International Journal of Engineering &Technology, 7 (1.8) (2018) 62-66

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A novel k-nearest neighbor distance based under sampling for improved opinion mining on skewed data using random forest

P Ratna Babu [1] *, Dr. Bhanu Prakash Battula [2]

[1] Sri Chundi Ranganayakulu Engineering College, AP, India
[2] Tirumala Engineering College, Narasaraopet, AP, India
**Corresponding author E-mail: ratnajoyal@gmail.com*

Abstract

In recent years, consumers are performing a pilot investigation using online resources before making any decision of purchase. One of the most popular social blogging online medium is twitter. The opinions collected from twitter at any point of frame in real world scenario are tending towards class imbalance in nature. The existing algorithms for opinion mining can work better on class balance nature, where opinions (positive and negative) are almost balance. In this paper, we propose a novel approach known as Improved Opinion Mining using Under Sampling (IOMUS) to efficiently summarize the reviews of class imbalance opinion mining corpus. The experimental set up is performed on the collection of opinion mining class imbalance dataset consisting of "1155" instances. The experimental results suggest that improved performance is obtained by the proposed IOMUS algorithm than the traditional approach.

Keywords: *Classification; Opinion Mining; Imbalanced Data; Under Sampling; IOMUS.*

1. Introduction

In many real time applications large amount of data is generated with skewed distribution. A data set said to be highly skewed if sample from one class is in higher number than other [1-2]. In imbalance data set the class having more number of instances is called as majority class while the one having relatively less number of instances is called as minority class [2]. Applications such as medical diagnosis prediction of rare but important disease is very important than regular treatment. Similar situations are observed in other areas, such as detecting fraud in banking operations, detecting network intrusions [3], managing risk and predicting failures of technical equipment. Studies on class imbalance classification have grown more emphasis only in recent years [4]. Reported works in classifications for class imbalance distribution come in many ranges of domain applications like fault diagnosis [5-6], anomaly detection [7], medical diagnosis [8-9], detection of oil spillage in satellite images [10], face recognition [11], text classification [12], protein sequence detection [13] and many others. The significant challenges of the class imbalance problem and its repeated incidence in practical applications of pattern recognition and data mining have engrossed many researchers that two workshops dedicated to research efforts in addressing the class imbalance problems were held at AAAI 2000 [14] and ICML 2003 [15] respectively.

Opinion mining has become a very active research area in natural language processing (NLP) and has attracted increasing interest in data mining, Web mining, and text mining. The aim of Opinion mining is to analyze people's opinions, evaluations, attitudes, appraisals, and emotions towards particular entities. These entities might be services, organizations, products, events, topics, individuals, issues, or their attributes. Opinion mining includes several tasks such as opinion extraction, subjectivity classification, polarity determination, affect analysis, review mining, etc. The research interest in Opinion mining can be attributed to several reasons. First of all, it has a wide range of applications, and is applicable in several domains, such as branding and product analysis, expressive text-to-speech synthesis, question answering, analysis of political debates, tracking sentiment timeliness in online forums and news, and conversation summarization. Second, there are still several gaps and challenging research problems that require further studies to build more reliable and effective systems. Third, Opinion mining is a helpful and useful tool to analyze the rapid growth of user-generated contents which are expressed in several online media such as blogs, wikis, web forums and social networks. Through these platforms or environments, users can express their opinions, post information, share knowledge, and get feedback from each others. In this research, we propose to present a framework for efficient opinion mining analysis for class imbalance data using under sampling strategy.

The arrangement of paper is follows as. We exhibit in Sec. 2 the recent approaches in opinion mining summarization. In Section 3, we present the proposed approach. Section 4 presents the dataset and the assessment criteria's designed for class imbalance learning of twitter dataset. Test results are accounted for in Section 5. In section 6 of conclusion, we finish up where we talk about real open issues and upcoming work.

2. Related work

There is vast literature published for opinion mining in recent years. We have selectively taken a very few contributions which are under the domain of class imbalance learning.

