Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 112 (2017) 1711–1720

Procedia Computer Science

www.elsevier.com/locate/procedia

21th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France

Harnessing Ratings and Aspect-Sentiment to Estimate Contradiction Intensity in Temporal-Related Reviews

Ismail Badache[a,*], Sébastien Fournier[a], Adrian-Gabriel Chifu[a]

[a]*LSIS UMR 7296 CNRS, University Aix-Marseille, Marseille, France*

Abstract

Analysis of opinions (reviews) generated by users becomes increasingly exploited by a variety of applications. It allows to follow the evolution of the opinions or to carry out investigations on products. The detection of contradictory opinions about a web resource (e.g., courses, movies, products, etc.) is an important task to evaluate the latter. This paper focuses on the problem of detecting contradictions in reviews based on the sentiment analysis around specific aspects of a resource (document). In general, for web resources such as online courses (e.g. on *Coursera* or *edX*), reviews are often generated during course sessions. Between each session users stop reviewing on the course, and this course may have updates. So, in order to avoid the confusion of contradictory reviews coming from two or more different sessions, the reviews related to a given resource should be firstly grouped according to their session. Secondly, certain aspects are extracted according to the distributions of the emotional terms in the vicinity of the most frequent names in the reviews collection. Thirdly, the polarity of each review segment containing an aspect is identified. Then taking only the resources containing these aspects with opposite polarities (positive, negative). Finally, we propose a measure of contradiction intensity based on the joint dispersion of the polarity and the rating of the reviews containing the aspects within each resource. The evaluation of our approach is conducted on the Massive Open Online Courses (MOOC) collection containing 2244 courses and their 73,873 reviews, collected from *Coursera*. The results of experiments revealed the effectiveness of the proposed approach to capture and quantify contradiction intensity.

Peer-review under responsibility of KES International

Keywords: Sentiment Analysis ; Aspect Extraction ; Rating ; Review ; Time ; Contradiction Intensity

1. Introduction

Nowadays, web 2.0 has become a participatory platform where people can express their opinions by leaving traces (e.g., review, rating, like) on web resources. Many services, such as blogs and social networks, allow the generation of these traces. They represent a rich source of social information, which can be analyzed and exploited in various applications and contexts [1,2]. In particular, the sentiment analysis [9], for example, to know a customer's attitude towards a product or its characteristics, or to reveal the reaction of people to an event. Such problems require rigorous analysis of the aspects covered by the sentiment to produce a representative and targeted result.

Another issue concerns the diversity of opinions on a given topic. Some works address it in the context of different fields of research. For example, *Wang and Cardie* [25] aim to estimate the sentiments of a sentence expressed

* Corresponding author.
E-mail address: Ismail.Badache@lsis.org

1877-0509
Peer-review under responsibility of KES International
10.1016/j.procs.2017.08.197

during a discussion and using them as characteristics in a classifier that predicts dispute in the discussion. *Socher et al.*[18] automatically identify debates between users from textual content (interactions) in forums, based on latent variable models.There are other studies in the analysis of user interactions, for example, extracting the *agreement* and *disagreement* expressions[15] and deducing the user relations by looking at their textual exchanges[8].

This paper investigates the subjects (e.g. aspects, topics) which the contradictions can occur in the reviews associated with a web resource (e.g. movies, courses, etc.) and how to quantify their intensity. We state 3 hypothesis:

(H1). *Reviews are related in time. The resource can be updated (e.g. corrected, changed), and these updates will be made after each session for the case of the MOOCs (Massive Open Online Courses) that are particularly the subject of our experiment. After each session, users stop reviewing (silence) until the next session. Therefore, temporal-related reviews mean the reviews generated during a specific period (called in this paper: session).*

(H2). *A contradiction in reviews related to a web resource means contradictory opinions expressed about a specific aspect, which is a form of diversity of sentiments around the aspect for the same resource.*

(H3). *An aspect with a negative sentiment in a review with a positive rating (and vice-versa) has a more important impact on the contradiction intensity than an aspect with a positive sentiment in a review with a positive rating.*

Moreover, a contradiction may occur in a review when an author presents different opinions on the same aspect, or through several reviews when different authors express different opinions on the same aspect. In order to design our model of automatic contradiction detection, fundamental tasks are applied: first, by automatic identification of aspects characterizing these reviews. Second, by detecting opposing opinions around each of these aspects through a model of sentiment analysis. Third, estimate the intensity of the contradiction in the reviews for each resource, using a measure of dispersion. Finally, tests carried out on a set of real data, as well as a user study, demonstrate that our approach is able to identify effectively and significantly the contradictions and their intensity. The research questions addressed in this paper are the following:

- **RQ1:** How to identify a contradiction in reviews?
- **RQ2:** How to estimate the intensity of contradiction between the reviews?
- **RQ3:** What is the impact of the joint consideration of the polarity and the rating of the reviews on the measurement of the intensity of the contradiction?

