Scientific VS Non-Scientific Citation Annotational Complexity Analysis using Machine Learning Classifiers

Hassan Raza[1], M. Faizan[2], Naeem Akhtar[3], Ayesha Abbas[4], Naveed-Ul-Hassan[5]
School of Computer Sciences
National College of Business Administration and Economics
Lahore, Pakistan

***Abstract*—This paper evaluates the citation sentences' annotation complexity of both scientific as well as non-scientific text related articles to find out major complexity reasons by performing sentiment analysis of scientific and non-scientific domain articles using our own developed corpora of these domains separately. For this research, we selected different data sources to prepare our corpora in order to perform sentimental analysis. After that, we have performed a manual annotation procedure to assign polarities using our defined annotation guidelines. We developed a classification system to check the quality of annotation work for both domains. From results, we have found that the scientific domain gave us more accurate results than the non-scientific domain. We have also explored the reasons for less accurate results and concluded that non-scientific text especially linguistics is of complex nature that leads to poor understanding and incorrect annotation.**

***Keywords*—Classification; machine learning; sentimental analysis; scientific citations; non- scientific citation**

I. Introduction

The popular research area in this era is sentiment analysis [14]. Researchers widely used different types of textual data to perform sentiment analysis. Every business and organization need their clients to review for the betterment of their products and services. To analyze the opinion, perception, mindset, and experience of the user is known as sentiment analysis. Judging the sentiments of citing paper' writer about cited paper is termed as sentiment analysis [17]. From the literature work, it has been identified that no work has been done on the problem of evaluating the annotation complexity of both scientific as well as non-scientific text related articles. To perform this work we are needed to prepare experimental data sets of both domains. To prepare scientific corpus we selected Elsevier Computer & Operations Research Journal and prepared a corpus consisted of 5161 citation sentences extracted from 262 research papers published in 2015-2019. On the other hand, we selected SJR Applied Linguistics Journal to prepare a nonscientific corpus consisted of 4989 citation sentences extracted from 250 research papers in 2015-2019. Different machine learning classification algorithms e.g. Naïve-Bayes (NB), Neural Network (NN), Support Vector Machine (SVM), Logistic Regression (LR), Gradient Boosting (GB), Decision Tree (DT), K-Nearest Neighbor (KNN), and Random Forest (RF) are implemented. Using evaluation metrics e.g. f-score, and accuracy score, the system' accuracy is evaluated and improved using different data processing features selection techniques e.g. Lemmatization, NGrams, Tokenization, Case Normalization, and Stop Words Removal.

II. Literature Review

The current state of the domain is analyzed by conducting a literature review in this research work. With the passage of time, researchers' interest has been aroused towards sentiment analysis. The major attention of this domain is towards the construction of framework, extraction of features, and determination of polarities.

Mainly supervised and unsupervised learning approaches are used for sentiment analysis [10]. In a supervised learning mechanism, classifiers' training needs annotated data. To prepare annotated data we need some annotation guidelines. Labeled data is beneficial for a supervised learning approach. Classifiers are trained by this labeled data and also testing of classifier's accuracy is performed. Another approach is unsupervised learning, in this approach data doesn't need to be labeled while there is a need for sentiment lexicons and considered as difficult as it needs various types of lexicons for various genres.

Sentiments are often not well expressed in scientific citation [3]. This may be due to the overall strategy of avoiding critique because of the citation's sociological aspect [12]. [25] mentioned that many works of "politeness, nationalism, or piety" are cited. Negative feelings, still available as well as observable to humans, are articulated in intricate positions and maybe suppressed, particularly when they cannot be explained quantitatively [9]. In scientific literature, citation sentences are often neutral in terms of opinion, either because they critically define an algorithm, strategy or technique, or because they favor a fact or argument [3]. [13] have worked on Sentiment Analysis of Roman Urdu. Most of the research works have been done on different subjects like "English", and "Chinese" etc. No work has been done on sentiment analysis of non-scientific literature because non-scientific literature is totally different from other literatures. Non-scientific citations are very difficult to understand for a non-linguistic person because most of the unfamiliar words are used. So we decided to go for the evaluation of both scientific as well as non-scientific articles' citation sentences' annotation complexity.