Lincy Meera Mathews et al., [16] have proposed an improved Nearest Neighbor Classifier for a two class imbalanced data using three oversampling techniques for generation of artificial instances for the minority class for balancing the distribution among the classes. Sadam Al-Azani et al., [17] compared the performance of

different classifiers for polarity determination in highly imbalanced short text datasets using features learned by word embedding rather than hand-crafted features.

Farrukh Ahmed et al., [18] have proposed on generating a mining table by aggregating information from multiple local tables and external data sources and automatically generating potentially discriminant features. They also prevented leakage of the class information by avoiding features built after the knowledge of the class label. Jerzy Stefanowski[19] have discuss different challenging tasks in machine learning and data mining such as the data difficulty factors which deteriorate classification performance: decomposition of the minority class into rare sub-concepts, overlapping of classes and distinguishing different types of examples.

TengNiu et al., [20] have introduced a multi-view sentiment analysis dataset (MVSA) including a set of image-text pairs with manual annotations collected from Twitter. The dataset can be utilized as a valuable benchmark for both single-view and multi-view sentiment analysis. With this dataset, many state-of-the-art approaches are evaluated. More importantly, the effectiveness of the correlation between different views is also studied using the widely used fusion strategies and an advanced multi-view feature extraction method. Vasileios Athanasiou et al., [21] have proposed a novel ensemble algorithm with gradient boosting technique that can learn with different loss functions providing the ability to work efficiently with high dimensional data. Moreover, the algorithm is build to work o class imbalance issues since the distribution of sentiments in real-world applications often displays issues of inequality.

Michael Crawford et al., [22] have conducted a survey using prominent machine learning techniques that have been proposed to solve the problem of review spam detection and the performance of different approaches for classification an detection of review spam. The majority of current research has focused on supervised learning methods, which require labelled data, a scarcity when it comes to online review spam. Research on methods for Big Data are of interest, since there are millions of online reviews, with many more being generated daily. V. Gopalakrishnan et al., [23] have proposed a modification in ensemble based bagging algorithm and also in sampling method used for data distribution, so as to solve class imbalance problem to improve the classification performance.

BartoszKrawczyk[24] have discussed open issues and challenges such as classification, regression, clustering, data streams, big data analytics and applications, e.g., in social media and computer vision that need to be addressed to further develop the field of imbalanced learning. Julien Ah-Pine et al., [25] have proposed a novel method for twitter opinion mining using machine learning method such as synthetic oversampling techniques for imbalanced learning using tweet-polarity classification. Troy Raeder et al., [26] have discussed on fundamental issue that, it is not sufficiently to consider the sensitivity of classifiers both to class imbalance as well as to having only a small number of samples of the minority class.

After analysing the above recent contributions, we concluded that there is great scope for investigation and research in the opinion mining. We propose to present a framework for opinion mining which can efficiently perform knowledge discovery for the class imbalance corpus of twitter social database.

1) Framework of Improved Opinion Mining using Under Sampling (IOMUS) Algorithm

Process of Sentiment Analysis

Sentiment Analysis is a complicated procedure which has five phases for analysis of sentiment data. The phases are:

Data collection: This is the first stage of Sentiment Analysis which comprises the collection of data from user-generated content present in blogs, forums or social networks. The data is not organized, conveyed in various manners through usage of various vocabularies, slang or contexts.

Text preparation: This comprises of cleaning of the extracted data prior to analysis. Non-textual as well as non-relevant content are detected and discarded.

Sentiment detection: Extracted sentences are analysed and those with subjective opinion are retained while the remaining are eliminated. Sentiment detection is done at different levels either single term, phrases, complete sentences or complete document with commonly used techniques.

Sentiment classification: Here, subjective sentences may be carried out through usage of several points. Presentation of output: The primary aim of opinion mining is the conversion of non structured data into useful information. When the analysis is over, text results are illustrated on graphs such as pie or bar charts and even line graphs. The different components of our new proposed framework are elaborated in the next subsections.