The rest of this paper is structured as follows: Section 2 presents some related work and the background. Section 3 details our approach for detecting contradictions. Then, Section 4, reports on the results of our experimental evaluation. Finally, Section 5 concludes this paper by announcing perspectives.

2. Background and Related Work

The detection and measurement of contradiction is a complex process that requires the use of several state of art methods (aspects detection, sentiment analysis). However, to our knowledge, very few studies treat the detection and the measurement of the intensity of contradiction. This section briefly presents some approaches of detecting controversies close to our work and then presents the approaches related to the detection of aspects and the sentiment analysis, which are useful for introducing our approach.

2.1. Contradiction and Controversy Detection

Several studies are close to the work presented in this paper. The works that are most related to our approach include *(Harabagiu et al., 2006)*[7], *(De Marneffe et al., 2008)*[5], *(Tsytsarau et al., 2010)*[22] and *(Tsytsarau et al., 2011)*[23], which attempt to detect contradiction in text. There are two main approaches, where contradictions are defined as a form of textual inference (e.g., entailment identification) and analyzed using linguistic technologies.

Harabagiu et al.[7] proposed an approach for contradiction analysis that exploits linguistic features (e.g., types of verbs), as well as semantic information, such as negation (explicit contradiction, e.g., "I love you - I do not love you") or antonymy (words that have opposite meanings, i.e., "hot-cold" or "light-dark"). Their work defined contradictions as textual entailment, when two sentences express mutually exclusive information on the same topic. Further improving the work in this direction, *De Marneffe et al.*[5] introduced a classification of contradictions consisting of seven types that are distinguished by the features that contribute to a contradiction, e.g., antonymy, negation, numeric mismatches which may be caused by erroneous data: "there are 7 wonders of the world - the number of wonders of

the world are 9". They defined contradictions as a situation where "two sentences are extremely unlikely to be true when considered together". *Tsytsarau et al.*[22,23] proposed an automatic and scalable solution for the contradiction detection problem. In their work, they studied the contradiction problem using sentiments analysis. The intuition of their contradiction approach is that when the aggregated value for sentiments (on a specific topic and time interval) is close to zero, while the sentiments diversity is high, the contradiction should be high.

Other common themes to our work concern the detection of controversies (dispute, controversy). Among these studies, several treat the controversy on Wikipedia and particularly in the case of the comments that surround the modifications of Wikipedia pages[25]. Other studies try to detect controversies on specifics domains, for example in the news[21] or in the debate analysis[18]. Other studies try to be more generic to detect the controversy on the web[11].

Our work has a similar motivation as those previous efforts, i.e., harnessing sentiment analysis around specific aspects (topics, subjects) to detect contradictions in text. However, to our knowledge none of these studies attempt to quantify the intensity of contradiction or controversy. Our goal is to measure contradiction intensity in reviews generated during a specific session, by exploiting their ratings and polarities around the aspects.

2.2. Aspect Detection Approaches

The first attempts to detect aspects were based on information extraction approach using the frequent nominal sentences[10]. Such approaches work well in the detection of aspects that are in the form of single name, but are less useful when the aspects have low frequency. Similarly, other studies use Conditional Random Fields (CRFs) or Hidden Markov Models (HMMs)[6]. Other methods are unsupervised and have proven their effectiveness by building a Multi-Grain Topic Model[20] and HASM[12] (unsupervised Hierarchical Aspect Sentiment Model) which allows to discover a hierarchical structure of the sentiment based on the aspects in the unlabelled online reviews. In our work, the explicit aspects are extracted using the unsupervised method presented by *Poria et al.*[17]. This method, based on the use of extraction rules for product reviews, corresponds to our experimental data (*Coursera* reviews).

2.3. Sentiment Analysis Approaches

Sentiment analysis has been the subject of much previous research. As in the case of the aspects detection, the supervised and unsupervised approaches each have their solutions. Thus, some unsupervised approaches are based on lexicons such as the approach developed by *Turney*[24] or corpus-based methods such as in *(Mohammad, 2013)*[14]. *Pang et al.*[16] proposed supervised approaches, which perceive the task of sentiment analysis as a classification task and therefore use methods such as SVM or Bayesian networks. Other recent studies are based on the RNN (Recursive Neural Network), such as in *(Socher et al., 2013)*[19]. In our work, sentiment analysis is only a part of the process of contradiction detection, it is inspired by *Pang et al.*[16] work using a Bayesian classifier. Naive Bayes is a probabilistic model that gives good results in the classification of sentiments and generally takes less time for the training compared to models like SVM (Support Vector Machines).

3. Detection of Contradictions

Our approach is based on both automatic detection of aspects within reviews as well as sentiment analysis of these aspects. In addition to the contradiction detection, our goal is also to estimate the intensity of these contradictions. To measure these contradictions, two dimensions are jointly exploited: the polarity around the aspect as well as the rating associated with the review. In order to consider these dimensions, reviews are modeled as a scatter plot using a dispersion function, whose coordinates are polarities and ratings.

