Leverage Social Media Information to find leads for Credit Cards and Mortgages

Tarini. E & Purba. H. Rao
Great Lakes Institute of Management, Chennai

Abstract : The credit card industry is a multi-billion dollar industry. Credit card is one of the most significant sources of consumer credit. In recent decades, the increasing competitions in the credit card industry and enhanced regulations of interest rate caps have led to wide variations of products offered by credit card companies to their customers. Every year, a typical credit card company makes a huge number of credit offers, both to attract new customers and to encourage existing customers to increase their borrowings. However, the offers made are very random and thus a very few customers convert.

This study explores the possibilities of integrating user generated social media information from Facebook and Twitter using Radian6 into a single data mart and leverage it to identify selling opportunities for credit cards and mortgage. Credit card companies can thus target potential customers rather than making random offers. This increases the response rate.

Users post on Twitter and Facebook about various events in their lives. These tweets are extracted based on keywords that are identified based on the following two broad categories.

i) Life events such as marriage, childbirth, home, travel, job etc. (The assumption made here is that certain life events trigger a need for financial aid)
ii) Customers dissatisfied with competitors
iii) Text mining is the process of deriving high quality information from text. Categorization of text will be done using various algorithms such as SVM, Random forest, Maximum entropy, GLMNET etc.
iv) The meaning of these tweets and posts must be deciphered to identify leads who can be targeted for banking opportunities

Keywords: Credit Cards, Mortgages, Personal Loan.

Literature Review

Millions of people across the world participate actively on social network websites such as Facebook, Twitter and LinkedIn. Businesses have discovered that these social network sites offer channels to reach customers as well as prospects.

Data generated in social media can not only be used to detect fraud and understand consumer behavior, but it can also be used these days to find potential leads using text mining and logistic regression techniques.This paper joins a large empirical literature related to Text Analytics in the Banking sector, specifically Credit card industry. There has been a lot of work in the field of fraud detection and individual and group behavior of credit card holders. However, there has not been much research on targeting customers for credit cards using data mining techniques.

Credit card fraud detection has drawn a lot of interest in the research field and a number of techniques, with special emphasis data mining and distributed data mining have already been suggested.

Shailesh S Dhok and Dr. G R Bamnote, 2012,Credit Card Fraud Detection Using Hidden Markov Model, International Journal of soft computing and engineering

The paper by Shailesh S Dhok and Dr. G R Bamnote modelled the sequence of operations in credit card transaction processing using the Hidden Markov Model (HMM) technique. HMM was initially trained with the normal behavior of a cardholder. If an incoming credit card transaction was not accepted by the trained HMM with significantly high probability, it is considered to be fraudulent. Hidden Markov Model helps to obtain a high fraud coverage along with a low false alarm rate. It has the system is also scalable for handling large volumes of transactions.

Joseph Pun, Yuri Lawryshyn, **2012,** Improving Credit Card Fraud Detection using a Meta-Classification Strategy, International Journal of Computer Applications. In this study over 1million unique credit card transactions (11 months) from a large Canadian bank was analyzed. A meta-classifier model was applied to the transactions after being analyzed by the Bank's existing neural network based fraud detection algorithm. The meta-classifier model consisted of 3 base classifiers constructed using the decision tree, naive Bayesian, and k-nearest neighbor algorithms. The naive Bayesian algorithm was also used as the meta-level algorithm to combine the base classifier predictions to produce the final classifier. Results from the research showed that when a meta classifier was deployed, there was an improvement up to 28%.

TCS - White Paper, 2011 Leveraging Unstructured Text Data for Banksthe white paper published by TCS, authors **Lipika Dey and Sandeep Saxena** provided an imperative for adopting unstructured data mining by using some use cases and examples. It dealt with some unstructured data mining techniques that can be leveraged to achieve the bank's business objectives. The paper particularly discussed how advancements in Big Data technologies have now enabled banks to process large amounts of unstructured text data to meet their various objectives. Insights from such data can be used to understand customers, competitors, and also to improve the efficiency of the existing processes. It detailed about the need for banks to adopt text mining techniques to gain useful insights for improving risk management, customer relationship and improving performance management. Some of the methods dealt are Context based analysis, Free Text analysis and Statistical text analysis.