Step 1: Preparation of Majority and Minority Subset

The datasets is partitioned into majority and minority subsets. As we are concentrating on under sampling, we will take majority data subset for further analysis to generate synthetic instances.

Step 2: Selection of novel subset of Instances

Majority subset can be further analyzed to find the missing or noisy instances so that we can eliminate those instances to improve the quality of the subset. For finding noisy, boarder line and missing value instances for generating improved majority subset one of the ways is to go through a preprocessing process and to apply distance measure.

Step 3: Performing under sampling from majority subset

The majority instances, which are excess in percentage than the minority subset are reduced by following the intelligent inexact technique for removal. In this technique the influential features are selected and retrieved for further utilization. The weak or less influential features are selected for removal of instances which are in the border line and range of misclassification. The process of finding such instances can be done by applying techniques of polarity finding in a semi group of instances using KNN (K Nearest Neighbor) searching algorithm. The main principle of investigation in the KNN approach is to find the percentage of opposite polarity instances in the group for identification of mostly misclassified or outlier instances.

Step 4: Forming the Strong Dataset

The under sampled majority subset and the minority subset are combined to form a strong and almost balance dataset, which is used for learning of a base algorithm. In this case again we have used Random Forest [27] as the base algorithm. Our method will be superior to other under sampling methods since our approach performed under sampling using the instance specific technique for instance removal.

3. Dataset and evaluation criteria's

4.1. Opinion mining with twitter datasets

The twitter datasets considered for analysis consists of 1155 opinions, in which 944 are positive opinions and 208 are negative opinions is show in the table 1. The imbalance ratio (IR) of the considered dataset is 5.52.

Table 1: The Twitter Datasets and Their Properties

S. No	Dataset	Instances	Missing	Features	IR
1	Twitter	1155	No	993	5.52

The twitter opinion mining dataset sample instances with features and class can be seen below,

Twitter Datasets:
@relation Twitter @attribute Twitter numeric @attribute body string @attribute class {pos,neg} @data 1229709107,'anyone feel motivated the fri afternoon prior to a holiday? wanted to get lots done... but i want jammies and judge judy... \"SIR!\" < 3 her ',pos 1231217680,'I had the same issue with dominions site. Fixed it by using internet explorer ', neg

In most of the cases, the analysis of the twitter dataset was done assuming it as a balance dataset. We propose to analyze the twitter dataset as an imbalance dataset, the reason is, almost all the real world datasets are in imbalance nature. The existing algorithms are not efficient in discovering the hidden knowledge from the imbalance twitter dataset. We proposed a novel IOMUS algorithm for efficient knowledge discovery from the imbalance twitter dataset.

4.2. Preparation of the dataset

Take the twitter imbalance dataset and convert the string to vector by following morphological approach. After the morphological conversion of dataset, the numbers of features generated are very high. The most important features required for the further analysis should be identified. In this work the approach used to identify the important feature subset is by considering the feature to feature correlation and feature to class correlation. The dataset with important subset of features is considered for further analysis using our proposed IOMUS approach.

Pre Processed Twitter Datasets:

```
@relation Twitter
@attribute = numeric
@attribute About numeric
@attribute Agis numeric
@attribute Alt numeric
@attribute Although numeric
@attribute Amazing numeric
@attribute And numeric
@attribute Are numeric
@attribute As numeric
@attribute August numeric
@attribute BTW numeric
@attribute Beach numeric
@attribute Beatz numeric
@attribute Beautiful numeric
@attribute Been numeric
@attribute Behaviour numeric
@attribute Bella numeric
@attribute Best numeric
@attribute class {pos,neg}
@data
{0 1229709107,6 1,19 1,186 1,233 1,241 1,251 1,253 1,293 1,357 1,390
1,407 1,419 1,455 1,464 1,470 1,485 1,491 1,492 1,528 1,574 1,649 1,747
1,764 1,803 1,804 1}
{0 1231217680,114 1,294 1,443 1,483 1,675 1,698 1,747 1,792 1,834
1,875 1,917 1,921 1,941 1,942 1,992 neg}
{0 1229063765,233 1,269 1,402 1,483 1,521 1,605 1,656 1,745 1,747
1,764 1,877 1,889 1,896 1,897 1,905 1,944 1,950 1,985 1,987 1,992 neg}
```

