A Conceptual Data Modelling Framework for Context-Aware Text Classification

Nazia Tazeen[1]*
PhD Scholar, Department of Computer Science and Engineering, SPMVV Tirupati, India

K. Sandhya Rani[2]
Professor, Department of Computer Science
SPMVV Tirupati, India

***Abstract*—Data analytics has an interesting variant that aims to understand an entity's behavior. It is termed as diagnostic analytics, which answers "why type questions". "Why type questions" find their applications in emotion classification, brand analysis, drug review modeling, customer complaints classification etc. Labeled data form the core of any analytics' problem, leave alone diagnostic analytics; however, labeled data is not always available. In some cases, it is required to assign labels to unknown entities and understand its behavior. For such scenarios, the proposed model unites topic modeling and text classification techniques. This combined data model will help to solve diagnostic issues and obtain meaningful insights from data by treating the procedure as a classification problem. The proposed model uses Improved Latent Drichlet Allocation for topic modeling and sentiment analysis to understand an entity's behavior and represent it as an Improved Multinomial Naïve Bayesian data model to achieve automated classification. The model is tested using drug review dataset obtained from UCI repository. The health conditions with their associated drug names were extracted from the reviews and sentiment scores were assigned. The sentiment scores reflected the behavior of various drugs for a particular health condition and classified them according to their quality. The proposed model performance is compared with existing baseline models and it is proved that our model exhibited better than other models.**

Keywords—Text classification; topic modeling; natural language processing; sentiment analysis; drug dataset; context-aware model; diagnostic analytics; feature extraction

I. Introduction

Data analytics is a branch of data mining, that deals with extracting useful information from the data. There are four types of analytics, Predicitve, Prescriptive, Diagnostic and Descriptive analytics. The concern of this work is centered on the concepts related to Diagnostic analytics. It provides answers to why something happened?. Why a person likes a particular product?, why a particular drug is harmful?, why a particular event occurred? are some issues that can be solved using this analytics.The application of such analytics operations in Natural Language datasets pose a greater challenge. There are three reasons for that, availability of labeled data regarding the entity of interest, extracting topics or coherent terms from documents and size of varying topics (saliency). To solve this, a system needs to include contextual information of texts [11]. The context-aware systems are more concerned about, topic relevance, term relevance, and topic labels when compared to bag-of words approach. Such a system capable of revealing systematic use cases of the business problems through conceptual models help to arrive at solutions easily [14]. The proposed conceptual framework is designed in such a way that, the model not only supports Natural Language Understanding(NLU) but also enables Natural Language Processing(NLP) in unstructured data environments. To identify entities of interest, a topic modeler is used and to study its behaviour to predict future scenarios, a machine learning data modeler is used. The combined model is built as a classifier, since almost all analytical problems can be represented as a classification problem and also building a classifier model will make the system generic for all similar use-cases in future. For topic modeling and labeling uknown entities, the famous conventional Latent Drichlet Allocation technique is improved. It is enhanced to handle sparse features, extracts latent semantic relationships from the data, keywords for topics of all sizes and minimize polysemy issues.

Also, to answer the why type questions taken for consideration, the setimental value of text is also measured. Sentiments could represent the emtional reaction of a person, an operational failure of a component, a side effect of a drug based on the application area. Using the sentiment analysis, the entity's behaviour could be reasoned out. This additional information when used along with a Machine Learning(ML) data model, can further improve the accuracy of classification results. The proposed model improves the exisiting Multinomial Naïve Bayes using sentiment analysis scores and and latent topics and builds a classifier to automate analysis.

The result of the analysis could be used for efficient decision making in medical diagnosis, manufacturing error probes, automobile efficiency monitioring etc. The proposed system is tested using drug review dataset obtained from UCI repository. The health conditions with their associated drug names were extracted from the reviews and sentiment scores are assigned. The sentiment scores reflected the behavior of various drugs for a particular health condition and classified them according to their quality. Quality is decided based on its positive and negative side effects.

The aim is to extract disease names mentioned in drug reviews and find out the most suitable drug names for each ailment. Understanding the drugs based on their positive and negative feedbacks reflect the reason why a particular drug is suitable for a particular health condition and thus satisfies the aim of the study. The proposed model was compared with existing baseline models and found that our model exhibited better performance.

Corresponding Author

II. Literature Review

The advent of e-commerce not only replaced the physical shopping experience but also altered the word of mouth spread regarding goods of interest. People started commenting about their most favorite and not so favorite items. User reviews have almost become store house of user preferences. This work uses topic modeling with LSTM to understand review comments that improves the traditional topic clustering. However, the usage of smaller sliding windows failed to extract meanings of longer sentences leading to ambiguity [12]. Lexical Selection applications are concerned about context information of sentences and topics present in documents separately. This work combines them to achieve machine translation by exploiting term correlations [8]. It uses statistical probability based Gibbs sampling technique to achieve hidden variable prediction using source information. The context level features are split into local and global to learn topic distributions in data. But the model did not take into consideration, the phrase level information and word level attributes for classification accuracy improvements.