Our approach consists in exploiting *reviews* and *ratings*, as source of evidence for detecting contradictions in reviews containing some specific aspects for a web resource. This social information can be represented by a triplet $< R_1, R_2, R_3 >$ where $R_1 = \{D_1, D_2, ...D_h\}$, $R_2 = \{re_1, re_2, ...re_m\}$ and $R_3 = \{rat_1, rat_2, ...rat_m\}$ are finite sets of instances: *Resources (documents)*, *Reviews* and *Ratings*, respectively. Resource D can be a traditional document such as a web page or a web 2.0 resource such as a video or any other similar entity. Each review is associated to a note generated by a user. The rating is a note on a discrete scale from 1 to a max value of 5, where for example 3 means "average" and 5 means "excellent".

3.1. Pre-processing

The pre-processing module is a key step before the contradiction intensity estimation and it consists of three main stages. First, the reviews are clustered according to their session. Second, aspects are extracted from the reviews. Third, sentiment analysis of the text related to these aspects is done. We detail these steps in the following.

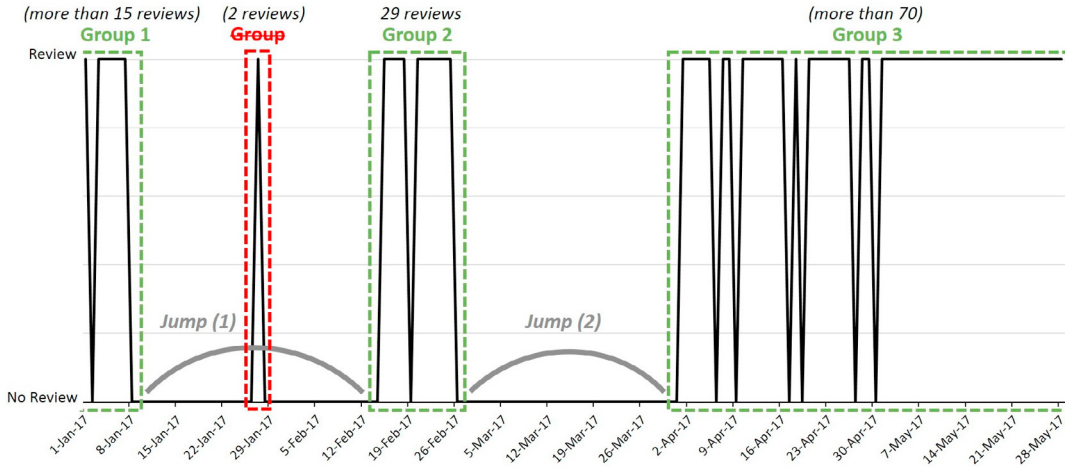

Figure 1: Distribution of some reviews in time for the course "Engagement and Nurture Marketing Strategies"

To obtain these reviews groups, we propose the (Algorithm 1) that brings together the reviews composing a session.

Algorithm 1: Grouping reviews of a resource according to their temporal session

Input: Days_Threshold (DsT), List_Reviews (LRs)
Output: Groups_of_Reviews (GRs)

```
GRs⟵ ∅ ;        // Initializing Output list of Groups of Reviews related to a given resource (e.g.  course)
GRTemp ⟵ ∅ ;          // Initializing Temporary list saving each Group of Reviews belonging to a session
List_GRTemp ⟵ ∅ ;       // Initializing Temporary list saving the Groups of Reviews belonging to a session
List_Number_Reviews_per_Session(LNRpS) ⟵ ∅ ; // Initializing the Reviews number for each Group per session
K_Clusters=2 ;    // K-Means parameter representing 2 types of clusters (sufficient/deficient Reviews Group)
Target_Cluster ⟵ ∅ ;          // Initializing the list saving only true Reviews Group identified by K-Means
// Constructing of Reviews Groups according to Days_Threshold (DsT) (session duration)
for i=0; i<size(LRs) – 1; i++ do
    if |LRs(i).Date – LRs(i+1).Date| < DsT then
        GRTemp.add(LRs(i));
    else
        GRTemp.add(LRs(i)) ;
        List_GRTemp.add(GRTemp);
        Temp ⟵ ∅ ;
    end
end
// Counting the number of reviews in each reviews group saved in List_GRTemp
foreach gr ∈ List_GRTemp do
    LNRpS.add(size(gr));
end
// Applying K-Means algorithm to identify two types of Reviews Groups (sufficient/deficient) i.e.
[C1, C2, Cluster1, Cluster2] =K-Means(K_Clusters, LNRpS) ;                                    // K-Means algorithm
// C1 and C2 are the centroids of each of the k types of clusters (Cluster1 and Cluster2)
if C1>C2 then
    Target_Cluster=Cluster1;
else
    Target_Cluster=Cluster2;
end
// Counting the number of reviews in each reviews group saved in List_GRTemp
foreach gr ∈ List_GRTemp do
    if size(gr) ∈ Target_Cluster then
        GRs.add(gr);
    end
end
```

3.1.1. Grouping Temporal-Related Reviews

Reviews are generated on resources chronologically, but some breaks have been observed on certain resources such as on the courses of "coursera". These disruptions represent a silence of reviews generation by the users. Analyzing these breaks, we observed that the reviews are temporally related to the evolution of the resource, because this resource is often updated after each 35 days (in average) for the case of "coursera" courses, for example. In order to properly handle the contradictions between the reviews, these reviews are grouped according to their session in time. These groups are defined after each 35-days jump without reviews or without significant number of reviews (see figure 1).
Remark. Only the groups (clusters) of reviews containing sufficient number of reviews are considered, i.e. for example, the groups of reviews containing 1 or 2 reviews are ignored, hence, the using of K-Means [13].

