Text Classification/regression – peer reviews

Ikechukwu Godwin Amasiatu
Applied Machine Learning
ID:00118304

***Abstract*—Text classification is widely used and is regarded as an important way to manage and process many documents in digital format that are constantly increasing. This report shows the ML text classification based on a paper peer review from the International Conference on Learning Expressions (between 2017 and 2020). In reviewing, two analytical algorithm was used to experiment or analyse the dataset: support vector machine (SVM) and logical regression. Machine learning models is categorised into supervised, unsupervised and semi-supervised learning. This aspect is a supervised learning model where classification and regression are used in this analysis. From the experiment, It was discovered that the classification rate had a poor accuracy. That indicates there was an in imbalance on the system. I chose random-under sampling was used to correct this and get this to near accuracy.**

I. Introduction and related work

Recent advance in Machine Learning, Deep Learning with the help of Neural Networks and easy to use models in python has opened the doors into data insight, making computers understand the complex human Language. This Report deals with text classification a subset of machine learning which is focused on natural language processing.

Text Classification consists of providing input to a text document to a set of pre-defined classes, using a machine learning technique. The classification is normally carried out on the basis of selected documents and features using text documents. Text Classification do not depend on rules that have been manually established. It learns to classify text based on previous observations, typically using training data for pre-labeled examples. Text Classification algorithms can discover the many correlations between distinct parts of the text and the predicted output for a given text or input. In highly complicated tasks, the results are more accurate than human rules, and algorithms can incrementally learn from new data

However, the classes are selected before the experimental analysis, which is called supervised machine learning operation. This is used to attribute a label or a probable value to an instance. In some cases, there can be variations in text classifications which can allow multiple assignment of labels. The supervised classification works on training and testing principles of the algorithm. The labeled dataset when trained is fed to the algorithms to work on and gives the pre-defined categories which is the output. At the testing phase, the algorithm is assigned to unobserved data and its categorized based on the testing phase. For the purpose of this task, is to gain insight, analyze and predict the accuracy of the datasets about the paper peer review from the international conference on learning expressions and also annotating and classifying this user -generated data set. I explored the text characteristics (features) that are potentially useful in distinguishing between acceptance status categories while predicting and reviewing the scores by applying the two machine-learning algorithms using supervised learning.

One example of text classification supervised learning is email spam screening. The content of each incoming email is automatically classified. supervised systems are used for language detection, intent, emotion, and sentiment analysis. It may be used for a variety of purposes, such as detecting emergency situations by analyzing millions of Pieces of web data. This is difficult task when trying to sort manually. To detect such scenario in a case of a designed smart public transit system. The classifier was trained with high accuracy to recognize emergency situations from millions of internet conversations. In solving this, it requires unique loss functions, sampling during training, and strategies such as building a stack of many classifiers, each refining the findings of the preceding one. Other similar projects include the Amazon text reviews, the Armed Conflict Location and Social Conflict Analysis Database Review.

II. Ethical discussion

Data ethics is the moral problems related to data (including generation, collection, processing, storage, sharing, use) and algorithm. Whilst there are no existing global standards, there are a number of recurrent themes. The biggest ones are: transparency, equality and data control

TRANSPARENCY: Individually, data processing processes and automated choices must make sense. They must be completely transparent and comprehensible. Individuals must have a clear awareness of the aim and interests of data processing in order to recognise dangers, as well as social, ethical, and societal repercussions. DATA CONTROL: Individuals should be in control of their data and have the power to use it to their advantage. People's autonomy should be a priority in data processes, and they should be involved in decisions about data about them. The individual is in control of how their data is used, the context in which it is processed, and how it is activated. EQUALITY:Democratic data processing is based on an awareness of the societal power relations that data systems sustain, reproduce or create. When processing data, special attention should be paid to vulnerable people, who are are particularly vulnerable to profiling that may adversely affect their self-determination and control or expose them to discrimination or stigmatisation, for example due to their financial, social or health related conditions. Paying attention to vulnerable people also involves working actively to reduce bias in the development of self-learning algorithms

III. Dataset preparation

Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. This is an important step in any data transformation process. Its basically involves transforming raw data into an understandable format for NLP models. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. This was initiated by importing pandas framework and other necessary libraries for exploratory processes in order to visualize the data. There was some missing data and as well empty columns. In cleaning the data, The missing values and empty rows were taken out and also duplicates were checked and removed to avoid overfitting of the dataset. Other data cleaning steps were applied ,which include removal of Subwords like pronouns was also expunged and special character like commas were removed. Tokenization: this separates the texts into smaller chunks called tokens. This was done with the TFIDF vectorizer with sklearn built in tokenizer. It further discarded tokens that appeared in more than half of the document and standardized the data for the next phase of preparing the train and test data set. At this next step, the data split into two data sets, Training and Test. The training data was now used to fit the model while the predictions will be performed on the test data set. This is done through the train test-split from the sklearn library. The Training Data was assigned 70 percent while the Test data had the remaining thirty percent for test size parameters. The main objective is for the trained model to transform unseen text into a vector, extract its relevant features, and make a prediction. As the texts is been transformed into vectors, they are fed into a machine learning algorithm together with their expected output to create a classification model that can choose what features best represent the texts and make predictions

