**Mining Churning Factors in Telecommunication Sector of India Using Social Media Analytics**

Nitish Varshney, S.K. Gupta  
[nitish.mcs12@cse.iitd.ac.in](mailto:nitish.mcs12@cse.iitd.ac.in), [skg@cse.iitd.ac.in](mailto:skg@cse.iitd.ac.in)   
Department of Computer Science and Engineering, Indian Institute of Technology Delhi, India

*Abstract—* **As the subscriber base of telecommunication services reaches a saturation level, churning becomes a challenging problem with serious impact on revenues. We explore techniques like social media analytics and association rule mining to understand causes of churning in Indian context which may assist in churn reduction.**

**Enormous feeds are available on social media indicating a subscriber’s satisfaction or dissatisfaction. These user's opinions include various parameters which point towards churning and can be effectively analysed for understanding causes of churning. In this paper our experiments are based on data taken from twitter. In the first phase, only telecom specific tweets are pulled from twitter, which are further cleaned for misspelled words. Stemming is then performed to tackle ambiguity. After transforming tweets into relational format we classify them using lexicon based classifier. Association rule mining is then applied to find the dominant churn factor out of a selected few factors as determined by domain expert.**

*Index Terms*— **Twitter, sentiment analysis, social media analysis, telecommunication, churn pertaining factors, data mining.**

# Introduction

First mobile telephone service in India started in 1995. In recent years, mobile service usage 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 telephone users (both fixed and mobile phone) [1] 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 [2] and penetration rate was about 71%. Out of these 899.86 million subscribers, about 97 % utilize wireless services. We therefore focus our attention to wireless telecom services and mobile service providers as they are predominant in numbers.

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:

• Retention of customer subscriber’s base: As cost of acquisition of a new customer can substantially exceed the cost of retaining the existing customer [4].

• Increase in average revenue per customer.

There is a trade off in these business drivers. Customer retention depends on factors like call rate and quality of services, a service provider provides. 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. In 1996, Reichheld [3] estimated that, with an increase in customer retention rate by 5%, average net present value of a customer increases from 35% to 95% in different domains.

In telecommunication, customer movement from one service provider to other service provider is defined by term churn rate. Churn rate is the percentage of subscribers who discontinue services with a service provider and change their service provider willingly. Customer churn rate 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 [2]. This is relatively very high, specially when the aim is to retain the existing customers. In order to better manage customer churn, companies need to fully understand the factors leading to the customer churn. These problems affecting churn have not been fully addressed in the literature. In this paper, we present a data mining based approach to determine factors affecting churn in Indian Telecom sector.

# Indian Market

The telecommunications market in India is unlike any other country. There are about 15 mobile carriers. 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.

Indian market also differs from other developed markets as 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). 70.7 million multiple SIM handsets were shipped to India in first half of 2011, which accounts for a 69.1% market share of total handsets shipped [5]. Such handset provides a choice of up to eight mobile carriers within a single device. Mobile subscribers often switch service providers by changing SIM card. In INDIA Mobile Number Portability (MNP) was licensed in April 2009, and was launched in January 2011. It enabled subscribers to retain their telephone numbers when switching from one mobile service provider to another. It has the potential to deliver benefit to a consumer by loosening barriers to competition. Government has put up a nominal fee of Rs 19 for porting. Additionally, the process for porting is made quite simple. It has also been found that above cost is getting paid by those service providers in which customer is porting to. Such a situation leads to high willingness of a customer to churn.

# RELATED WORK

Existing literature on churn factors can be classified into two categories. Both of these categories attempt to predict customers getting ready to switch, understand why and connect with them to offer incentives to mitigate churn [6,8,14,10].

The first approach is based on large-scale actual customer transaction and billing data. This is proprietary data of a subscriber. Such studies use various machine learning techniques like Support Vector Machine [11], Regression [6,12], Decision Trees [6,12,14] etc. They rely on the rich subscriber call data records which were available inhouse. Such dataset describe calling behavior of a customer by providing information such as their voicemail plan, call lengths and usage patterns. This led to determination of customers 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 [10,17,28] avoiding privacy issues. These consider consumer’s perceptions of service experiences and intention not to churn. However, the survey data may not fully represent the customer’s actual future continued patronage decision.