The experimental methodology used for experimental simulation is 10 fold cross validation. In 10 fold cross validation the data source is divided into 10 equal partitions. In each run, one of the folds is used for testing and remaining folds are used for training the model. The mean of 10 runs are used for computing of evaluation metrics such as accuracy, AUC, TP rate, TN rate etc…

We performed the implementation of our new algorithms within the Weka [28] environment on windows 7 with i5-2410M CPU running on 2.30 GHz unit with 4.0 GB of RAM. The validation of the results is done using 10 fold cross validation, in which the dataset is split into 10 subsets and in each run nine subset are used for training and the remaining subset is used for testing. In 10 runs, the testing subset is altered and average measures for the 10 runs are generated.

4. Experimental results

We used the positive/negative polarity of the opinions from Twitter dataset. The dataset is highly imbalanced; the majority class is "positive" with 944 opinions, while the minority class is "negative" with 208 opinions. In our experiments, 10 fold cross validation technique is used for experimental validation. We evaluate our proposed approach with decision tree evaluator C4.5 and REP.

Table 2 summarizes the results obtained using C4.5, REP and the proposed IMOUS. We evaluated seven measures: accuracy, AUC, precision, recall, f-score, FP rate and FN rate. F-score is a more informative score since it considers both precision and recall measures. The evaluation metrics used in the paper are detailed below,

Accuracy is the percentage of correctly classified instances. AUC can be computed simple as the micro average of TP rate and TN rate when only single run is available from the classification algorithm.

The Area under Curve (AUC) measure is computed by,

$$AUC = \frac{TP_{RATE} + TN_{RATE}}{2} \quad (1)$$

The Precision measure is computed by,

$$\text{Pr}ecision = \frac{TP}{(TP)+(FP)} \quad (2)$$

The Recall measure is computed by,

$$\text{Re}call = \frac{TP}{(TP)+(FN)} \quad (3)$$

The F-score value is computed by,

$$F-score = \frac{2 \times \text{Pr}ecision \times \text{Re}call}{\text{Pr}ecision + \text{Re}call} \quad (4)$$

Table 2 provides both the numerical average performance (Mean) and the standard deviation (SD) results. If the proposed IMOUS technique is better than the compared technique then '●' symbol appears in the column. If the proposed technique is not better than the compared technique then '○' symbol appears in the column. The mean performances were significantly different according to the T-test at the 95% confidence level.

Table 2 presents the performance of C4.5 [29], REP [30] and proposed approach on class imbalance twitter dataset. The values in the table give a general view of the performance of C4.5, REP and proposed method IMOUS using each of the seven performance metrics. From table 2, it can be noticed that IOMUS learning method have performed robustly and the overall average results for all metrics are improved. The accuracy of IOMUS is 97.91±1.26, while it is 95.99±1.89 for C4.5 and 93.80±2.63 for REP algorithm. This also holds for AUC whose value is 0.990±0.015 for IOMUS, the value of 0.935±0.046 for C4.5 and 0.881±0.065 for REP. For precision the measures IOMUS is 0.980±0.013, which is an improvement from 0.964±0.019 for C4.5 and 0.956±0.021 for REP.

