Augmented Machine Learning Ensemble Extension Model for Social Media Health Trends Predictions

Sonia Saini, S. P. Singh, Ruchi Agarwal

Abstract*:*Social Networks are the source of rich, interactive, textual, and other media.Users of the social media generate data at a tremendous pace.This data consisting of user opinions and attitudes is so large that it has necessitated automated methods to analyze and extract knowledge from the same. Social networks have been studied and analyzed using various graph-based analysis techniques. Prominent analysishas centered on features like ego-networks, distance, centrality, sub-networks etc. The areas of study for social media analysis have been centered around populations, boundaries, Cohesion, Centrality and Brokerage, Prestige and Ranking. In the past several models have been propounded for various machine learning based analytics for the Social Networks study but there is a perceived need for studying social networks for health data using Ensemble Learning wherein an array of various Machine Learning techniques can be employed to achieve better classification or clustering results. We introduce an Analytical Model which will identify most discussed terms/ topics of health/ healthcare on social networks to predict the emerging health trends. The model is to use temporal datasets to deduce multi-label classification of health-related topics. The Model employs the technique of Temporal Clustering (using Machine Learning) on the Topic Classification done on datasets using Ensemble Machine Learning to deduce the most discussed topics. Using this model, we will see how Ensemble Machine Learning based Analytical Model for analyzing social network data for health topics is efficient than traditional Machine Learning technique(s).

Index Terms: Augmentation Analytical Model, Ensemble Learning, Machine Learning, Social Media Data

I. INTRODUCTION

Social Networks present myriad complexity with respect to performing precise analytics. The varying and diffused data, when considered with a temporal perspective, presents challenge with respect to predictive analysis. To consider the problem scope of analyzing what health topics are most discussed in different chronological eras, we need more than just performing simple topic classification using machine learning. The Social Network Analysis space is dominated by perspectives such as the socio-centered perspective as well as the ego-centered perspective. While the socio-centered perspective analyses complete network structure by looking for patterns of ties thus identifying cohesive social groups, central actors which could be of importance to the integration of the social network, and asymmetries that may reflect social prestige or social stratification. Recent advances are found primarily in the technique of blockmodeling. The ego-centered perspective works by analysis on the composition of local network structure. It studies the problem space that whether actors influence one another through their network ties, also known as social influence model, and/or actors adjust their ties to the characteristics of their peers and to their ties with them, also known as social selection model.

II. THE NEED FOR ENSEMBLE MACHINE LEARNING IN SOCIAL NETWORK ANALYSIS

A. *Ensemble Machine Learning*

Ensemble Machine Learning is a relatively new concept of combining various machine learning techniques to do several fold iterations over datasets to achieve more efficiency in the said machine learning task of text classification, image classification, clustering or deep learning. Some of the applications of ensemble learning are classification, clustering, collaborative filtering, and anomaly detection. If a one-pass anomaly detection rate is80 percent, then the goal of using ensemble learning is to do a 10-fold iteration over the dataset to improve the anomaly detection up to 90 percent or even more. Ensemble machine learning can also be used in predictive analytics, wherein the predictive accuracy is substantially improved when blending multiple predictors.

B. *Need for Ensemble Machine Learning*

Some of the reasons for using Ensemble Machine Learning are related to dataset characteristics and training confidence. Some of the reasons are as follows.

The dataset is too large or too small. For a large dataset, it clearly cannot be trained by a single model and we need to create several small subsets to train different models and use average of all as final prediction. For small datasets, which cannot be used to train a single model, bootstrap methods can be used to create random subsamples of data.

Revised Manuscript Received on July 5, 2019

Sonia Saini, Department of Computer Science & Engineering, BIT Mesra (Noida Extension center)Noida ,India,

S. P. Singh, Department of Computer Science & Engineering, BIT Mesra (Noida Extension center)Noida ,India

Ruchi Agarwal, BCA Department ,JIMS Engineering Management Technical Campus, Greater Noida, India

Complex (nonlinear) data, when a single model does not define the class boundary clearly resulting in underfitting of the model and we can use ensemble learning to predict distinct boundaries. High confidence, training on multiple classifiers to get correlated output does ensure a higher prediction rate.

