup is described in Section 3. In Section 4, we present and discuss the obtained results. Concluding remarks are presented in Section 5.

2 Methodology

The proposed process is depicted in Figure 1. The steps are summarized in the following steps:

1. A corpora for a specific domain, containing positive and negative opinions is acquired from the Web.
2. Each document is pre-processed to get text, remove HTML tags and scripts.
3. Opinion adjectives and nouns are extracted using POS-Tagging and the Window Size algorithm.
4. The correlation score of a given word with a seed word and domain is computed using $AcroDef_{MI3}$ and TrueSkill. lexicons are inferred based on these correlation scores that identify semantic orientation for each extracted word. High correlation score words are selected.

We perform experiments over two domains: Agricultural domain (opinions extracted from Twitter) and a Movie domain[1] (data set introduced in (Pang et al., 2002)). Further details are given in the next sections.

2.1 Corpus Acquisition

Some words can express neutral, positive or negative opinion in specific domain such as:

Neutral → I attend *scientific conferences*.
Positive → The list shows the *scientific discoveries*.
Positive → He made a good *scientific discovery*.

[1]http://www.cs.cornell.edu/People/pabo/movie-review-data/

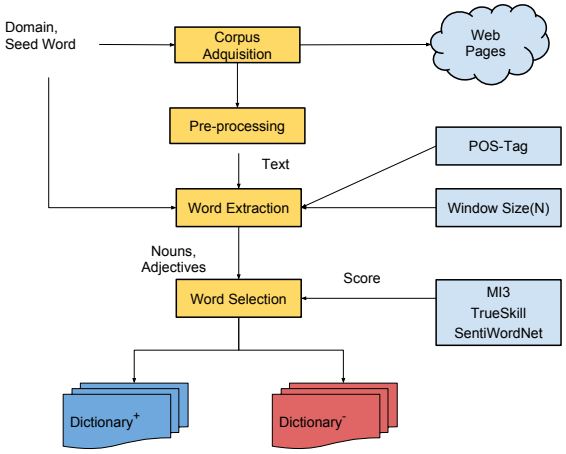


Figure 1: Lexicons are inferred from Web pages correlated to seed words and domains, via extraction and process of candidate words.

These examples show that a given word, for instance *scientific*, can be highly correlated to a particular domain (Harb et al., 2008). The first example is considered a neutral opinion. Conversely, the second example is considered a positive opinion. The third example is also a positive opinion because of the word *good*. Thus, some words are useful to learn opinion words related to a given domain. We can define a seed word, such as *good*, that can help us to find others opinion words.

Lexicons are built using selected words from web page corpus. Web pages are retrieved using Bing search engine. Queries used to retrieve this web pages combine seed words and domain keywords. We have positive and negative seed words, P = {*good*, *nice*, *excellent*, *positive*, *fortunate*, *correct*, *superior*} , Q = {*bad*, *nasty*, *poor*, *negative*, *unfortunate*, *wrong*, *inferior*}, respectively.

A positive (negative) seed word ensure a positive (negative) web page about a query domain, due to all opposite seed words are excluded from that query. For example, the following query can be used for retrieving positive pages: $\text{query}^{+} = +opinion + review + gmo + good - bad - nasty - poor - negative - unfortunate - wrong - inferior$

Thus, we have positive and negative web pages denoted by $\text{corpus}^{+}, \text{corpus}^{-}$ respectively. Each corpus is related to a seed word and a given domain. In the next section we will extract words near seed words for each web page corpus using POS-Tagging and the Window Size algorithm.

2.2 Word extraction

Opinion words near a seed word can have the same polarity (Roche and Prince, 2007; Harb et al., 2008). The same approach has been used to extract candidate opinion words. To identify opinion words (nouns and adjectives) in web page corpus, TreeTagger[2] has been used. Previously, HTML tags, scripts, blank spaces and stop words[3] were removed from web pages. In order to get near words for each seed word a Window Size algorithm has been used (Algorithm 1). The Window Size Algorithm looks for opinion words in both left and right sides of a seed word given a K distance. This distance is the number of left (right) opinion words of a seed word given a web page corpus. This process is shown in Algorithm 1.