Another work uses Gaussian Mixture Neural Topic Model for extracting contextual information by using multi vector clusters [13]. The terms in each topics and sentences are jointly modeled. This order sensitive and context aware system can extract effective topics along with appropriate features. It does not consider polysemy of words and also manually detect topics which are time consuming. Topic identification in dialog systems is an interesting application of Topic Modeling [2]. It is applied in human and Chabot systems. It uses contextual key words related to each topic and the conversational features to achieve topic clusters with Deep Average and Contextual Attention Deep Average Networks. The coherent dialog process is easily achieved using this combined technique to annotate topics of interest. Unsupervised variant was not tested using this method.

Another interesting direction in modeling sentiment analysis is establishing relationship between ratings and sentiments of yelp reviews [6]. This helps to achieve a quantitative value for comments in better classifying favorite restaurants. The model was tested using Support Vector Machine algorithm in a binary classification setup. The model however, failed to model other rating values such as 3, 4, and 5 which is the norm in most rating systems. Also multiclass classification was not possible. To identify really helpful reviews from the available reviews is a challenging task for users. Most users rely on ratings to filter useful reviews [SA2]. The model uses TF-IDF for feature extraction which does not consider order of terms related to topics. Hence, the obtained topics will not be effective.

The Review Rating Prediction and Review Text Content analysis were combined to predict user preferences. The role of non-rated reviews is analyzed in this work [7]. It uses sentiment analysis of aspects to predict the ratings for all such non- rated reviews. This helps to identify most liked products in a quantitative manner. Usage of sentiments extracts contextual information also. To achieve this the model used Sentiment Base Conditional Random Fields to measure term co-occurrence. The model requires more training data for effective functioning which is not always possible. Another similar work analyses the multi aspect labeling of sentences and rating prediction for topic modeling [1]. It employs a semi-supervised approach for Perceptron Ranking technique. It is however, a weak prediction model not applicable for other stronger prediction models like support vector regression. In one more work related to rating analysis, physiological signals and reviews obtained from other data sources were used [10]. The global rating was assigned to global reviews using NLP techniques along with Electro encephalogram signals were recorded for each product. Sentiment analysis, Random forest with regression and Artificial Bee Colony algorithm is used to model global and local rating. But real emotions of patients were not considered. The work models the most subjective sentences in the document and measures its aspect sentiment orientation [17]. WordNet based clustering helps to analyze sentiment of clusters. Usage of hand coded rules, bigger training data are drawbacks of the model. Aspect is termed as topics.

An information retrieval based text classification is tried in this work to predict the rating value using the user comments [5]. By involving overall satisfaction of the customers through ratings and a clear description of why they have reacted in such a way to a product a justifiable decision making process is obtained. The classifier is modeled as supervised with vector space model and sentiment analysis to predict ratings on the scale of one to five. No usage of advanced NLP techniques was found. One major drawback in analyzing the sentiments of reviews, lies in the technique of modeling long term dependencies which are so complicatedly found in the short text reviews [19]. A deep sequential model is built for sentiment analysis using Gated RNN for dependency between sentences and LSTM to vectorize the sentences. It works on the basis of principle of compositionality which states that summarized meaning of reviews reflect actual concepts mentioned in the reviews. The model was not tested for document level analysis and its applicability for longer reviews was not available.

The mining of electronic patient records for decision-making and diagnostics requirements is a prominent issue in text classification [18]. The discharge summaries of patients were morphologically analyzed for extracting features with the help of two-dimensional attribute mapping through correspondence analysis. The class labels were assigned after obtaining the keywords and rank them accordingly. The distance between the data points and class labels are measured and those with shortest distance were assigned the corresponding label. It was found that decision trees though capture the structure of data, performed poor.

Another similar work on the same kind of data used statistical techniques, rules of associations and Extreme learning machine encoder for extracting patient specific features [16]. It works on the principle of sentence level sentiment analysis. However, it requires the sentences to be subjective and availability for a lexicon specific dictionary is scarce. In vehicle fault diagnosis applications, the text classification system can be used along with machine learning and search-prompt techniques [20]. The diagnostic codes are integrated with term weight matrix to obtain similarity scores between documents and labels. Latent Semantic Indexing is found not so suitable for this kind of analysis. The diagnostic codes have to manually obtained and integrated with the system. In the place of diagnostic codes, the feedback or review data could be used which will be more effective.

Problems to address in the proposed work:

- Extracting topics present in random texts suffers from, high-frequent keyword elimination, subjectivity of sentences and lack of quantitative value of texts.
- Sentiment Analysis without rating information leads to imprecise classification results.
- Use of conventional TF-IDF is not suitable for opinion mining due to long term dependency in sentences and shorter expressions.

