Vector Space Model of Text Classification Based on Inertia Contribution of Document

Demba Kandé[1], Fodé Camara[2(✉)], Reine Marie Marone[1,2], and Samba Ndiaye[1]

[1] Department of Mathematics, Cheikh Anta Diop University, Dakar, Senegal
demba4.kande@ucad.edu.sn
[2] Department of Mathematics, Alioune Diop University, Bambey, Senegal
fode.camara@uadb.edu.sn

Abstract. The use of textual data has increased exponentially in recent years due to the networking infrastructure such as Facebook, Twitter, Wikipedia, Blogs, and so one. Analysis of this massive textual data can help to automatically categorize and label new content. Before classification process, term weighting scheme is the crucial step for representing the documents in a way suitable for classification algorithms. In this paper, we are conducting a survey on the term weighting schemes and we propose an efficient term weighting scheme that provide a better classification accuracy than those obtening with the famous TF-IDF, the recent IF-IGM and the others term weighting schemes in the literature.

Keywords: Vector space model · Classification · Text mining
Term weighting scheme

1 Introduction

In the recent years, web users generated a large amount of various and useful text information. This textual data from Facebook, Twitter, Wikipedia, Blogs, and so one can be analyzed to identify most informative comments, to get users' opinions from comments, to recognize a potentially spam content, etc.

Before classification, text documents must be represented in a way suitable for data mining algorithms. Thus, several term weighting schemes (also called vector space models) have been developed in the literature to improve the performance of text classification algorithms. These techniques can be divided into two approaches, unsupervised and supervised term weighting methods, depending on the use of the class label in training corpus. The pioneer works are the unsupervised weighting scheme, binary and the popularly-used TF-IDF [3]. The binary method tells when a term appears in a document, and TF-IDF determines terms that are frequent in the document, but infrequent in the corpus.

However, the traditional unsupervised weighting scheme is not really useful for text classification tasks. As an alternative, various works have been done on weighting models based on the known class label, including, the recent TF-IGM scheme [9]. TF-IGM adopts a new statistical model to measure a term's class distinguishing power. To the best of our knowledge, it is the most efficient term weighting scheme.

R. Zitouni and M. Agueh (Eds.): AFRICATEK 2018, LNICST 260, pp. 155–165, 2019.
https://doi.org/10.1007/978-3-030-05198-3_14

This paper challenges TF-IGM [9], and introduce a new and efficient supervised term weighting scheme based on inertia contribution of document. Our weighting scheme has the benefit because it affects positively the classification performance. The experimental results show that our algorithm outperforms the famous TF-IDF, and the recent and efficient TF-IGM.

The rest of the paper is organized as follows. Section 2 discusses related works. In Sect. 3, we give the details of our proposition. In Sect. 4, we evaluate the performance of our algorithm. Section 5 concludes the paper and gives some future works.

2 Analyses of Current Term Weighting Schemes

In the literature, various term weighting schemes have been proposed for text categorization (TC), and thus for optimizing the classifier accuracy. We have focused on the limitations of TF-IDF [3] and TF-IGM [9] and others, which are respectively the most used and the most efficient term weighting schemes.

We can explore the literature, through a simple example. Let's consider the following corpus, denoted *d*:

Table 1. An simple example of corpus d

Id document	Document contain	Class
d_1	"The sky is blue"	Negeative
d_2	"The sun is bright today"	Positive
d_3	"The sun in the sky is bright"	Positive
d_4	"We can see the shining sun, the bright sun"	Positive

Then, its dictionary is {'blue', 'sky', 'bright', 'sun', 'today', 'can', 'see', 'shining'}.

2.1 Traditional Term Weighting Schemes

Traditional term weighting schemes are Binary (or Boolean), TF and TF-IDF weighting [2], which are originated from information retrieval. As the weight of a term, the term frequency (TF) in a document is obviously more precise and reasonable than the binary value, 1 or 0, denoting term presence or absence in the document because the topic terms or key words often appear in the document frequently and they should be assigned greater weights than the rare words. But term weighting by TF may assign large weights to the common words with weak text discriminating power.

To offset this shortcoming, a global factor, namely inverse document frequency (IDF), is introduced in the TF-IDF scheme.

$$w(t_j) = tf_{ij} \times log\left(\frac{N}{df_j}\right) \tag{1}$$

Where tf_{ij} denotes the frequency of term j in document i and N is the total number of documents and df_j is the number of documents that contains the term j.