However the above two approaches focus on predictive accuracy rather than descriptive explanation or reasons thereof. However our idea is it will be of interest to know what went wrong. Also such studies can not be directly useful in making a decision support system. For example, if Vodafone reduces 3G usage charges to 2paisa/10KB from 10paisa/10KB. 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.

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 best model to achieve above two objectives.

# Social Media as Sentiment Analyzer

The growing availability of the internet has marked the birth of a new type of society, where people can freely communicate and exchange ideas and opinions. It has led to development of a big database corpus which incorporates various data sources like social media (Facebook, Twitter, MySpace, Google+), news, blogs and etc. Because of free format of messages and an easy accessibility of microblogging platforms, Internet users tend to shift from traditional communication tools (such as blogs, mailing lists etc.). Among various chatter that people share and exchange on microblogging sites, there are a lot of personal thoughts to public statements about telecommunication services used by people. One of the most promising example of this approach can be “In 2011, President Obama used Twitter to ask Americans which government programs he should cut as he sought to reduce the federal deficit. Over the course of 2011, the White House account received a total of 125,832 Twitter replies from 42,902 people” [16]. It was very successful use of opinion mining. Opinion mining or sentiment analysis is an area within the field of text mining, to provide better method in analysing text in such dataset (data is unstructured). With the growing popularity and availability of such data sources, there is a surge of interest in the field that deals directly with labelling of opinions in positive and negative labels. Some examples of such applications are generating a summary of the important factors mentioned in the reviews of a product, comparing similar products, mining aggregate sentiments about politician to predict poll rating, forecast box office revenue of movies and predict actual financial performance of products in the market.

More precisely, we investigate the following problems:

* Does the churning in telecommunication sector is related to positive and negative sentiments present in social media feeds?
* Can churn determinants be mined from social media feeds?

# METHDOLOGY

## Data Corpus Creation

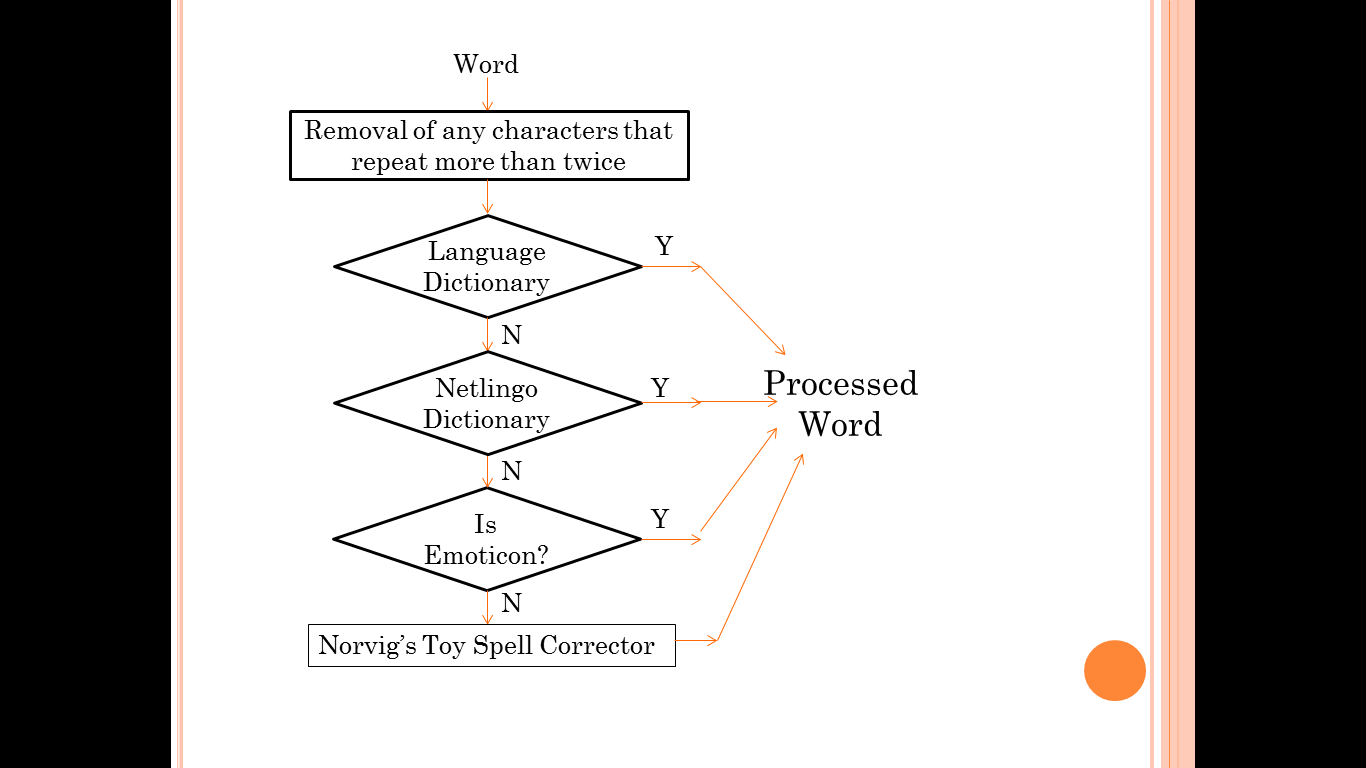
Twitter, the most popular social media platform, is used to build data corpus. It is publicly available and is very rich in content. So the collected corpus can be arbitrarily large. Additionally, twitter audience represents users from different social and interest groups providing a good sample.

