ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

Churn Prediction with Sentiment Analysis using NLP and Machine Learning Techniques for online Customer Reviews

Danave Bharati Murlidhar

Asst. Professor, New Arts, Commerce and Science College, Ahmednagar, Maharastra, India

ABSTRACT: Today, communication technologies are a dynamic environment. Consumer churn is the major concern affecting virtually all the world's telecommunications industries now. The telecommunications industry, in particular, is facing complex challenges because of a variety of vibrant innovative service providers. But keeping current customers is becoming very difficult for them. Since the cheaper than attracting is much greater than the price of maintaining potential customers, it is time for the telecoms industries to take the required steps to retain customers in order to maintain their market value. Several data mining techniques have been proposed in the literature in the last decade to predict the redemptions using interconnected customer information. This paper discusses the various types of customer data available in open databases, make predictions and quality measures used in telecom industry churn prediction in various researchers. We design and develop system for churn prediction using NLP and machine learning techniques using sentiment analysis approach. The data collection has done from twitter customer reviews using Twitter API of telecom services users. The NLP has used to data pre-processing, lemmatization and feature extraction as well as feature selection etc. Finally system works with training and testing with various cross validation using numerous machine learning algorithms. The experimental analysis demonstrates effectiveness and accuracy of system.

KEYWORDS: Churn prediction, feature extraction, NLP, feature selection, Sentiment analysis, Machine learning, classification, telecom reviews.

I.INTRODUCTION

It is found that data mining techniques are more effective in predicting consumer churn from the research conducted over the past few years. Creating an efficient churn prediction model is an essential activity requiring a lot of work right from determining appropriate predictor variables (features) from the large volume of available customer data to choosing an effective predictive data mining technique suitable for the feature set. Telecom Industries collect a large amount of customer-related data such as customer profiling, calling pattern, and democratic data in addition to the network data they generate. Based on the customer's history of calling behaviour and behaviour, there is a possibility to classify their attitude of either going away or not. Data mining techniques are found to be more effective in predicting churn from the research done over the past decade. The predictive modelling techniques in churn prediction are also considered to be more accurate. Churn prediction systems and sentiment analysis using classification as well as clustering techniques to classify churn customers and the reasons behind the churning of telecom customers. In telecom industry should we generate large amount of data on daily basis, it is very tedious task to mine such a kind of last data using specific data mining techniques, while hard to interpret the prediction on classical techniques. Various researchers already described search a work to eliminate churn from large data sets fusion static as well as dynamic approaches, but still such systems are facing many problems actual identification of churn. Sometime such telecommunication data may be containing some churn and, it is much necessary to identify search problems. To successful identification of churn from large data is providing effectiveness to customer relationship management (CRM).

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

In this research we proposed churn identification as well as prediction from large scale telecommunication data set using Natural Language Processing (NLP) and machine learning techniques. First system deals with strategic NLP process which contains data pre-processing, data normalization, feature extraction and feature selection respectively. Feature extraction techniques have been proposed like TF-IDF, Stanford NLP and occurrence correlation techniques. Where machine learning classification algorithms are has used to train and test the entire module.

II.LITERATURE SURVEY

In the area of customer churn prediction modelling there are a lot of research being carried out. In this section, we survey some of the research done in the last few years in this area. According to [1] the phases of a general model of churn prediction such as data collection , preparation, classification and prediction are depicted. It also explains the major effect of determining the correct grouping of variables in improving the percentage of accurate predictions (TP). A churn prediction model has been proposed[1] that works in 5 steps: I problem identification; ii) data set selection; iii) data set investigation; iv) classification; v) clustering; and vi) knowledge-based investigation. It's basically a total pattern. Classification methods are employed to differentiate between Churners. Clustering is used to test models. Decision tree, Help vector machine and Neural Network are used for classification. And Simple K-Means was used for the clustering. It concluded that in distinguishing churners from non-churners SVM was the best among the three methods.

In recent years , the development of an effective customer churn prediction model using different techniques has become an important topic for business and academics[2]. For the past few years, the identification of why customers abandon their relationships has been the focus of marketing research[3]. Because of the huge growth in customer-related data and call-detail data collected and maintained by companies in recent years, more sophisticated metrics have evolved to describe customer behavior and better understand how behavioral characteristics can be linked to customer retention and firm performance[4].