Finally, Ensemble Machine Learning is essentially used to improve the algorithm accuracy and the robustness of the model. Ensemble methods vie to achieve this by trying to manage the trade-off analysis between bias and variance errors with respect to any model's complexity and errors produced thereof [10].

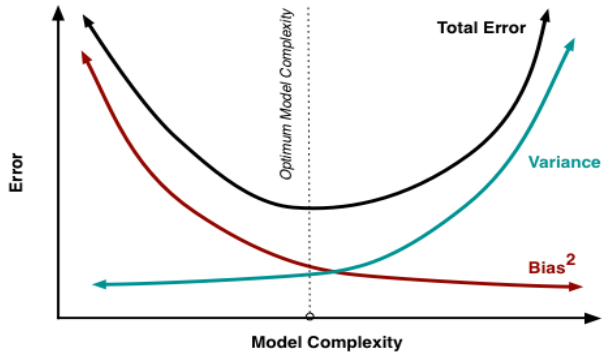

Figure 1: The Bias and Variance Tradeoff and optimum model complexity

In the above Figure 1, the optimum model complexity can be arrived at intersection of bias and variance, both of which need to be reached as low as possible, and thus the overall error is also at the lowest implying a better classification or ensemble result accuracy.

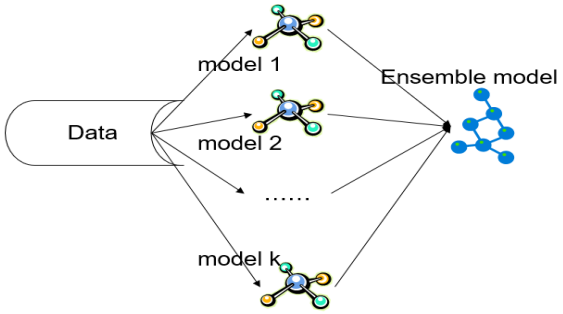

Figure 2: An Ensemble Model

Figure 2 illustrating an ensemble model, comprising of an assemblage of two or more machine learning models. This ensemble model can be classification ensemble or clustering ensemble.

III. RELATED WORK

More recently, the problem of multi-label classification has become a keen topic of research. In this kind of classification, each sample is associated with a set of class labels. Ensemble machine learning comprises of taking one or more supervised learning algorithms which are like base-level algorithms and aggregate the outcomes of these base-level algorithms to make a more informed estimation. Basic ensemble would take algorithms and apply them to random subsets of input corpus and do a vote for choosing the best fit or average out the results from various base-level algorithms.There have been efforts of using supervised as well as unsupervised learning methods for tweet classification of target audience while using minimal annotation efforts [9]. In this research there was an automatic discovery of Topic domains from content shared by Twitter account owners and Latent Dirichlet Allocation (LDA) technique was used. A Support Vector Machine (SVM) ensemble was then trained using contents from different account owners of the various topic domains identified by Twitter LDA. Results of these methods have demonstrated identification with high accuracy.

In an earlier work an Importance Sampled Learning Ensemble approach was espoused wherein prediction risk and large value coefficient penalty was considered for base learners in machine learning ensemble [7].

In the area of Multi-Label classification, earlier work done involves using Random K-Label sets for common multi-labels across various domains demonstrated better accuracy when compared with other multi-label classification approaches [8].

One recent research demonstrated combining classification and clustering algorithms for Tweet Sentiment Analysis using an SVM classifier along with a clustering ensemble to achieve better classification accuracies by espousing a C3E-SL algorithm which refines tweet classifications from additional information provided by clusterers [12].

Data Augmentation can be done prior to putting the corpus through machine learning models. Earlier works on data augmentation to reduce class imbalance have met with fair success. An approach of augmentation of text data done at paragraph level achieved a 10 percent classification accuracy wherein the paragraph order switch does not change the semantic meaning of the text. [14].

A. *Ensemble methods for multi-label classification*

Ensemble methods have been shown to be an effective tool for solving multi-label classification tasks. Ensemble learning techniques have already been observed as effective in supervised learning [1], [2],[3] and unsupervised learning [4],[5],[6] machine learning tasks. Among various methods used for ensemble learning, two prominent methods used in ensemble learning are Bagging and Boosting along with bootstrapping as a pre-cursor step which may be required based on the dataset. There is also the stacked generalization technique. Bagging, also known as Bootstrap aggregation, is a three stage process wherein we bootstrap our data, aggregate (model fit) and finally combine predictions from different models.