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 these service providers name as keyword. REST API is used to pull tweets and retweets. The parameter “time” used to paginate through the results.

## 

## Data Cleansing

Twitter provided enormous amount of tweets when we queried it for above mentioned service providers using their names as keyword. However while fetching all such tweets, we found that a service provider can be active in different domains. So feeds needs to be filtered out.



Tweets also often have lots of grammatically misspelled words due to 140 character limit. And we are proposing to apply lexicon based text categorization method on such feeds which requires us to overcome such errors. However, most popular Norvig’s toy spell corrector [22] can’t be applied in its row form in our corpus as feeds often incorporates internet lingo, colloquial expressions and emoticons. As toy spell corrector’s base file ‘big.txt’ rarely contains any such word. It leads us to build 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 from Netlingo[[1]](#footnote-1), NoSlang[[2]](#footnote-2) and Webopedia[[3]](#footnote-3) (retrieved on February 10th 2013) for matching. Further emoticon list provided by DataGenetics[[4]](#footnote-4) is used. If there is no match found for the word then big.txt has been used to calculate edit distance as explained by Norvig [22]. Still our approach is not able capture phenomena such as sarcasm, irony, humor etc., but overall, data captured in such a manner is quite reasonable. Also our approach in current form works with English language only. However one can utilize other languages dictionary to enhance the model.

## Data Preprocessing

In addition to these, NLTK wordnet lemmatizer is used to stem and lemmatize each word. Also stop words from the feeds are removed to retrieve better picture from the data. Further to structure text and handle moderately-sized dataset, feeds are pushed into a MySQL RDBMS. MySQL has been used to store the data because of familiarity of writer with the language. In such a way we retrieved around 75K feeds.

## Constructing Knowledge Base (Aspects)

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 has been considered, which can influence a customer to churn to start with [20,21]. Price aspect incorporates call rates, pricing options. Service aspect incorporates call quality, coverage, billing and customer service. Satisfaction aspect incorporates already customer churned. Every other churn indicators had been put up in miscellaneous aspect. Miscellaneous aspect also incorporates “insulting” effects on the reputation of the company. If miscellaneous comes out to be most pertaining factor, then we have looked for trending topic among all tweets talking about miscellaneous aspect, for mining cause of unexpected event. Each aspect has some feature indicators like price aspect has rate, price, charge and tariff as feature indicators. These features are manually identified.

Unigrams, bigrams and trigrams are created manually to build a knowledge base from a bag of sentiment expressive words specially in telecommunication sector. Bigrams are of the form string-string like “tariff slashed”, “shittiest network”, “too-much interrupt”. Trigrams are constructed to capture cases like “ported to bsnl”. Wordnet database is also used to get sets of cognitive synonyms (synsets) of lexicons used in building unigrams, bigram and trigrams.

Hu and Liu [32] have provided a list of domain independent strongly positive and negative words. It contains few words which are not strongly positive / negative for telecommunication sector for example pretty, cheap, problem, tough etc. All such words are removed from the list to make it context specific. Few domain dependent strongly positive / negative word are also added to above list for example highspeed, lightning, slashes, flop etc. Further these words are also added to knowledge base.