The proposed classifier model attempts to solve the above problems using a combined approach of topic modeling and text classification based on sentiment value of documents. The detailed approach is explained in the following sections.

III. Proposed Methodology

The proposed model is built as a two stage classifier for multi class classification in opinion mining problems. The aim is to extract entities of interest from the 'review' or 'comments' data. Then based on the subjectivity, the sentiment values are assigned to sentences that in turn make up the topics. The knowledge about sentiments related to topics reiterates the originality of ratings provided by the people. The proposed framework consists of a topic modeler, that extracts latent topics from the reviews and its behavior is studied using the sentimental values and ratings provided by the user. In the next step, a ML classifier data model is built for effective classification requirements off the data. The proposed framework is given in Fig. 1.

According to Fig. 1, the drug reviews dataset was retrieved from UCI machine learning repository. The dataset is a collection of names of drugs, its related conditions, review regarding the side effects and ratings. The aim is to extract unique drug names and health conditions using topic modeling techniques and in the next phase assign sentiment tags to each drug. The name of drugs mentioned in the dataset is not related to one particular health condition or disease, but generic opinions of patients that took the drugs. If the system could filter each disease along with its side effects, name of drug and sentiment value, it would assist the medical practitioners is swift diagnosis of patients.

In the initial data pre-processing stage, the dataset is thoroughly studied. The exploratory data analysis was done to extract complete information of the drug reviews which is spread over 2,15,0633 instances containing attributes such as names of drugs, type of health conditions, review of patients, ratings, date of review and count of users that found review helpful.

A. Preprocessing

The records were checked for null entries and it was not found anywhere in the dataset. The identifiers given to each reviews are unique. Regular expression rules were manually constructed to remove 'non-English characters', 'symbols', 'stop words' and 'upper case' characters. The average length of reviews is found to be around 500. The reviews were cleaned and stored in a separate column; this preprocessing is enough for the sentiment tagging stage. For topic modeling, more steps were carried out in cleaning such as stemming, lemmatization and tokenization. The corrupted reviews that were of no use for others were removed and accounted for 1171 reviews. Averagely, the number of reviews for a particular drug is found to be 3658 and that for a particular health condition is around 836. The top ten most reviewed drug names are plotted as a bar graph and can be found in Fig. 2.

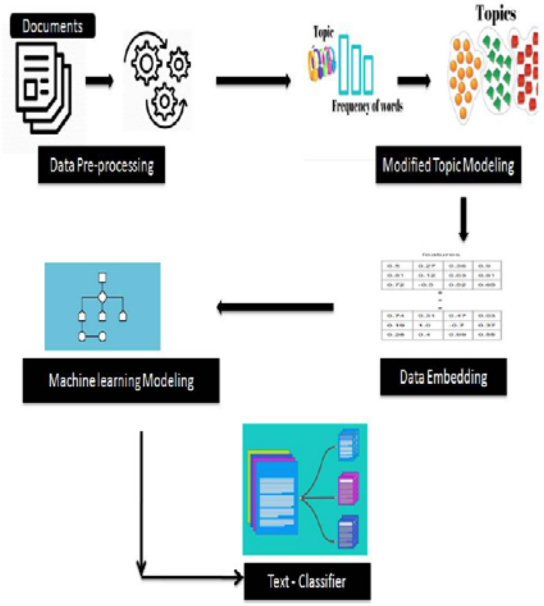

Fig. 1. Improved Multinomial Naive Bayesian Framework for Text Classification.

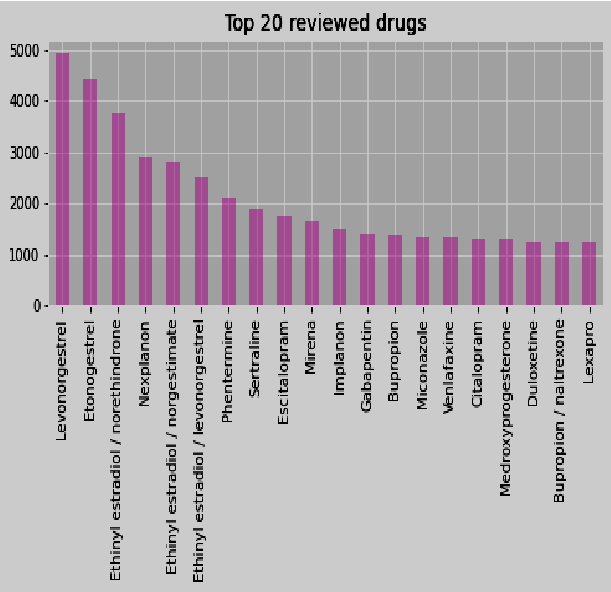

Fig. 2. Most Reviewed Drug Names.