Algorithm 1 The Window Size Algorithm

Require: *seed words*, *corpus*, K
Ensure: opinion words
1: *words* ← TreeTagger to each $corpus$
2: *words* ← filter adjectives and nouns
3: **for** *index*= 0 until *total of words* **do**
4: **if** *words*{*index*} in *seed word* **then**
5: **for** $k = 1$ until K **do**
6: *left word* ← *words[index - k]*
7: *right word* ← *words[index + k]*
8: opinion words ← *left word* and *right word*

In Figure 2 adjectives (JJ) and nouns (NNS) are retrieved using *TreeTagger*. The *good* word is a positive seed word and its nearest adjective is *safe* given $k = 1$ distance. Likewise, *scientific* and *studies* words are retrieved with distance $k = 2$. In addition, *safe* is a positive opinion word candidate because it occurred near a positive seed word (*good*). In this sense, we can have a set of opinion words (positive and negative), that can be candidates to include into the resulting lexicon. To get the correlation score of each extracted word given a seed word, two measures are employed: $AcroDef_{MI3}$ and TrueSkill which are described in the next section.

2.3 Word Selection

As seen in our previous example (Figure 2), the *scientific* word was retrieved using window size distance $= 2$. However, specific words, such as *gmo*, can be used to express a domain opinion. Hence, we need to measure the correlation of a given extracted word with *domain* and *seed word* to build a lexicon. In order to get candidate opinion words we propose to use the statistical measure $AcroDef_{MI3}$ (equation 2) (Roche and Prince, 2007). Moreover, we also propose a novel probabilistic measure based on the TrueSkill Algorithm (Herbrich et al., 2007) (Algorithm 3).

The $AcroDef_{MI3}$ measure takes each word extracted using the Window Size algorithm and computes the following equation 2, which is based on web mining.

The total web page results, based on queries that combine candidate words, seed words and domain keywords, are used in the $AcroDef_{MI3}$ measure to get the correlation score for each extracted word.

$$AcroDef_{MI3} = \log\left(\frac{(nb(sw\ word \text{ AND } domain)+ nb(word\ sw \text{ AND } domain))^3}{nb(sw \text{ AND } domain) \cdot nb(word \text{ AND } domain)}\right) \quad (2)$$

where sw is a seed word, *nb(x)* function is the number of total result pages, x is the query used to retrieve pages in the search engine, and $word$ is the word extracted using the Window Size algorithm. This process is detailed in Algorithm 2.

Algorithm 2 Word selection algorithm using $AcroDef_{MI3}$

Require: *corpus*, *seed words = P*, *keywords of domain*
Ensure: correlation score values for each *word*
1: **for** each $corpus$ **do**
2: $words^+$ = window size($corpus^+, P$)
3: **for** word in $words^+$ **do**
4: given each seed word and keywords of domain compute correlation score:
5: $score \leftarrow max(AcroDef_{MI3})$

Unlike $AcroDef_{MI3}$, in the TrueSkill approach words are extracted using the Window Size algorithm and the measure function is applied. Furthermore, words are extracted for each positive (negative) page against k random negative (positive) pages and then their score words are computed. Thus, TrueSkill configures a match between positive pages words against negative pages

[2] http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/
[3] http://www.ranks.nl/stopwords

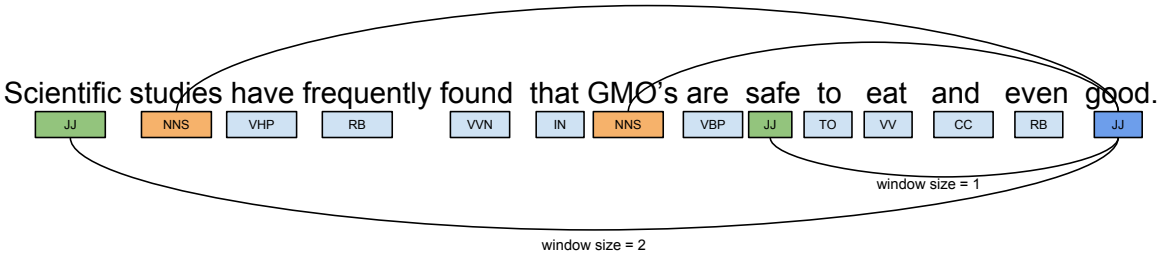


Figure 2: Window size sample for *good* seed word.

words. The process is detailed in Figure 3, where

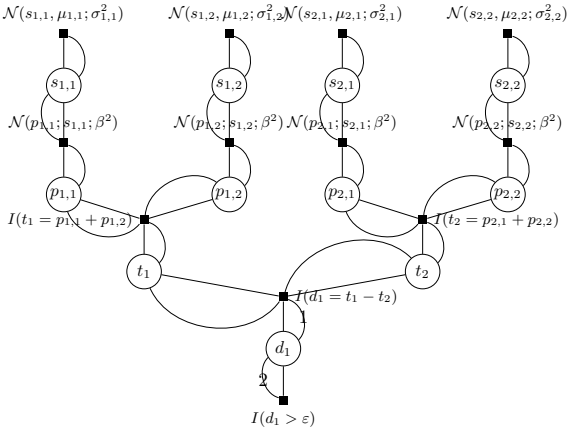


Figure 3: TrueSkill Model, learning score for each word selected given the positive and negative corpus.

