Addressing the Problem of Unbalanced Data Sets in Sentiment Analysis

Asmaa Mountassir, Houda Benbrahim and Ilham Berrada
ALBIRONI Research Team, ENSIAS, Mohamed 5 University, Souissi, Rabat, Morocco

Keywords: Sentiment Analysis, Opinion Mining, Unbalanced Data Sets, Machine Learning, Text Classification, Natural Language Processing, Arabic Language.

Abstract: Sentiment Analysis is a research area where the studies focus on processing and analysing the opinions available on the web. This paper deals with the problem of unbalanced data sets in supervised sentiment classification. We propose three different methods to under-sample the majority class documents, namely Remove Similar, Remove Farthest and Remove by Clustering. Our goal is to compare the effectiveness of the proposed methods with the common random under-sampling. We use for classification three standard classifiers: Naïve Bayes, Support Vector Machines and k-Nearest Neighbours. The experiments are carried out on two different Arabic data sets that we have built and labelled manually. We show that results obtained on the first data set, which is slightly skewed, are better than those obtained on the second one which is highly skewed. The results show also that we can rely on the proposed techniques and that they are typically competitive with random under-sampling.

1 INTRODUCTION

Nowadays, the web is no longer just a source of information for internet users; it represents also a space where simple users can provide information. With the emergence of social media (such as social networking sites, online news sites, online web forums, personal blogs and online review sites), internet users are more and more invited to express their opinions, post comments or share experiences about any topic. Therefore, the online opinion has become an important currency for many researches especially in the field of Opinion Mining (OM) and Sentiment Analysis (SA).

SA is a subfield of Text Mining that gives interest to process and analyse opinions expressed, by different kinds of authors, on the web. There are several tends in this area. Some studies deal with Subjectivity Analysis where the classification classes are OBJECTIVE vs. SUBJECTIVE (Abdul-Mageed et al., 2011), while others focus on Sentiment Classification (i.e. classification by polarity), where classification classes are POSITIVE vs. NEGATIVE (Pang et al., 2002). We can find also some studies about opinion summarization (Zhuang et al., 2006).

For most studies in SA, we note that the problem of unbalanced data sets (UD) is not tackled. It is often assumed that positive and negative classes are balanced. After building and labelling of the data set, this one is often equalized so to have the same number of documents in each class. Otherwise, if document collection is based on a rating system, we gather the same number of documents for each class (Pang et al., 2002; Rushdi-Saleh et al., 2011a). Nevertheless, this assumption may not hold in the real world since we cannot always have the same number of positive documents as the number of negative ones for a given subject. Hence the importance of addressing the problem of UD in SA.

To resolve the problem of UD, there are generally two approaches (Japkowicz, 2003). The first approach tends to modify the classifier, such as using the cost-sensitive learning (Brank et al., 2003) and modifying the classifier to handle UD (Wu and Chang, 2003). The second approach deals with the modification of the data set itself. It consists of two common methods, the first focuses on under-sampling (Kubat and Matwin, 1997) while the second deals with over-sampling (Chawla et al., 2002). The under-sampling method seeks to reduce the number of majority class members in the training set. While the over-sampling method seeks to increase the number of minority class members in

Mountassir A., Benbrahim H. and Berrada I..
Addressing the Problem of Unbalanced Data Sets in Sentiment Analysis.
DOI: 10.5220/0004142603060311
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2012), pages 306-311
ISBN: 978-989-8565-29-7
Copyright © 2012 SCITEPRESS (Science and Technology Publications, Lda.)

the training set. Note that majority class refers to the class with more documents, and minority class denotes the class with fewer documents.

Existing works that address UD in SA have mostly used random under-sampling. Li et al. (2011) have tested many methods for UD and found that random under-sampling is the most effective. Then they applied this technique together with a semi-supervised method to classify four different unbalanced data sets for OM. Their study cannot be directly comparable to ours since they are interested in semi-supervised approach while our work focuses on supervised learning. Rushdi-Saleh et al. (2011b) used, among others, an English highly unbalanced data set of product reviews. They have used Support Vector Machines as a classifier. They achieved high results on the unbalanced data set. However, they did not give details about how they handle the skewed data. Burns et al. (2011) have performed a comparative study between balanced and unbalanced sentiment classification for customer reviews. Likewise, they used random-sampling to balance their data sets. They employed Naïve Bayes and Language Model (Carpenter, 2005) as classifiers. They show that a realistic unbalanced data set can achieve substantially better results. However, their results could not be reliable since they used the accuracy as performance measure. According to Kubat and Matwin (1997), the classifier's performance in applications of this kind cannot be expressed in terms of the average accuracy. We will give more details about this problem in section 4.