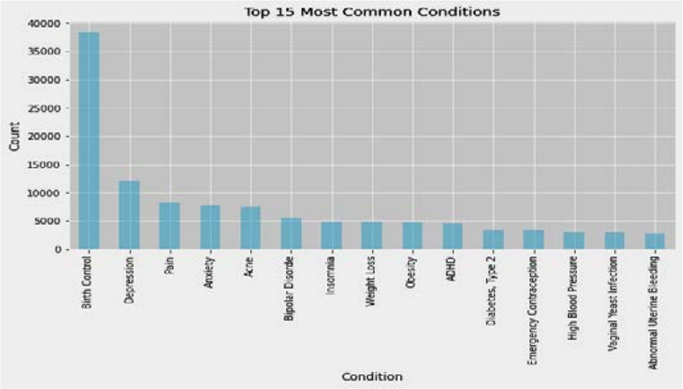

Fig. 3. Most Reviewed Health Conditions.

The statistics were extracted as part of the exploratory data analysis using count vectorizer function. The dataset was analyzed to obtain the most common health conditions of people. From the Fig. 3, we could see that birth control and depression are most talked about health conditions of people.

B. Modified Topic Modeling

The topic modeling is carried out using Latent Drichlet Allocation (LDA) algorithm as it does not require prior annotations. A group of words describes each topic or health condition. The application of LDA for text classification is a rare scenario especially in healthcare. The proposed LDA is an improvement over the existing model, wherein, sentiment analysis and rating information is included to model the topics apart from the general approach to topic modeling. The latent nature of topics could be substituted by using this technique to provide more accurate results. From the exploratory data analysis, it was identified that, on an average, there exist 3658 reviews for a particular drug and 836 reviews per health condition. The conditions for which only one drug name was given were eliminated. The unwanted 1171 corrupted reviews were also removed. The aim is to extract disease names mentioned in drug reviews and find out the most suitable drug names for each ailment.

All the disease names were selected as topic labels. Some of the topic labels are, 'Birth Control', 'Depression', 'Acne', 'Pain', 'Anxiety', 'Bipolar Disordered. For each of these diseases it was found that on an average 830 reviews existed. The topics were modeled as clusters and for each clusters 50 terms were set to maximum. The LDA generates probabilistic topic terms based on the concepts they share. The primary assumption is regarding the topics that are latent in the dataset, denoted by some constant say 'T'.

There exist 'N' number of documents for each latent topic which is nothing but a polynomial distribution over its constituent terms. It is assumed again that the 'T' latent topics generate the documents. It is important to extract the information that are hidden in the documents. This information is termed as, composition data of the unknown topic denoted by (α, x). The parameters of Richet can be mentioned as, β and λ. The distribution of random variables in a given document 'Doc' can be computed using the probability distribution,

$$P(\alpha, x, t \mid \beta, \lambda) = P(\alpha, \beta)\prod_{i=1}^{Q_e} P(x|\alpha)P(t|\alpha,\beta) \qquad (1)$$

The probabilistic values are approximated using Gibbs Sampling. The next step is topic or word embedding, since words or terms constitute topics. Instead of using TF-IDF which removes repeated terms irrespective of its contribution to a particular topic, we have used Topic2vec[11].Topic2vec is used with a modification in this improved-LDA, where, the high frequent words in a particular topic are replaced with their single entry obtained from the initial iterations and assigned to topics with the same probability values. However, there is presence of sparse features due to unequal topic representation. It can be removed by appending a medical thesaurus with the topic dictionary through an index. Such a thesaurus is constructed using medical articles scraped from web and tokenizing words related to topics found in our dataset. This not only removes sparse feature issues but also minimizes polysemy. When using Topic2vec model, the keywords of all sizes a treated as a fixed length vector, thereby solving the lengthy keyword problem. Thus, using these techniques, the various disease names are obtained as topics along with their constituent terms.

Followed by this, the Sentiment analysis is carried out to identify positive and negative reviews for a particular drug. Using the sentiment polarity of each review, the corresponding drug name is chosen as the best and worst drug for a particular disease obtained in the previous stage. Based on the review comments' polarity and corresponding rating values, the most common health conditions of people are identified. Here, the sentimental value of reviews helps to identify both a particular health condition and its appropriate drug. The useful count of reviews was included to analyze the sentiments. The Harvard polarity dictionary was used to assign polarity to each review using useful count. The drugs for each health condition were discovered using a review's useful count and sentiment value obtained.

The rating information is appended to the drug and health condition names along with its sentiment to form a semantic knowledge base for the subsequent processing. After identifying the disease names, respective drugs and the rating information from the reviews are indexed in a topic dictionary. This dictionary helps to identify drugs for various diseases quantitatively. Using this knowledge, the topics and their most likely terms are represented as a dataset for further processing. From the topic dictionary, the top-10 drugs for the sample health conditions, 'Birth Control', 'Depression', 'Acne', 'Pain', 'Anxiety' and 'Bipolar Disorder' are given in Table I. The novelty is to extract health conditions from drug reviews and identifying the corresponding drug names through topic modeling. Understanding the drugs based on their positive and negative feedbacks reflect the reason why a particular drug is suitable for a particular health condition and thus satisfies one of the aims of the study.