Most of the times data presents imbalanced distribution of classes and thus poses obstacle for supervised as well as semi-supervised class learning. Research has also focused on use of ensembles of self-trained classifier to address imbalance of data [13].

IV. PROBLEM FORMULATION

The problem under study is the analysis of social media data to arrive at any emerging trends of health topics. By proposal of this model, we try to understand the efficiency of such an ensemble machine learning model. The social network data is a rich media amalgamation that has differences in population and temporal data sets from a social network such as Twitter serve as a conducive platform for employing ensemble learning in the machine learning techniques.

The quantification of performance of the classifier can be ascribed to achieving a balance between bias and variance with a small sample size being the source of variance, whereas a large sample size may still produce inaccurate results, but variance of predictions would be dramatically reduced.

The tradeoff is well represented by the following mathematical formula:

$$Err(x) = \left(f(x) - \frac{1}{k}\sum_{i=1}^{k} f(x_i)\right)^2 + \frac{\sigma_\epsilon^2}{k} + \sigma_\epsilon^2 \quad (1)$$

$$Err(x) = (Bias)^2 + \text{variance} + \text{irreducible error}$$

By reducing bias and variance, we can achieve more accuracy in the predictions.

For disparate temporal data from social network, an ensemble of classification models and an ensemble of clustering models can be conjugated for reducing the bias and variance.

We introduce an Augmented Machine Learning Ensemble Extended (AMLEE) Model, the goal of which is to reduce the Bias Variance as well as enhance the prediction accuracy using a composite ensemble of classification and clustering. This model is trained as a final model, which is trained on all available data (both training as well as test data) and then this model is used to make predictions on the new data. Theaugmentation aspect of the model relates to data augmentation, the purpose of which is to reduce overfitting. Data Augmentation works by addressing the class imbalance, by down-sampling the majority class, and up-sampling the minority class. Two popular implementations used for text augmentation are Part-of-Speech (POS)tag replacement and threshold-based replacement. Augmentation can be performed on data such as synonym replacement. While the most common applications of data augmentation can be perceived in the Computer Vision fields but social media with its vastly varying data also poses an interesting proposition for the same.

The following two figures 3 and 4 delineate the difference between traditional machine learning technique and the approach employing the augmented machine learning ensemble. Whereas the conventional machine learning technique does not readily account for class imbalance, the augmented machine learning model takes care of the same. The resultant linearly balanced corpus is further passed through an ensemble model of classification techniques.

Figure 3: Conventional machine learning approach for text classification.

EM_1 and EM_2 in the Figure 4 below are two distinct ensemble machine learning models for multi-label text classification. CEM_1, and CEM_2 are ensemble models for clustering which will cluster the topicwise multi-label classification by taking into account the time of the tweets.

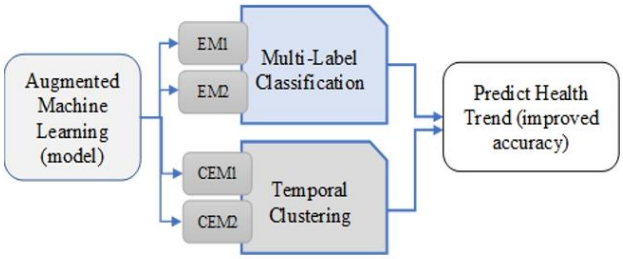

Figure 4: Illustration of Augmented Machine Learning Ensemble Extended (AMLEE) model.

Figure 4 illustrates how the Augmented Machine Learning Ensemble Extended (AMLEE) model can utilize both an Augmentation as well as an Ensemble approach to improve prediction accuracy.

V. THE PROPOSED METHOD

In this study, an ensemble learning method is proposed for improving multi-label classification evaluation criteria. We have compared our method with well-known base-level algorithms on some data sets. Experiment results show the proposed approach outperforms the entry level well-known classifiers when done for multi-label classification problem [11].

A. *Task for the proposed analytical model*

The Ensemble Machine Learning model must do machine learning tasks such as Multi-Label Topic classification, and Clustering. Given the Ti = 1 to ntemporal datasets, analyze the datasets, and classify the various labels per document the various labels further topic classification.