Financial Institutions reduce fraud risk with social media, 2011, Infosys insights documents publication dealt with leveraging social media information to reduce credit risk. The social graph was used to determine if prospective customers were connected to individuals or communities with a good history of credit. The graph was evaluated and analyzed to determine if the customer can be a trusted partner or not. The paper also discussed about leveraging information on social sites to find a customer's most recent location and venue detail using a mobile application by selecting them from a list of venues. Location information can enable a banker to detect fraud, especially in lost or cloned credit and debit cards cases.

Article – Banking on Customer Behavior, The article was an overview about "Analytics in the Banking sector". It studied how customer data in social media can be used to predict individual and group behavior. It also suggested that this data can be used in real time and automatically sending alerts when a pattern change indicates fraud, thus reducing risk. The paper specifically talked about predicting customer attrition, segmenting and targeting customers in new ways, identifying high value customers, migrating customer's to more profitable schemes, and personalize banking offers, services and rewards that match customer preference.

The Asian Banker – White paper, 2013, Leveraging Data and Analytics for customer centricity and Innovation, Asian Bank Research survey on use of data and analytics in banks in Asia pacific Asian Banker Research conducted a survey of senior bankers in Asia Pacific (Feb – March) 2013 on the use of analytics. The purpose of the survey was to determine how banks utilized analytics to gain customer insights across various sectors of banking. The survey covered mature markets such as Australia, Singapore, Hong Kong, and emerging markets such as Malaysia and Indonesia. It detailed how banks in the Asia Pacific region are increasing their focus on data and analytics, driven by the goal of achieving a competitive edge through greater understanding of cost behavior, customer needs, and better risk management. Superior analytical capabilities can help banks differentiate themselves through better customer insight, and there is an increasing focus on creating an intelligent and integrated multichannel platform. The White Paper represented the views of key financial services practitioners on the use of data and analytics in the banking and financial sectors.

Guangliat al, 2011, Credit card churn forecasting by logistic regression and decision tree two data mining algorithms were applied to build a churn prediction model using credit card data from a real Chinese bank. The contribution of four variable categories: customer information, card information, transaction activity and risk information were examined. The paper analyzed a process of dealing with the variables when data is obtained from a database instead of a survey. All the 135 variables were not considered in the model directly. Certain variables were selected from the perspective of correlation and economic sense. The paper also designed a misclassification cost measurement by taking into account the two types error and the economic sense, which was more suitable to evaluate the credit card churn prediction model. The algorithms used in this study include logistic regression and decision tree which were proven mature and powerful classification algorithms. The test result showed that regression performs a little better than decision tree.

Syeda, M., Zhang, Y. Q., and Pan, Y., **2002** Parallel Granular Networks for Fast Credit Card Fraud Detection, Proceedings of IEEE International Conference on Fuzzy Systems, pp. 572-577 (2002)used parallel granular neural networks to increase the speed of data mining and the knowledge discovery process in credit card fraud detection. A complete system was implemented for this purpose.

Stolfo, S. J., Fan, D. W., Lee, W., Prodromidis, A., and Chan, P. K., 2000. Cost-Based Modeling for Fraud and Intrusion Detection: Results from the JAM Project, Proceedings of DARPA Information Survivability Conference and Exposition, vol. 2 (2000), pp. 130-144suggested a credit card fraud detection system using meta learning techniques to learn the models of fraudulent credit card transactions. Meta learning is a generic strategy which provides means for combining and integrating a number of separately built models or classifiers. A Metaclassifier is thus trained on the correlation of the predictions of the base classifiers. The same group also worked on a cost-based model for intrusion and fraud detection.

Ghosh, S., and Reilly, D.L., **1994**. Credit Card Fraud Detection with a Neural-Network, 27th Hawaii International l Conference on Information Systems, vol. 3 (2003), pp. 621-630proposed credit card fraud detection using a neural network. They built a detection system, which was trained on a large sample of labeled credit card account transactions. These transactions contained example fraud cases due to lost cards, stolen cards, application fraud, mail-order fraud, counterfeit fraud and no received issue (NRI) fraud.