TABLE I. TOP-10 DRUGS REVIEWED FOR MOST COMMON HEALTH CONDITIONS

BirthControl	Depression	Acne	Pain
Plan B	Niravam	Milk of Magnesia	Ketoprofen
Femcon Fe	Serzone	Benzaclin	Acetaminophen / phenyltoloxamine
Ortho-Novum 7 / 7 / 7	Alprazolam	Magnesium hydroxide	Dolophine
Kyleena	Parnate	Benzoyl peroxide	Fentora
Ortho Evra	Xanax	Phillips' Milk of Magnesia	Lodine
Levonorgestrel	Xanax XR	Benzoyl peroxide / erythromycin	Methadone Diskets
Lybrel	Nefazodone	Retin-A	Proctofoam

C. Improved Naïve Bayesian Classifier

The model is further enhanced as a generic text classifier capable of solving any type of opinion based text classification problem by introducing a machine learning data model. To achieve this, we use a variant of Bayesian algorithm, Multinomial Naïve Bayes (MNB). This algorithm suffers from, independency between the features of its classes. The context information of classes is disregarded in the process [15]. To avoid this, the Bayesian procedure has to include some form of dependency between its features since, text data is heavily depended on its neighborhood information. To attain this functionality, the information obtained from the I-LDA phase is used for training the MNB to finally derive Improved MNB. This technique is named as I-MNB throughout the article.

The probability distribution of topic terms is already available from the improved LDA output. For each topic, here we have the name of diseases as our topic. For each topic, the topic terms probability is taken from the previous phase, its again affixed with sentiment scores for each topic. These scores are adjusted relative to the intra topic terms. For example, the sentiment value for a particular drug that receives highest positive score will be the most suitable drug for that disease. The rest of the drugs will be scored using relative scoring technique to order them in descending order according to the relative scores they obtained. This relative score establishes a dependency among terms in the topics, thereby capturing the optimal Bayesian network. The classification result is improved through this representation, since; it gives a mathematical explanation to the classification results. In other words, it tells why a certain drug is prescribed over the others.

Probability of each term per class is give as:

$$\mathrm{P}(\alpha, \mathrm{x}, \mathrm{t} \mid \beta, \lambda) = \mathrm{P}(\alpha, \beta)\prod_{i=1}^{Q_e} P(x|\alpha)P(t|\alpha,\beta) \quad (2)$$

The sentiment weighted terms for a class is given as

$$\mathrm{P}(\alpha, \mathrm{x}, \mathrm{t} \mid \beta, \lambda) = \mathrm{P}(\alpha, \beta)\prod_{i=1}^{Q_e} P(x|\alpha)P(t|\alpha,\beta) + |\mathrm{senti}(\mathrm{T}_i)| \quad (3)$$

$|\mathrm{senti}(\mathrm{T}_i)|$ denotes the absolute value of sentiment score for a given class, it is calculated using subjectivity analysis. Here, we have not included the rating information since, ratings are not always available for all entities. The sentiment of the drug reflects the sentiment of the opinions of the patient or users that consumed them. Also, the aggregated sentiment value of a review is the sentiment of the drug, therefore the neighborhood information of drug reviews decides the overall sentiment of the drug. This captures the dependency among features related to a particular drug. The dependency between other drugs for the dame disease can be obtained through comparing the sentiment scores of each drug and ranking them relatively.

$$\mathrm{senti(T)}=[\mathrm{senti}(\mathrm{t}_1),\mathrm{senti}(\mathrm{t}_2),\mathrm{senti}(\mathrm{t}_3),\ldots\mathrm{senti}(\mathrm{t}_\mathrm{n})] \quad (4)$$

Relative Scoring,

$$\max\,[\mathrm{senti(T)}] = \mathrm{senti}(\mathrm{t}_1) > \mathrm{senti}(\mathrm{t}_2) > \mathrm{senti}(\mathrm{t}_3) > \ldots\ldots\ldots\ldots\ldots\ldots\ldots > \mathrm{senti}(\mathrm{t}_\mathrm{n}) \quad (5)$$

The sentiment scores are normalized using laplace smoothing.

$$\mathrm{senti(i|j)} = \frac{term_{ij}+\theta}{topic_j + |senti|+1}, \theta = 0.001 \quad (6)$$

senti is the sentiment score of all terms in the topic. Now the combined probability distribution of a topic 'T' can be obtained using the distribution over its terms.

For a given topic j and term i, at the term frequency denoted by tf,

$$P(j) \propto \mu_j \prod_{i=1}^{|senti|} p(i|j)^{tf_i} \quad (7)$$

Taking log on the probability distribution will avoid term overflow for topics

$$P(j) \propto log\ (\ \mu_j \prod_{i=1}^{|senti|} p(i|j)^{tf_i}) \quad (8)$$

$$P(j) = log\mu_j + \sum_{i=1}^{|senti|} tf_i \log(P(i|j)) \quad (9)$$

Thus, the optimal I-MNB model can be obtained as,

$$P(j) = log\mu_j + \sum_{i=1}^{|senti|} \log(1+tf_i) \log(P(i|j)) \quad (10)$$

The proposed model with the necessary improvements, have given impressive results in the experimental analysis, which will be discussed in the next section.