The weight is composed of two factors: the local factor TF (for Term Frequency) metric that calculates the number of times a word appears in a document; and the global factor IDF (Inverse Document Frequency) term is computed as the logarithm of the number of the documents in the corpus divided by the number of documents that are specific to the term. The basic idea of TF-IDF is to determine term weight that are frequent in the document (using the TF metric), but infrequent in the corpus (using the IDF metric).

The term frequency (i.e., TF) for sky in d_1 is then 1. The word sky appears in two documents. Then, the inverse document frequency (i.e., IDF) is calculated as $\log\left(\frac{4}{2}\right) = 0.301$. Thus, the TF-IDF weight is the product of these quantities: $1 \times 0.301 = 0.301$.

The main drawback of TF-IDF is the fact that it unsupervised method; it does not take into account the distribution of class label.

Since the traditional TF-IDF (term frequency-inverse document frequency) is not fully effective for text classification. Several various of TF-IDF based on supervised methods have been proposed in the literature. These variants introduce a new statistic model: feature selection models to measure the term's distinguishing power in a class.

2.2 Supervised Methods Term Weighting

By considering the deficiencies of TF-IDF, researchers have proposed supervised term weighting schemes (STW) [4]. Otherwise, weighting a term by using an information known by the classes. The distribution of a term in different category is described with a contingency table shown in Table 2.

Table 2. The contingence table information

Class	c_k	$\bar{c}_k$
t_j	AA	B
$\bar{t}_j$	CC	D

A denotes the number of documents belonging to category c_k where the term t_j occurs at least once; B denotes the number of documents not belonging to category c_k where the term t_j occurs at least once; C denotes the number of documents belonging to category c_k where the term t_j does not occur; D denotes the number of documents not belonging to category $\bar{c}_k$ where the term t_j does not occur. The contigence table shows that:

- if term t_j is highly relevant to category c_k only, which basically indicates that it is a good feature to represent category c_k, then the value of $\frac{A}{B}$ tends to be higher.
- if the value of $\frac{A}{C}$ is larger, which means that the number of documents where term t_j occurs are greater than the documents where term t_j does not occur in class c_k.

- if term t_j is highly relevant to category $\bar{c}_k$ only, which basically indicates that it is a good feature to represent category $\bar{c}_k$, then the value of $\frac{B}{A}$: tends to be higher.
- if the value of $\frac{B}{D}$ tends to be higher, which means the number of documents where term t_j occurs are greater than the documents where term t_j does not occur in class $\bar{c}_k$.
- The product of $\frac{A}{B}$ and $\frac{A}{C}$ indicates terms t_j's relevance with respect to a specific category c_k. On the other hand, the product of $\frac{B}{A}$ and $\frac{B}{C}$ indicates terms t_j's relevance with respect to a specific category $\bar{c}_k$.

In [4], combining the term frequency and $\chi 2$ statistic, authors introduce the TF-Chi2 weight of term t_j:

$$w(t_j, c_k) = tf_{ij} \times \frac{N \times (A \times D - B \times C)^2}{(A+B) \times (C+D) \times (A+C) \times (B+D)} \tag{2}$$

In TF.Chi2, the weight of a term is specific to the c_k category, i.e. it depends on the contribution of the term in the c_k category. But, the size of the positive class is often smaller than that of the negative counterpart. The Chi2 statistic is limited in the case of multi-class classification, because it is a bi-class schema, hence causes performance loss of classifier. In addition to the drawbacks listed above, the terms informations in the corpus have not been considered [3].

The Measure of Relevance and Distinction with the AD metric [5] is frequently used as a criterion in the field of machine learning. It is based on the notion of relevance of characteristic from the distribution of terms in the category c_k. The more a term contributes to the distinction of category c_k, the higher its relevance is in c_k. AD of a feature t_j toward a category c_k can be defined as follows:

$$w(t_j, c_k) = \frac{A}{B} \times \frac{A}{C} \times \left(\frac{A}{B} \times \frac{A}{C} - \frac{B}{A} \times \frac{B}{C}\right) \tag{3}$$

In AD metric, only the known information of the category is considered, it ignores the contribution of the terms in the corpus [4] and constitutes a method to bi-class. In the case of multi-class classification some category may not be taken into account because are all group in c_k.

