# Mining Churning Factors in Indian Telecommunication Sector Using Social Media Analytics

Nitish Varshney and S.K. Gupta

Department of Computer Science and Engineering,
Indian Institute of Technology Delhi, India
{nitish.mcs12,skg}@cse.iitd.ac.in

**Abstract.** In this paper we address the problem of churning in the telecommunication sector in Indian context. Churning becomes a challenging problem for telecom industries especially when the subscriber base almost reaches saturation level. It directly affect the revenue of the telecom companies. A proper analysis of factors affecting churning can help the telecom service providers to reduce churning, satisfy their customers and may be design new products to reduce churning. We use social media analytics, in particular twitter feeds, to get opinion of the users. The main contribution of the paper is feasibility of data mining tools, in particular association rules, to determine factors affecting churning.

**Keywords:** Sentiment analysis, Social data analysis, Data mining applications: telecommunication, Churn pertaining factors.

## 1 Introduction

Mobile service usage in India has increased rapidly following the reduction in call cost and emerging use of new mobile phone technologies. Currently India's telecommunication network is the second largest in the world in terms of total number of users (both fixed and mobile phone) [14] and it has one of the lowest call tariffs enabled by the mega telephone network operators and hyper-competition among them. On 30th September, 2013, country's telecom subscriber's base was as huge as 899.86 million [13] and penetration rate was about 71%. Out of these 899.86 million total telecom subscribers, about 97 % utilize wireless services. Due to such large number, our focus is on wireless telecom services and mobile service providers as they are predominant in numbers.

There are about 15 mobile carriers in India. Most of them are quite stable and have pan India presence. It is estimated that about 96 % of all mobile subscribers opt for a prepaid service and there is a fiercely competitive dynamic environment leading to luring customers from one service provider to another.

Above situation depicts a condition where market is almost saturated and telecom service providers are stable. It leads to intensification of competition among existing mobile service providers in order to maintain their subscriber base. In such a situation, the significant business drivers would be customer subscriber base retention and increase in average revenue per customer [5].

L. Bellatreche and M.K. Mohania (Eds.): DaWaK 2014, LNCS 8646, pp. 405–413, 2014.

There is a trade off in between these business drivers. Customer retention depends on factors like call rate and quality of services provided by a service provider. A superior quality of service imposes heavy implementation cost which has to be passed on to the customer. This has twin fall back of either higher call rate charges or subscriber churning. Hence, telecom service providers would want optimal values for both.

However, in Indian market most of the devices have multi-SIM card capability and it is easier to switch the service provider by getting a new SIM or using Mobile Number Portability (MNP). MNP enabled subscribers to retain their telephone numbers when switching from one service provider to another. It made it difficult for service providers to retain customers.

In telecommunication, customer movement from one service provider to other service provider is termed as churning. Churn rate is the percentage of subscribers who discontinue services with a service provider and change their service provider as per their choice. Customer churning is of great concern for any service provider. However, according to statistical information provided by Telecom Regulatory Authority of India (TRAI) already 100+ million users have utilized mobile number portability service [13]. This is relatively very high, specially when the aim is to retain the existing customers. Companies need to fully understand the factors leading to customer churn. These problems have not been fully addressed in the literature. In this paper, we present a data mining based approach to determine factors affecting churn in the Indian telecom sector.

## 2 Related Work

Existing literature on churn factors can be classified into two categories. Both categories attempt to predict customers getting ready to switch, understand why and connect with them to offer incentives to mitigate churn [9,15].

The first approach is based on large-scale actual customer transaction and billing data. This is proprietary data of a service provider. Such studies use various machine learning techniques like Support Vector Machine [2], Regression [6], Decision Trees [6,9] etc. They rely on the rich subscriber call data records which are available inhouse. Such dataset describe calling behavior of a customer by providing information such as their voicemail plan, call lengths and usage patterns. This leads to determination of patterns like:

- Patterns followed by churning customers.
    - If I am calling more than X minutes, then I will churn.
    - If I am calling to customer care more than Y times, then I will churn.
    - If my in net call duration is low, then I will churn.
- Potential value of a customer.

The second approach avoids the proprietary nature of actual customer call record data, and deals with consumer survey data [7,11] avoiding privacy issues. These consider consumers perceptions of service experiences and intention to churn. However, the survey data may not fully represent the customers actual future continued patronage decision.