A negated context also been taken care of as each word and associated with it polarity in a negated context become negated (e.g. “I like this book” and “I don’t like this book” are considered to be very similar by most commonly-used similarity measures, the only differing token, the negation term, forces the two sentences into opposite classes.)

## Aspect Based Text Classification

The collected dataset is cross-checked against knowledge base to extract aspect and sentiment. In sentiment extraction, we specifically tackle ternary classification task, which is given a tweet xi, the classifiers task is to predict yi {NEGATIVE, NEUTRAL, POSITIVE}. Neutral class has been incorporated in model to classify tweets that is not assigned positive / negative class.

Classification task can be broadly broken into two phases. In the first phase lexicons referring to target and aspect’s feature indicators are matched with words appearing in a tweet. The second phase calculates overall sentiment score based on summation of individual lexicon sentiment retrieved in first phase. Negating words are capable of reversing polarity of sentiment again. Our model assumes that customers providing negative tweets would lead to churn in few months (ranging from 1 to 3). To collect these sentiments, we followed the same procedure as in Taboada et al. [29]. Following are few examples of tweets fetched. Lexicons referring to target are *italicized* and aspect’s feature indicators are in **bold.**

aircel i *hate* u, no **wifi** since morning.

*pathetic* connection of airtel, planning to **switch to** BSNL.

bsnl **3g** is *decent* here speed wise in Kerala.

bsnl *launches* new call **tariff** *reduction* stvs for **2g 3g** prepaid customers bsnl telecom.

In first phase, lexicons referring to target are lexicons in knowledge base. And presence of aspect feature indicator indicates presence of aspect in a tweet. A tweet can have multiple aspects or it may have none. After annotation, tweets having positive or negative polarity and having price, service, satisfaction or miscellaneous aspect are used, other tweets are ignored.

Above described baseline classifier is not able to properly assign sentiment to tweets having multiple service providers. As inclusion of multiple service providers in a tweet presents semantic issues. For example tweets in figure 4 can be misclassified due to (shit, virgin) and (crap, bsnl) matching as a unigram-unigram pair. To tackle the issue our classifier has been modified to use a simple neighbourhood proximity based approach. For each occurrence of lexicon referring to target, each target words word-distance has been measured. Lexicon referring to target is attached to the target word occurring closer to the target. The sentiment classification is given by the sum of the weighted individual word sentiments for each target occurrence. Table 1 shows utilizing proximity improved accuracy of our model.

idea and bsnl both are *shit* i am pretty sure about this need to **shift to** virgin.

@yearning4d\_sky airtel is also *crap* i **mnped from** airtel **to** bsnl one of the best decisions taken in telecom industry.

## Association Rule Mining for most prevalent churning factors

Association rule mining is one of the most common data mining techniques, which is used to discover rules among multiple independent elements that co-occur frequently [27]. We found out that it can be used to mine out co-occurrence between an aspect and sentiment. Previous steps can be considered as pre-processing steps for mining most prevalent Association Rule 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.

Mostly Association Rule Mining has been done by varying support and confidence values. However such methodology suffers from either generating too many results or too few results neglecting valuable information. Taking this into consideration we have used Top-k Association Rule mining technique [31]. In particular only top most rule(s) has been considered for mining most pertaining churning factor(s). In this paper, we only report the best results that are based on the K value = 2.

# Experimental Evaluations

In this section we shall be looking at the results obtained after applying above approach on data collected in the previously specified time. In order to evaluate the model, we randomly selected 200 tweets and manually annotated them. Table 1 compares the F-value of the model by varying proximity techniques.

Table 2 shows the results .

# Validation of Results

There are several methods documented for validating such a model. Out of them we have used 70-30 cross validation method for setting relationship with TRAI data and our result And trending term analysis for

# CONCLUSION

In this paper we presented a way which tries to mine churning factors in telecommunication sector of India using social media analytics. Model mines reason of churning in the following stages. First, only telecom specific tweets are pulled from twitter so as to target the problem, which are further cleaned for misspelled words. Stemming is then performed to tackle ambiguity. Data in text form is then transformed in relational format which can be input for further analysis. Then the tweets are classified into three categories using lexicon based classifier. Finally Association rule mining is applied to find the dominant churn factor out of a selected few factors as determined by domain expert. Our evaluation showed an overall precision of %.

Write about decision making. Data remains constant unless there are certain forces that make it change.