$S = \{s_{1,1}, s_{1,2}, , s_{1,n}\}$ and $S = \{s_{2,1}, s_{2,2}, , s_{2,n}\}$, s are learning values for each word in positive and negative web page respectively. p is the learning performance for each word, t is the sum of total performance for each word in corpus.

As $TrueSkill$ learns s according its match outcome, we set a high punctuation for $corpus_+$, and less punctuation for $corpus\_$. Therefore, we have $d = t_1 - t_2$. Due to difference (d) is important, we set $t_1 = 1$ to a positive corpus and $t_2 = 2$ to a negative corpus, where 1 denotes first. This process is detailed in Algorithm 3.

The following example shows how TrueSkill measures two collected web pages:

$corpus^+$= By the way a New York Times $\cdots$ *excellent* job $\cdots$ *bioengineered* food $\cdots$.

$corpus^-$= Roundup Ready cotton $\cdots$ *wrong* solution $\cdots$ at any *economic* advantage.

Algorithm 3 Word selection algorithm using TrueSkill

Require: *corpus*, *seed words(P, Q)*
Ensure: correlation score values for each *word*

```
k = 10 number of match for each corpus.
for each corpus do
    words+ = window size(corpus+, P)
    for k random corpus− do
        words− = window size(corpus−, Q)
        given each word compute correlation score:
        score ← TrueSkill(words+, words−, t = [1, 2])
```

Team	Words	S^i	S^{i+1}
$word^+$	bioengineered	$22,738$	$22,809$
$word^-$	economic	$0,001$	$0,022$

Where: S^i denotes current correlation score for each word, and S^{i+1}, the updated value after matching pages (positive against negative page), *bioengineered* is a word near *excellent*, a seed word $\in P$, and *economic* is near *wrong*, seed word $\in Q$ when the Window Size algorithm has distance $k = 1$. Thus, when the same $corpus^+$ has a match with other $corpus^-$:

$corpus^-$ = Various studies $\cdots$ *poor agricultural* income $\cdots$.

Team	Words	S^i	S^{i+1}
$word^+$	bioengineered	$22,738$	$28,023$
$word^-$	agricultural	$-0,108$	$-4,764$

It is worth noting that *agricultural* becomes a more negative word than *economic* because its value decreases more after the match using the same positive word: *bioengineered*. On one hand,

if a word is often found in a $corpus^-$ its value tends to decrease. On the other hand, if it is in a $corpus^+$ its value will increase. If the word is found in both corpus it tends to be constant. In the next section, experiments results are showed.

3 Experiments

In order to validate our approach experiments over two data sets were conducted. The polarity of each opinion from domains (Agricultural tweets and Movie reviews) is predicted using the inferred lexicons, $AcroDef_{MI3}$ and TrueSkill measures. *Precision*, *recall* and *f-score* were measured in order to compare to the SentiWordNet approach. Data sets used are described in the next section.

3.1 Datasets

The domains keywords used in queries were: Agriculture domain = {*gmo*, *agricultural biotechnology*, *biotechnology for agriculture*}, and Movie domain = {*cinema*, *film*, *movie*}. In order to test the agricultural domain, tweets using these keywords were collected and manually classified. There were 50 positive and 61 negative tweets. The Movie domain [4] is based on (Pang and Lee, 2004). The number of positive and negative is respectively 1000 and 1000.

A simple classification procedure was used. In order to do so, the number of positive and negative words in each tweet or review is computed using the inferred lexicons. If the difference is greater than zero then it is classified as positive, otherwise is negative. The following kind of lexicons were used to sentiment classification:

- $MI3$: seed words + WS with $AcroDef_{MI3}$.
- TS: seed words + WS with $TrueSkill$.
- SWN: SentiWordNet.

where WS denotes words extracted with window size. Finally, the number of web pages retrieved during the corpus acquisition for each seed word was $k = 20$.

In the next, we show word distributions for each type of lexicon.