The work in [6], proposed a term frequency based on weighting scheme using naïve bayes (TF-RTF). It considered the binary text classification case (for a document, d, and its label, c_k, let $c_k = 1$ denote the positive class, and $\bar{c}_k = 0$ the negative one) and calculated the weight of a term from the posterior probability of each class:

$$w(t_j, c_k) = Nu * \left| log \frac{(M1u+1)}{(M0u+1)} + log \frac{(M0+p)}{(M1+p)} \right| \tag{4}$$

Where N_u is the term frequency of a word w_u in the document; M_{1u}, M_{0u} are the term frequencies of w_u respectively in the positive class and negative class; M_1, M_0 are respectively the total term frequencies in the positive class and negative class;

$\log\frac{(M_0+p)}{(M_1+p)}$, is the ratio of total term frequencies. Like all probability patterns, *TF-RTF* can cause a loss of information in multi-class categorization.

As others proposed metrics, the Information Gain [7] of a given feature t_j with respect to class c_k is the reduction in uncertainty about the value of t_j when we know the value of c_k. The more Information Gain is high for a feature, the more important a feature is for the text categorization. Information Gain of a feature t_j toward a category c_k can be defined as follows:

$$w(t_j, c_k) = \sum_{c\epsilon\{c_k\bar{c}_k\}} \sum_{t\epsilon\{t_j\bar{t}_j\}} P(t_j, c_k)\log\frac{P(t_j, c_k)}{P(c_k)P(t_j)} \tag{5}$$

Where $p(c_k)$ is the fraction of the documents in category cover the total number of documents, $p(t_j, c_k)$ is the fraction of documents in the category c_k that contain the word t over the total number of documents. $p(t_j)$ is the fraction of the documents containing the term t_j over the total number of documents.

The work presented in [8] (TF-BDC), the relevance of a term in a category is defined from the value of entropy. More the entropy is high, more it appears in several categories, and less discriminating they are. However, higher the concentration of the feature in a c_k category is, more important its discriminating power is. Conversely, a term with a more or less distribution uniform in the different categories has often-smaller entropy.

$$w(t_j, c_k) = 1 + \frac{\sum_{k=1}^{|c|}\frac{p(t_j|c_k)}{\sum_{k=1}^{|c|}p(t_j|c_k)}log\frac{p(t_j|c_k)}{\sum_{k=1}^{|c|}p(t_j|c_k)}}{log(|C|)} \tag{6}$$

With $p(t_j, c_k) = \frac{f(t_j,c_k)}{f(c_k)}$, where $f(t_j, c_k)$ denotes the frequency of term *tj* in category c_k and $f(c_k)$ denotes the frequency sum of all terms in category c_k.

Example: *in Table 1, the term "sky" has an entropy more higher than the term "sun", but "sun" has a higher discriminant power because it is specific to the category "positive".*

Like all feature selection methods, TF-BDC ignores the contribution of terms in the document collection.

In order to overcome the shortcomings of the bi-class schemes, Chen and al. propose Inverse Gravity Moment –*TF-IGM* [9] in order to explore both the contribution of terms in the classification and the provision of information in corpus. It is defined by:

$$w(t_j, c_k) = tf_{ij} * (1 + \lambda \cdot igm(t_j)) \tag{7}$$

Where $1 + \lambda \cdot igm(t_j)$ denotes the igm based global weighting factor of term t_j in document d_i, and $\lambda\epsilon[5; 9]$ is an adjustable coefficient for keeping the relative balance between the global and the local factors in the weight of a term. The $igm(t_j)$ is defined as follows:

$$\frac{f_{j1}}{\sum_{r=1}^{m} f_{jr}} \tag{8}$$

Where the frequency f_{jr} (r = 1, 2, …, m) usually refers to the class-specific document frequency of the term and f_{j1} the maximal frequency of the term of the class m (sort in descending order). TF-IGM is a supervised term weighting system (STW) because the global *IGM* weighting factor depends only on known class information, and the contribution of terms on the corpus is ignored.

Like all the supervised methods studied in this paper, only class information is used to determine the overall factor. However, the relevance of a document *di* depends on its position relative to the center of gravity *Gi*. Hence the importance of the terms that constitute it.