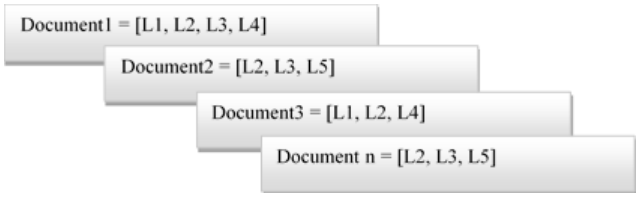

Figure 5: A Temporal dataset comprising of distinct labels

Figure 5 shows temporal dataset of n documents, with each having x distinct Labels(L) which the classification machine learning algorithm will classify.

A label $L_i$is a health term label such as Diabetes, Stroke, Alzheimer etc. For example L1 can be Diabetes,L2 can be stroke and L3 can be Alzheimer etc.

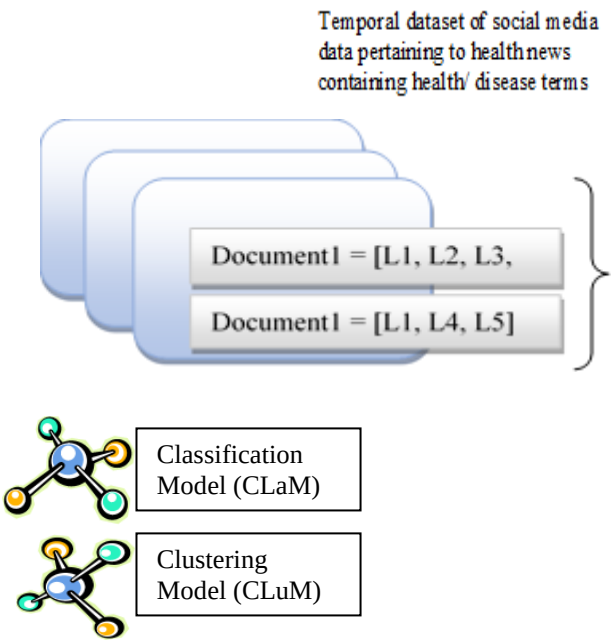

Figure 6: Classification and clustering model Ensemble operating on document corpus.

Above Figure 6 depicts multiple temporal datasets, on which multi-label classification is run and after that clustering of the labels is done to identify the emerging most mentioned or most discussed health terms and trends in social networks.

The classifier in the ensemble model achieves the task of classifying tweet terms into some pre-defined health related classes.

VI. EXPERIMENT AND RESULT

The experiment undertaken used the Augmented data and performed multi-label classification task, wherein number of tweets belonging to one or more categories were identified. A corpus of 42410 tweets pertaining to news about health, and lifestyle were collected from the Twitter micro-blogging site. Data cleaning methods were employed to remove the text corpus of noise such as acronyms, emoticons, stop words, and other such unwanted data. The experiment was run using Python 3.7 and the sci-kit learn Python library.

The temporal clustering yielded 15097 tweets for year 2013, 19296 tweets for year 2014, and 8017 tweets for year 2015.

Further, using the multi-label classification, an ensemble of Naïve Bayes and Support Vector Classifier (SVC), a higher accuracy was observed. The results in the figures below demonstrate that while a large number of the corpus indicates one category per tweet, the count of 1, 2, 3 or even 4 categories per tweet was also observed. Below results show the temporal (year wise) category distribution of the tweets with respect to the disease terms, and count of tweets wherein multiple disease terms occur. Since the main topic of the corpus was "health" so health appears prominently, with Cancer and Diabetes also showing relative prominence as per the indicated result.