3.2 Seed words

Table 1 shows the number of occurrences for each seed word in web pages.

[4] http://www.cs.cornell.edu/People/pabo/movie-review-data/

Seed Word	Domain	
	Agricultural	Movie
superior	42	10
good	406	178
positive	54	17
fortunate	23	4
excellent	47	20
correct	24	7
nice	40	23
poor	58	14
negative	65	25
wrong	64	43
bad	98	39
unfortunate	22	27
nasty	23	15
inferior	23	11

Table 1: Seed words(SW) frequency for Agricultural Domain

3.3 Window size

Using web pages number $k = 20$, a high number of low frequency adjectives are retrieved as shown in Figure 5a. To get a word near a seed word with window size$= 1$, the maximum distance allowed is 10 words per window size.

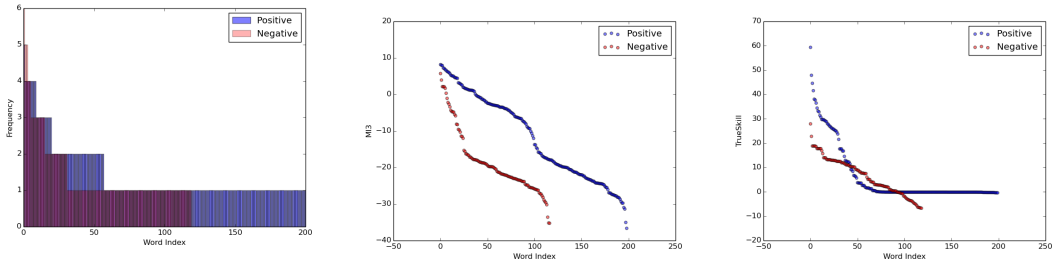
3.4 Measure function ($AcroDef_{MI3}$, TrueSkill)

Figures 4, 5 show words scores obtained using the measures proposed. It can be observed that words better discriminate than frequencies of Window Size Algorithm as shown in Figure 5a. Table 2, Table 3 show the top 5 words of inferred lexicons.

3.5 SentiWordNet

SentiWordNet[5] is a lexical resource for opinion mining. It assigns to each synset of WordNet three sentiment scores, positive, negative and neutral. We compute differences between positive and negative scores. If the result is greater than zero then the polarity of the word is positive, otherwise negative. SentiWordNet assigns a different score for each word according its context. As context is not considered, higher positive and negative word scores are obtained. Finally, SentiWordNet comprises 21479 adjectives and 117798 nouns.

[5] http://sentiwordnet.isti.cnr.it/

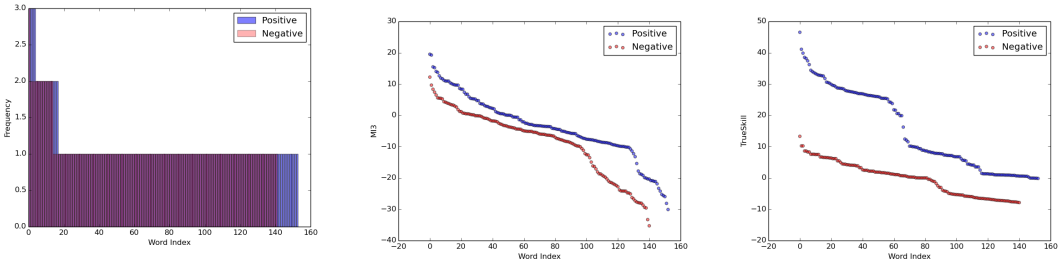


(a) Word frequency using Window Size (WS)

(b) MI3

(c) TS

Figure 4: Adjective words for Agricultural domain



(a) Word frequency using Window Size (WS)

(b) MI3

(c) TS

Figure 5: Adjective words for Movie domain

Adjective Words		
WS	*MI3*	*TS*
Positive		
dark	cheap	qualified
daily	fat	inconclusive
active	coconut	ideal
favorite	false	fresh
full	probiotic	active
Negative		
stunning	rural	devastating
german	chemical	irreversible
hungry	standard	sick
wealthy	brutish	general
medical	hungry	chemical

Table 2: Top 5 adjectives for Agricultural domain

Noun Words		
WS	*MI3*	*TS*
Positive		
flavor	luck	note
fit	night	commitment
movie	morning	judgment
opportunity	source	continent
job	vodka	jihad
Negative		
farmer	regulation	farmer
debate	bread	regulation
cost	guy	group
intensity	gmos	problem
gmos	soil	tomato