IV. PERFORMANCE EVALUATION

The model is tested using drug review dataset. The two-phased approach is built as a multi class text classifier to be generically able to deal with opinion mining problems in similar applications. This model is compared with other benchmark algorithms in text data classification. Some dataset-oriented changes were made appropriate to each of the benchmark algorithms, such as, discrete to continuous features, binary to multiclass, numerical categorical conversions, normalizations etc. The benchmark algorithms taken for the experimental study are, Linear Support Vector Classification [9], Logistic Regression [3] and Random Forest classifier models [4].

The performance is validated through the score of accuracy of the model. It is the measure of ratio between total predictions obtained correct and sum of predictions made in the entire dataset. Accuracy is not suitable for problems with imbalanced class distributions. Hence, along with accuracy, it is required to measure other metrics such as 'Precision', 'Recall' and 'F1-score.

A. *Precision*

Precision verifies the number, percentage, or value of True Positive (TP) predictions for all classes to the True Positive and False Positive (FP) predictions of all classes in the dataset [15]. It is denoted by the formula,

$$\text{Precision} = \text{Sum}(TP(cs)) / \text{Sum}(TP(cs) + FP(cs)) \qquad (11)$$

Where, cs denotes classes in the dataset

B. *Recall*

Recall verifies the number, percentage, or value of True Positive (TP) predictions for all classes to the True Positive and False Negative (FN) predictions of all classes in the dataset [15]. It is denoted by the formula,

$$\text{Recall} = \text{Sum}(TP(cs)) / \text{Sum}(TP(cs) + FN(cs)) \qquad (12)$$

C. *F1-Score*

To improve precision and recall, the tweaking of one measure might increase or decrease another. To avoid this F1-score is used. It summarizes the overall performance of a system, by deriving the harmonic mean of both the precision and recall results [15].

$$\text{F1-score} = (2 * \text{Pre} * \text{Rec}) / (\text{Pre} + \text{Rec}) \qquad (13)$$

The combined Precision, Recall and F1-scores obtained for the proposed model is compared with the benchmark algorithms and plotted as bar graph.

From the Fig. 4, we can see that the performance of our proposed I-MNB is higher than the other algorithms in all three metrics. This is due to better knowledge representation rendered by the I-LDA phase. The Random Forest Classifier though almost matches the proposed model's perfromance, it falls short in Recall values drastically. According to Eqns (10, 11 and 12) it is also proved that our model did not overfit the data by showing near perfect perfromance values.

D. *Accuracy*

Accuracy is the ratio of True Positive results and the entire number of results obtained for the given dataset [15]. It is denoted by,

$$\text{Acc} = (TP+TN)/(TP+FP+FN+TN) \qquad (14)$$

From the Fig. 5, we can see that the performance of our proposed I-MNB is higher than the other models in accuracy. The accuracy achieved for our system is 91%. It can be noted that, the accuracy of Linear SVC falls way below other models, stating that this algorithm is not suitable for opinion mining problems.

E. *AUC –ROC*

Receiver Operating Curve and the relative Area Under Curve (ROC-AUC) state the separation between positive and negative classes [15]. The number of positive classes is known as True Positive Rate(TPR) and that of the negative classes are known as False Positive Rate(FPR) and range of separation is measured using various threshold values. It is a plot between TPR and FPRin other words sensitivity and 1-specificity denotes.

$$\text{Sensitivty} = \text{Recall} = \text{Sum (TP (cs)) / Sum (TP (cs) + FN (cs))}$$
$$\text{1- Specificity} = FP/(TN+FP)$$

From the Fig. 6 we can see that the pink lines represented by our proposed model I-MNB outperformed the others. Though the models LinearSVC and Random Forest Classifier converge with our model it is mainly with increasing thresholds which is not advisable since, it will drastically decrease the FP rate and increase FN rate simultaneously. The increase is not gradual. Also, the area occupied under the curve is still higher for our proposed model.

From the Fig. 7 we can see that AUC-ROC values obtained for some of our sample classes by the proposed I-MNB. The model has shown impressive results in all the six classes with the 'Pain' class topping the accuracy. It can also be seen that 80% of classes obtained better results and the remaining classes obtained satisfactory results. This is due to the people reviewing the particular disease and its constituent term ambiguity. Overall the model shows that the area occupied under the curve is still commendable for a multiclass text classifier.

