# Finding and Quantifying Temporal-Aware Contradiction in Reviews

Ismail Badache, Sébastien Fournier, and Adrian-Gabriel Chifu
LSIS UMR 7296 CNRS, University Aix-Marseille, France
{ismail.badache, sebastien.fournier, adrian.chifu}@lsis.org

**Abstract.** Opinions (reviews) on web resources (e.g., courses, movies), generated by users, become increasingly exploited in text analysis tasks, the detection of contradictory opinions being one of them. This paper focuses on the quantification of sentiment-based contradictions around specific aspects in reviews. However, it is necessary to study the contradictions with respect to the temporal dimension of reviews (their sessions). In general, for web resources such as online courses (e.g. *coursera* or *edX*), reviews are often generated during the course sessions. Between sessions, users stop reviewing courses, and there are chances that courses will be updated. 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 corresponding session. Secondly, aspects are identified according to the distributions of the emotional terms in the vicinity of the most frequent nouns in the reviews collection. Thirdly, the polarity of each review segment containing an aspect is estimated. Then, only resources containing these aspects with opposite polarities are considered. Finally, the contradiction intensity is estimated based on the joint dispersion of polarities and ratings of the reviews containing aspects. The experiments are conducted on the Massive Open Online Courses data set containing 2244 courses and their 73,873 reviews, collected from *coursera.org*. The results confirm the effectiveness of our approach to find and quantify contradiction intensity.

**Keywords:** Sentiment analysis,Aspect detection,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 data, which can be exploited in various contexts [1], [2]. We mention in particular the field of sentiment analysis [9], where traces are exploited with the purpose of identifying a customer's attitude towards a product or its characteristics, or revealing 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 opinion diversity on a given topic. Some work addresses it in different fields of research. For example, Wang and Cardie [24] aim to identify the sentiments of a sentence expressed during a discussion and

using them as characteristics in a classifier that predicts dispute in the discussion. Socher et al.[19] automatically identify debates between users from textual content in forums, based on latent variable models. Other studies analyze 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 entities (e.g. aspects) for which the contradictions can occur in the reviews associated with a web resource (e.g. movies, courses) and how to estimate their intensity. A contradiction can occur when there are conflicting opinions for a specific aspect, which is a form of sentiment diversity. Moreover, this contradiction can occur throughout a specific time period (session).

To design our approach, different fundamental tasks are aggregated: 1) clustering reviews according to their session; 2) identifying aspects characterizing these reviews; 3) analyzing sentiments of reviews to capture opposing opinions around each aspect; 4) using a measure of dispersion to estimate the intensity of these contradictory opinions. Furthermore, tests carried out on a set of real data (*coursera.org*), as well as a user study, demonstrate that our approach is able to identify effectively and significantly the contradictions and their intensity. The main contributions of this work can be summarized as it follows:
**(C1).** We present an approach for contradiction detection, which is based on sentiment-aspect extraction during each specific session of reviews. Therefore, we group reviews according to their sessions.
**(C2).** We formally estimate the contradiction intensity, and further describe two variations of the solution: *Averaged centroid* and *Weighted centroid*.
**(C3).** We experimentally evaluate the proposed approach by creating our own data set collected from *coursera.org*. In addition, we perform a user study.

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 contradiction and compute its intensity. Section 4 reports on the results of our evaluation. Finally, Section 5 concludes this paper and announces our perspectives.

## 2 Background and Related Work

Contradiction detection is a complex process that requires the use of several state of the art methods (aspect detection, sentiment analysis). Moreover, to the best of our knowledge, very few studies treat the estimation of contradiction intensity. This section briefly presents the approaches related to aspect detection and sentiment analysis, which are useful for introducing our approach. Then, it presents some approaches of detecting controversies that are close to our work.