Once the reviews are grouped according to their session (see Algorithm 1), the pre-processing continues for the second step which consists in extracting the characterizing aspects of all these groups of reviews.

3.1.2. Extraction of Aspects

In our study, an aspect is a frequently occurring nominal entity in reviews and it is surrounded by emotional term. In order to extract the aspects from the reviews text, the following instructions are applied:

1. Terms frequency calculation of the reviews corpus,
2. Terms categorization (part-of-speech tagging) of reviews using *Stanford Parser*[1],
3. Selection of terms having nominal category without considering stopwords,
4. Selection of nouns with emotional terms in their five-neighborhoods (using *SentiWordNet*[2] dictionary),
5. Extraction of the most frequent (used) terms in the corpus among those selected in the previous step. These terms will be considered as aspects.

Example: Let *re* be a review associated to a document *D*. Table 1 summarizes the applying of the 5 steps described above to extract aspects from the review *re*.

Table 1: Steps to extract the aspects of a review

Step	Description
(1)	course : 44219, material : 3286, assignments : 3118, content : 2947, lecturer : 2705,.....term$_i$
(2)	The/DT **lecturer**/NN was/VBD an/DT annoying/VBG **speaker**/NN and/CC very/RB repetitive/JJ ./. I/PRP just/RB could/MD n't/RB listen/VB to/TO him/PRP .../: I/PRP 'm/VBP sorry/JJ ./. There/EX was/VBD also/RB so/RB much/JJ about/IN human/JJ **development**/NN **etc**/NN that/IN I/PRP started/VBD to/TO wonder/VB when/WRB the/DT **info**/NN about/IN **dogs**/NNS would/MD start/VB .../: ./. I/PRP found/VBD the/DT **formatting**/NN so/RB different/JJ from/IN other/JJ **courses**/NNS I/PRP 've/VBP taken/VBN ,/, that/IN it/PRP was/VBD hard/JJ to/TO get/VB started/VBN and/CC figure/VB **things**/NNS out/RP ./. Adding/VBG to/TO that/DT ,/, was/VBD the/DT constant/JJ **interruption**/NN of/IN the/DT "/" paid/VBN **certificate**/NN "/" **page**/NN ./. If/IN I/PRP answer/VBZ "/" no/UH "/" once/RB ,/, please/VB leave/VB me/PRP alone/RB !/. I/PRP also/RB think/VBP it/PRP 's/VBZ a/DT bit/RB suspect/JJ for/IN a/DT **prof**/NN to/TO be/VB plugging/VBG his/PRP$ own/JJ **book**/NN for/IN one/CD of/IN these/DT **courses**/NNS ./.
(3)	lecturer, speaker, development, dogs, formatting, courses, interruption, certificate, page, prof
(4)	lecturer, speaker
(5)	lecturer

First, terms frequencies are computed from the set of reviews (e.g., "course", "material", "assignments", "content", "lecturer" appear 44219, 3286, 3118, 2947 , 2705, respectively). Secondly, each word is labeled grammatically (e.g., "NN", "NNS" mean name in singular and plural, respectively[3]). Thirdly, after removing stopwords, only termes having nominal category ("NN", "NNS") are selected. Fourthly, the nouns surrounded by emotional terms (using *SentiWordNet* dictionary) in their 5-neighborhoods are extracted (The *lecturer* was an annoying *speaker* and very repetitive). Finally, only the names that are among the most frequent names in the corpus of reviews are considered as useful aspects (*lecturer*).

Once we have defined the list of useful aspects that characterize our data collection, we estimate the polarity of sentiment around these aspects. The following section presents our model of sentiment analysis.

[1] http://nlp.stanford.edu:8080/parser/
[2] http://sentiwordnet.isti.cnr.it/
[3] http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

3.1.3. Sentiment Analysis

The sentiments are a real number in the range $[-1, 1]$ which indicates the polarity of the opinion expressed in the review segment with respect to an aspect (called review-aspect *ra*). Negative and positive values respectively represent negative and positive opinions.