3 Our Proposed Term Weighting Scheme: TF-ICD

In this section, we propose a so-called *ICD* (inertia contribution document) model to measure the class distinguishing power of a term and then put forward a new term weighting scheme, TF-ICD, by combining term frequency (TF) with the ICD measure.

3.1 Problem Definition and Motivations

Let d be a set of labeled documents d_i, in which class is of a finite number of discrete symbols, each representing a class of the classification problem to be addressed. A document d_i is represented as a vector of terms $d_i = \{t_{1i}, \ldots, t_{ri}\}$ where r is the cardinality of the dictionary $\{t_1, \ldots, t_n\}$, and $0 < t_{\mathrm{ij}} < 1$ represents the contribution of term t_j to the prediction of class. Thus, d_i is represented by a matrix t_{ij}. Non-zero t_{ij} indicates that term t_{j} is contained in d_i.

The aim of our proposition is to transform the initial corpus d into matrix t_{ij} such as t_{ij} outperforms the state-of-the-art term weighting scheme by giving better classifier accuracy:

$$tf - icd(d) = \text{matrix } t_{ij}/f : \{T_1, \ldots, T_n\} \rightarrow \textit{class is better}.$$

Where *icd* represents our statistical model that measures the information quantity of a document, which reflects the term's class distinguishing power.

3.2 Analyzing the Discriminating Power of a Document

From the multidimensional statistical models a corpus can be presented as an individual-variable as described in the Fig. 1.

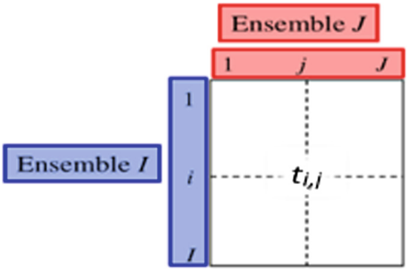

Fig. 1. Matrix t_{ij}/f: $\{T_1, .., T_n\}$

Where I is all individuals (documents), J is set of variables (terms), and t_{ij} is frequency of the term j in the document i.

By replacing the contingency table with the probability table, we obtain (Fig. 2):

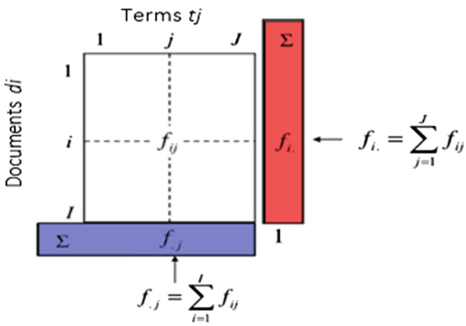

Fig. 2. Matrix f_{ij}/f: $\{T_1, .., T_n\}$

From its average conditional distribution ($\frac{f_j}{n}$ likelihood of using the t_j term). The higher the independence gap, the lower its weight is and its high inertial contribution $\lambda_{(di)}$.

3.3 Inertial Contribution of a Document–ICD

The inertial contribution is the amount of information that a document provides in a corpus, it depends on the product of two measures: (i) the weight of a document d_{i}; (ii) and its difference to independence.

The weight of a document is the probability of obtaining the document $\mathrm{d_i}$ belonging to the category c_k and is defined by

$$\frac{f_i}{n}. \tag{9}$$

The relevance of a document relies to its distance to the origin of the center of gravity described in Fig. 3.

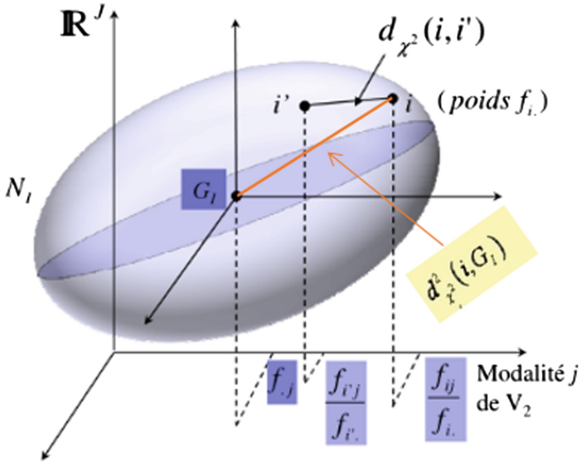

Fig. 3. Distance of a document from the center of gravity.