The authors in [5] suggested a Random Forest approach for the extraction of features based on a decision tree. Actually N samples with $q < Q$ are randomly selected from the original data with Q features for each tree to form a forest. The number of trees depends on the number of randomly combined features from all of the Q features for each decision tree. The creation of a predictive model focused on a data-centered approach to detect early warning signs of churn and the creation of a Churn Score to classify subscribers likely to terminate their relationship with the business was addressed in[6], where the calling behavior of the consumer played a major role in predicting churn.

An overview of socially grouped users and their behavioural pattern are elaborately identified in [7]. It also explores the impact of centrality features among customers. It concludes that when a customer in a group leaves the network, there is a high probability of others in that group to leave the network. It classifies the feature variables in two types: i) dependent variables (call duration of in-degree and out-degree), ii) independent variables (social forum like services and the customer's involvement in the forum). The level of customer's active participation was used as the measure of probability of churn. Customers who are members of multiple community forums are at high risk than those who are members in less number of forums.

Gavril et al. [8] presented an advanced methodology of data mining to predict churn for prepaid customers using dataset for call details of 3333 customers with 21 features, and a dependent churn parameter with two values: Yes/No. Some features include information about the number of incoming and outgoing messages and voicemail for each customer. The author applied principal component analysis algorithm "PCA" to reduce data dimensions. Tree machine learning algorithms were used: Neural Networks, Support Vector Machine, and Bayes Networks to predict churn factor. The author used AUC to measure the performance of the algorithms. The AUC values were 99.10%, 99.55% and 99.70% for Bayes Networks, Neural networks and support vector machine, respectively. The dataset used in this study is small and no missing values existed.

Huang et al. [9] studied the problem of customer churn in the big data platform. The goal of the researchers was to prove that big data greatly enhance the process of predicting the churn depending on the volume, variety, and velocity of the data. Dealing with data from the Operation Support department and Business Support department at China's

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

largest telecommunications company needed a big data platform to engineer the fractures. Random Forest algorithm was used and evaluated using AUC.
Various researches studied the problem of unbalanced data sets where the churned customer classes are smaller than the active customer classes, as it is a major issue in churn prediction problem. Amin et al. [10] compared six different sampling techniques for oversampling regarding telecom churn prediction problem. The results showed that the algorithms (MTDF and rules-generation based on genetic algorithms) outperformed the other compared oversampling algorithms.

III.PROPSED SYSTEM DESIGN

In the proposed research work to design and develop an approach for churn prediction using NLP and machine learning approaches to enhance the system accuracy which is illustrated in figure 1. Then we identify the customer changing behavior pattern during prediction. We also evaluate the factor which mostly influences to reduce accuracy of churn prediction and finally evaluate and calculate churn rate for month wise as well as day wise, which useful for enhance the service quality of system.