Pang's researches[16] indicate that standard machine learning methods perform very well. Therefore, in order to estimate the sentiment of the review-aspect *ra*, we used Naïve Bayes algorithm. After several empirical experiments, the review-aspect *ra* is defined by an excerpt of 5 words before and after the aspect in the review *re*. Our supervised sentiment model take into account:

(a) Negation handling (word preceded by *"no", "not", "n't"*). Our algorithm uses a state variable (Negative) to store the negation state. It transforms a word preceded by *"no", "not"* or *"n't"* into *"not_"+word*. Whenever the negation state variable is verified, read words are treated as *"not_"+word*. The state variable is reset when a punctuation mark ("?.,!:;") is encountered or when there is a double negation. The negative forms with respect to the normal forms of the same words are balanced during the training. This is to ensure that the number of *"not_"* forms is sufficient for the classification;

(b) Combinations (*bigrams*) of adjectives with other words such as intensifiers and adverbs (e.g. "*very* bad" and "*absolutely* recommended").

3.2. Measure of Contradiction

A typical Contradiction Analysis application needs to follow the same steps we identified for Opinion Mining, namely, topic identification and sentiment extraction. For certain techniques of Contradiction Analysis it is possible to rely directly on the output of Opinion Mining, thus simplifying the entire workflow. Then, we need to have a contradiction detection step, where individual sentiments are processed in order to reveal contradictions.

In the contradiction detection step, the goal is to efficiently combine the information extracted in the previous steps, in order to determine the topics and time intervals in which contradictions occur. In this step, statistical methods can be used, as well as clustering, or other unsupervised methods. The contradiction detection step requires efficient data mining methods, which will enable the online identification of contradictions, and will have the ability to work on different time resolutions.

The main research problem addressed in this paper is related to the effective detection of contradictory opinions in reviews related to specific aspects, as well as their contradiction intensity.

A review on a given resource (e.g. courses, movies, media) covers one or more specific aspects (e.g. *lecturers* of courses, *actors* of movies, etc). For each review some sentiments are expressed around these aspects. Our goal is first to identify and record the polarities of these sentiments as described previously. Then, according to these polarities (positive or negative) the contradictory reviews-aspect ra_i are identified for each resource.

Definition 1. *There is a contradiction on an aspect between two portions of reviews containing this aspect (with* $ra_1, ra_2 \in D$*), where* $pol(ra_1) \cap pol(ra_2) = \phi$*, when the opinion around the aspect are opposite.*

The degree of contradiction around an aspect between the reviews is estimated using two dimensions: the polarity *pol* of the review-aspect *ra* and its rating *rat*. We assume that the greater the distance (i.e. dispersion) between these values related to each review-aspect ra_i of the same document D, the degree of contradiction is more important.

Let $ra(pol_i, \widehat{rat_i})$ be a point on the cloud (plane). The dispersion indicator with respect to the centroid $ra_{centroid}$ is defined as follows:

$$Disp(ra^{pol_i}_{rat_i}, D) = \frac{1}{n}\sum_{i=1}^{n} Distance(pol_i, \widehat{rat_i}) \quad (1)$$

with:

$$Distance(pol_i, \widehat{rat_i}) = \sqrt{(pol_i - \overline{pol})^2 + (\widehat{rat_i} - \overline{rat})^2} \quad (2)$$

$Distance(pol_i, \widehat{rat_i})$ is the distance between the point ra_i of the cloud and the centroid $ra_{centroid}$, and n is the number of points ra_i of the cloud. The two quantities pol_i and rat_i have different scale, it is essential to normalize them. The polarity pol_i is a probability, but the ratings rat_i values can be normalized as follows: $\widehat{rat_i} = \frac{rat_i - 3}{2}$ $(\widehat{rat_i} \in [-1, 1])$.

By assigning each point of the cloud having the same mass $1/n$, the indicator $Disp(ra^{pol_i}_{rat_i}, D)$ represents the inertia of the cloud with respect to the centroid $ra_{centroid}$ (see figure 2).

- *Disp* is positive or zero; $Disp = 0$ means that all points in the cloud are merged into $ra_{centroid}$ (no dispersion).
- *Disp* increases when a point in the cloud is moved away from $ra_{centroid}$ (i.e. when the dispersion is increased).

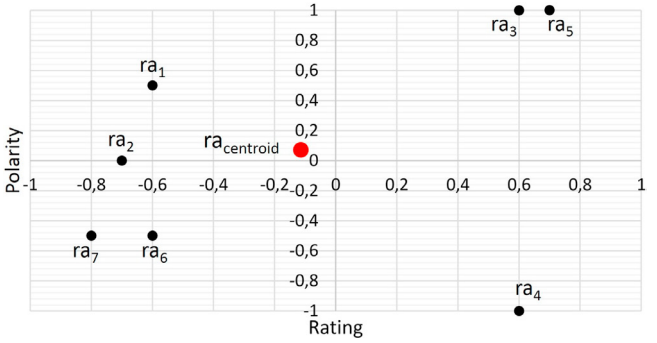

Figure 2: Dispersion of reviews-aspect ra_i in the cloud (plane)

The coordinates $(\overline{pol}, \overline{rat})$ of the centroid $ra_{centroid}$ can be calculated in two different ways. A simple way is to calculate the average of the points, in this case the centroid $ra_{centroid}$ corresponds to the average point of the coordinates $ra_i(pol_i, \widehat{rat_i})$. Another finer way is to weight this average by the difference in absolute value between the two values of the coordinates (dimensions: polarity and rating).