The above two approaches focus on predictive accuracy rather than descriptive explanation or reasons thereof. Also such studies can not be directly useful in making a decision support system. To illustrate it, suppose if Vodafone reduces 3G usage charges to 2paisa/10KB from 10paisa/10KB, then should Idea also reduce their 3G usage charges? Here mining the impact of not reducing the charges is one of the most important parameters in decision making. The above approaches would not provide any information about such queries. So there were problems with such works as they can not pinpoint driving forces and used to measure their impact. We therefore focus our attention to know what went wrong and what would be the impact.

This paper has two distinct research objectives compared with the previous approaches. The first objective is to identify factors pertaining to churn which may help decision makers improve operations in terms of their marketing strategy, specifically customer churn prevention programs. The second objective is to develop a comprehensive model which can help in taking business decisions. Specifically, this research uses data mining techniques to find a model to achieve above two objectives.

## 3 Methodology

Figure 1 shows methodology applied to find dominant churning factor.

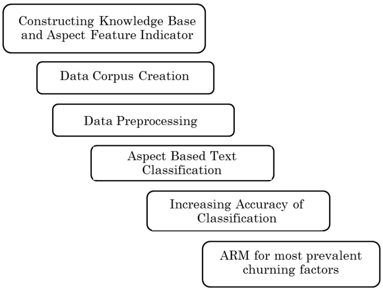

Constructing Knowledge Base and Aspect Feature Indicator
Data Corpus Creation
Data Preprocessing
Aspect Based Text Classification
Increasing Accuracy of Classification
ARM for most prevalent churning factors

**Fig. 1.** Methodology applied to find dominant churning factor

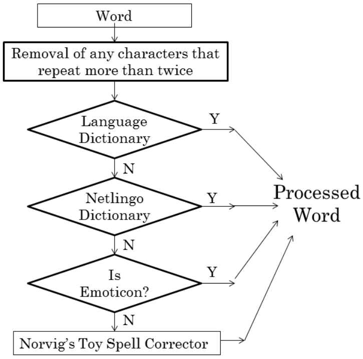

Word
Removal of any characters that repeat more than twice
Language Dictionary
Y
N
Netlingo Dictionary
Y
N
Is Emoticon?
Y
N
Norvig's Toy Spell Corrector
Processed Word

**Fig. 2.** NLP based spell corrector which takes a word as input and tries to correct spell errors if any.

### 3.1 Constructing Knowledge Base and Aspects Feature Indicators

An aspect (target) based extraction model has been developed, which looks for certain words in tweets using lexicon matching approach. Three major aspects (Price, Service and Satisfaction) in telecom business have been considered, which can influence a customer to churn [8]. Price aspect incorporates call rates and pricing options. Service aspect incorporates call quality, coverage, billing and customer service. Satisfaction aspect incorporates customers who have already churned. Every other churn indicators have been put up in miscellaneous aspect. Miscellaneous aspect also incorporates insulting effects on the reputation of the company. Each aspect has some feature indicators. For example, price aspect has rate, price, charge and tariff as feature indicators.

A knowledge base having unigram, bigram and trigram features are created manually. Bigram features are of the form string-string like tariff slashed, too-much interrupt. Trigram features are constructed to capture cases like ported to bsnl. In our approach, bigram and trigram features are not required to co-occur consecutively. Additionally, our approach partly utilizes a list of domain independent strongly positive and negative words provided by Hu and Liu [4], while building knowledge base. All words which are not strongly positive / negative in context of telecom business, are removed from the list. Few domain dependent strongly positive / negative word are also added to above list for example highspeed, slashes, flop etc. Features in the knowledge base are termed as sentiment describing terms and aspect feature indicators are termed as topical terms. These features are manually identified.

### 3.2 Data Corpus Creation

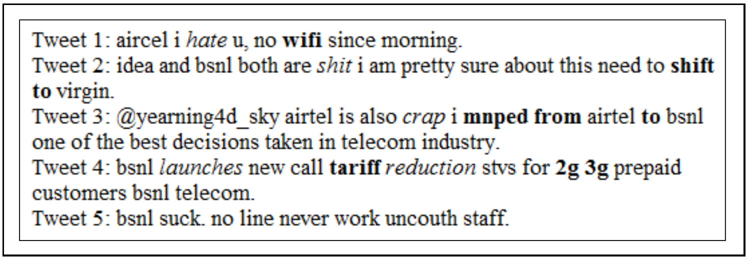
People often share and exchange ideas on social media platform. Also, there are a lot of personal thoughts to public statements about telecommunication services used by people on such platforms. The key idea is to utilize these feeds to determine factors leading to churn.