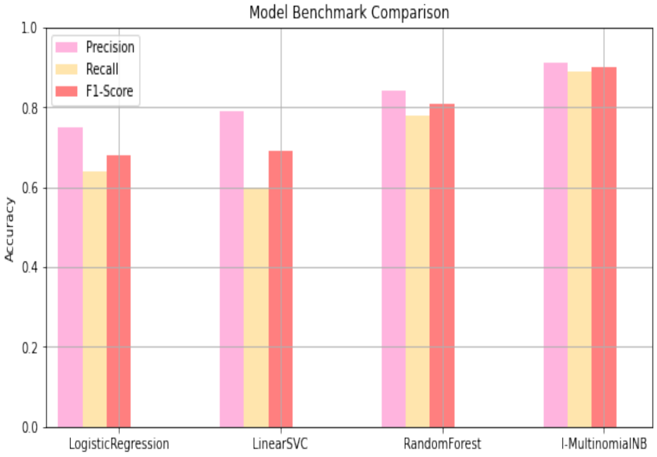

Fig. 4. Model Benchmark Comparison Results.

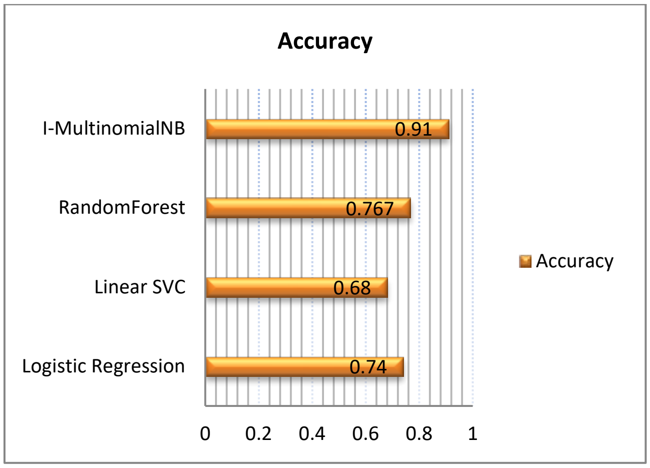

Fig. 5. Model Benchmark Comparison Results for Accuracy.

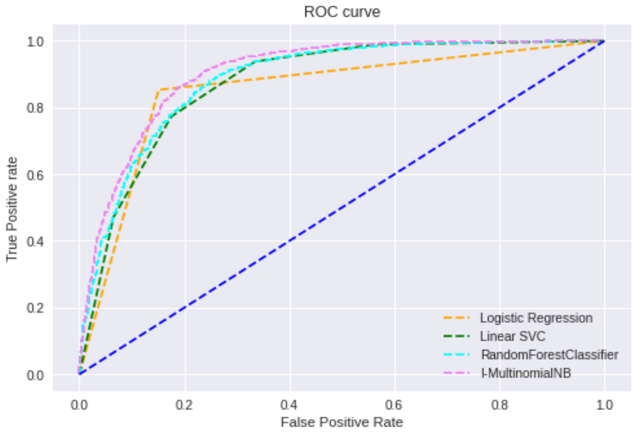

Fig. 6. AUC-ROC of Benchmark Models.

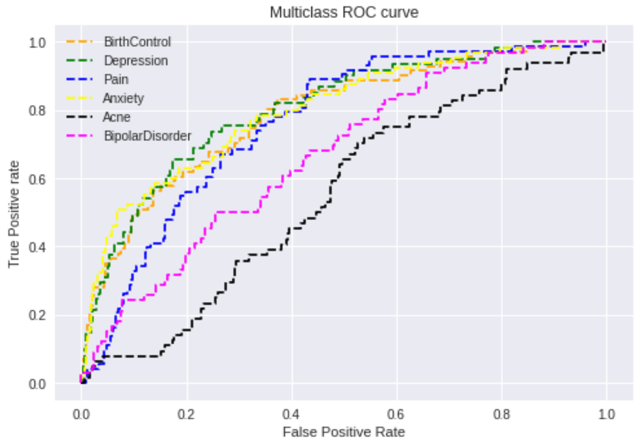

Fig. 7. AUC-ROC of Classes in I-MNB.

V. Discussion

The proposed model is built as a two phase classifier for multi class classification in opinion mining problems. The aim is to extract entities of interest from the 'review' or 'comments' data. Then based on the subjectivity, the sentiment values are assigned to sentences that in turn make up the topics. All the disease names were selected as topic labels. Some of the topic labels are, 'Birth Control', 'Depression', 'Acne', 'Pain', 'Anxiety', 'Bipolar Disordered. For each of these diseases it was found that on an average 830 reviews existed. Instead of using TF-IDF in LDA which removes repeated terms irresepctive of its contribution to a particular topic, we have used Topic2vec to improve LDA. The repplacement of high frequency words with their base words, sparse features minimization using medical thesaurus, polysemy representation and semantic modelling of knowledge base of classifier have shown remarkable perfromance compared to other baseline models. Theknowledge about sentiments related to topics reiterates the originality of ratings provided by the people in identifying the best drug for a health condition. The novelty in combining LDA and MNB and using sentiment scores to represent class features with neighborhood dependency improved the existing MNB to suit the needs of opinion mining problems. The accuracy obtained through various metrics is around 91%, which is remarkable, given the complex nature of test data and presence of multiple classes.