Negative negative kahan kahan htaya aand kyun

Improvement after proximity

##### References

1. “Telecommunications in India”, In Wikipedia. Retrieved January 24, 2014, from <http://en.wikipedia.org/wiki/Telecommunications_in_India>.
2. TRAI, “Telecom Subscription Data” as on 30th september, press Release No. 78/2013.
3. Abdul Latif Sultan, September 2011 “The Impact Of Customer Relationship Marketing the luxury retail market of kuwait:A thematic analysis”, Master’s Thesis, University of Gloucestershire.
4. Shin-Yuan Hung , David C. Yen , Hsiu-Yu Wang, “Applying data mining to telecom churn management”, Expert Systems with Applications 31 (2006) 515–524.
5. Apurva Chaudhary, “Indian Mobile Handset Shipments At 102.43M During Jan-Jun 2012: Report”, Published October 30, 2012.
6. Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., & Kaushansky, H. (2000). “Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry”. IEEE Transactions on Neural Networks, 11(3), 690–696.
7. .
8. Wei, C. P., & Chiu, I. T. (2002). “Turning telecommunications call details to churn prediction: A data mining approach. Expert Systems with Applications” , 23 (2), 103–112.
9. .
10. Weerahandi, S., & Moitra, S. (1995). “Using survey data to predict adoption and switching for services. Journal of Marketing Research”, 32 (1), 85–96.
11. Cheung, K. W., Kwok, J. T., Law, M. H., & Tsui, K. C. (2003). Mining customer product ratings for personalized marketing. Decision Support Systems, 35, 231–243.
12. Hwang, H., Jung, T., & Suh, E. (2004). An LTV model and customer segmentation based on customer value: A case study on the wireless telecommunication industry. Expert Systems with Applications, 26, 181–188.
13. .
14. Kim, S. Y., Jung, T. S., Suh, E. H., & Hwang, H. S. (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. Expert Systems with Applications, 31 , 101–107.
15. .
16. Klint Finley, “Out in the Open: You Too Can Use Obama’s Secret Social Media Weapon”, Published October 14, 2013 on <http://www.wired.com/wiredenterprise/2013/10/thinkup/>.
17. Keaveney, S. M. (1995). Customer switching behavior in service industries: An exploratory study. Journal of Marketing , 59 (2), 71–82
18. .
19. .
20. Lee, G.H., Shin, S.M., & Park, Y.W. (2003). Studies on the determinants of subscribers’ churning in the high speed internet service market . KISDI report.
21. Kim, H. S., & Yoon, C. H. (2004). Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. Telecommunications Policy , 28 (9/10), 751–765.
22. Norvig, Peter. "How to write a spelling corrector.” Web Page, Visited February 8th 2013."
23. .
24. .
25. .
26. Jay M. Ponte and W. Bruce Croft. 1998. “A language modeling approach to information retrieval”. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '98). ACM, New York, NY, USA, 275-281.
27. Rakesh Agrawal, Tomasz Imielinski, Arun N. Swami: Mining Association Rules between Sets of Items in Large Databases. SIGMOD Conference 1993: 207-216
28. Oghojafor, Benjamin, et al. "Discriminant Analysis of Factors Affecting Telecoms Customer Churn." *International Journal of Business Administration* 3.2 (2012).
29. Maite Taboada , Julian Brooke , Milan Tofiloski , Kimberly Voll , Manfred Stede, Lexicon-based methods for sentiment analysis, Computational Linguistics, v.37 n.2, p.267-307, June 2011
30. Minqing Hu , Bing Liu, Mining and summarizing customer reviews, Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, August 22-25, 2004, Seattle, WA, USA
31. Fournier-Viger, Philippe, Cheng-Wei Wu, and Vincent S. Tseng. "Mining top-k association rules." *Advances in Artificial Intelligence*. Springer Berlin Heidelberg, 2012. 61-73.
32. Hu, Minqing, and Bing Liu. "Mining and summarizing customer reviews." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.

1. <http://www.netlingo.com> [↑](#footnote-ref-1)
2. <http://www.noslang.com/spelling.php>, [↑](#footnote-ref-2)
3. <http://www.webopedia.com/quick_ref/textmessageabbreviations.asp> [↑](#footnote-ref-3)
4. <http://datagenetics.com/blog/october52012/index.html> [↑](#footnote-ref-4)