$$d^2_{i^2}(i, GI) = \sum_{j=1}^{j=J} \frac{(f_{ij} - f_{.j})^2}{f_{.j}} \tag{10}$$

We thus obtain the inertia contribution of a document d_i in the corpus, defined by

$$\lambda(d_i) = \frac{fi.}{n} \cdot \sum_{j=1}^{j=J} \frac{(f_{ij} - f_{.j})^2}{f_{.j}} \tag{11}$$

The Table 3 presents the inertia distribution by categories and by term.

Table 3. Intrtia distribution by categories and by term

Class	c_1	c_2	…	c_k
$\lambda(c_k)$	$\sum_{di\epsilon c1} \lambda(d_i)$	$\sum_{di\epsilon c2} \lambda(d_i)$	…	$\sum_{di\epsilon ck} \lambda(d_i)$
$\lambda(t_{ij\epsilon d})$	$\sum_{\{d_i\epsilon c_1\}} \lambda(d_i)$	$\sum_{\{d_i\epsilon 2\}} \lambda(d_i)$	…	$\sum_{\{d_i\epsilon k\}} \lambda(d_i)$

$$ICD(t_j, c_k) = log_2\left(1 + \frac{\sum_{\{d_i \epsilon c_k\}\{t_{ij \neq 0}\}} \lambda(d_i)}{N_j}\right) \quad (12)$$

Where $\sum_{\{d_i \epsilon c_k\}\{t_{ij \neq 0}\}} \lambda(d_i)$ is the sum of the inertia of documents d_i of category c_k containing t_j and N_j is the number of documents d_i of category c_k containing t_j.

3.4 Term Weighting by TF-ICD

The weight of a term in a document should be determined by its importance in the corpus and its contribution to text classification, which correspond respectively to the local and global weighting factors in term weighting. A term's contribution to text classification depends on its class distinguishing power, which is reflected by its contribution of documents inertia. Higher the inertia is, greater term weighting is important. This last can be measured by the ICD metric.

Hence, instead of the traditional IDF factor, a new global factor in term weighting is defined based on the ICD metric of the term, as shown in (12). Therefore, the TF-ICD weight of term t_j in document d_i is the product of the TF-based local weighting factor and the ICD-based global weighting factor, i.e., $w(t_j, c_k) = tf_{ij} \times ICD(tj, c_k)$, which is expressed as (13).

$$w(t_j, c_k) = tf_{ij} \times log_2\left(1 + \frac{\sum_{\{d_i \epsilon c_k\}\{t_{ij \neq 0}\}} \lambda(d_i)}{N_j}\right) \quad (13)$$

4 Experiments

4.1 Datasets

In order to evaluate the performance of the proposed method, we used the Spam collection [10]. In data preprocessing, all words are converted to lower case, punctuation marks are removed and we used stop lists and no stemming algorithm.

The sms spam collection is composed by 4,827 legitimate messages and 747 mobile spam messages, a total of 5,574 short messages. Table 4 shows its basic statistics.

Table 4. Basic statistics.

Class	Amount	%
Hams	4,827	86.60
Spams	747	13.40
Total	5,574	100

4.2 Results

After applying our term weighting scheme, we have tested three well-known data mining algorithms on the transformed corpus. Table 5 shows the effectiveness of our term weighting algorithm for text classification. The classification accuracies obtained by successively applying SVM, DT and LR algorithms on our term weighting representation are better than those obtained on TF-IDF and TF-IGM.

Table 5. Basic statistics.

	Classification accuracy		
Algorithm	*tf-icd*	*tf-igm*	*tf-idf*
SVM	**0.8829**	0.8779	0.8756
DT	**0.9354**	0.9292	0.9297
LR	**0.8836**	0.8787	0.8763

5 Conclusion and Perspectives

In this paper, we studied the term weighting scheme issue. We proposed an efficient term weighting scheme based on inertia contribution of a document.

The test results of text classification show their convincible efficiency. We plan in our future work to conduct our algorithm on others benchmarks data sets.