Twitter, the most popular social media platform, is used to build data corpus. It is publicly available and is very rich in content. The collected corpus can be arbitrarily large. Additionally, twitter audience represents users from different social and interest groups providing a good sample without bias.

We collected data over a span of 9 months (1 August, 2012 to 31 April, 2013) for three major service providers (BSNL, Aircel, Tata Indicom/Docomo) in India. We queried twitter setting service providers name as keyword. REST API is used to pull tweets and retweets using time parameter for pagination.

### 3.3 Data Pre-processing

Tweets often have lots of grammatically misspelled words due to 140 character limit. We propose to apply lexicon based text categorization method on such feeds which require us to overcome such errors. However, most popular Norvigs toy spell corrector [10] cannot be applied in its raw form in our corpus as feeds often incorporate internet lingo, colloquial expressions and emoticons (as toy spell correctors base file big.txt rarely contains any such words). It leads us to build a NLP based spell corrector, as described in Figure 2. It first tries to remove any characters that repeat more than twice making a word cool to appear as coooool as people sometimes repeat characters for added emphasis. Second it matches words in tweet with an English dictionary. Then to handle netlingo words, slangs and abbreviations our model incorporated words taken



Tweet 1: aircel i *hate* u, no **wifi** since morning.
Tweet 2: idea and bsnl both are *shit* i am pretty sure about this need to **shift to** virgin.
Tweet 3: @yearning4d_sky airtel is also *crap* i **mnped from** airtel **to** bsnl one of the best decisions taken in telecom industry.
Tweet 4: bsnl *launches* new call **tariff** *reduction* stvs for **2g 3g** prepaid customers bsnl telecom.
Tweet 5: bsnl suck. no line never work uncouth staff.

**Fig. 3.** Sample Tweets

from Netlingo[1], NoSlang[2] and Webopedia[3] for matching. Further, emoticon list provided by DataGenetics[4] is used. If no match is found for the word then big.txt is used to calculate edit distance as explained by Norvig [10]. Still our approach is not able to capture phenomena such as sarcasm, irony, humor etc., but overall, data captured in such a manner is quite reasonable.

In addition to these, NLTK wordnet lemmatizer was used to lemmatize each word. Further to structure text and handle moderately-sized dataset, feeds are pushed into a MySQL RDBMS. We retrieved around 75K feeds in such a way.

### 3.4 Aspect Based Text Classification

The collected dataset is cross-checked against knowledge base to extract aspect and sentiment. In sentiment extraction, ternary classification task were specifically designed to classify a tweet $x^i$, to a class $y^i \in$ {NEGATIVE, NEUTRAL, POSITIVE}. Neutral class has been incorporated in the model to classify tweets that are not assigned a positive / negative class.

Classification task can be broadly broken into two phases. In the first phase, sentiment describing terms and topical terms are matched with words appearing in a tweet. To collect these sentiments, we followed the same procedure as described in Taboada et al. [12]. The second phase calculates overall sentiment score based on summation of individual lexicon sentiment retrieved in the first phase. Our model assumes that customers providing negative tweets would lead to churn in few months. Figure 3 shows some examples of tweets fetched, sentiment describing terms are *italicized* and topical terms are in **bold**. A tweet can have multiple aspects or it may have none like Figure 3, tweet 4 has multiple aspects tariff, 2g and 3g. Tweets having positive or negative polarity and having price, service, satisfaction or miscellaneous aspect are used, other tweets are ignored.