III. Methodology

Fig. 1 highlights the purposed methodology adopted in this research work. First of all, in order to analyze the citation sentences' annotation complexity of both scientific as well as non-scientific text related articles we prepared our own data sets of separate domains mentioned in section IV. As we are following the supervised learning approach so there is a need for labeled data sets. For labeling the data we developed some annotation guidelines mentioned in section 4(B) and performed the annotation procedure with the help of human annotators. The annotators classified the citation sentences into 3-classes positive, negative, and neutral. After data is completely labeled, we developed a classification system using python based library named "Sickit-Learn". Test Train Split method is used to divide the data randomly through 60 percent of training data and 40 percent of test data. Experiments are conducted in two phases. In the first phase, we just applied uni-gram, bi-gram, and tri-gram features on data and computed F-scores and Accuracy Scores. Additionally, to boost the quality of the evaluations, we applied different features selection techniques (punctuations and stop words removal, lemmatization, case normalization, etc.) along with n-grams and then computed the above-mentioned metrics again. The later approach helped out in minimizing noise and data complexity. In order to calculate average results, thirty iterations of each experiment were carried out and a total of six experiments were carried out. Finally, we have explored the reasons for less accurate results for non-scientific data classification and concluded that non-scientific text especially linguistics is of complex nature that leads to poor understanding and incorrect annotation process.

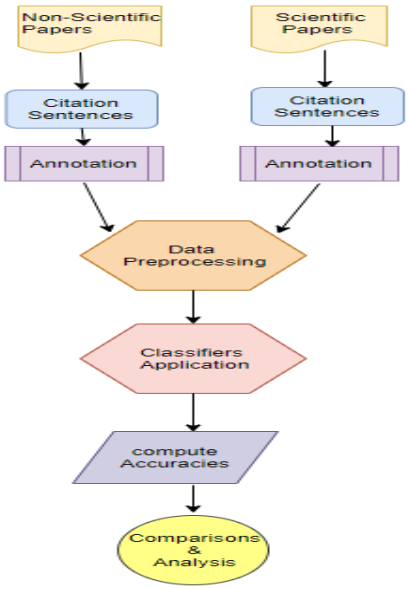

Fig. 1. Step by Step Process Working Stream.

IV. Corpus Construction

To evaluate the citation sentences' annotation complexity of both scientific as well as non-scientific text related articles we need corpora of these domains. We developed two different corpora. To prepare the scientific citations' corpus we choose a science-related journal named "Elsevier Computers and Operations Research Journal" and developed a data set consisted of 5161 citing sentences retrieved from 262 research articles published in 2015 – 2019. On the other hand to prepare non-scientific citations' corpus we specifically choose a non-scientific domain-related journal named "SJR Applied Linguistics" and extracted 4989 citation sentences from 250 linguistics research papers published in 2015-2019.

A. Citation Sentiment Annotations

After preparing the data set the next step was to label the data using a data annotation procedure. We executed this process by applying our own defined guidelines. Citation sentences are categorized by three separate positive, negative, and neutral classes. Annotation guidelines used are as follows:

B. Annotation Guidelines

We have developed some annotation rules according to different scenarios and categorize them as follows.

a) Positive: All those citation sentences which based on words that express *attitude of writers* contains the feeling of "compatibility", "appreciation", "positivity", "excellence", "interest", "admiration", "proposed", "introduced", "analysis", "refers", "thankful" regarding cited paper will be annotated as ''positive". Citations that contain *positive terms* like "outperformed", "accurate", "better", "fast", "favorable", "high quality", and "excellent" etc. Citation sentences that just contain positive terms except the negation terms that reverse the meaning of a sentence like "no", "not", "never", "neither", "nor", and "none" etc.

b) Negative: All those citations sentences based on words that express the *attitude of writers* contain the feeling of "negativity", "doubt", "ambiguity", "criticism", "un-clarity", "degrade" regarding cited paper will be annotated as ''negative". Citation sentences based on *negative terms* like "burden", "complicated", "inability", "lack", "poor", "unclear", and "unexplored" etc. Citation sentences just contain negative terms except for negation terms that reverse the meaning of a sentence like "no", "not", "never", "neither", "nor", and "none" etc.

c) Neutral: All sentences that not contain any positive word and negative words considered as neutrals like "This work was done and evaluated".

C. Statistics of Annotated Corpus

Scientific citation' corpus consists of 5161 and non-scientific citation consists of 4989 sentences. These data sets were annotated using the own defined categories mentioned in Section 4(B). Here are the statistics of the annotated scientific and non-scientific citation sentences' corpus in Table I and Table II.