Acknowledgment. We would like to express our sincere thanks to the CEA-MITIC (Centre d'Excellence Africain en Mathématiques, Informatique et Tic) who financed our research by paying the publication fees of the 2papiers that we published in Africatek2018. The CEA-MITIC, located at the UFR of Applied Sciences and Technology (UFR SAT) of the Gaston Berger University (UGB) of Saint-Louis in Senegal, is a consortium of university institutions in Senegal and subregion of Senegal, research institutions and national, regional and international companies involved in the ICT sector.

References

1. Dejun, X., Maosong, S.: Chinese text categorization based on the binary weighting model with non-binary smoothing. In: Sebastiani, F. (ed.) ECIR 2003. LNCS, vol. 2633, pp. 408–419. Springer, Heidelberg (2003). https://doi.org/10.1007/3-540-36618-0_29
2. Lan, M., Tan, C.L., Jian, S., Yue, L.: Supervised and traditional term weighting methods for automatic text categorization. IEEE Trans. Pattern Anal. Mach. Intell. **31**(4), 721–735 (2009)
3. Deng, Z.-H., Tang, S.-W., Yang, D.-Q., Zhang, M., Li, L.-Y., Xie, K.-Q.: A comparative study on feature weight in text categorization. In: Yu, J.X., Lin, X., Lu, H., Zhang, Y. (eds.) APWeb 2004. LNCS, vol. 3007, pp. 588–597. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24655-8_64
4. Debole, F., Sebastiani, F.: Supervised term weighting for automated text categorization. In: Proceedings of the 2003 ACM Symposium on Applied Computing, pp. 784–788 (2003)

5. Yang, J., Wang, J., Liu, Z., Qu, Z.: A term weighting scheme based on the measure of relevance and distinction for text categorization. In: International Conference on Advanced Computing Technologies and Applications, ICACTA-2015, pp 13–22. https://doi.org/10.1016/j.procs.2015.03.074
6. Feng, G., Wang, H., Sun, T., Zhang, L.: A term frequency based weighting scheme using naïve bayes for text classification. J. Comput. Theor. Nanosci. 319–326 (2016). https://doi.org/10.1166/jctn.2016.4807
7. Wang, T., Cai, Y., Leung, H., Cai, Z., Min, H.: Entropy-based Term Weighting Schemes for Text Categorization in VSM. In: 2015 IEEE 27th International Conference on Tools with Artificial Intelligence, 12 p. https://doi.org/10.1109/ictai.2015.57
8. Yoshida, T.M.M.K.K.: Term weighting method based on information gain ratio for summarizing documents retrieved by IR systems. J. Nat. Lang. Process. **9**(4), 3–32 (2001)
9. Chen, K., Zhang, Z., Long, J., Zhang, H.: Turning from TF-IDF to TF-IGM for term weighting in text classification. J. Expert Syst. Appl. **66**(C), 245–260 (2016)
10. Cormack, G.V., Gómez Hidalgo, J.M., Puertas Sánz, E.: Spam filtering for short messages. In: Proceedings of the Sixteenth ACM Conference on Conference on information and Knowledge Management, Lisbon, Portugal, 06–10 November 2007, CIKM 2007, pp. 313–320. ACM, New York (2007). http://doi.acm.org/10.1145/1321440.1321486
11. Geng, J., Lu, Y., Chen, W., Qin, Z.: An improved text categorization algorithm based on VSM. In: 2014 IEEE 17th International Conference on Computational Science and Engineering. https://doi.org/10.1109/cse.2014.313
12. Wu, H., Gu, X., Gu, Y.: Balancing between over-weighting and under-weighting in supervised term weighting. Inf. Process. Manag. **53**, 547–557 (2017). https://doi.org/10.1016/j.ipm.2016.10.003
13. Karisani, P., Rahgozar, M., Oroumchian, F.: A query term re-weighting approach using document similarity. Inf. Process. Manag. **52**(3), 478–489 (2016). https://doi.org/10.1016/j.ipm.2015.09.002
14. Haddoud, M., Mokhtari, A., Lecroq, T., Abdeddaïm, S.: Combining supervised term-weighting metrics for SVM text classification with extended term representation. Knowl. Inf. Syst. (2016). https://doi.org/10.1007/s10115-016-0924-1
15. Pang, B., Lee, L.: A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the ACL 2004. http://www.cs.cornell.edu/people/pabo/movie-review-data