Contradiction in Reviews: is it Strong or Low?

Ismail Badache, Sébastien Fournier, and Adrian-Gabriel Chifu

LIS UMR 7020 CNRS, Aix-Marseille University, France
{ismail.badache, sebastien.fournier, adrian.chifu}@lis-lab.fr

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 web resource (e.g. courses, movies, products). The detection of contradictory opinions is an important task to evaluate the latter. This paper focuses on the problem of detecting and estimating contradiction intensity based on the sentiment analysis around specific aspects of a resource. Firstly, certain aspects are identified, according to the distributions of the emotional terms in the vicinity of the most frequent names in the whole of the reviews. Secondly, the polarity of each review segment containing an aspect is estimated using the state-of-the-art approach *SentiNeuron*. Then, only the resources containing these aspects with opposite polarities (positive, negative) are considered. Thirdly, a measure of the intensity of the contradiction is introduced. It is based on the joint dispersion of the polarity and the rating of the reviews containing the aspects within each resource. The evaluation of the proposed approach is conducted on the Massive Open Online Courses collection containing 2244 courses and their 73,873 reviews, collected from Coursera. The results revealed the effectiveness of the proposed approach to detect and quantify contradictions.

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. Social web (e.g. social networks) allow the generation of these traces. They represent a rich source of social information, which can be analysed and exploited in various applications [1] [2] [3]. For example, opinion mining or sentiment analysis [12], 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. For example, Wang and Cardie [31] aim to identify the sentiments of a sentence expressed during a discussion and they use them as features in a classifier that predicts dispute in discussions. Qiu et al. [22] 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 [18] and deducing the user relations by looking at their textual exchanges [11].

This paper investigates the entities (e.g. aspects, topics) for which the contradictions can occur in the reviews associated with a web resource (e.g. movies, courses) and how to estimate their intensity. The interest of estimating contradiction intensity depends on application framework. For example, following controversial political events/crises such as United States recognition of Jerusalem as capital of Israel. This has generated contradictory (diverse) opinions (reviews), in social networks, between different communities around the world. Estimating the intensity of this conflict may be useful for better analyzing the trend and the consequences of this political decision. In social information retrieval, for some users' information needs, measuring contradiction intensity can be useful to retrieve and rank the most controversial documents (e.g. news, events, etc). In our case, knowing the intensity of conflicting opinions on a specific aspect (e.g. speaker, slide, quiz) of an online course, may be helpful to know if there are certain elements for this course that need to be improved. Table 1 presents an instance of contradictory reviews about a "speaker" of a given coursera course.

Resource	Review (left)	Aspect	Review (Right)	Polarity	Rating
Course[1]	The lecturer was an annoying	speaker	and very repetitive.	-0.9	1
	Passionate	speaker	and truly amazing things to learn	+0.7	4

Table 1: Example of contradictory opinions about a "speaker" of a *coursera* course

Therefore, measuring the intensity of contradiction is for a better nuanced understanding of the diversity (dispersion) of opinions around a specific aspect. In order to design our approach, fundamental tasks are performed. First, aspects characterising these reviews are automatically identified. Second, opposing opinions around each of these aspects through a model of sentiment analysis are captured. Third, the intensity of contradiction in the reviews are estimated, using a measure of dispersion based on ratings and polarities of reviews containing an aspect. Finally, user studies experiments were conducted to evaluate the effectiveness of our approach, using a dataset collected from *coursera.org*. The main contributions addressed in this paper are twofold:

(C1). 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. But in addition to detecting the contradiction, it is desirable to estimate its intensity. Therefore, we try to answer in this paper the following research questions:

- **RQ1:** How to estimate the intensity of contradiction?
- **RQ2:** 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?

(C2). A development of a data collection collected from *coursera.org* which is useful for the evaluation of contradiction intensity measurement systems. Our experimental evaluation is based on user study.

The rest of this paper is structured as follows: Section 2 presents related work and background. Section 3 details our approach for detecting contradiction and estimating the intensity. Section 4 reports the results of our experiments. Section 5 concludes this paper and launches perspectives.