### 2.1 Aspect Detection Approaches

The first attempts to detect aspects were based on classical information extraction approaches using the frequent nominal sentences [10]. Such approaches work well for 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, such as [20], that built a Multi-Grain Topic Model, and [12] that proposed HASM (unsupervised Hierarchical Aspect Sentiment Model) which allow 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 in [17]. Poria et al., [17] proposed a rule-based approach that exploits common-sense knowledge and sentence dependency trees to detect both explicit and implicit aspects in product reviews.

### 2.2 Sentiment Analysis Approaches

Sentiment analysis has been the subject of much previous research. As in the case of the aspects detection, the supervised and the unsupervised approaches each propose their solutions.Thus, some unsupervised approaches are based on lexicons, such as the approach developed by [23] or corpus-based methods such as in [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 (Support Vector Machines) or Bayesian networks. Other recent studies are based on the RNN (Recursive Neural Network), such as in [19]. In our work, sentiment analysis is only a part of the contradiction detection process, it is inspired by the work of [16] 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 as SVM or RNN.

### 2.3 Contradiction and Controversy Detection Approaches

The most related works to ours include [7],[3],[21]and[22]. They attempt to detect contradiction in text. There are 2 main approaches, where contradictions are defined as a form of textual inference 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. [3] 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. [21], [22] proposed a 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 sentiment diversity is high, the contradiction should be

high. Another theme related to our work concern the detection of controversies and disputes. Among these studies, several treat the controversy on Wikipedia and particularly in the case of the comments that surround the modifications of Wikipedia pages [24]. Other studies try to detect controversies on specific domains, for example in news or in debate analysis [18]. Other studies try to be more generic and detect the controversy on web [11]. Ennals et al. [5] addressed the problem as a search of conflicting topics on the web through text patterns like "It is not correct that...". In addition, there is also a line of investigation known as controversy research. The aim is to identify whether Web contents deal with controversial topics (e.g. abortion, religion, same-sex marriage) and notify the user when the topic that they are searching is controversial [4].

Our work also has a certain proximity to previous efforts concerning the detection of contradiction in text. However, unlike previous works such as [3], [7], [5] and [4], that defined contradiction based on linguistic features and numeric mismatches, our work defines contradiction as sentiment-based conflicting opinions for a specific aspects, which is a form of sentiment diversity. This kind of contradiction can occur at one specific point of time or throughout a certain time period. In addition, to the best of our knowledge, none of previous studies attempt to quantify the intensity of contradiction or of controversy. Our main goal is to measure contradiction intensity in reviews generated during specific period (session), by exploiting their ratings and polarities around the aspects.

## 3 Time-Aware Contradiction Intensity

To measure contradiction intensity during a session, two dimensions are jointly exploited: the polarity around the aspect as well as the rating associated with the review. We used a dispersion function, based on these dimensions, that estimates the intensity between contradictory opinions (called after: reviews-aspect).

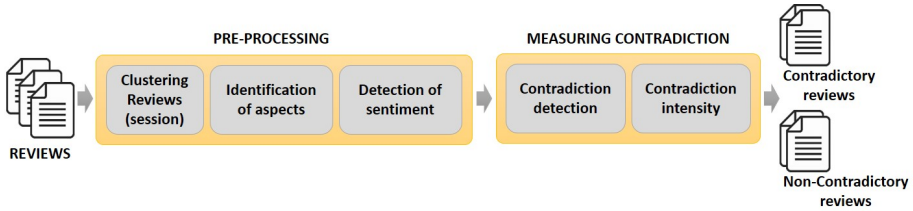


**Fig. 1:** Temporal sentiment-based contradiction intensity framework

### 3.1 Pre-processing

The pre-processing module consists of 3 main steps: 1) clustering reviews according to their session; 2) aspects extraction from reviews; and 3) sentiment analysis of the text related to these aspects. We detail these steps in the following.