Table 2: Summary of Tenfold Cross Validation Performance for Accuracy on the Twitter Datasets

Measure	C4.5	REP	IOMUS
Accuracy	95.99±1.89●	93.80±2.63● 97.91±1.26	Accuracy
AUC	0.935±0.046●	0.881±0.065●	0.990±0.015
Precision	0.964±0.019●	0.956±0.021●	0.980±0.013
Recall	0.990±0.011●	0.973±0.019●	0.998±0.006
F-Score	0.977±0.011●	0.964±0.015●	0.989±0.007
FP Rate	0.209±0.111●	0.252±0.124○	0.239±0.164
FNRate	0.010±0.011●	0.027±0.019●	0.002±0.006

●Bold dot indicates the win of IOMUS on C4.5 and REP algorithm;

(A)

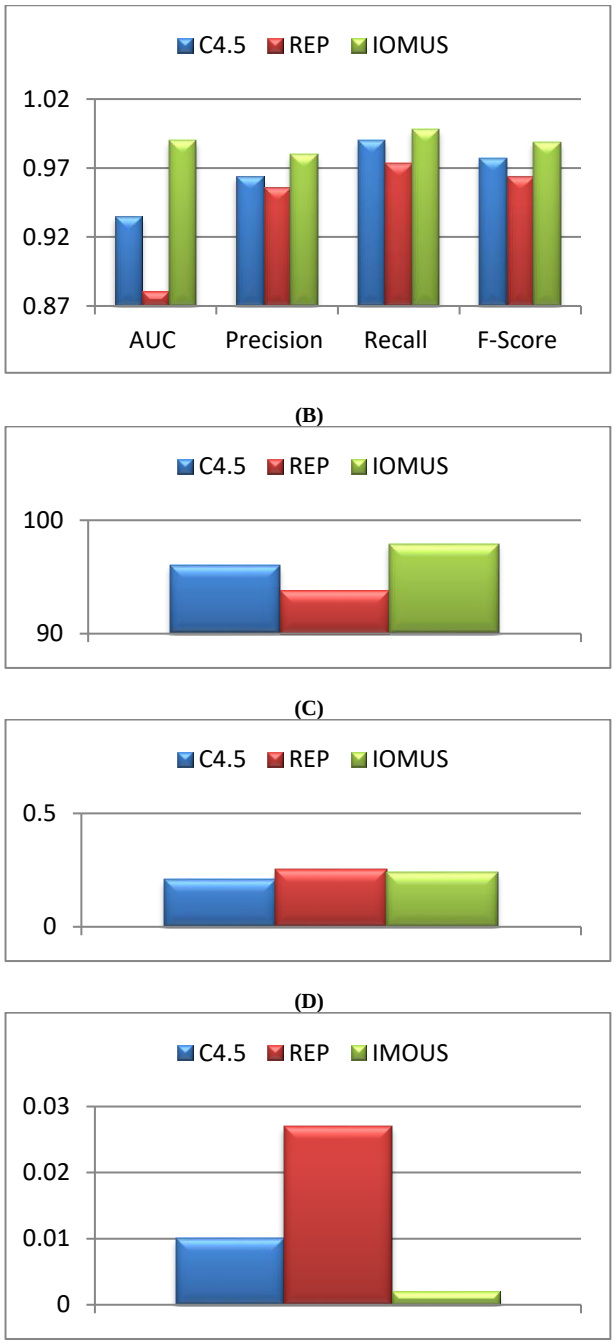

Fig. 1: (A) – (D): Presets Graphical Representation of All the Eight Validation Measures.

The value of recall of IMOUS i.e 0.998±0.006 is higher than for C4.5 i.e0.990±0.011 and 0.973±0.019 for REP. It is, also, clear that IOMOUS value has relatively better performed than C4.5 and REP in terms of f-score from 0.989±0.007 and 0.964±0.015 to 0.977±0.011. The FP rate ad FN rate measures are also improved for IOMUS than C4.5 and REP.

Table 2 Summary of tenfold cross validation performance for Accuracy on the Twitter datasets

Fig. 1 a–d Test results on AUC, precision, recall, F-score, accuracy, FP Rate and FN Rate between C4.5, REP verses IOMUS on imbalance twitter datasets.