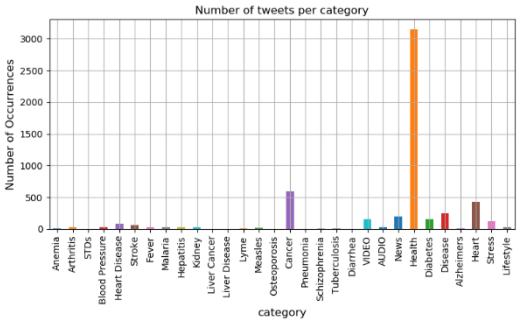

Figure 7: Tweets category-wise classification for year 2013

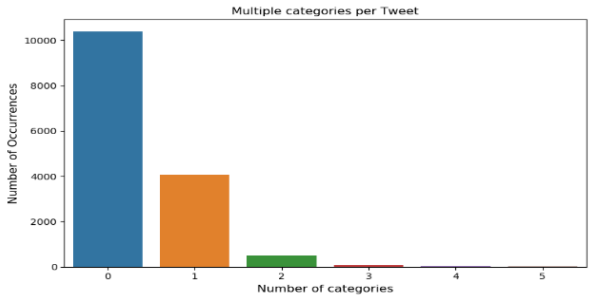

Figure 8: Multi-category tweets count for year 2013

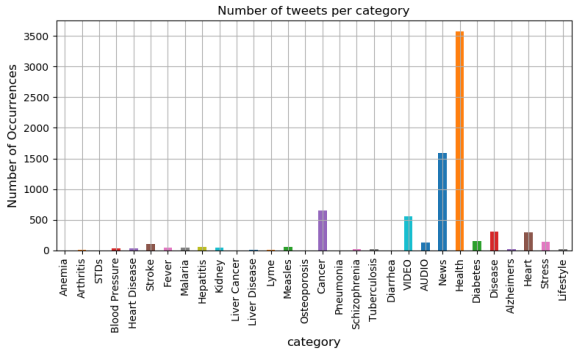

Figure 9: Tweets category-wise classification for year 2014

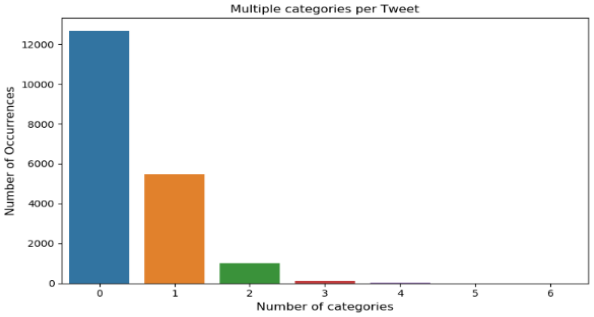

Figure 10: Multi-category tweets count for year 2014

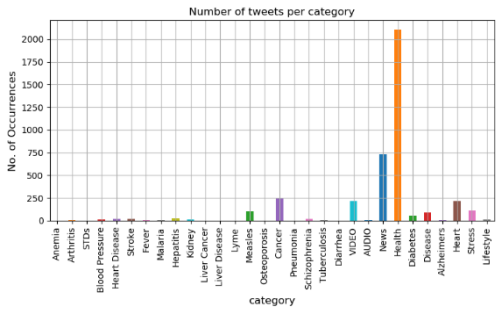

Figure 11: Tweets category-wise classification for year 2015

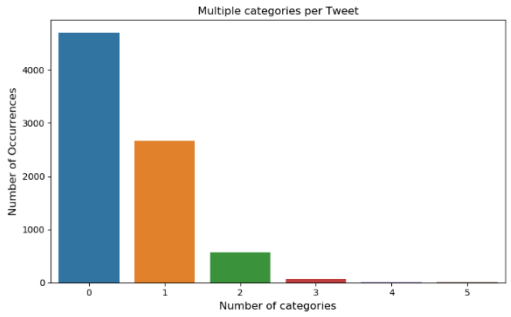

Figure 12: Multi-category tweets count for year 2015

VII. CONCLUSION

In this paper we looked at various methods of ensemble learning and how ensemble learning can be used to improve the accuracy of classification/ clustering tasks which are done on datasets as a part of data analysis. We also propose an analytical model for which works on temporal data to deduce emerging health trends and this model can be used on social network data as well as medical data such from hospital/ clinics etc. This augments traditional approaches such as label classification accuracy using just a single classifier with an improved machine learning ensemble and reduction of class imbalance by employing data augmentation technique like semantic text swap, and word cosine similarity threshold as the hyper parameter. This text augmented data may still produce better accuracy when put through an ensemble of machine learning algorithms for clustering or classification in order to address relevant problem space scenarios. We have provided a baseline of such as composite model for strengthening the accuracy of machine learning algorithms.