**a) Session-Based Clustering of Reviews.** Generally, reviews are chronologically generated on resources, but some breaks (jumps) have been observed. These jumps (see figure 2) represent a silence of reviews generation by users during a specific period. Analyzing these jumps, we observed that the reviews

are temporally related to the resource evolution, because this resource is often updated after each specific period. In order to properly handle the contradictions between the reviews, these reviews should be grouped according to their session. The sessions are defined for each resource according to specific X-days jump (silence) without reviews or without significant number of reviews (see figure 2).

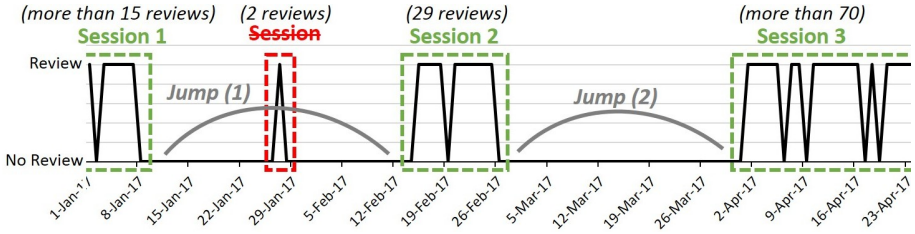


**Fig. 2:** Distribution of reviews in time of "Engagement & Nurture Marketing Strategies" course

To obtain these review groups, the following treatments are applied:

1. A threshold representing the jump duration is calculated for each course based on the mean of distances between reviews (e.g. the jump for the course "Engagement & Nurture Marketing Strategies" is: 35-days),
2. Grouping reviews according to the threshold for each course,
3. Identification of significant sessions by eliminating groups (false sessions) that contain insufficient number of reviews (no-dense sessions).

*Remark.* Only the groups (clusters) of reviews containing sufficient number of reviews are considered, i.e. for example, in figure 2 the group of reviews containing 2 reviews is ignored, hence, the using of K-Means [13]. In other words, k-means clustering is useful to partition the reviews into $k = 2$ sessions (dense or no-dense sessions (clusters) in terms of reviews quantity).

**b) Extraction of Aspects.** In our study, an aspect is a frequently occurring nominal entity in reviews and it is surrounded by emotional terms. In order to extract the aspects from the reviews' text, we were inspired by the work of Poria et al., [17]. This method corresponds to our experimental data (*coursera* reviews). Additionally, the following treatments are applied:

1. Term frequency calculation of the reviews corpus,
2. Term categorization (part-of-speech tagging) of reviews using *Stanford Parser*[1],
3. Selection of terms having nominal category (NN, NNS)[2],
4. Selection of nouns with emotional terms in their five-neighborhoods (using *SentiWordNet*[3] 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.

**c) Sentiment Analysis.** The sentiments are represented by 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 $ra$).

[1] http://nlp.stanford.edu:8080/parser/
[2] https://cs.nyu.edu/grishman/jet/guide/PennPOS.html
[3] http://sentiwordnet.isti.cnr.it/

Pang's research [16] indicates that standard machine learning methods perform very well, even definitively outperforming human classifiers. Therefore, in order to estimate the sentiment of the review-aspect $ra$, we used Naive Bayes algorithm to predict sentiment polarity [16]. After several empirical experiments, the review-aspect $ra$ is defined by an excerpt of 5 words before and after the aspect in review $re$. Our supervised sentiment model take into a account also:

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

### 3.2 Measuring Contradiction Intensity

Our application of time-aware contradiction analysis needs to follow the same steps that we previously identified for opinion mining, namely, session-based clustering of reviews, aspect identification and sentiment analysis.