In this paper, we focus on under-sampling methods in supervised sentiment classification in an Arabic context. We propose three different techniques, namely Remove Similar (RS), Remove Farthest (RF) and Remove by Clustering (RC). The idea behind these methods is that we seek to keep, among majority class documents, only those that can be representative for their class. Our initial hypothesis is that targeted removal, from majority class, may be more effective than random removal. This is why we compare the effectiveness of these techniques with that of random under-sampling that we call Random Removal (RR). We point out that these methods are independent from domain, language and classification technique. We have built and manually labelled two different sized Arabic data sets of two different domains. The first data set is slightly skewed, while the second is highly skewed. We use three standard classifiers, namely Naïve Bayes (Mitchell, 1996), Support Vector Machines (Vapnik, 1995) and k-Nearest Neighbours (Dasarathy, 1991). Our main goal is to evaluate the performance of the different under-sampling methods used in this study on two data sets with different unbalance percentage.

The remainder of this paper is organized as follows. The second section describes the three under-sampling methods that we propose. The third section presents the data collection that we have used. The fourth section describes the experimental environment and the performed experiments. Afterward, the obtained results are presented. We finish by a conclusion and future works in the last section.

2 UNDER-SAMPLING METHODS

In this section we describe each of the proposed under-sampling methods. These techniques are respectively Remove Similar, Remove Farthest and Remove by Clustering. In the following, we denote by C majority class.

2.1 Remove Similar (RS)

This method consists of eliminating from majority class the documents that may be very similar to other documents of the same class. The intuition behind this method is that similar documents may not give supplementary information to learn the classifier. We think that removing such documents may help to balance data sets without losing much information. The algorithm of this method is as follows.

a. Compute for each document of C its distance to each document of C.
b. Assign a score to each document of C. This score corresponds to the minimum distance of calculated distances in step a.
c. Remove from C the document with the smallest score.
d. If the desired number of documents to remove is achieved, end of the algorithm. Otherwise, return to step b (we do not take into account removed documents while assigning scores).

2.2 Remove Farthest (RF)

The principle of this method is to eliminate from majority class documents that are the farthest ones from the rest of majority class members. We think that the farthest documents might be a source of noise or might represent specific documents. In both cases, the removal of such documents may help to

balance data sets without losing much information. We specify that, in our study, the maximum value that can have a distance between two given documents is 1. This notion is used in the algorithm of this method where the steps are as follows.

a. Compute for each document of C its distance to each document of C.
b. Assign a score to each document of C. This score corresponds to the number of times where computed distances in step a had 1 as value.
c. Sort, in descending order, documents of C by their scores.
d. Remove from C the first n documents sorted in step c. The number n corresponds to the number of documents that we desire to remove.

2.3 Remove by Clustering (RC)

The principle of this method is different from that of RS and RF. RC is based on clustering while RS and RF are based on computing distances between documents. The intuition behind this method is that, if we apply a clustering algorithm on majority class documents, the selection of clusters' centres to represent majority class might be helpful to perform an optimal balancing of data sets. The steps of this method are described below.

a. Apply a clustering algorithm on C documents; the number of clusters to form corresponds to the number of C documents that we seek to keep.
b. For each cluster, compute for all its documents their scores. A document score is obtained by averaging its distances to the other members of the same cluster.
c. Select from each cluster the document with the smallest score. The selected documents are assumed to represent in a way clusters' centres.
d. Remove all C documents except those selected in step c.

3 DATA COLLECTION

We collected our data sets from online forums of Aljazeera's web site[1]. We built two data sets of two different domains. The first consists of 594 comments about movie reviews toward a movie that has generated much noise in the Arab-Muslim world, we called it DSMR. The second consists of 1082 comments about a political issue titled "Arab support for the Palestinian affair", we called it DSPo.

Since our study deals with supervised learning, we had to label manually our data sets. The categories to consider are POSITIVE, NEGATIVE, OBJECTIVE and NOT_ARABIC. POSITIVE (POS) category includes all comments that reflect a positive opinion regardless of the opinion object. Likewise, NEGATIVE (NEG) category contains comments with negative sentiments. OBJECTIVE (OBJ) category consists of non-opinionated comments (as in "The movie relates the war events"), comments that reflect neutral opinions (such as "I hope to know more details about this story") or comments that contain a mixture of positive and negative opinions (as in "I enjoyed the movie; however there was a falsification of the real story"). Finally, NOT_ARABIC (N_AR) category used to clean our data sets since this category includes all comments not written in Arabic. Table 1 gives an overview on the distribution of each category for the two data sets.