3.2.1. Centroid based on average of dimensions (polarities and ratings)

Let the statistical series with two variables (dimensions), *pol* and *rat*, where values are couples (pol_i, rat_i). The centroid (mean point of the series) based on the average of both polarities and normalized ratings is the point $ra_{centroid}$ in figure 2, which their coordinates are computed as follows:

$$\overline{pol} = \frac{pol_1 + pol_2 + ... + pol_n}{n}; \qquad \overline{rat} = \frac{\widehat{rat_1} + \widehat{rat_2} + ... + \widehat{rat_n}}{n} \tag{3}$$

3.2.2. Centroid based on the weighted average of dimensions

In this case, the coordinates of the centroid $ra_{centroid}$ are computed based on the weighted average of polarities and ratings as follows:

$$\overline{pol} = \frac{c_1 \cdot pol_1 + c_2 \cdot pol_2 + ... + c_n \cdot pol_n}{n}; \qquad \overline{rat} = \frac{c_1 \cdot \widehat{rat_1} + c_2 \cdot \widehat{rat_2} + ... + c_n \cdot \widehat{rat_n}}{n} \tag{4}$$

where n is the number of points ra_i in the space. The coefficient c_i is computed as follows: $c_i = \frac{|pol_i - \widehat{rat_i}|}{2n}$

In this two-dimensional vector representation, our hypothesis is that a point in this space is more important if the values of both dimensions are the most distant. We believe that a negative aspect in a review with a high rating has more weight and vice-versa. Consequently, a coefficient of importance for each point in space is calculated. This coefficient is based on the difference in absolute value between the values of the dimensions. The absolute value ensures that the coefficient is positive. The division by $2n$ represents a normalization by the maximum value of the difference in absolute value ($max(|pol_i - \widehat{rat_i}|) = 2$) and n. For example, for a polarity of -1 and a rating of 1, the coefficient is $1/n$ ($|-1-1|/2n = 2/2n = 1/n$), and for a polarity of 1 and a rating of 1, the coefficient is 0 ($|1-1|/2n = 0$).

4. Experimental Evaluation

In order to validate our approach, a series of experiments was carried out on reviews collected from the site of *Coursera*[4]. Our main objective in these experiments is to evaluate the impact of taking into account the sentiment analysis and rating on the detection of contradictions in the reviews around certain specific aspects identified automatically, as well as evaluating the impact of the averaged and weighted centroid on the contradiction intensity.

[4] https://www.coursera.org/

4.1. Description of Test Dataset

To the best of our knowledge, there is no standard dataset to evaluate the effectiveness of contradiction detection systems. Therefore, 2244 English courses are extracted from *coursera.org* via its API[5]. For each course, we have also collected its reviews and ratings via the *parsing* of the course web pages (see the statistics in the table 2).

Table 2: Data statistics of Coursera.org dataset

Field	Total Number
Courses	2244
Courses Rated	1115
Reviews	73873
Ratings	298326
Reviews ★	1705
Reviews ★★	1443
Reviews ★★★	3302
Reviews ★★★★	12202
Reviews ★★★★★	55221

Table 3 presents some aspects among 22 useful aspects captured automatically from the reviews. To obtain judgments of contradictions and sentiments for a given aspect: a) 3 users were asked to assess the sentiment class for each review-aspect; b) 3 other users assessed the degree of contradiction between reviews-aspect. In average 6 reviews-aspect per course are judged manually for each aspect (totally: 1320 reviews-aspect of 220 courses i.e. 10 courses for each aspect). We note that each aspect has been judged by 3 users.

Table 3: Statistics on the aspects extracted from the reviews of Coursera.org

Aspects	#Rating 1	#Rating 2	#Rating 3	#Rating 4	#Rating 5	#Negative	#Positive	#Review	#Course
Assignment	204	208	333	840	1726	1057	1763	2384	186
Content	176	179	341	676	1641	505	1496	1883	207
Exercise	29	46	94	290	693	195	531	673	58
Information	100	123	238	523	1389	299	1165	1359	143
Instructor	129	106	122	302	1514	295	1107	1322	140
Knowledge	74	72	121	400	1604	905	791	1243	178
Lecture	185	206	290	613	1762	763	1508	1988	208
Lecturer	32	41	48	85	461	55	193	236	39
Lesson	40	59	75	224	712	187	420	554	84
Material	191	203	328	722	2234	784	1693	2254	237
Method	19	23	40	125	404	53	187	224	31
Presentation	46	50	75	142	413	93	196	274	54
Professor	76	74	129	452	3001	331	2234	2369	151
Quality	55	53	51	110	372	113	170	262	54
Question	94	98	172	284	356	311	289	502	104
Quiz	151	155	221	401	581	481	475	824	128
Slide	56	64	81	121	115	131	102	192	47
Speaker	17	15	34	70	170	34	72	103	24
Student	140	105	171	383	1035	519	709	1066	172
Teacher	62	46	82	293	2180	248	1481	1642	119
Topic	67	89	176	437	1154	236	951	1066	130
Video	228	238	356	707	1614	941	1421	2058	245

To evaluate sentiments and contradictions in the reviews-aspect of each course, 3-points scale are used for sentiments: *Negative, Neutral, Positive*; and 5-points scale for contradictions: *Not Contradictory, Very Low, Low, Strong* and *Very Strong*.