A review on a given resource (e.g. courses, movies, media) and during a specific session 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. Then, we need to have a contradiction detection step, where individual sentiments (positive or negative) for each review-aspect $ra_i$ are processed in order to reveal contradictory reviews-aspect $ra_i$. In this step, the goal is to efficiently combine the information extracted in the previous steps, in order to determine the aspects and time intervals (session) in which contradictions occur.
**Definition.** *Contradiction exists between two portions of review-aspect* $ra_1$ *and* $ra_2$ *containing an aspect with* $ra_1$, $ra_2 \in D$ *(Document), when the opinions (polarities) around the aspect are opposite (i.e.* $pol(ra_1) \cap pol(ra_2) = \phi$*).*

The main research problem addressed in this paper is related to the effective estimation of the contradictory opinions intensity in reviews related to specific aspects. The degree of contradiction around an aspect between the reviews is estimated using two dimensions: the polarity $pol_i$ of the review-aspect $ra_i$ and its rating $rat_i$. 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 $ra_i$ be a point on the plane with coordinates $(pol_i, rat_i)$. The dispersion indicator with respect to the centroid $ra_{centroid}$ with coordinates $(\overline{pol}, \overline{rat})$ is defined as follows:

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

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

$Distance(pol_i, rat_i)$ represents the distance between the point $ra_i$ of the scatter plot and the centroid $ra_{centroid}$ (see figure 3), and $n$ is the number of $ra_i$. 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 values of the ratings $rat_i$ can be normalized as follows: $rat_i = \frac{rat_i - 3}{2}$ $(rat_i \in [-1, 1])$.

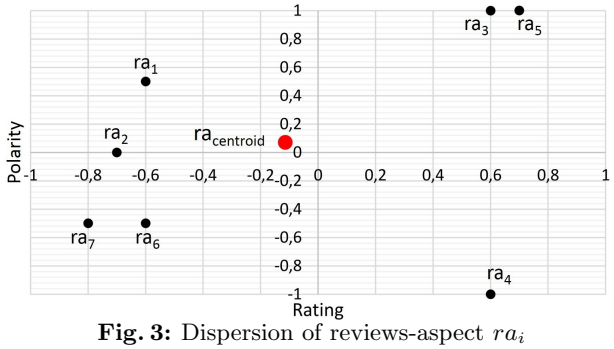


**Fig. 3:** Dispersion of reviews-aspect $ra_i$

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

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

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 $ra_i$, in this case the centroid $ra_{centroid}$ corresponds to the average point of the coordinates $ra_i(pol_i, rat_i)$. Another finer way is to weight this average by the difference in absolute value between the two values of the coordinates (dimensions: $pol_i$, $rat_i$).

**i) Averaged centroid** i.e. centroid based on average of dimensions (polarity and rating). Let the statistical series with two variables (dimensions), where values are couples $(pol_i, rat_i)$. The centroid (mean point of the series) based on the average of polarities and ratings is the point $ra_{centroid}$ in figure 3, which their coordinates are computed as follows:

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

**ii) Weighted centroid** i.e. 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:

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

where $n$ is the number of points $ra_i$. The coefficient $c_i$ is computed as follows:

$$c_i = \frac{|rat_i - pol_i|}{2n} \tag{5}$$

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 ($max(|rat_i - pol_i|) = 2$) 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 *coursera.org*. The objectives of these experiments are to:

1. evaluate the impact of sentiment-aspect on the detection of contradiction,
2. evaluate the impact of dispersion function based on polarity and rating to quantify contradiction *intensity*, using *averaged* and *weighted* centroid,
3. evaluate the effectiveness of contradiction detection based on reviews session.

### 4.1 Description of Test Data set

To the best of our knowledge, no standard or annotated data set is available to evaluate the intensity of contradiction. Therefore, 2244 English courses are extracted from *coursera.org* via its API[4]. For each course, we have also collected its reviews, dates of reviews and ratings via the *parsing* of the course web pages (see the statistics in the table 1).