The strength of our model is to remove the most weak examples recursively thereby strengthens the majority class. One more point to consider is our method tries to remove the most misclassified instances from majority subset. Firstly, the removal of some weak instances from majority set will not harm the dataset; in fact it will reduce the root cause of our problem of class imbalance as a whole by reducing majority samples in a small proportion.

Finally, we can say that proposed method is one of the best alternatives to handle class imbalance opinion datasets effectively. This experimental study supports the conclusion that the a prominent recursive under sampling approach can improve the class imbalance behaviour when dealing with imbalanced datasets, as it has helped the proposed approach to be the best performing algorithm when compared with C4.5 and REP algorithm.

5. Conclusion

Opinion mining is the process of analyzing the opinions to provide a recommendation to the user. The class imbalance opinion minion datasets are of critical in nature to analyze. In this paper, we propose a novel approach known as Improved Opinion Mining using Under Sampling (IOMUS) to efficiently summarize the reviews of class imbalance opinion mining corpus. The experimental set up is performed on the collection of opinion mining class imbalance dataset consisting of "1155" instances. The experimental results suggest that improved performance is obtained by the proposed IOMUS algorithm than the traditional approach.

References

[1] Shuo Wang, Member, and Xin Yao, "Multiclass Imbalance Problems: Analysis and Potential Solutions", IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, Vol. 42, No. 4, August 2012 https://doi.org/10.1109/TSMCB.2012.2187280.

[2] Nitesh V. Chawla, Nathalie Japkowicz, AleksanderKolcz "Special Issue on Learning from Imbalanced Data Sets" Volume 6, Issue 1 - Page 1-6.

[3] MikelGalar,Fransico, "A review on Ensembles for the class Imbalance Problem: Bagging,Boosting and Hybrid Based Approaches" IEEE Transactions On Systems, Man, And Cybernetics—Part C: Application And Reviews, Vol.42,No.4 July 2012

[4] Kotsiantis, S., D. Kanellopoulos, and P. Pintelas, Handling imbalanced datasets: a review. GESTS International Transactions on Computer Science and Engineering, 2006.Vol 30(No 1): p. 25-36.

[5] Yang, Z., et al., Association rule mining-based dissolved gas analysis for fault diagnosis of power transformers. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 2009. 39(6): p. 597-610. https://doi.org/10.1109/TSMCC.2009.2021989.

[6] Zhu, Z.-B. and Z.-H. Song, Fault diagnosis based on imbalance modified kernel Fisher discriminant analysis. Chemical Engineering Research and Design, 2010. 88(8): p. 936- 951. https://doi.org/10.1016/j.cherd.2010.01.005.

[7] Tavallaee, M., N. Stakhanova, and A.A. Ghorbani, toward credible evaluation of anomaly based intrusion-detection methods. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 2010. 40(5): p. 516-524. Aida Ali et al. 196

[8] Mazurowski, M.A., et al., Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance. Neural networks: the official journal of the International Neural Network Society, 2008. 21(2-3): p. 427-436.

[9] Soler, V., et al. Imbalanced Datasets Classification by Fuzzy Rule Extraction and Genetic Algorithms. in Data Mining Workshops, 2006. ICDM Workshops 2006.Sixth IEEE International Conference on. 2006.

[10] Kubat, M. and S. Matwin. Addressing the curse of imbalanced training sets: one-sided selection in ICML. 1997.

[11] Yi-Hung, L. and C. Yen-Ting. Total margin based adaptive fuzzy support vector machines for multi view face recognition. in Systems, Man and Cybernetics, 2005 IEEE International Conference on. 2005.

[12] Li, Y., G. Sun, and Y. Zhu. Data imbalance problem in text classification in Information Processing (ISIP), 2010 Third International Symposium on. 2010. IEEE.

[13] Al-Shahib, A., R. Breitling, and D. Gilbert, Feature selection and the class imbalance problem in predicting protein function from sequence. Applied Bioinformatics, 2005. 4(3): p. 195-203. https://doi.org/10.2165/00822942-200504030-00004.