REFERENCES

1. Dimitriadou E., Weingessel A., Hornik K,Voting-Merging: An Ensemble Method for Clustering. In: Dorffner G., Bischof H., Hornik K. (eds) Artificial Neural Networks — ICANN 2001. ICANN 2001. Lecture Notes in Computer Science, vol 2130. Springer, Berlin, Heidelberg,2001,ch.31.
2. A. Strehl and J. Ghosh, "Cluster ensembles — a knowledge reuse framework for combining multiple partitions," The Journal of Machine Learning., vol. 3, pp. 583–617, March.,2003. Available: http:/ https://dl.acm.org/citation.cfm?id=944935/
3. X. Z. Fern and C. E. Brodley, "Solving cluster ensemble problems by bipartite graph partitioning." in ICML, C. E. Brodley, Ed., vol. 69. ACM, 2004. [Online]. Available: http://dblp.uni-trier.de/db/conf/icml/icml2004.html#FernB04.
4. L. Breiman, "Bagging predictors," Machine Learning, vol. 24, no. 2, pp. 123–140, 1996.
5. R. E. Schapire, "A brief introduction to boosting," in IJCAI, Stockholm, Sweden, 1999, pp. 1401–1406.
6. T. Chakraborty, D. Chandhok, and V. S. Subrahmanian, "MC3: A multi-class consensus classification framework," in PAKDD, Jeju, South Korea, 2017, pp. 343–355.
7. J. Friedman and B. Popescu, "Predictive learning via rule ensembles," Annals of Applied Statistics, vol. 3, no. 2, pp. 916–954, 2008.
8. Tsoumakas, Grigorios; Vlahavas, Ioannis , "Random k-labelsets: An ensemble method for multilabel classification.",2007
9. Lo SL, Chiong R, Cornforth D ,"Using Support Vector Machine Ensembles for Target Audience Classification on Twitter". PLoS ONE 10(4): e0122855. doi: 10.1371/journal.pone.0122855, 2015
10. Scott Fortmann-Roe.(2012,month).Understanding the Bias-Variance Tradeoff.Available:http://scott.fortmann-roe.com/docs/BiasVariance.html
11. Zheng Fang and Zhongfei (Mark) Zhang. 2012. Simultaneously Combining Multi-view Multi-label Learning with Maximum Margin Classification. In Proceedings of the 2012 IEEE 12th International Conference on Data Mining (ICDM '12). IEEE Computer Society, Washington, DC, USA, 864-869.
12. Coletta, Luiz & Felix, Nadia & Hruschka, Eduardo & Hruschka, Estevam." Combining Classification and Clustering for Tweet Sentiment Analysis" published in Brazilian Conference on Intelligent Systems. DOI: 10.1109/BRACIS.2014.46 .
13. Ana Stanescu and Doina Caragea, Ensemble-based semi-supervised learning approaches for imbalanced splice site datasets 2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM),UKDOI: 10.1109/BIBM.2014.6999196
14. "Shuffling Paragraphs - using data augmentation in NLP to increase accuracy." https://medium.com/bcggamma/shuffling-paragraphs-using-data-augmentation-in-nlp-to-increase-accuracy. [Accessed: 2019-01-29].

AUTHORS PROFILE

Ms. Sonia Saini is pursuing Ph.D. from BIT Mesra Ranchi. She is working as Assistant Professor in Amity Institute of Information Technology, Amity University, Noida.She has done M. Tech (IT) and MPhil (Computer Science). She has more than 15 years of teaching experience. She has published several research papers in various international conferences& journals in India. Her research area includes Data Mining, Big Data Analytics, Healthcare analytics

Dr. S. P. Singh is Ph.D. in Computer Science. Presently he is working as Assistant Professor in the Department of Computer Science & Engineering, BIT Mesra (Noida Extension center). He has more than 16 years of teaching experience. He has published many research papers in National/International Journals of repute. His research area includes Data Mining & Warehousing, Big Data Analysis, and Software Engineering etc.

Dr.Ruchi Agarwalis an Associate Professor and HOD of BCA department of JIMS Engineering Management Technical Campus, Greater Noida. She has done Ph.D. (Computer Science) from Birla Institute of Technology (BIT), Mesra, Ranchi in the field of data mining. She has more than 15 years of teaching experience. She has published variouspapers in international journals and conference proceedings. She has guided various B Tech andM. Tech level projects. She isguiding Ph.D. students in the area of Big Data Analytics.Her research interest areas are Big Data Analytics, Data Mining and Customer Analytics