VI. Conclusion

The aim of the work is to extract entities of interest from the 'review' or 'comments' data. Based on the subjectivity, the sentiment values are assigned to sentences that in turn make up the topics. Some of the topic labels are, 'Birth Control', 'Depression', 'Acne', 'Pain', 'Anxiety', 'Bipolar Disordered. The replacement of high frequency words with their base words, sparse features minimization using medical thesaurus, polysemy representation and semantic modelling of knowledge base of classifier have shown remarkable perfromance compared to other baseline models. The knowledge about sentiments related to topics reiterates the originality of ratings provided by the people in identifying the best drug for a health condition. The novelty in combining LDA and MNB and using sentiment scores to represent class features with neighborhood dependency improved the existing MNB to suit the needs of opinion mining problems. The accuracy obtained through various metrics is around 91%, which is remarkable, given the complex nature of test data and presence of multiple classes. Also the AUC, ROC, Precision, Recall, F1-score values for the proposed system is obtained in the range above 90% signifying 10 to 12% improvement over the similar benchmark models. In future, other opinion or review mining problems can be considered with ensemble algorithms and different applications.

References

[1] Kaporo, H. (2019, April). Cross-collection Multi-aspect Sentiment Analysis. In Computer Science On-line Conference (pp. 107-118). Springer, Cham.

[2] Ahmadvand, A., Choi, J. I., & Agichtein, E. (2019, July). Contextual dialogue act classification for open-domain conversational agents. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1273-1276).

[3] Wu, T., Huang, Q., Liu, Z., Wang, Y., & Lin, D. (2020, August). Distribution-balanced loss for multi-label classification in long-tailed datasets. In European Conference on Computer Vision (pp. 162-178). Springer, Cham.

[4] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., & Androutsopoulos, I. (2019). Large-scale multi-label text classification on eu legislation. arXiv preprint arXiv:1906.02192.

[5] E. M. Alshari, A. Azman, and N. Maluksstheaeprh, “Prediction of Rating from Comments based on Information Retrieval and Sentiment Analysis,” p. 5, 2016.

[6] Yadav, A., & Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: a review. Artificial Intelligence Review, 53(6), 4335-4385.

[7] Xu, H., Liu, B., Shu, L., & Yu, P. S. (2019). Bert post-training for review reading comprehension and aspect-based sentiment analysis. arXiv preprint arXiv:1904.02232.

[8] J. Su et al., “A Context-Aware Topic Model for Statistical Machine Translation,” in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, China, 2019, pp. 229–238, doi: 10.3115/v1/P15-1023.

[9] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. Information, 10(4), 150.

[10] Kumar.S, Yadav.M & Roy. P.P (2018). Fusion of EEG response and Sentiment Analysis of Product Reviews to Predict Customer Satisfcation. Information Fusion.https://doiorg/10.1016/j.inffus. 2018.11.001.

[11] Li, W., Matsukawa, T., Saigo, H., & Suzuki, E. (2020, May). Context-Aware Latent Dirichlet Allocation for Topic Segmentation. In Pacific-

Asia Conference on Knowledge Discovery and Data Mining (pp. 475-486). Springer, Cham.

[12] M. Jin, X. Luo, H. Zhu, and H. H. Zhuo, “Combining Deep Learning and Topic Modeling for Review Understanding in Context-Aware Recommendation,” p. 10.

[13] M. Yang, T. Cui, and W. Tu, “Ordering-Sensitive and Semantic-Aware Topic Modeling,” p. 7.

[14] Nalchigar, S., & Yu, E. (2018). Business-driven data analytics: a conceptual modeling framework. Data & Knowledge Engineering, 117, 359-372.

[15] Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

[16] S. A. Waheeb, N. A. Khan, B. Chen, and X. Shang, “Machine Learning Based Sentiment Text Classification for Evaluating Treatment Quality of Discharge Summary,” p. 16, 2020.

[17] S. Gojali and M. L. Khodra, (2016) “Aspect Based Sentiment Analysis for Review Rating Prediction,” p. 6.

[18] S. Tsumoto, T. Kimura, H. Iwata, and S. Hirano, “Mining Text for Disease Diagnosis,” Procedia Computer Science, vol. 122, pp. 1133–1140, 2017, doi: 10.1016/j.procs.2017.11.483.

[19] S. Verma, M. Saini, and A. Sharan, “Deep Sequential Model for Review Rating Prediction,” p. 6, 2017.

[20] Y. L. Murphey, L. Huang, H. X. Wang, and Y. Huang, “Vehicle Fault Diagnostics Using Text Mining, Vehicle Engineering Structure and Machine Learning,” International Journal of Intelligent Information Systems, p. 13.