The banking and finance sector has undergone various changes in the way they conduct business and focus on modern technologies to compete in the market. The banking sector has now realized the importance of creating database and its utilization for the benefits of the bank in the area of strategic planning to survive. In the modern era, the technologies are advanced and it facilitates to generate, capture and store data have increased enormously. Data is a very valuable asset, especially in financial industries. The value of this asset can be evaluated only if the organization extracts the knowledge hidden in the raw data. The increase in the volume of data as a part of day-to-day operations and through other internal and external sources makes it mandatory to use mining to derive insights from data. Data mining technology also provides the facility to access the right information at the right time from huge volumes of raw data. In the banking sector, data mining technologies are adopted in various sectors especially in fraudulent transaction detections, customer segmentation and profitability, risk predictions, predictions on prices/values of different investment products, default prediction on pricing. However, the gap seems to be in identifying potential leads for credit cards and mortgages. Thus, the objective of this paper is to explore Text Mining as a technique to target potential customers.

Theory

Text mining, also referred to as text analytics, refers to the process of deriving high-quality information from text. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output.

- Users post tweets on Twitter everyday about various events in their lives
- Tweets will extracted based on keywords that are identified based on the following two broad categories
    - Life events such as Marriage, Childbirth, home, travel, job etc. (The assumption made here is that certain life events triggers a need for financial aid)
    - Customers dissatisfied with competitors

The various life event keywords are further subcategorized, for example

- The tweets consisting of words like marriage, married, marry, wedding, weds, engagement, engaged are grouped into marriage
- Tweets having words like pregnant, child birth, adoption, child care are grouped into child

Posts captured are grouped into one 18 categories for identification of leads

S.No.	Life Event(s)	Category	Example	CC	M	D&T	PL
1	Dissatisfied with Competitors	Indirect Need	Dear @ bankwest, you seriously serious faily at Customer service, SERIOUSLY	✓	✓	✓	✓
2	Express Need for Credit Card	Direct Need - CC	I want a credit card. I seriously need to go online shopping. Like now.	✓	×	×	×
3	Pregnant, Child Birth, Adoption, Child Care	Child	My wife is pregnant again and I'm surprised how excited I am. It's hard to convey this emotion in a tweet.	✓	×	×	✓
4	Getting Engaged Married	Marriage	Can anyone suggest a make-up artist/hair stylist for my wedding? I'm pretty fussy and don't want to end up looking overdone!	✓	✓	✓	✓
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
16	Got First Job / Promotion / New Job / Partime Job	Job	Got my first job today! MONEY MONEY!	✓	×	✓	×
17	Going on Oversease Holiday / Honeymoon	Travel	4 months untill I am in Cambodia:) Honestly I am a little worried because it is my first time goint overseas	✓	×	×	✓
18	Buying / Moving House, Relocating, Renovating	Home	Buying a house with @ mungusi	×	✓	×	×

* CC – Credit card, PL – Personal Loan, M – Mortgage, D&T – Deposits & Transaction. The meaning of these tweets must be deciphered to identify tweets that can be targeted for banking opportunities