Table 1: Number of documents per category for each data set.

	POS	NEG	OBJ	N_AR
DSMR	184	284	106	20
DSPo	149	462	383	88

Since the present study focuses on sentiment classification, the categories we are interested in are POSITIVE and NEGATIVE ones. Documents of OBJECTIVE and NOT_ARABIC categories were ignored. We will return back to objective documents in next studies since they represent an important percentage (up to 35.4% of the whole data set). Table 2 illustrates the final structure of the used data sets.

Table 2: Structure of the final data sets.

	POS	NEG	Total
DSMR	184 (39.4%)	284 (60.6%)	468
DSPo	149 (24.4%)	462 (75.6%)	611

As we can see from table 2, both DSMR and DSPo are unbalanced. We can see also that DSMR is slightly skewed since it contains 39.4% of positive documents versus 60.6% of negative documents. However, DSPo is highly skewed in comparison with DSMR. DSPo contains 24.4% of positive documents versus 75.6% of negative documents. We specify that, for our study, majority class is the

[1] http://www.aljazeera.net

NEGATIVE category; minority class is the POSITIVE one. This unbalance reflects one of the main characteristics of comments derived from Aljazeera's site: positive comments are largely dominated by negative ones. This is due to the nature of discussed issues in the forums of Aljazeera's web site which often deal with problems of the Arab-Muslim world. Finally, we note that, while labelling the data, we noticed that, for both data sets, there are a great number of comments that are off-topic. This type of comments can be a source of noise. Moreover, the more their percentage is large the more the data set becomes heterogeneous. This lack in document homogeneity may have a great impact on classification performance.

4 EXPERIMENTS

In this section, we present the experiments that we have performed for this study. Firstly, we describe the experiments' environment, i.e. some details about the pre-processing techniques, the classifiers, the validation method and the performance measure that we have used. Secondly, we present the obtained results then we discuss the different findings.

4.1 Experimental Design

4.1.1 Pre-Processing

As a pre-processing step, we have removed from textual documents all characters which are not Arabic letters. We have also removed stop words. As stemming process, we have applied light-stemming method (Khoja and Garside, 1999). Moreover, we have eliminated, from feature space, terms that occur once in the data set. This elimination may help us to clean data sets from typing errors made by document authors who are simple internet users. For weighting scheme, we have used a binary weighting which is based on term presence. Finally, documents were considered as bags-of-words.

4.1.2 Classification and Algorithms

In this study, we are interested in a single-label binary classification where each document is assigned one of the two categories POSITIVE vs. NEGATIVE.

For tasks of pre-processing and classification, we have used the data mining package Weka[2] (Witten and Frank, 2005). We have used three standard classifiers; namely Naïve Bayes (NB), Support Vector Machines (SVM) and k-Nearest Neighbours (k-NN). For the NB classifier, we have used a kernel estimator, rather than a normal distribution, for numeric attributes (John and Langley, 1995). Concerning SVM classifier, we have used a normalized polynomial kernel with a Sequential Minimum Optimization (SMO) (Platt, 1999). For the k-NN classifier, we have used a linear search with a cosine-based distance (Salton and McGill, 1983).

For our proposed under-sampling techniques, we have used cosine-based distance for RS and RF methods, and k-means (Hartigan, 1975) as clustering algorithm for RC method.

4.1.3 Validation Method

Both DSMR and DSPo were randomly split into two sets: a training set representing 75% of the data set, and a test set representing 25% of the data set. We specify that this split is stratified, i.e. we keep the same category distribution in both training and test sets as in the initial data set. This process was repeated 25 times to obtain, as output, 25 samples for each data set.

Each experiment to perform on a data set is run on all its 25 samples. The final result to consider, for a given data set, results from averaging its samples' results. We point out that all under-sampling methods are applied only on training sets; test sets are let as they are. Training sets need to be balanced in order not to bias classifiers' learning. But test sets are not modified so as to test classifiers on sets representing the reality.

4.1.4 Performance Measure

To evaluate our proposed under-sampling methods, we have adopted the popular g-performance which is obtained as follows:

$$\text{g-performance} = \sqrt{Acc_+ * Acc_-} \quad (2)$$

where Acc_+ and Acc_- denote the accuracy for positive and negative class respectively. These two accuracies are obtained from the confusion matrix that presents all possible category assignments for a binary classification. It is illustrated in table 3.