**Table 1:** Statistics on coursera data set

| Field | Total Number |
|---|---|
| Courses | 2244 |
| Courses Rated | 1115 |
| Reviews | 73873 |
| Reviews ★ | 1705 |
| Reviews ★★ | 1443 |
| Reviews ★★★ | 3302 |
| Reviews ★★★★ | 12202 |
| Reviews ★★★★★ | 55221 |

**Table 2:** List of detected aspects

| | | |
|---|---|---|
| Assignment | Content | Exercise |
| Information | Instructor | Knowledge |
| Lecture | Lecturer | Lesson |
| Material | Method | Presentation |
| Professor | Quality | Question |
| Quiz | Slide | Speaker |
| Student | Teacher | Topic |
| Video | | |
| 22 aspects | | |

| Aspects | #Rat 1 | #Rat 2 | #Rat 3 | #Rat 4 | #Rat 5 | #Negative | #Positive | #Review | #Course |
|---|---|---|---|---|---|---|---|---|---|
| Content | 176 | 179 | 341 | 676 | 1641 | 505 | 1496 | 1883 | 207 |
| Lecture | 185 | 206 | 290 | 613 | 1762 | 763 | 1508 | 1988 | 208 |
| Video | 228 | 238 | 356 | 707 | 1614 | 941 | 1421 | 2058 | 245 |

**Table 3:** Statistics on some aspects extracted from the reviews of coursera.org

Table 3 presents some aspects among 22 useful aspects captured automatically from the reviews. To obtain judgments of contradictions and sentiments for

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

a given aspect: a) 3 assessors were asked to assess the sentiment class for each review-aspect; b) 3 other assessors 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). To evaluate sentiments and contradictions in the reviews-aspect of each course, 3-levels are used for sentiments: *Negative, Neutral, Positive*; and 5-levels for contradictions: *Not Contradictory, Very Low, Low, Strong* and *Very Strong*.

We analyzed the agreement degree between assessors for each aspect using Kappa Cohen measure $k$. This indicator takes into account the proportion of agreement between the assessors and the proportion of agreement expected between the assessors by chance. The Kappa measure is equal to 1 if the assessors completely agree, 0 if they agree only by chance. $k$ is negative if the agreement between assessors is worse than random. In our case, the $k$ is 0.76 for sentiment assessors and $k$ is 0.68 for contradiction assessors, which corresponds to a substantial agreement.

### 4.2 Results and Discussions

To evaluate the performance of our approach, correlation study was conducted (official measure on SemEval tasks[5]), by using the correlation coefficients of *Pearson* and *Spearman*, 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[6] (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%.

| Measure | Config (1): averaged centroid | Config (2): weighted centroid |
|---|---|---|
| **WITHOUT Considering Reviews Session** | | |
| **(a) between contradiction judgments and approach results (sentiment accuracy:** 79%**)** | | |
| Spearman | 0.42 | 0.49 |
| Pearson | 0.45 | 0.51 |
| **(b) between contradiction judgments and approach results (sentiment accuracy:** 100%**)** | | |
| Spearman | 0.65 | 0.79 |
| Pearson | 0.68 | 0.87 |

**Table 4:** Correlation results (WITHOUT Considering Reviews Session)

| Measure | Config (1): averaged centroid | Config (2): weighted centroid |
|---|---|---|
| **WITH Considering Reviews Session** | | |
| **(a) between contradiction judgments and approach results (sentiment accuracy:** 79%**)** | | |
| Spearman | $0.58^*$ | $0.69^*$ |
| Pearson | $0.61^*$ | $0.71^*$ |
| **(b) between contradiction judgments and approach results (sentiment accuracy:** 100%**)** | | |
| Spearman | $0.70^*$ | $0.87^*$ |
| Pearson | $0.73^*$ | $0.91^*$ |

**Table 5:** Correlation results (WITH Considering Reviews Session)

Tables 4 and 5 summarize the correlation values obtained by taking into account the *averaged centroid* (Config (1)) and the *weighted centroid* (Config

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

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

(2)) *WITH* and *WITHOUT* considering reviews session. In order to check the significance of the results (*WITH*) compared to (*WITHOUT*), we conducted the Student's t-test. The asterisk * is attached to the performance number of each row in table 5 when *p-value<0.05*. The results are discussed in the following.