Table 3: Top 5 nouns for Agricultural domain

3.6 Classification

In order to classify opinions the inferred lexicons are used. We have positive and negative lexicons (dictionary) for each data sets (Agricultural, Movie), as shown in Table 7. In the Agricultural domain 32 new words have been learned that do not appear in SentiWordNet. Likewise, in the Movie domain 20 new words that do not appear in SentiWordNet have been learned. Table 6 shows top 10 new words ordered by their correlation score value. In order to validate the algorithms we calculate *recall*, *precision* and *f-score*. Figures 7, 6 show the *recall*, *precision* and *fscore* using each word type(*noun*, *adjectives*), and the results using *MI3*, SentiWordNet and TrueSkill.

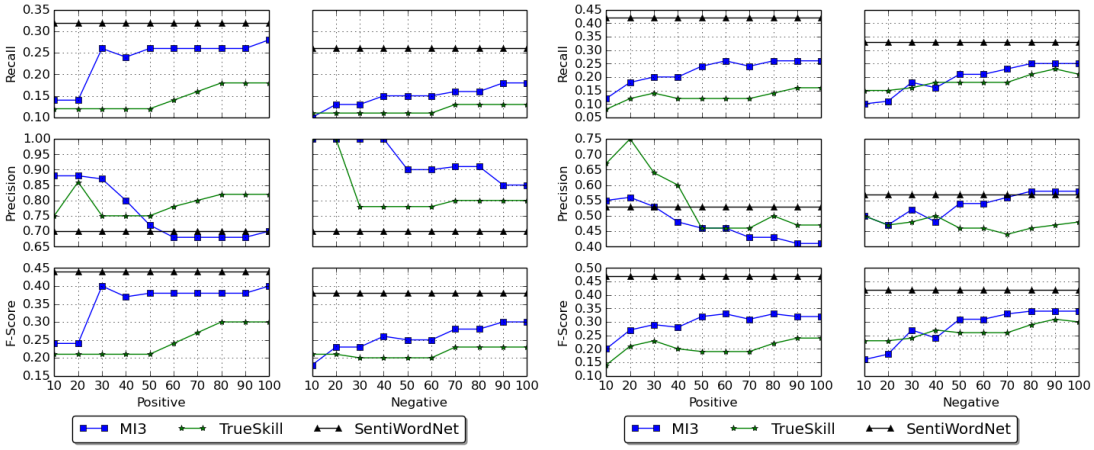


Figure 6: Tweet classification, left with adjectives, right with nouns.

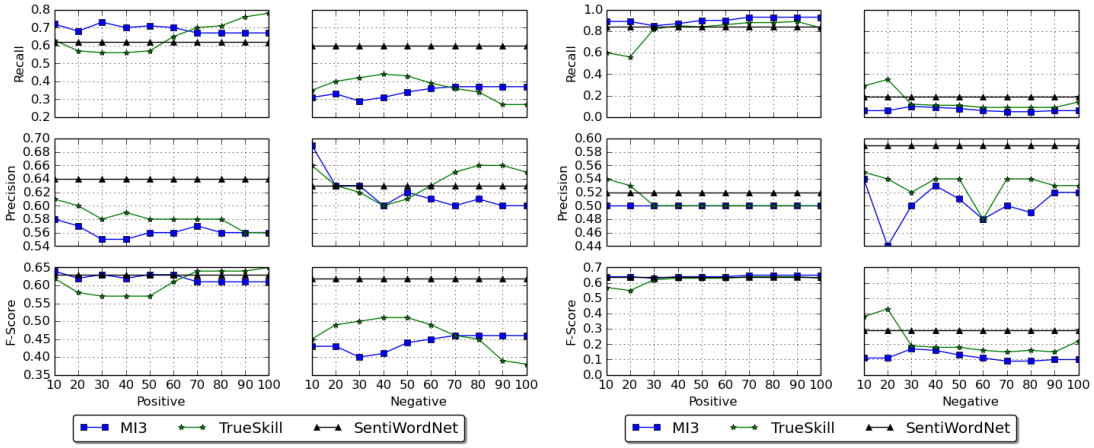


Figure 7: Classification using Movie Reviews, left with adjectives, right with nouns.