[1] http://www.netlingo.com

[2] http://www.noslang.com/spelling.php

[3] http://www.webopedia.com/quick_ref/textmessageabbreviations.asp retrieved on February 10th 2013

[4] http://datagenetics.com/blog/october52012/index.html

### 3.5 Increasing Accuracy of Classification

Aspect based text classifier is not able to properly assign sentiment to tweets having multiple service providers because inclusion of multiple service providers in a tweet presents semantic issues. For example, tweet 2 and tweet 3 in figure 3, would lead to misclassification due to <shit, virgin> and <crap, bsnl> matching as a bigram feature. To tackle the issue, the classifier has been modified to use a simple neighbourhood proximity based approach. Sentiment describing terms are attached to those aspect feature indicators that occur closest to them. Experiments show that utilizing proximity improved accuracy of our model from 69% to 79.5%.

Negating words are capable of reversing polarity of sentiment again. Experiments show that if negative context lies in proximity of antonym of sentiment term and topical term, it does not change polarity of a tweet. For example tweet1 and tweet 5 in figure 3, should be classified as negative due to presence of bigram features <hate, wifi>, <suck, staff>, <uncouth, staff>. However our classifier would end up classifying such tweets as positive due to presence of negative words <no>, <never>. In an experiment we have manually annotated 200 tweets having negative words in the proximity of sentiment word and topical term, 85.5% of them are found to be wrongly classified using aspect based text categorization approach. Therefore we modified our approach to classify a tweet negative even in presence of negating words.

### 3.6 Association Rule Mining for Most Prevalent Churning Factors

Association rule mining (ARM)[1] is one of the most common data mining techniques, which is used to discover rules among multiple independent elements that co-occur frequently. We found out that it can be used to mine out co-occurrence between set of aspects and sentiments. Previous steps are considered as preprocessing steps for mining association rule with maximum confidence within a time window, say month. Top most of all such rules can be used to denote most prevailing churning factor within that time frame.

Association rules mined in such a manner are implication of the form A → B, where $A \subseteq I$, $B \subseteq I$, $A \cap B = \Theta$ and I be a set of literals called items. For our analysis, I = {positive, negative, price, service, satisfaction, miscellaneous}. Classification output obtained after applying aspect based text classification has been used to fill positive / negative item. Hence, in a tuple, either 'positive' or 'negative' item can have value 1. Aspect feature indicators obtained in the process are used to fill price, service, satisfaction, miscellaneous. In such a way, a binary table is obtained, similar to to transaction table generally used for ARM. Such a table for tweets in Figure 3 is shown in Table 1. We use Top-K Association Rule mining technique [3] as otherwise we are flooded with many association rules. In particular, we find that K=2 serves our purpose.

Top-K rules obtained after applying ARM algorithm on this dataset denote most pertaining churning factor, if $\{negative\} \subseteq A$ in association rule found. To illustrate it, suppose we obtain a rule negative → price in a time frame for

**Table 1.** Tweets in relational format for Binary Association Rule Mining

| TID | Positive | Negative | Price | Service | Satisfaction | Miscellaneous |
|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| 2 | 0 | 1 | 0 | 0 | 1 | 0 |
| 3 | 1 | 0 | 0 | 0 | 1 | 0 |
| 4 | 1 | 0 | 1 | 1 | 0 | 1 |
| 5 | 0 | 1 | 0 | 1 | 0 | 0 |

a service provider, it means within that time frame tariff charges of services provided by the service provider are not competitive and high in comparision to other service providers. Service provider needs to reduce tariff of the service otherwise customers are going to churn from it. All such rules obtained for Aircel service provider have been shown in table 2.

## 4 Experimental Evaluations

In this section we will be looking at the results obtained after applying above approach on data collected within a time frame. Table 2 shows top-k association rules mined, with k=2 and total customers for aircel service provider. We could interpret these results in the light of sentiments derived from tweets. Top-k association rules obtained are given in first row. Rules are presented in the form {A,B x} A→B with x confidence. Pos, Neg represents positive and negative sentiment respectively. Details of total customers of the service provider provided by