**1) WITHOUT Considering Reviews Session**

**Config (1): averaged centroid.** Table 4 show that the dispersion measurement based on the averaged centroid provides a positive correlation with judgments, Spearman: 0.42, 0.65 and Pearson: 0.45, 0.68, for the both cases: (a) 79% and (b) 100% sentiment accuracy, respectively. Indeed, the more polarities between the reviews-aspect are opposite, the more the set of reviews-aspect diverge from the centroid, hence the increased intensity dispersion. Moreover, the results obtained using the manual sentiments judgments (table 4 (b)) surpass those obtained using our sentiment analysis model (table 4 (a)) approximately with 50% (Spearman: 0.42 *Vs* 0.65 and Pearson: 0.45 *Vs* 0.68). Therefore, losing 21% in sentiments accuracy involves a 50% loss in detecting contradictions performance.

**Config (2): weighted centroid.** The results are also positive (Spearman: 0.49, 0.79 and Pearson: 0.51, 0.87). The results obtained by considering the coefficient $c_i$ for each point of the space (review-aspect *ra*) are better compared to those obtained when this coefficient is ignored. These improvements are 16% (Spearman) using our sentiment model (table 4 (a)) and 22% (Spearman) using manual sentiment judgments (table 4 (b)). Indeed, the more divergent values of rating and polarity for every review-aspect, the higher is the impact on contradiction intensity. Also, the results of Config (2) presented in table 4 (b) are much better (Spearman: 0.79) than those presented in table 4 (a) (Spearman: 0.49). Therefore, the sentiment analysis model is an important factor that impacts the detection and the measurement of contradictions.

**2) WITH Considering Reviews Session**

As previously, table 5 shows a positive correlation for Config (1) and (2), with both assumptions in terms of sentiment accuracy (79% and 100%). The sentiment analysis model is always a factor that influences the results of the contradictions. Indeed, losing 21% in sentiments accuracy involves in average 23.5% loss in detecting contradictions performance. However, the results "WITH Considering Reviews Session" (see table 5) show that the correlations values are better compared to those obtained when reviews session is ignored "WITHOUT Considering Reviews Session" (see table 4). The comparative discussion is below.

**Config (1): averaged centroid.** The results (in table 5 (a) and (b)) show a significant improvement compared to those in table 4 (a) and (b). Indeed, when the *time* dimension (i.e. reviews are grouped by session) is taken into account, an improvement of 38% (Spearman) is recorded using our sentiment model in contradiction estimation (table 5 (a)) and 7% (Spearman) using manual sentiment judgments (table 5 (b)). From this comparison, we conclude that grouping reviews according to their session contributes to the effective contradiction detection. Moreover, the intensity of contradiction is estimated finely by taking into account only the reviews related to the specific session.

**Config (2): weighted centroid.** Using both the weighted centroid and the review session allows to improve the results even better than all previous runs. Compared to Config (2) in table 4, the improvements are 49% (Spearman) using our sentiment model (table 5 (a)) and 10% (Spearman) using manual sentiment judgments (table 5 (b)). The difference comes from the advantage of the consideration of centroid based on the weighted average of dimensions (polarity and rating), as well as the clustering pre-processing of reviews (session).

Finally, we observe in all cases that our contradiction analysis approach, in terms of detection and intensity estimation, provides good results. The best results are obtained by Config (2) which takes into account the weighted centroid with considering temporal factor (session of reviews). According to t-test, the results show a statistically significant improvement. We believe that these improvements comes from the 3-steps pre-processing. Specifically, the grouping reviews according to their corresponding resources sessions, this contribute significantly to these well results. 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.