Analyzing the agreement degree between assessors for each aspect using Kappa Cohen measure k[4]. The $k = 0.76$ for sentiment assessors and $k = 0.68$ for contradiction assessors, which corresponds to a substantial agreement. The measure of the agreement varies from 0.41 to 0.88.

[5] `https://building.coursera.org/app-platform/catalog`

4.2. Results and Discussions

To evaluate the performance of our approach, an experimentation was conducted (official measure on SemEval tasks[6]), by using the correlation coefficients of *Pearson* and *Spearman*[3], between the contradiction judgments given by the assessors and our obtained results.

Remarks: First, our sentiment analyzer takes as a training set 50,000 reviews of *IMDb* movies[7] (Due to the similarity of the vocabulary used in the reviews on *IMDb* and *coursera*), and as a test set our reviews-aspect of *coursera*. Second, our sentiment analysis system provides an accuracy of 79% according to the correlation study. Third, assessors' judgments on sentiments are considered as perfect (reference) results and represent an accuracy of 100%.

Table 4 shows the correlation values obtained by the Config (1) presented in 3.2.1 (centroid based on average of ratings and polarities) and the Config (2) presented in 3.2.2 (centroid based on the weighted average of ratings and polarities). The results are discussed in the following.

Table 4: Results of correlations between contradiction judgments and the results of our approach

Correlation Measure	Config (1): centroid based on average of ratings and polarities	Config (2): centroid based on weighted average of ratings and polarities
(a) Correlation between contradiction judgments and the results of our approach (with sentiment analysis accuracy of 79%**)**		
Spearman	0.58	0.69
Pearson	0.61	0.71
(b) Correlation between contradiction judgments and the results of our approach (with sentiment analysis accuracy of 100%**)**		
Spearman	0.70	0.87
Pearson	0.73	0.91

1) Centroid based on the average of dimensions. The results show that the dispersion measurement based on the averaged centroid provides a positive correlation with judgments, Spearman: 0.58, 0.70 and Pearson: 0.61, 0.73. Indeed, the more polarities between the reviews-aspect related to their session are opposite, the more the cloud points diverge from the centroid, hence the increased intensity dispersion. In addition, the results obtained using the manual sentiments judgments (table 4 (b)) surpass those obtained using our sentiment analysis model (table 4 (a)) with an approximate rate (Spearman and Pearson) of 20%. Therefore, losing 21% in sentiments accuracy involves a 20% loss in detecting contradictions performance.

2) Centroid based on the weighted average dimensions. The configuration (2) results are also positive (Spearman: 0.66, 0.87 and Pearson: 0.71, 0.91). The results obtained by considering the importance coefficient for each point of the space (review-aspect *ra*) are better compared to those obtained when this coefficient is ignored. These improvements are 14% (Spearman) using our sentiment model (table 4 (a)) and 25% (Spearman) using manual sentiment judgments (table 4 (b)). Indeed, the more divergent values of rating and polarity for every review-aspect, the higher the impact on contradiction intensity. Also, the results of configuration (2) presented in table 4 (b) are much better (Spearman: 0.87) than those presented in table 4 (a) (Spearman: 0.66). Therefore, the sentiment analysis model is an important factor that impacts the detection and the measurement of contradictions.

Finally, table 4 shows that the best results are obtained by configuration (2) which takes into account the importance coefficient c_i. The dispersion formula measuring the intensity of contradiction becomes more effective when combined with an effective sentiment analysis model, which leads to a significant improvement of the results. Noting also that grouping reviews according to their corresponding resources sessions, contribute significantly to these well results.

5. Conclusion

This paper introduced an approach that aims at estimating contradiction intensity, drawing attention to aspects in which users have contradictory opinions during a specific session. The intuition behind the proposed contradiction measure is that when the jointly dimensions (polarities and ratings) associated to reviews (on a specific aspect and session interval) are divergent (dispersed), while the sentiments diversity is high, then the contradiction should be high. Our study shows that contradiction exists if the sentiments around these reviews-aspect for the same resource are diverse (**H2**). Concerning our first hypothesis (**H1**), the formation of groups of reviews, with respect to their sessions of appearance, is benefit to avoid fake contradictions. In other words, clustering the reviews by sessions allow an effective treatment of contradictions in the reviews that are generated for a specific state of a resource (corresponding

[6] http://alt.qcri.org/semeval2016/task7/

[7] http://ai.stanford.edu/~amaas/data/sentiment/

session). Additionally, to quantify the contradiction, reviews-aspect are modeled as a scatter plot using dispersion function, where more the coordinates polarities and ratings are opposite (divergent) the more the impact is important on the contradiction intensity (**H3**). The validation of our overall assumptions was examined on the data collection of *Coursera.org*. The obtained results reveal the effectiveness of our approach. Particularly, the best results have been obtained using the centroid based on weighted average method, which verifies the third hypothesis.