IV. Methods

There are many machine learning algorithms used in text classification. The mostly used are the linear regression, Support Vector Machines (SVM), and Logistic regression. For this classification, logistic regression and the SVM algorithms was implemented. Logistic regression is a supervised learning method usually employed to handle binary classification tasks. In essence, it takes any real-valued integer and translates it to a value between 0 and 1. It is used for predicting the categorical dependent variable using a given set of independent variables. It is mostly suitable where the dependent variables or the target group is a categorical variable with two outcomes. Strengths of model is that Logistic regression can be updated after is has already been trained, meaning that new reviews can be used to teach the algorithm after it has already been trained . to this end ,this is a Justification in its ability to update itself after its initial training period . So this Classifier is best suited for determining the outcome of the model. A weighted average of about 73 percent was arrived at. Support Vector Machine (SVM) is a supervised machine learning algorithm which suitable for both classification or regression challenges. The model extracts a best possible hyper-plane/line that determine the two classes .The algorithm determines the optimal limit between a vector belonging to a specific category and a vector not belonging to it. The strength lies in when there is a clear margin of separation between classes as in this case. In a nutshell, the support vector machine classifier was chosen basically, after transforming the text it determines the process to separate the data at the output .

V. Experiments and evaluation

From the summary results from the classifiers, it can deduced that the logistic regression classifier has the better result of about 74-percent though not outstanding but performs better than the support vector machine algorithm . Using the logistic regression, the data was divided into the training set and the testing set. This was done in a way that the training set is bigger that the testing set. With this, I f the model on the training text and then evaluated the performance by using the sklearn metrics that is the classification report over accuracy score to check for the performance of the model. However, the SVM algorithm output show that there is an imbalance in the system. For inbalance systems , accuracy might not be truly used to get the best prediction. However, further evaluation on the system on classification report can say the same for the precision, recall, f1-score having an overall weighted average of about 74-percent which output on the Logistic regression as our best model. This implies that 26-percent offset of accuracy or loss of accuracy which might not be comfortable to the user.

VI. Discussion and future work

The system was best determined by the outcome of then logistic regression model which predicted a near outcome with a accuracy of a weighted average score of 73 percent which is almost efficient and can be valid for the predictions. Whereas, the support vector machine model had a low average, reason because the low data was not efficient for the model. That show there was an imbalance. However, this can be resolved by using any of the over sampling or under sampling method to get a better robust system but not so reliable . In the future and in order to improve on our accuracy ; -a larger set will be used to test the model to get a better prediction, -Perform more future engineering and explore more additional pre-processing on the dataset .

VII. Conclusions

The task for the implementation of then of the text classification of a paper review from the international conference of the learning representation in the years between 2017 and 2020. Th major aim was to to find the best model on predicting the accuracy of the review status. The experiment shows the weekness of the algorithm when applied to a highly inbalanced dataset. Our experiments have shown that more work is required to find a universal approach to solving the imbalanced distribution problem in this data. However, we did train a logistic classifier that achieved a high accuracy. I also discovered that that large dataset is all but required to achieve a designed result

.

References

[1] Argamon, S., Saric, M., and Stein, S. (2003). Learning Algorithms and Features for Multiple Authorship discrimination. Proceedings of IJCAI '03 Workshop on Computational Approaches to Style Analysis and Synthesis. Acapulco, Mexico, 10 August 2003.

[2] Craig, H. (1999). Authorial attribution and computational stylistics: If you can tell authors apart, have you learned anything bout them? Literary and Linguistic Computing, 14(1): 103–13.

[3] Forman, G. (2003). An extensive empirical study of feature selection metrics for text categorization. Journal of Machine Learning Research, 3: 1289–305.

[4] Liaw A. Weiner M. Classification and Regression by randomForest. R News. 2002;Vol 2(2):18- 22

[5] N. Cherniavsky, I. Laptev, J. Sivic, A. Zisserman Semi-supervised learning of facial attributes in video Proceedings of the European Conference on Computer Vision (2010), pp. 43-56

[6] N. Kumar, A.C. Berg, P.N. Belhumeur, S.K. Nayar,(2011). Describable visual attributes for face verification and image search IEEE Trans. Pattern Anal. Mach. Intell.

[7] Ni Zhuang, Yan Yan, Si Chen, Hanzi Wang, Chunhua Shen, Multilabel learning based deep transfer neural network for facial attribute classification, Pattern Recognition, Volume 80,2018. Pages 225- 240, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2018.03.018

[8] P. Luo, X. Wang, X. Tang A deep sum-product architecture for robust facial attributes analysis Proceedings of the IEEE International Conference on Computer Vision (2013), pp. 2864-2871

[9] Taigman, Y.; Yang, M.; Ranzato, M.; and Wolf, L. 2014. Deepface: Closing the gap to human-level performance in face verification. CVPR

[10] S. Kang, D. Lee, C.D. Yoo Face attribute classification using attributeaware correlation map and gated convolutional neural networks. Proceedings of the IEEE International Conference on Image Processing (2015), pp. 4922- 4926.

[11] Wang, G., and Forsyth, D. 2009. Joint learning of visual attributes, object classes and visual saliency. CVPR