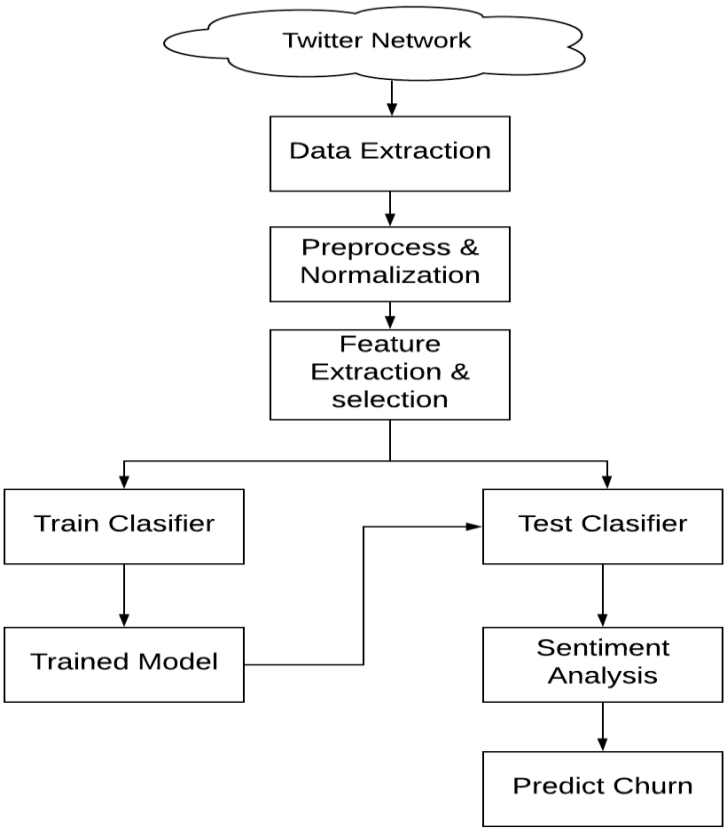

Figure 1 : Proposed System Architecture

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

ALGORTHMS DESIGN

1 : Stop word Removal Approach

Input: Stop words list L[], String Data D for remove the stop words.

Output: Verified data D with removal all stop words.

Step 1: Initialize the data string S[].

Step 2: initialize a=0,k=0

Step 3: for each(read a to L)

If(a.equals(L[i]))

Then Remove S[k]

End for

Step 4: add S to D.

Step 5: End Procedure

2 Stemming Algorithm.

Input : Word w

Output : w with removing past participles as well.

Step 1: Initialize w

Step 2: Intialize all steps of Porter stemmer

Step 3: for each (Char ch from w)

If(ch.count==w.length()) && (ch.equals(e))

Remove chfrom(w)

Step 4:if(ch.endswith(ed))

Remove 'ed' from(w)

Step 5: k=w.length()

If(k (char) to k-3 .equals(tion))

Replace w with te.

Step 6: end procedure

3 TF-IDF

Input : Each word from vector as Term T, All vectors V[i…n]

Output : TF-IDF weight for each T

Step 1 : Vector = {c1, c2, c3….cn}

Step 2 : Aspects available in each comment

Step 3 : D = {cmt1, cmt2, cmt3, cmtn}

and comments available in each document

Calculate the Tf score as

Step 4 :tf (t,d) = (t,d)

t=specific term

d= specific document in a term is to be found.

Step 5 :idf = t → sum(d)

Step 6: Return tf *idf

4 Classification

Input : Training Rules Tr[], Test Instances Ts[], Threshold T.

Output : Weight w{0-1}

Step 1 : Read each test instance from (TsInstnace from Ts)

Step 2 : TsIns $= \sum_{k=0}^{n} \{Ak \ldots An\}$

Step 3 : Read each train instance from (TrInstnace from Tr)

Step 4 :$TrIns = \sum_{j=0}^{n} \{Aj \ldots \ldots Am\}$

Step 5 :*w = WeightCalc(TsIns, TrIns)*

Step 6 :if (w >= T)

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

Step 7 : Forward feed layer to input layer for output OutLayer[] ← {Tsf,w}
Step 8 :optimized feed layer weight, Cweigt←OutLayer [0]
Step 9 : Return Cweight

IV,RESULTS AND DISCUSSIONS

The below analysis is the system classification graph. The graphs display how system classify the overall inputs into classification various instances. The proposed system is implemented with RNN combination, which gives all results with satisfactory level. For performance evaluation 5000instances given for training and 1500reviews given for testing with different cross validation. Here system compares the proposed results with two different existing systems.