[1] `https://www.coursera.org/learn/dog-emotion-and-cognition`

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 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

The studies that are most related to our approach include [10], [5], [28] and [29], 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. [10] proposed an approach for contradiction analysis that exploits linguistic features (e.g. types of verbs), as well as semantic information, such as negation (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 7 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. [28], [29] proposed an automatic and scalable solution for the contradiction detection problem. 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. In the literature, the detection of controversies has been addressed both by supervised methods as in [20], [4] and [32] or by unsupervised methods as in [7], [6], [8] and [15]. To detect controversial events on Twitter (e.g., David Copperfield's charge of rape between 2007 and 2010)[2], Popescu and Pennacchiotti [20] proposed a decision-tree classifier and a set of features such as discourse parts, the presence of words from opinion or controversial lexicons, and user interactions (*retweet* and *reply*). Balasubramanyan et al. [4] extended the supervised LDA model to predict how members of a different political communities will emotionally respond to the same news story. Support vector classifiers and

[2] http://www.foxnews.com/story/2009/08/20/magician-david-copperfield-accused-raping-woman-on-private-island.html

logistic regression classifiers have also been proposed in [32] and [31] to detect disputes in Wikipedia page discussions. For example in the case of the comments that surround the modifications of Wikipedia pages.

Other works have also exploited Wikipedia to detect and to identify controversial topics on the web [7], [6], [14] and [15]. Dori-Hacohen and Allan in [7], [6] and Jang and Allan in [14] proposed to align web pages to Wikipedia pages on the assumption that a page deals with a controversial topic if the Wikipedia page describing this topic is itself controversial. The controversial or non-controversial nature of a Wikipedia page is automatically detected based on the metadata and discussions associated with the page. Jang et al. [15] constructed a controversial topics language model learned from Wikipedia articles and then used to identify if a web page is controversial.

Detection of controversies in social networks was also discussed without supervision based on interactions between different users [8]. Garimella et al. [8] proposed alternative measurement approaches based on the network, such as the *random walk* and the *betweenness centrality* and the low-dimensional embeddings. The authors tested simple content-based methods and noted their inefficiency compared to user graph-based methods. Other studies try to detect controversies on specific domains, for example in news [27] or in debate analysis [22]. However, to the best of our knowledge, none of the state-of-the-art works attempt to estimate, explicitly and concretely, the intensity of the contradiction or controversy. In this paper, unlike previous work, rather than only identifying controversy in a single hand-picked topic (e.g., aspect related to political news), we focus also on estimating the intensity of contradictory opinions around specific topics. We propose to measure the intensity of contradiction using some characteristics of the opinion (e.g. rating, polarity).

2.2 Aspect Detection

The first attempts to detect aspects were based on the classical information extraction approach using the frequent nominal sentences [13]. Such approaches work well for the detection of aspects that are in the form of a single name, but are less useful when the aspects have low frequency. Similarly, other studies use Conditional Random Fields (CRF) or Hidden Markov Models (HMM) [9]. Other methods are unsupervised and have proven their effectiveness, such as [26] that built a Multi-Grain Topic Model and [16] that proposed HASM (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 in [21]. This method, based on the use of extraction rules for product reviews, corresponds to our experimental data (coursera).

2.3 Sentiment Analysis

Sentiment analysis has been the subject of much previous research.As in the case of aspect detection, the supervised and unsupervised approaches both propose

their solutions. Thus, some unsupervised approaches are based on lexicons, such as the approach developed by [30], or corpus-based methods, such as in [17]. *Pang et al.* [19] proposed supervised approaches, that 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 RNN (Recursive Neural Network), such as in [24]. In our work, sentiment analysis is only a part of contradiction detection process, we were inspired by [19] using Bayesian classifier as baseline. Naïve Bayes is a probabilistic model that gives good results in the classification of sentiments and generally takes less time for training compared to models like SVM or RNN.

3 Intensity of Contradiction

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 the contradictory opinions intensity, two dimensions are jointly exploited: the polarity around the aspect as well as the rating associated with the review. The dimensions associated to the contradictory opinions (called in this paper: reviews-aspect) are represented using a dispersion function (see figure 1).