Table 3: Confusion matrix.

	Positive prediction	Negative prediction
Positive class	True Positive (TP)	False Negative (FN)
Negative class	False Positive (FP)	True Negative (TN)

TP and TN denote the number of correctly classified positive and negative documents, while FP and FN denote the number of misclassified positive and negative documents, respectively. The positive and negative accuracies are respectively defined as follows:

$$Acc_{+} = TP / (TP+FN) \quad (3)$$

$$Acc_{-} = TN / (TN+FP) \quad (4)$$

This metric is called g-performance because it corresponds to the geometric mean of the positive and negative accuracies. G-performance is suitable for unbalanced classification since it maximizes the accuracy of the two classes in order to balance both classes at the same time (Kubat and Matwin, 1997).

4.2 Results

The experiments that we have carried out consist on equalizing the two classes of both DSMR and DSPo by the application of the four under-sampling methods that we study (RS, RF, RR and RC). Afterward, we employ each of the three classifiers (NB, SVM and k-NN) to classify the balanced data sets.

Tables 4 and 5 present the results in g-performance obtained respectively on DSMR and DSPo. For each classifier, we show the result of classification after equalization of the two classes by each of the four under-sampling methods. The best result for each classifier is marked in bold.

Table 4: Results in g-performance by classifier on DSMR.

	RR	RS	RF	RC
NB	**71.9**	65.5	60.9	70.6
SVM	**72.1**	68.1	**72.1**	69.2
kNN	66.6	60	**67**	64.5

Table 5: Results in g-performance by classifier on DSPo.

	RR	RS	RF	RC
NB	**65.7**	60.9	51	65
SVM	53.1	58.8	**61.4**	53.6
kNN	**62.9**	56.7	61.8	60.3

As a comparison between the two data sets, we can see that results on DSMR typically outperform those obtained on DSPo. Results range from 60% to 72.1% for DSMR; and from 51% to 65.7% for DSPo. Recall that DSPo is highly skewed with regard to DSMR. This could explain the difference in results. Hence, the more the data set is unbalanced the more the results are not promising.

To compare between the four under-sampling methods, we can see from tables 4 and 5 that, generally and by taking into account the standard deviation, the four methods yield near results. However, it is clear that RR gives mostly the best results for the four under-sampling methods. Moreover, RF seems to be competitive with RR for SVM on both DSMR and DSPo. We can also see from tables 4 and 5 that RF is not recommended when we use NB. Likewise, RS does not suit with kNN. This can be understandable since kNN is based on neighbourhood to perform classification. By removing similar documents, we do not help this classifier to correctly classify documents.

We can explain the fact that the three proposed methods did not outperform random under-sampling by the nature of our data sets which are not homogeneous (see section 3). Indeed, when the documents of a data set are not homogeneous, this means that, generally, all documents are far from each other; so neither removing similar nor removing the farthest could be meaningful in that case. It is the same for removal by clustering, since even we choose a centre for each cluster, this centre may not be able to represent all documents of its cluster. The reason is that we specify in the beginning the number of clusters to form.

5 CONCLUSION AND FUTURE WORKS

The present study deals with the problem of unbalanced data sets in Sentiment Analysis. We propose three different under-sampling methods that we called respectively Remove Similar, Remove Farthest and Remove by Clustering. Our main goal, by this study, is to evaluate the effectiveness of these three methods. Furthermore, we compare them with the random under-sampling which is commonly used in the literature of unbalanced classification. We focus on supervised classification of Arabic documents. We use three standard classifiers known by their effectiveness: Naïve Bayes, Support Vector Machines and k-Nearest Neighbours. We use as data collection two unbalanced Arabic data sets (DSMR and DSPo) that we have built internally and labelled manually. These data sets are different in many aspects, namely document size, domain and unbalance degree. DSMR contains 468 documents; 39.4% of them are positive and 60.6% are negative. So, DSMR is slightly skewed. DSPo contains in total 611 documents, where 24.4% are positive and 75.6% are negative. DSPo is therefore highly skewed. The majority class in our study is the NEGATIVE category, while the POSITIVE one

presents the minority class.

Our experiments consist on balancing the two classes of each data set by the use of the four studied under-sampling methods, i.e. RR, RS, RF and RC. Then we evaluate the performance of the three classifiers on the balanced data sets.