**Table 2.** Association Rules Mined for Aircel service provider with actual customers Aircel had during the period

| | **Aug, 2012** | **Sep, 2012** | **Oct, 2012** | **Nov, 2012** | **Dec, 2012** |
|---|---|---|---|---|---|
| Top-K Rules | {Pos,Mis 54.1},<br>{Pos,Ser 36.67} | {Pos,Mis 47.4},<br>{Pos,Ser 34.54} | {Pos,Mis 48.5},<br>{Pos,Ser 38.37} | **{Neg,Ser 38.2}**<br>,{Pos,Mis 34.15} | **{Neg,Ser 50.3}**,<br>{Pos,Mis 44.18} |
| Total Customers | 65952244 | 66607361 | 66786295 | 65323317 | 63347284 |
| Change in Customers | 793717 | 655117 | 178934 | **-1462978** | **-1976033** |

| | **Jan, 2013** | **Feb, 2013** | **Mar, 2013** | **Apr, 2013** | |
|---|---|---|---|---|---|
| Top-K Rules | **{Neg,Ser 47.6}**,<br>{Pos,Ser 46.86} | **{Neg,Ser 57.96},**<br>**{Neg,Sat 42.7}** | {Pos,Mis 38.1},<br>{Pos,Ser 30.51} | {Pos,Mis 68},<br>**{Neg,Sat 35.32}** | |
| Total Customers | 61571291 | 60872785 | 60071967 | 60080216 | |
| Change in Customers | **-1775993** | **-698506** | **-800818** | **8249** | |

TRAI [13] are shown in next line. Few interesting rules obtained are in **bold**. For example, Top-K rule for Aircel in Nov,2012 , Dec,2012 , Jan,2013 and Feb,2013 months have higher negative confidence, due to which during these months total customers of Aircel might be dropped. Hence, churning factors during these months are {Service}, {Service}, {Service, Satisfaction} and {Satisfaction}. Similar, results are obtained for other service providers.

## 5 Conclusion and Future Work

In this paper we presented a technique which attempts to mine churning factors in telecommunication sector of India using social media analytics. In the preprocessing stage telecom specific tweets are pulled, cleaned for misspelled words, stemming is performed and data is translated into relational format. Further tweets are classified into three categories using lexicon based classifier. Finally ARM is applied to find the dominant churn factor out of a selected few factors as determined by domain expert. Results obtained are helpful in interpreting the customer satisfaction and also knowing the reason of customer dissatisfaction.

The results can be improved further by considering availability of inhouse data and performing deeper analysis of tweets for genuineness.

## References

1. Agrawal, R., Imieliski, T., Swami, A.: Mining association rules between sets of items in large databases. ACM SIGMOD Record 22(2), 207–216 (1993)
2. Cheung, K.W., Kwok, J.T., Law, M.H., Tsui, K.C.: Mining customer product ratings for personalized marketing. Decision Support Systems 35, 231–243 (2003)
3. Fournier-Viger, P., Wu, C.-W., Tseng, V.S.: Mining top-k association rules. In: Kosseim, L., Inkpen, D. (eds.) Canadian AI 2012. LNCS, vol. 7310, pp. 61–73. Springer, Heidelberg (2012)
4. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM (2004)
5. Hung, S.Y., Yen, D.C., Wang, H.Y.: Applying data mining to telecom churn management. Expert Systems with Applications 31(3), 515–524 (2006)
6. Hwang, H., Jung, T., Suh, E.: An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. Expert Systems with Applications 26(2), 181–188 (2004)
7. Keaveney, S.M.: Customer switching behavior in service industries: An exploratory study. Journal of Marketing 59(2), 71–82 (1995)
8. Kim, H.S., Yoon, C.H.: Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. Telecommunications Policy 28(9/10), 751–765 (2004)
9. Kim, S.Y., Jung, T.S., Suh, E.H., Hwang, H.S.: Customer segmentation and strategy development based on customer lifetime value: A case study. Expert Systems with Applications 31, 101–107 (2006)
10. Norvig, P.: How to write a spelling corrector, `http://norvig.com/spell-correct.html` (visited February 8, 2013)

11. Oghojafor, B., et al.: Discriminant Analysis of Factors Affecting Telecoms Customer Churn. International Journal of Business Administration 3(2) (2012)
12. Taboada, M., et al.: Lexicon-based methods for sentiment analysis. Computational Linguistics 37(2), 267–307 (2011)
13. Telecom Regulatory Authority of India, Telecom Subscription Data as on 30th September, Press Release No. 78/2013
14. Telecommunications in India, In Wikipedia, `http://en.wikipedia.org/wiki/Telecommunications_in_India` (retrieved January 24, 2014)
15. Wei, C.P., Chiu, I.T.: Turning telecommunications call details to churn prediction: A data mining approach. Expert Systems with Applications 23(2), 103–112 (2002)