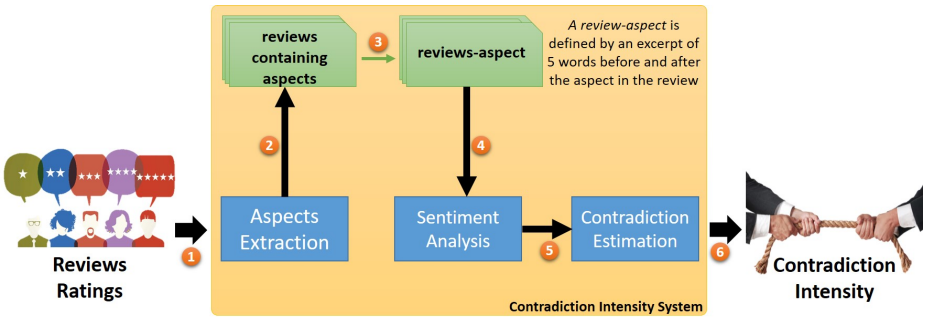

Fig. 1: Sentiment-based contradiction intensity framework

3.1 Pre-processing

The pre-processing module consists of two main stages: 1) extraction of aspects from the reviews; and 2) sentiment analysis of the text related to these aspects.

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., [21]. This method corresponds to our experimental data (*coursera*). Additionally, the following steps are applied (see an example in table 2).

1. Term frequency calculation of the reviews corpus,
2. Part-of-speech tagging of reviews using *Stanford Parser*[3],

[3] `http://nlp.stanford.edu:8080/parser/`

3. Selection of terms having nominal category (NN, NNS)[4],
4. Selection of nouns with emotional terms in their five-neighborhoods (using *SentiWordNet*[5] 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.

Step	Description
(1)	course : 44219, material : 3286, assignments : 3118, content : 2947, speaker : 2705,.....$term_i$
(2)	re = The/DT **lecturer**/NN was/VBD an/DT annoying/VBG **speaker**/NN and/CC very/RB repetitive/JJ ./. 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 ./.
(3)	lecturer, speaker, formatting, things
(4)	lecturer, speaker
(5)	speaker

Table 2: Steps to extract the aspects of a review re

Once the list of aspects is defined, the sentiment polarity around these aspects must be estimated. The following section presents sentiment analysis models.

Sentiment Analysis. The sentiment of the review on aspect (review-aspect) is estimated using two approaches: first, Naive Bayes algorithm [19] which treats: a) Negation (word preceded by *no, not, n't*). 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* and *trigrams*) of adjectives with other words such as adverbs *"very bad"* and *"absolutely recommended"*. Second, an unsupervised *SentiNeuron*[6] model proposed by Radford et al. [23] to detect sentiment signals in reviews. The model consisted of a single layer multiplicative long short-term memory (mLSTM) cell and when trained for sentiment analysis it achieved state of the art on the movie review dataset[7]. They also found a unit in the mLSTM that directly corresponds to the sentiment of the output. *SentiNeuron* provides very good results compared to several models of the state of the art. Especially in the case of IMDb reviews as well as our case (*coursera* reviews).

3.2 Measure of Contradiction

Definition. There is a contradiction between two portions of review-aspect ra_1 and ra_2 containing an aspect, where ra_1, $ra_2 \in D$ (Document), when the opinions (polarities) around the aspect are opposite (i.e. $pol(ra_1) \cap pol(ra_2) = \phi$). We note that after several empirical experiments, the review-aspect ra is defined by an excerpt of 5 words before and after the aspect in review re.

Contradiction intensity is estimated using 2 dimensions: polarity pol_i and rating rat_i of the review-aspect ra_i. Let each ra_i be a point on the plane with coordinates (pol_i, rat_i). Assuming, the greater is the distance (i.e. dispersion) between these values related to each review-aspect ra_i of the same document

[4] https://cs.nyu.edu/grishman/jet/guide/PennPOS.html
[5] http://sentiwordnet.isti.cnr.it/
[6] https://github.com/openai/generating-reviews-discovering-sentiment
[7] https://www.cs.cornell.edu/people/pabo/movie-review-data/

D, the contradiction intensity is more important. The dispersion indicator with respect to the centroid $ra_{centroid}$ with coordinates $(\overline{pol}, \overline{rat})$ is as follows:

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

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

$Distance(pol_i, rat_i)$ represents the distance between the point ra_i of the scatter plot and the centroid $ra_{centroid}$, 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: $rat_i = \frac{rat_i - 3}{2}$ $(rat_i \in [-1, 1])$. 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 weigh this average by the difference in absolute value between the two coordinate values (polarity and rating).

a) Centroid based on average of dimensions. In this case, the coordinates of the centroid $ra_{centroid}$ are computed based on the average of polarities and ratings as follows:

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

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

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

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 division by $2n$ represents a normalisation 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, experiments were carried out on reviews collected from the site of *coursera.org*. Our main objective in these experiments is to evaluate the impact of considering the sentiment analysis and the rating on the contradiction detection 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 estimation.