Our results show that performance obtained on DSMR is better than that obtained on DSPo. This proves that the more the data set is unbalanced the more the results are bad.

As a comparison between under-sampling methods, we can say that, generally, the four methods give near results. But iQn most of cases RR yields the best results. RF is not recommended for NB, it is rather recommended for SVM. For kNN, we do not recommend to use RS.

As future works, we look for performing the same experiments on unbalanced data sets that are more homogeneous so as to validate our hypothesis about the impact of heterogeneity on the performance of the proposed techniques. We will also study the effectiveness of the four under-sampling methods by decreasing progressively majority class size. On one hand, we aim to see whether it is necessary to achieve a balance of 50%-50% to have the best results. On the other hand, we aim to observe the behaviour of our classifiers, by using the different under-sampling methods, toward the different steps of majority class decreasing. Finally, we have as perspective too the study of feature selection techniques on unbalanced data sets of SA.

REFERENCES

Abdul-Mageed, M., Diab, M.T., Korayem, M., 2011. Subjectivity and Sentiment Analysis of Modern Standard Arabic. *In Proc. ACL (Short Papers)*. pp.587-591.

Brank, J., Grobelnik, M., Milić-Frayling, N, Mladenić, D., 2003. Training text classifiers with SVM on very few positive examples. *Technical report*, MSR-TR-2003-34.

Burns, N., Bi, Y., Wang, H., Anderson, T., 2011. Sentiment Analysis of Customer Reviews: Balanced versus Unbalanced Datasets. *KES 2011*, Part I, LNAI 6881, pp. 161-170.

Carpenter, B., 2005. Scaling High-Order Character Language Models to Gigabytes. *In: Workshop on Software. Association for Computational Linguistics, Morristown*. pp. 86–99.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer. 2002. W. P. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research (JAIR)*, Volume 16, pp. 321-357.

Dasarathy, B. V., 1991. Nearest Neighbor (NN) *Norms: NN Pattern Classification Techniques*. McGraw-Hill Computer Science Series. Las Alamitos, California: IEEE Computer Society Press.

Hartigan, J., 1975. *Clustering Algorithms*. John Wiley & Sons, New York, NY.

Japkowicz, N., 2003. Class Imbalances: Are we Focusing on the Right Issue? *In Proc. Of ICML'03*.

Khoja, S., Garside, R., 1999. Stemming Arabic text. Computer Science Department, Lancaster University, Lancaster, UK.

Kubat, M., Matwin, S., 1997. Addressing the Curse of Imbalanced Data Sets: One-Sided Sampling. *In Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 179-186.

Li, S., Wang, Z., Zhou, G., Lee, S. Y. M., 2011. Semi-Supervised Learning for Imbalanced Sentiment Classification. *In Proc. Of the Twenty-Second International Joint Conference on Artificial Intelligence*, pp.1826-1831.

Mitchell, T., 1996. *Machine Learning*. McCraw Hill.

Pang, B., Lee, L., Vaithyanathain, S., 2002. Thumbs up? Sentiment classification using machine learning techniques. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*. pp.79-86.

Platt, J., 1999. Fast training on SVMs using sequential minimal optimization. *In Scholkopf, B., Burges, C., and Smola, A. (Ed.), Advances in Kernel Methods: Support Vector Learning*, MIT Press, Cambridge, MA, pp.185-208.

Rushdi-Saleh, M., Martin-Valdivia, M. T., Urena-Lopez, L. A., Perea-Ortega, J. M., 2011a. Bilingual Experiments with an Arabic-English Corpus for Opinion Mining. *In Proc. Of Recent Advances in Natural Language Processing, Hissar, Bulgaria.* pp.740-745.

Rushdi-Saleh, M., Martin-Valdivia, M. T., Urena-Lopez, L. A., Perea-Ortega, J. M., 2011b. Experiments with SVM to classify opinions in different domains. *Expert Systems with Applications 38*, pp.14799-14804.

Salton, G., McGill, M., 1983. *Modern Information Retrieval*. New York: McGraw-Hill.

Vapnik, V., 1995. *The Nature of Statistical Learning*. Springer-Verlag.

Witten, I. H., Frank, E., 2005. *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco, California.

Wu, G., Chang, E., 2003. Class-Boundary Alignment for Imbalanced Dataset Learning. *In Proc. Of ICML'03*.

Zhuang, L., Jing, F., Zhu, X., 2006. Movie Review Mining and Summarization. *In CIKM'06*. Virginia, USA.