The major weakness of this approach is its dependence on the quality of sentiment analysis model. As the training set (IMDb reviews) is different from the test set (Coursera reviews), if a word in the training set appears only in one class and does not appear in any other class, in this case, the classifier will always classify the text to that particular class. Moreover, the sentences are not processed, only predefined window of 5 words before and after the aspect is considered. Further scale-up experiments on other types of datasets are also envisaged. Even with these simple elements, the first results obtained encourage us to invest more in this track.

Acknowledgements

The project leading to this publication has received funding from Excellence Initiative of Aix-Marseille University - A*MIDEX, a French "Investissements d'Avenir" programme

References

1. Ismail Badache and Mohand Boughanem. Emotional Social Signals for Search Ranking. In *ACM SIGIR*, 2017.
2. Ismail Badache and Mohand Boughanem. Fresh and diverse social signals: any impacts on search? In *ACM CHIIR*, 2017.
3. Sorana-Daniela Bolboaca and Lorentz Jäntschi. Pearson versus spearman, kendall's tau correlation analysis on structure-activity relationships of biologic active compounds. *Leonardo Journal of Sciences*, 5(9):179–200, 2006.
4. J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37, 1960.
5. Marie-Catherine De Marneffe, Anna N Rafferty, and Christopher D Manning. Finding contradictions in text. In *ACL*, volume 8, pages 1039–1047, 2008.
6. Hussam Hamdan, Patrice Bellot, and Frederic Bechet. Lsislif: Crf and logistic regression for opinion target extraction and sentiment polarity analysis. In *SemEval*, page 753758, 2015.
7. Sanda Harabagiu, Andrew Hickl, and Finley Lacatusu. Negation, contrast and contradiction in text processing. In *AAAI*, volume 6, pages 755–762, 2006.
8. Ahmed Hassan, Amjad Abu-Jbara, and Dragomir Radev. Detecting subgroups in online discussions by modeling positive and negative relations among participants. In *EMNLP*, pages 59–70, 2012.
9. Amal Htait, Sébastien Fournier, and Patrice Bellot. LSIS at semeval-2016 task 7: Using web search engines for english and arabic unsupervised sentiment intensity prediction. In *International Workshop on Semantic Evaluation*, pages 469–473, 2016.
10. Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, 2004.
11. Myungha Jang and James Allan. Improving automated controversy detection on the web. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 865–868, 2016.
12. Suin Kim, Jianwen Zhang, Zheng Chen, Alice H Oh, and Shixia Liu. A hierarchical aspect-sentiment model for online reviews. In *AAAI Conference on Artificial Intelligence (AAAI-13)*, 2013.
13. James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, pages 281–297, 1967.
14. Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *SemEval*, 2013.
15. Arjun Mukherjee and Bing Liu. Mining contentions from discussions and debates. In *ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 841–849, 2012.
16. Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP*, pages 79–86, 2002.
17. Soujanya Poria, Erik Cambria, Lun-Wei Ku, Chen Gui, and Alexander Gelbukh. A rule-based approach to aspect extraction from product reviews. In *Proceedings of the second workshop on natural language processing for social media (SocialNLP)*, pages 28–37, 2014.
18. Minghui Qiu, Liu Yang, and Jing Jiang. Modeling interaction features for debate side clustering. In *ACM Conference on information & knowledge management*, pages 873–878, 2013.
19. Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, volume 1631, page 1642, 2013.
20. Ivan Titov and Ryan McDonald. Modeling online reviews with multi-grain topic models. In *WWW*, pages 111–120, 2008.
21. Mikalai Tsytsarau, Themis Palpanas, and Malu Castellanos. Dynamics of news events and social media reaction. In *KDD*, 2014.
22. Mikalai Tsytsarau, Themis Palpanas, and Kerstin Denecke. Scalable discovery of contradictions on the web. In *Proceedings of the 19th international conference on World wide web*, pages 1195–1196. ACM, 2010.
23. Mikalai Tsytsarau, Themis Palpanas, and Kerstin Denecke. Scalable detection of sentiment-based contradictions. *DiversiWeb, WWW*, 2011.
24. Peter D Turney. 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*, pages 417–424, 2002.
25. Lu Wang and Claire Cardie. A Piece of My Mind: A Sentiment Analysis Approach for Online Dispute Detection. In *Annual Meeting of the Association for Computational Linguistics*, pages 693–699, 2014.