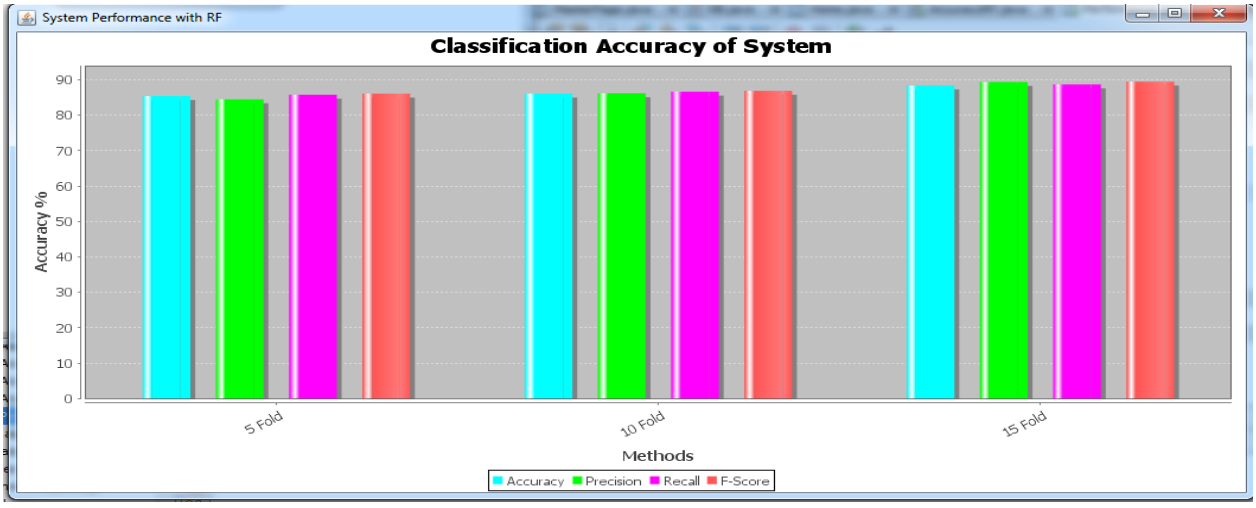

Figure 3 : Classification accuracy of RF

Table 1 : Classification accuracy of RF

(RF)	5-Fold	10-Fold	15-Fold
Accuracy	85.40	86.10	89.40
Precision	84.40	86.50	89.40
Recall	85.70	86.65	89.70
F-Score	86.50	86.90	89.50

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

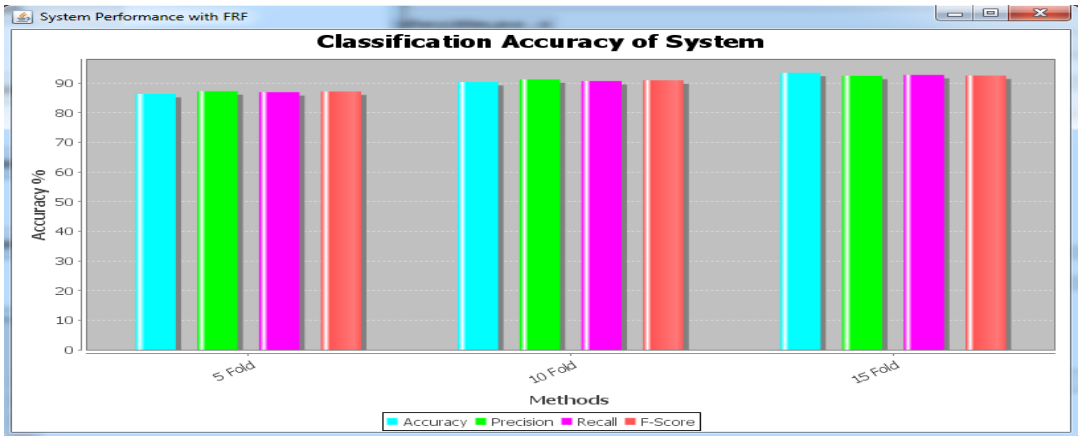

Figure 4 : Classification accuracy of FRF

Table 2 : Classification accuracy of FRF

(FRF)	5-Fold	10-Fold	15-Fold
Accuracy	86.30	90.30	92.60
Precision	87.20	91.00	91.20
Recall	86.90	90.25	91.00
F-Score	87.10	90.80	91.60