## 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. Clustering the reviews by sessions allow an effective treatment to avoid fake contradictions. Additionally, to quantify the contradiction, review-aspects are exploited using dispersion function, where more the coordinates polarities and ratings are opposite the more the impact is important on the contradiction intensity. The validation of our overall assumptions was examined on the data collection of *coursera.org*. The obtained results reveal the effectiveness of our approach. Finally, we note that we are aware that our approach of detecting contradiction is still limited. The major weakness of our approach is its dependence on the quality of sentiment analysis and aspect models. 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 data are also envisaged. Even with these simple elements, the first results obtained encourage us to invest more in this track.

**Acknowledgement.** 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. I. Badache and M. Boughanem. Social priors to estimate relevance of a resource. In *IIiX*, pages 106–114, 2014.
2. I. Badache and M. Boughanem. Fresh and diverse social signals: any impacts on search? In *CHIIR*, pages 155–164, 2017.
3. M-C. De Marneffe, A. Rafferty, and C. Manning. Finding contradictions in text. In *ACL*, volume 8, pages 1039–1047, 2008.
4. S. Dori-Hacohen and J. Allan. Automated controversy detection on the web. In *ECIR*, pages 423–434, 2015.
5. R. Ennals, D. Byler, J.M. Agosta, and B. Rosario. What is disputed on the web? In *WICOW*, pages 67–74, 2010.
6. H. Hamdan, P. Bellot, and F. Bechet. Lsislif: Crf and logistic regression for opinion target extraction and sentiment polarity analysis. In *SemEval*, page 753758, 2015.
7. S. Harabagiu, A. Hickl, and F. Lacatusu. Negation, contrast and contradiction in text processing. In *AAAI*, volume 6, pages 755–762, 2006.
8. A. Hassan, A. Abu-Jbara, and D. Radev. Detecting subgroups in online discussions by modeling positive and negative relations among participants. In *EMNLP*, 2012.
9. A. Htait, S. Fournier, and P. Bellot. Using web search engines for english and arabic unsupervised sentiment intensity prediction. In *SemEval*, 2016.
10. M. Hu and B. Liu. Mining and summarizing customer reviews. In *KDD*, 2004.
11. M. Jang and J. Allan. Improving automated controversy detection on the web. In *SIGIR*, pages 865–868, 2016.
12. S. Kim, J. Zhang, Z. Chen, A. Oh, and S. Liu. A hierarchical aspect-sentiment model for online reviews. In *AAAI*, 2013.
13. J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Berkeley symposium on mathematical statistics and probability*, 1967.
14. S.M Mohammad, S. Kiritchenko, and X. Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *SemEval*, 2013.
15. A. Mukherjee and B. Liu. Mining contentions from discussions and debates. In *KDD*, pages 841–849, 2012.
16. B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP*, pages 79–86, 2002.
17. S. Poria, E. Cambria, L. Ku, C. Gui, and A. Gelbukh. A rule-based approach to aspect extraction from product reviews. In *SocialNLP*, 2014.
18. M. Qiu, L. Yang, and J. Jiang. Modeling interaction features for debate side clustering. In *CIKM*, pages 873–878, 2013.
19. R. Socher, A. Perelygin, J.Y Wu, J. Chuang, C.D Manning, A.Y Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, volume 1631, page 1642, 2013.
20. I. Titov and R. McDonald. Modeling online reviews with multi-grain topic models. In *WWW*, pages 111–120, 2008.
21. M. Tsytsarau, T. Palpanas, and K. Denecke. Scalable discovery of contradictions on the web. In *WWW*, pages 1195–1196. ACM, 2010.
22. M. Tsytsarau, T. Palpanas, and K. Denecke. Scalable detection of sentiment-based contradictions. *DiversiWeb, WWW*, 2011.
23. Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *ACL*, pages 417–424, 2002.
24. L. Wang and C. Cardie. A Piece of My Mind: A Sentiment Analysis Approach for Online Dispute Detection. In *ACL*, pages 693–699, 2014.