TABLE. I. Scientific Citations' Statistics

Polarities	Notations	Total Count	Percentage
Positive	P	2014	39.02%
Negative	N	272	5.27%
Neutral	O	2875	55.71%
Total		**5161**	**100**

TABLE. II. NON-SCIENTIFIC CITATIONS' STATISTICS

Polarities	Notations	Total Count	Percentage
Positive	P	2616	52.4
Negative	N	201	4.0
Neutral	O	2172	43.6
Total		**4989**	**100**

V. CLASSIFICATION PROCESSING

This section briefly explains the process of classification applied in this research work. This process consists of various sub-processes including pre-processing data, features' application, classifiers' application, and evaluation metrics.

A. Pre-Processing Data

Data preprocessing is a technique of data mining involving the transformation of raw data into a concise format. Real-world data is often incomplete, contradictory, and lacking in certain habits or patterns, and is likely to contain several mistakes. Preprocessing data is a proven way to solve these problems. Preprocessing the data allows raw data to be processed further. Citations sentences are annotated using 3-classes (Target attributes). Whole data was split into training and testing data using 60:40 ratio randomly.

B. Features' Application

We implemented various features for data classification including N-Grams [16][17], Stop Words Removal [17], Lemmatization [17], Tokenization [17], and Case Normalization to clean down the data.

C. Classifiers' Application

In order to perform the classification procedure we have used different classification algorithms including NB[15][17],SVM[8][17][18][21],DT[2][17],RF[7][8][11][17][18], KNN[17][20][22], LR[5][17][19][24], GB[4][6][23], and NN[1].

D. Evaluation Metrics

To determine the accuracy of a classification we have preferred to use Accuracy score [17], and F-Score [17] evaluation metrics.

VI. RESULTS

Table III represents the evaluation scores of both scientific and non-scientific data sets' classification using the F-Score and Accuracy score. In the case of scientific citation' data set SVM using Uni-gram achieved highest F-score of 70.6% and Accuracy score of 70.6% as well. While in the case of non-Scientific citation' data set LR using tri-gram feature achieved the highest F-score of 65.3% and Accuracy score of 65.3% as well. The reasons for low evaluation scores in case of non-scientific data set is its complex annotation procedure. As human annotators faced much difficulty and complexity while annotating the non-scientific citation sentences due to its complex nature that leads to poor understanding and incorrect annotation. The major reasons we have found because of achieving low accuracy scores in case of non-scientific data set are language differences as linguistic research papers are related to different languages e.g; English, Dutch, French, and Chinese that leads to difficult understanding. Appearing complex terms inside citation sentences is another reason that is responsible for the poor annotation process. Most of the terms that authors found during the annotation procedure were unfamiliar, having different meanings as considered normally. Most of the citation sentences in which the writer's view was difficult to judge that leads to neutral sentiment. Lengthy citation sentences with complex orientation of terms also lead to difficult understanding and annotation process. These are the reasons that leads to complex annotation process and less accuracy scores of non-scientific citation' as compared to scientific citation'.

TABLE. III. HIGHEST SCORES AFTER THIRTY ITERATIONS

Data Set	N-Gram	Classifier	F-Score	Accuracy Score
Scientific	Uni-Gram	SVM	70.6%	70.6%
Non-Scientific	Tri-Gram	LR	65.3%	65.3%

VII. CONCLUSION

In this section, we conclude our work done. We have evaluated the citation sentences' annotation complexity of both scientific as well as non-scientific text related articles to find out major complexity reasons by performing sentiment analysis of scientific and non-scientific domain articles by using our own developed corpora of these domains separately. We prepared a science-related data set consisted of 5,161 citation sentences, we also prepared a non-scientific dataset consist of 4,989 citation sentences and applied polarities using our some rules that are mentioned above. We classified these data sets using different classifiers by applying different features. With the evaluation results, we reached a conclusion that in case of scientific data highest f-score of 70.6% and accuracy score of 70.6% using uni-gram feature is achieved while in case of non-scientific data set highest f-score of 65.3% and accuracy score of 65.3% using the tri-gram feature is achieved. We have concluded major reasons of low accuracy scores in case of non-scientific data set are linguistic differences, Complex words, unfamiliar terms, the neutrality of author's sentiment, and lengthy citations sentences with complex orientation of terms. These are the reasons that lead to difficult and complex annotation process leads to less accuracy scores as compared to scientific citation'.

ACKNOWLEDGMENT

The writers would like to begin by thanking Almighty Allah for His grace and blessings. The writers would also like to thank their parents, friends, and colleagues for their encouragement and support.

REFERENCES

[1] Acharya, U. R., Bhat, P. S., Iyengar, S. S., Rao, A., & Dua, S. (2003). Classification of heart rate data using artificial neural network and fuzzy equivalence relation. Pattern recognition, 36(1), 61-68.