Adjective Words		
WS	*MI3*	*TS*
Positive		
comfy	big	late
expensive	real	clear
late	natured	common
french	sound	french
infectious	easy	commercial
Negative		
makeshift	video	emotional
video	pretty	russian
lost	english	cartoonish
fast	acting	treacly
attentive	full	dull

Table 4: Top 5 adjectives for Movie domain

Noun Words		
WS	*MI3*	*TS*
Positive		
info	place	info
wife	day	place
people	food	staff
service	feel	credo
party	luck	city
Negative		
blood	thing	blood
rate	word	character
character	person	idea
interest	blood	progression
time	video	activity

Table 5: Top 5 nouns for Movie domain

4 Discussion of the results

When the inferred lexicon for the Movie domain is considered, TrueSkill performs better (Recall, Precision and F-Score) than SentiWordNet and $AcroDef_{MI3}$ for positive reviews using adjectives and nouns. When negative reviews are con-

Domain	
Agricultural	Movie
chocolaty	configurable
glyphosate	updated
phosphonic	readymade
carfentrazone	nature
sporogene	directorial
kalu	spendidly
protato	cartoonish
adeed	mic
phthalates	showreel
genotoxicity	coverup

Table 6: Top 10 words of inferred lexicons using $AcroDef_{MI3}$ and TrueSkill methods, which are not in $SentiWordNet$

Word	Positive	Negative
	Agricultural	
Adjective	200	119
Noun	314	189
	Movie	
Adjective	153	141
Noun	171	183

Table 7: Total of inferred lexicon words by domain.

sidered TrueSkill performs better using nouns than adjectives.

On the other hand, in the Agricultural domain, SentiWordNet performs better than $AcroDef_{MI3}$ and TrueSkill. This is due to the agricultural domain was collected from Twitter. Tweets are short texts that usually have more seed words and common words as shown in Table 1. The agricultural domain has frequent seed words.

5 Conclusion

Most of the dictionary-based algorithms for sentiment analysis consider word frequency in documents. However, this research has shown that collected corpus words with low frequencies can be useful to set polarities. Thus, We propose a dictionary-based algorithm for sentiment analysis that uses $AcroDef_{MI3}$ and TrueSkill methods so as to compute correlation word scores that allow us to differentiate between positive and negative polarities. This is particularly useful for low frequency words obtained from corpus. In addition, by using the Window Size Algorithm, it is possible to obtain new adjectives entries in both agricultural and movie domains when compared to SentiWordNet.

Acknowledgments

This work has been supported and financed by FONDECYT.

References

Abdelmalek Amine, Reda Mohamed Hamou, and Michel Simonet. 2014. Detecting opinions in tweets. volume abs/1402.5123.

Benjamin Duthil, François Trousset, Mathieu Roche, Gérard Dray, Michel Plantié, Jacky Montmain, and Pascal Poncelet, 2011. *Towards an Automatic Characterization of Criteria*, pages 457–465. Springer Berlin Heidelberg, Berlin, Heidelberg.

Ali Harb, Michel Plantie, Gerard Dray, Mathieu Roche, Francois Trousset, and Pascal Poncelet. 2008. Web opinion mining: How to extract opinions from blogs? In *Proceedings of the 5th International Conference on Soft Computing As Transdisciplinary Science and Technology*, CSTST 08, pages 211–217, New York, NY, USA. ACM.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. Trueskill(tm): A bayesian skill rating system. In *Advances in Neural Information Processing Systems 20*, pages 569–576. MIT Press, January.

Edison Marrese-Taylor, Juan D. Velsquez, Felipe Bravo-Marquez, and Yutaka Matsuo. 2013. Identifying customer preferences about tourism products using an aspect-based opinion mining approach. *Procedia Computer Science*, 22(0):182 – 191. 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems - {KES2013}.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mathieu Roche and Violaine Prince, 2007. *Modeling and Using Context: 6th International and Interdisciplinary Conference, CONTEXT 2007, Roskilde, Denmark, August 20-24, 2007. Proceedings*, chapter AcroDef: A Quality Measure for Discriminating Expansions of Ambiguous Acronyms, pages 411–424. Springer Berlin Heidelberg, Berlin, Heidelberg.

Peter D. Turney. 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA. Association for Computational Linguistics.

R. Varghese and M. Jayasree. 2013. Aspect based sentiment analysis using support vector machine classifier. In *Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on*, pages 1581–1586, Aug.

Guangwei Wang and Kenji Araki. 2007. Modifying so-pmi for japanese weblog opinion mining by using a balancing factor and detecting neutral expressions. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, NAACL-Short '07, pages 189–192, Stroudsburg, PA, USA. Association for Computational Linguistics.