4.1 Description of Test Dataset

DATA. To the best of our knowledge, there is no standard data set to evaluate the contradiction intensity. Therefore, 73,873 reviews and their ratings of 2244 English courses are extracted from *coursera* via its API[8] and web pages *parsing*. More details about the statistics on our *coursera* dataset are presented in table 3. Our full test dataset and its detailed statistics are publicly available[9]. Table 5 presents some stats on 4 aspects among 22 useful aspects, listed in table 4, captured automatically from the reviews.

Table 3: 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 4: 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
Lecturer	32	41	48	85	461	55	193	236	39
Material	191	203	328	722	2234	784	1693	2254	237
Quiz	151	155	221	401	581	481	475	824	128

Table 5: Statistics on some aspects extracted from the reviews of Coursera.org

User Study. To obtain contradiction and sentiment judgements for a given aspect, we conducted a user study as follows:

(a) 3 users were asked to assess the sentiment class for each review-aspect provided by our system (see section 3.1). The users must judge just its polarity;

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

[9] https://www.irit.fr/~Ismail.Badache/#projects

(b) 3 other users assessed the degree of contradiction between these reviews-aspect as shown in the figure 2.

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-points scale are used for sentiments: *Negative, Neutral, Positive*; and 5-points scale for contradictions: *Not Contradictory, Very Low, Low, Strong* and *Very Strong* (see figure 2). We computed the agreement degree between assessors for each aspect using Kappa Cohen measure k. Since we have 3 assessors, the Kappa value was calculated for each pair of assessors and then their average was calculated. The average k is 0.76 for sentiment assessors and 0.68 for contradiction assessors, which corresponds to a substantial agreement.

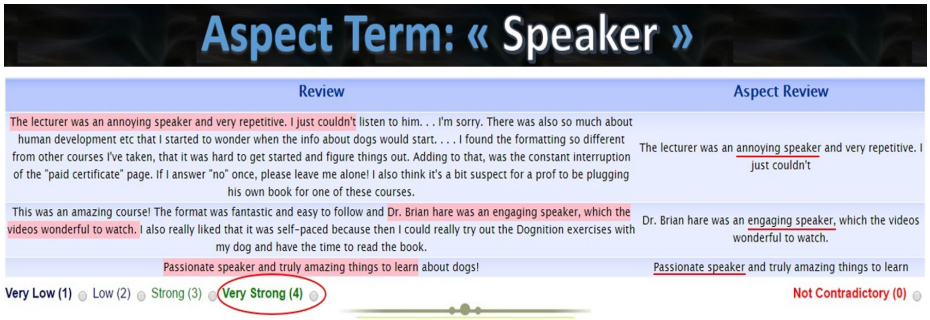

Fig. 2: Evaluation system interface

4.2 Results and Discussions

Correlation study was conducted (one of the official measures on SemEval tasks[10]), by using the coefficient of *Pearson*, between the contradiction judgements given by the assessors and our obtained results. In addition, the precision was computed for each configuration. The configuration that consider Naive Bayes-based sentiment analyser is considered as baseline in these experiments.

Remarks: First, the Naive Bayes sentiment analyser takes as a training set 50,000 reviews of *IMDb* movies[11] (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, this sentiment analysis system provides an accuracy of 79%. Third, assessors' judgements on sentiments are considered as perfect (reference) results and represent an accuracy of 100%.

In order to check the significance of the results compared to the baseline, we conducted the Student's t-test [25]. We attached * (strong significance against Baseline) and ** (very strong significance against Baseline) to the performance number of each row in the tables when $p\text{-}value<0.05$ and $p\text{-}value<0.01$ confidence level, respectively. We discuss in the following the results of each configuration we investigated (see table 6).