[14] Japkowicz, N. in Proc AAAI 2000 Workshop on Learning from Imbalanced Data Sets. 2000. AAAI Tech Report WS-00-05.

[15] Chawla, N.V., N. Japkowicz, and A. Kotcz.inProc ICML 2003 Workshop on Learning from Imbalanced Data Sets. 2003.

[16] LincyMeera Mathews, HariSeetha," On Improving the Classification of Imbalanced Data", CYBERNETICS AND INFORMATION TECHNOLOGIES, Volume 17, No 1,Sofia, 2017, BULGARIAN ACADEMY OF SCIENCES.
[17] Sadam Al-Azani, El-Sayed M. El-Alfy," Using Word Embedding and Ensemble Learning for Highly Imbalanced Data Sentiment Analysis in Short Arabic Text ", Procedia Computer Science 109C (2017) 359–366, the 8th International Conference on Ambient Systems, Networks and Technologies, ANT 2017.
[18] Farrukh Ahmed, Michele Samorani, Colin Bellinger, Osmar R. Za¨ıane," Advantage of Integration in Big Data: Feature Generation in Multi-Relational Databases for Imbalanced Learning",
[19] Jerzy Stefanowski," Dealing with Data Difficulty Factors while Learning from Imbalanced Data", S. Matwin and J. Mielniczuk (eds.), Challengesin Computational Statistics and Data Mining, Springer Studies in Computational Intelligence vol. 605, 2016, pp. 333-363. https://doi.org/10.1007/978-3-319-18781-5_17.
[20] TengNiu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik," Sentiment Analysis on Multi-View Social Data", Q. Tian et al. (Eds.): MMM 2016, Part II, LNCS 9517, pp. 15–27, 2016, https://doi.org/10.1007/978-3-319-27674-8.
[21] VasileiosAthanasiou and ManolisMaragoudakis," A Novel, Gradient Boosting Framework for Sentiment Analysis in Languages where NLP Resources Are Not Plentiful: A Case Study for Modern Greek, Algorithms 2017, 10, 34; https://doi.org/10.3390/a10010034.
[22] Michael Crawford, Taghi M. Khoshgoftaar, Joseph D. Prusa, Aaron N. Richter and Hamzah Al Najada," Survey of review spam detection using machine learning techniques", Crawford et al. Journal of Big Data (2015) 2:23, https://doi.org/10.1186/s40537-015-0029-9.
[23] V. Gopalakrishnan and C. Ramaswamy, Sentiment Learning from Imbalanced Dataset: An Ensemble Based Method, International Journal of Artificial Intelligence, vol. 12, no. 2, pp. 75-87, 2014,CESER Publications
[24] BartoszKrawczyk," Learning from imbalanced data: open challenges and future directions", ProgArtifIntell, DOI 10.1007/s13748-016-0094-0.
[25] Julien Ah-Pine and EdmundoPavel Soriano Morales," A Study of Synthetic Oversampling for Twitter Imbalanced Sentiment Analysis", In: P. Cellier, T. Charnois, A. Hotho, S. Matwin, M.-F. Moens, Y. Toussaint (Eds.): Proceedings of DMNLP, Workshop at ECML/PKDD, Riva del Garda, Italy, 2016.
[26] Troy Raeder, George Forman, and Nitesh V. Chawla," Learning from Imbalanced Data: Evaluation Matters", D.E. Holmes, L.C. Jain (Eds.): Data Mining: Found. &Intell. Paradigms, ISRL 23, pp. 315–331, 2012.
[27] Leo Breiman (2001). Random Forests. Machine Learning. 45(1):5-32. https://doi.org/10.1023/A:1010933404324.
[28] Witten, I.H. and Frank, E. (2005) Data Mining: Practical machine learning tools and techniques.2nd edition Morgan Kaufmann, San Francisco.
[29] J. R Quinlan, (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann, Los Altos.
[30] J. Quinlan. Induction of decision trees, Machine Learning, vol. 1, pp. 81C106, 1986.