[2] Al-Barrak, M. A., & Al-Razgan, M. (2016). Predicting students final GPA using decision trees: a case study. International Journal of Information and Education Technology, 6(7), 528.

[3] Athar, A. (2014). Sentiment analysis of scientific citations (No. UCAM-CL-TR-856). University of Cambridge, Computer Laboratory.

[4] Babajide Mustapha, I., & Saeed, F. (2016). Bioactive molecule prediction using extreme gradient boosting. Molecules, 21(8), 983.

[5] Bai, S. B., Wang, J., Lü, G. N., Zhou, P. G., Hou, S. S., & Xu, S. N. (2010). GIS-based logistic regression for landslide susceptibility mapping of the Zhongxian segment in the Three Gorges area, China. Geomorphology, 115(1-2), 23-31.

[6] Ganjisaffar, Y., Caruana, R., & Lopes, C. V. (2011, July). Bagging gradient-boosted trees for high precision, low variance ranking models. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (pp. 85-94). ACM.

[7] Gao, D., Zhang, Y. X., & Zhao, Y. H. (2009). Random forest algorithm for classification of multiwavelength data. Research in Astronomy and Astrophysics, 9(2), 220.

[8] Hasan, M. A. M., Nasser, M., Pal, B., & Ahmad, S. (2014). Support vector machine and random forest modeling for intrusion detection system (IDS). Journal of Intelligent Learning Systems and Applications, 6(01), 45.

[9] Hyland, K. (1995). The Author in the Text: Hedging Scientific Writing. Hong Kong papers in linguistics and language teaching, 18, 33-42.

[10] In European conference on machine learning (pp. 4-15). Springer, Berlin, Heidelberg.

[11] Lin, W., Wu, Z., Lin, L., Wen, A., & Li, J. (2017). An ensemble random forest algorithm for insurance big data analysis. Ieee Access, 5, 16568-16575.

[12] MacRoberts, M. H., & MacRoberts, B. R. (1984). The negational reference: Or the art of dissembling. Social Studies of Science, 14(1), 91-94.

[13] Mehmood, K., Essam, D., & Shafi, K. (2018, July). Sentiment Analysis System for Roman Urdu. In Science and Information Conference (pp. 29-42). Springer, Cham.

[14] Moravcsik, M. J., & Murugesan, P. (1988). Some Results on the Function and Quality of Citations: Social Studies of Science. 研究 技術計画, 3(4), 538.

[15] Mukherjee, S., & Sharma, N. (2012). Intrusion detectionusing naive Bayes classifier with feature reduction. Procedia Technology, 4, 119-128.

[16] Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up? sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10(pp. 79-86). Association for Computational Linguistics.

[17] Raza, H., Faizan, M., Hamza, A., Mushtaq, A., & Akhtar, N. (2019). Scientific Text Sentiment Analysis using Machine Learning Techniques: International Journal of Advanced Computer Science and Applications(IJACSA), 10(12), 2019.

[18] Selvaraj, H., Selvi, S. T., Selvathi, D., & Gewali, L. (2007). Brain MRI slices classification using least squares support vector machine. International Journal of Intelligent Computing in Medical Sciences & Image Processing, 1(1), 21-33.

[19] Tsangaratos, P., & Ilia, I. (2016). Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size. Catena, 145, 164-179.

[20] Wang, J. S., Lin, C. W., & Yang, Y. T. C. (2013). A k-nearest-neighbor classifier with heart rate variability feature-based transformation algorithm for driving stress recognition. Neurocomputing, 116, 136-143.

[21] Widodo, A., & Yang, B. S. (2007). Support vector machine in machine condition monitoring and fault diagnosis. Mechanical systems and signal processing, 21(6), 2560-2574.

[22] Yu, X. G., & Yu, X. P. (2006, August). The Research on an adaptive k-nearest neighbors classifier. In 2006 International Conference on Machine Learning and Cybernetics (pp. 1241-1246). IEEE.

[23] Zhang, F., Du, B., & Zhang, L. (2015). Scene classification via a gradient boosting random convolutional network framework. IEEE Transactions on Geoscience and Remote Sensing, 54(3), 1793-1802.

[24] Zhu, J., & Hastie, T. (2002). Kernel logistic regression and the import vector machine. In Advances in neural information processing systems (pp. 1081-1088).

[25] Ziman, J. M. (1968). Public knowledge: An essay concerning the social dimension of science (Vol. 519). CUP Archive.