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

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

Measure	Config (1): Averaged Centroid	Config (2): Weighted Centroid
(Baseline) Using Naive Bayes: sentiment analysis accuracy of 79%		
Pearson	0.45	0.51
Precision	0.61	0.70
(a) Using SentiNeuron: sentiment analysis accuracy of 93%		
Pearson	0.61^{*}	0.80^{**}
Precision	0.75^{**}	0.88^{**}
(b) Using Users' Judgements: sentiment analysis accuracy of 100%		
Pearson	0.68^{**}	0.87^{**}
Precision	0.82^{**}	0.91^{**}

Table 6: Results of correlations and precision

Config (1): Averaged Centroid. The results show that the dispersion measurement based on the averaged centroid provides a positive correlation with judgements, Pearson: 0.45, 0.61, 0.68. 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. In addition, the results obtained using the users' sentiments judgements (table 6 (b)) surpass those obtained using the sentiment analysis models (table 6 (a) and (b)) with an approximate percentage of 35% for (a) (Pearson: 0.45 *Vs* 0.61) and of 50% for (b) (Pearson: 0.45 *Vs* 0.68). In terms of precision, compared to baseline, we record an improvement rate of 23% for (a) when *SentiNeuron* is used, and 34% for (b) when the users' sentiments judgements are used in the estimation of contradiction intensity. Therefore, losing 21% in sentiments (100% - 79%) involves a 34% loss in precision.

Config (2): Weighted Centroid. The configuration (2) results are also positive (Pearson: 0.51, 0.80, 0.87). The results obtained by considering the importance coefficient c_i for each point of the space (review-aspect ra_i) are better compared to those obtained when this coefficient is ignored. These improvements in terms of Pearson correlation value are 13% using Naive Bayes-based sentiment model (table 6 (Baseline)) and 31% using *SentiNeuron* (table 6 (a)), and 28% using manual sentiment judgements (table 6 (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 in terms of precision and correlations for configuration (2) presented in table 6 (b) are much better (Precision: 0.91) than (Baseline) (Precision: 0.70) and (a) when *SentiNeuron* is used (Precision: 0.88). Therefore, sentiment model is an important factor that impacts the estimation of contradictions.

Finally, table 7 shows the distribution of contradictions according to their level (*Very Low, Low, Strong* or *Very Strong*) as well as the number of detected and undetected contradictions for each configuration and for both systems (a) and (b). We notice that also these results show that the best results are obtained by configuration (2) which takes into account the weighted centroid. While we were pleasantly surprised by the efficacy of our approach, we did not use the best sentiment analysis model and aspect detection model of state-of-arts. We believe that improving these pre-processing models enhance our contradiction detection model significantly.

Level	Very Low	Low	Strong	Very Strong	Undetected	Detected
(a) Using SentiNeuron: sentiment analysis accuracy of 93%						
Config(1)	25	45	47	48	55	165
Config(2)	33	52	52	57	26	194
(b) Using Users' Judgements: sentiment analysis accuracy of 100%						
Config(1)	27	44	49	61	39	181
Config(2)	33	53	53	61	20	200

Table 7: Number of contradictions for each level

5 Conclusion

This paper introduced an approach that aims at estimating contradiction intensity, drawing attention to aspects in which users have contradictory reviews. Contradiction exists if the sentiments around these reviews-aspect for the same resource are diverse. Additionally, to quantify the contradiction, reviews-aspect are exploited using dispersion function, where more the dimensions polarities and ratings are opposite, the more the impact is important on the contradiction intensity. The experiments conducted on *coursera* data set reveal the effectiveness of our approach. Moreover, our dataset can be useful for the community.

The potential problem of our approach is its dependency on the quality of sentiment and aspect models. Moreover, the sentences are not processed, only a predefined window of 5 words before and after the aspect is considered. Further scale-up experiments on other types of data sets are also envisaged. A supervised approach based on the state-of-the-art learning approaches can improve significantly the prediction of contradiction intensity level. Even with these simple elements, the first obtained results encourage us to invest more in this track.