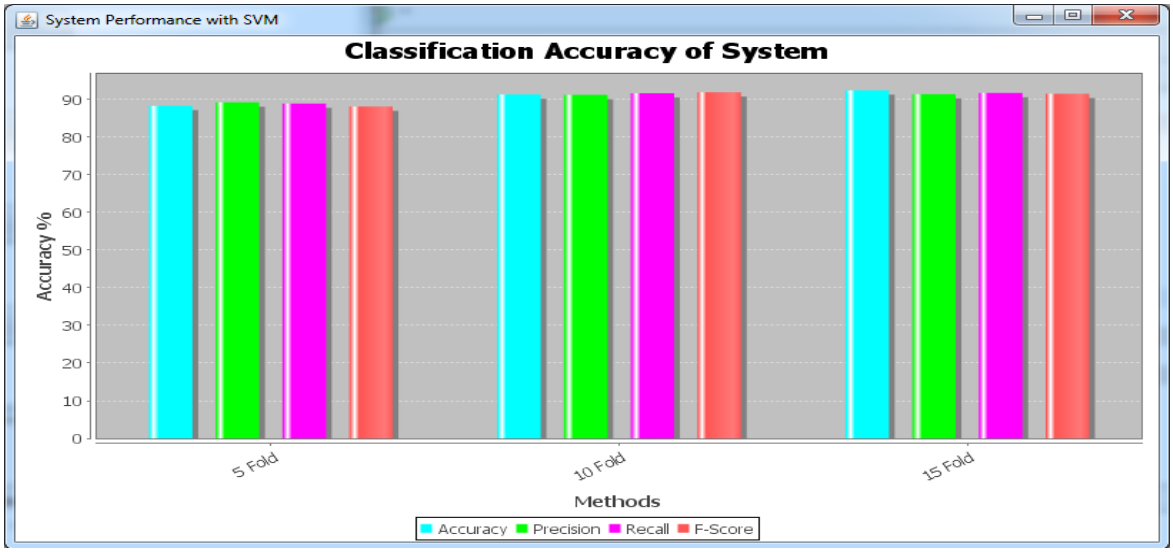

Figure 5 : Classification accuracy of SVM

Table 3 : Classification accuracy of SVM

(SVM)	5-Fold	10-Fold	15-Fold
Accuracy	88.80	91.10	92.40
Precision	89.00	91.15	91.40
Recall	89.30	91.20	91.70
F-Score	88.70	91.60	91.50

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

V.CONCLUSION AND FUTURE WORK

This research mainly focuses on identifying and detecting churn consumers from massive data set of telecommunications, state-of-the-art discusses churn prediction systems produced by different research. Some systems still face problems of conversion of linguistic data, which can occur at high error rate during execution. Many researchers have been putting forward Natural Language Processing (NLP) techniques as well as various machine learning algorithms such a combination is likely to generate good performance when structuring data. If any machine learning algorithm interacts with that kind of a method, it is necessary to test or confirm the entire data set with even sampling techniques that reduce data imbalance problems and provide reliable predictive flow of data. For future direction to implement a proposed system with various machine learning algorithm to achieve better accuracy, as well as the input data contains large size and volume, if we deal the proposed systems with HDFS framework and parallel machine learning algorithm which will provide better result in low computation cost.

REFERENCES

[1] Liao, Shu-Hsien, Pei-Hui Chu, and Pei-Yuan Hsiao. "Data mining techniques and applications–A decade review from 2000 to 2011." Expert Systems with Applications 39, no. 12 (2012): 11303-11311.
[2] Kamalraj, N., and A. Malathi. "A survey on churn prediction techniques in communication sector."International Journal of Computer Applications 64, no. 5 (2013).
[3] N.Hashmi, N.ButtandM.Iqbal. Customer Churn Prediction in Telecommunication A Decade Review and Classification. International Journal of Com-puter Science Vol.10(5),2013
[4] V. Umayaparvathi, K. Iyakutti, " Applications of Data Mining Techniques in Telecom Churn Prediction", International Journal of Computer Applications, Vol. 42, No.20, 2012 [5] V. Umayaparvathi, K. Iyakutti,, "Attribute Selection and Customer Churn Prediction in Telecom Industry", Proceedings of the IEEE International Conference On Data Mining and Advanced Computing, 2016 (to be appeared).
[6] Huang, Bingquan, MohandTaharKechadi, and Brian Buckley. "Customer churn prediction in telecommunications." Expert Systems with Applications 39, no. 1 (2012): 1414-1425
[7] Shaaban, Essam, YehiaHelmy, AymanKhedr, and Mona Nasr. "A proposed churn prediction model." IJERA 2 (2012): 693-697
[8] Brandusoiu I, Toderean G, Ha B. Methods for churn prediction in the prepaid mobile telecommunications industry. In: International conference on communications. 2016. p. 97–100.
[9] Huang F, Zhu M, Yuan K, Deng EO. Telco churn prediction with big data. In: ACM SIGMOD international conference on management of data. 2015. p .607–18.
[10] Amin A, Anwar S, Adnan A, Nawaz M, Howard N, Qadir J, Hawalah A, Hussain A. Comparing oversampling techniques to handle the class imbalance problem: a customer churn prediction case study. IEEE Access. 2016;4:7940–57