References

1. I. Badache and M. Boughanem. Harnessing social signals to enhance a search. In *IEEE/WIC/ACM, volume 1*, pages 303–309, 2014.
2. I. Badache and M. Boughanem. Emotional social signals for search ranking. In *SIGIR*, pages 1053–1056, 2017.
3. I. Badache and M. Boughanem. Fresh and diverse social signals: any impacts on search? In *CHIIR*, pages 155–164, 2017.
4. R. Balasubramanyan, W.W. Cohen, D. Pierce, and D.P. Redlawsk. Modeling polarizing topics: When do different political communities respond differently to the same news? In *ICWSM*, pages 18–25, 2012.
5. M-C. De Marneffe, A. Rafferty, and C. Manning. Finding contradictions in text. In *ACL*, volume 8, pages 1039–1047, 2008.
6. S. Dori-Hacohen and J. Allan. Automated controversy detection on the web. In *ECIR*, pages 423–434, 2015.
7. Shiri Dori-Hacohen and James Allan. Detecting controversy on the web. In *CIKM*, pages 1845–1848, 2013.
8. K. Garimella, G. D. F. Morales, A. Gionis, and M. Mathioudakis. Quantifying controversy in social media. In *WSDM*, pages 33–42, 2016.

9. 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.
10. S. Harabagiu, A. Hickl, and F. Lacatusu. Negation, contrast and contradiction in text processing. In *AAAI*, volume 6, pages 755–762, 2006.
11. A. Hassan, A. Abu-Jbara, and D. Radev. Detecting subgroups in online discussions by modeling positive and negative relations among participants. In *EMNLP*, pages 59–70, 2012.
12. A. Htait, S. Fournier, and P. Bellot. LSIS at semeval-2016 task 7: Using web search engines for english and arabic unsupervised sentiment intensity prediction. In *SemEval*, 2016.
13. M. Hu and B. Liu. Mining and summarizing customer reviews. In *KDD*, pages 168–177, 2004.
14. M. Jang and J. Allan. Improving automated controversy detection on the web. In *SIGIR*, pages 865–868, 2016.
15. M. Jang, J. Foley, S. Dori-Hacohen, and J. Allan. Probabilistic approaches to controversy detection. In *CIKM*, pages 2069–2072, 2016.
16. S. Kim, J. Zhang, Z. Chen, A. Oh, and S. Liu. A hierarchical aspect-sentiment model for online reviews. In *AAAI*, 2013.
17. S.M Mohammad, S. Kiritchenko, and X. Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *SemEval*, 2013.
18. A. Mukherjee and B. Liu. Mining contentions from discussions and debates. In *KDD*, pages 841–849, 2012.
19. B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP*, pages 79–86, 2002.
20. A.M. Popescu and M. Pennacchiotti. Detecting controversial events from twitter. In *CIKM*, pages 1873–1876, 2010.
21. S. Poria, E. Cambria, L. Ku, C. Gui, and A. Gelbukh. A rule-based approach to aspect extraction from product reviews. In *SocialNLP*, pages 28–37, 2014.
22. M. Qiu, L. Yang, and J. Jiang. Modeling interaction features for debate side clustering. In *CIKM*, pages 873–878, 2013.
23. A. Radford, R. Józefowicz, and I. Sutskever. Learning to generate reviews and discovering sentiment. *CoRR*, abs/1704.01444, 2017.
24. 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, pages 1631–1642, 2013.
25. Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908.
26. I. Titov and R. McDonald. Modeling online reviews with multi-grain topic models. In *WWW*, pages 111–120, 2008.
27. M. Tsytsarau, T. Palpanas, and M. Castellanos. Dynamics of news events and social media reaction. In *KDD*, 2014.
28. M. Tsytsarau, T. Palpanas, and K. Denecke. Scalable discovery of contradictions on the web. In *WWW*, pages 1195–1196, 2010.
29. M. Tsytsarau, T. Palpanas, and K. Denecke. Scalable detection of sentiment-based contradictions. *DiversiWeb, WWW*, 2011.
30. Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *ACL*, pages 417–424, 2002.
31. L. Wang and C. Cardie. A Piece of My Mind: A sentiment analysis approach for online dispute detection. In *ACL*, pages 693–699, 2014.
32. L. Wang, H. Raghavan, C. Cardie, and V. Castelli. Query-focused opinion summarization for user-generated content. In *COLING*, pages 1660–1669, 2014.