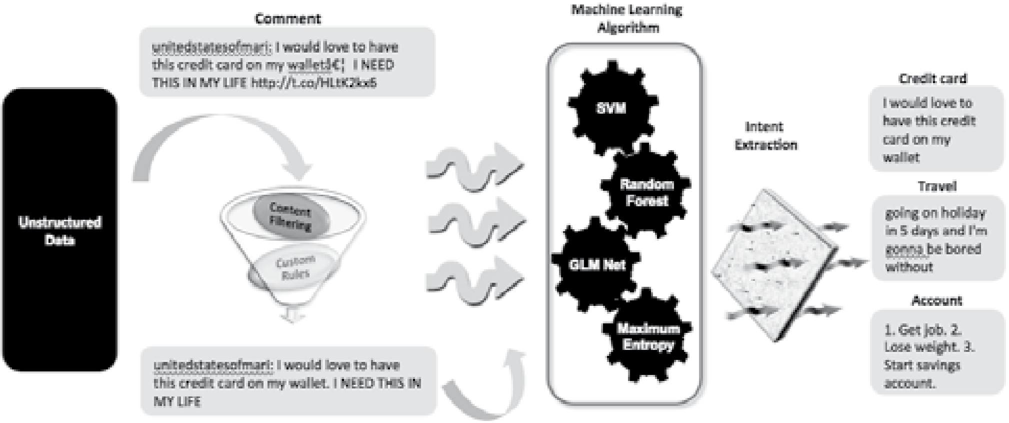
Text Categorization

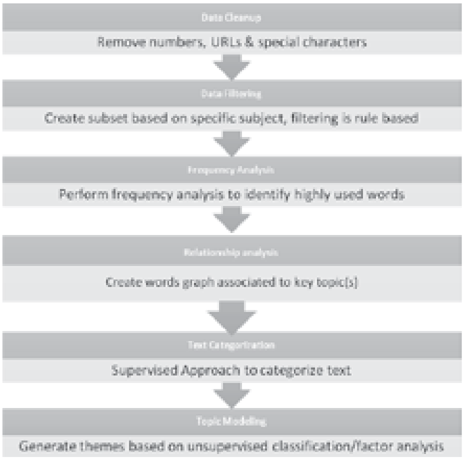
Text categorization is done in order to target customers under three different categories

– Loan
– Credit card
– Loan and Credit card

For Example:

Keywords pertaining to competitors will be mapped into credit cards with an assumption that the competitor doesn't have enough offers and the customers would be willing to change their credit cards. It is imperative to keep in mind that re tweets and tweets from people not directly related to the life event are not considered. The twitter profile details like name and location of the user is identified.These customers can then be targeted by re tweeting or by emails. If customers are existing customers, phone calls can also be made.


Comment
unitedstatesofmari: I would love to have this credit card on my walletâ€¦ I NEED THIS IN MY LIFE http://t.co/HLtK2kx6
Unstructured Data
Content Filtering
Custom Rules
unitedstatesofmari: I would love to have this credit card on my wallet. I NEED THIS IN MY LIFE
Machine Learning Algorithm
SVM
Random Forest
GLM Net
Maximum Entropy
Intent Extraction
Credit card
I would love to have this credit card on my wallet
Travel
going on holiday in 5 days and I'm gonna be bored without
Account
1. Get job. 2. Lose weight. 3. Start savings account.


Data Cleanup
Remove numbers, URLs & special characters
Data Filtering
Create subset based on specific subject, filtering is rule based
Frequency Analysis
Perform frequency analysis to identify highly used words
Relationship analysis
Create words graph associated to key topic(s)
Text Categorization
Supervised Approach to categorize text
Topic Modeling
Generate themes based on unsupervised classification/factor analysis

Method

Text Mining will be used to analyses the data. Detailed explanation of the methodology is as follows.Text mining tasks in the current scope include –

The tasks can be performed separately as per requirement and need not follow the flow as specified above.

Data Handling

Description

Data handling or cleanup or text pre-processing tasks include –

1. Removal of phone numbers
2. Removal of special characters (e.g. @, #, $, ~, &, *, +, -, \, “, etc.)
3. Removal of Stop words
4. Removal of URLs
5. Removal of white space
6. Removal of Email Address
7. Parts of Speech Extraction – Nouns , verbs , adjectives are extracted and grouped

Removal of phone numbers, emails, URLs

Regular expression has been created for the above and then matched using pattern matching to facilitate their removal from the text document.

Removal of Stop words

There is an inbuilt Stop word dictionary (English) available in R. So such words are searched throughout the text document and removed.

Stemming of words

Words are stemmed to their roots using Porter’s Stemming Algorithm.

Parts of Speech Extraction

The above process helps extract language entities viz. noun, verb, adjective, etc. from the text data. Apache OpenNLP library is being used to tag words with POS information.

EDA Text – Frequency Analysis

Description

Frequency Analysis does a detailed analysis of the data and performs some of the following action to the data

1. Removes sparse terms from it,
2. Consider words with a minimum threshold frequency for analysis
3. Finds the most frequently occurring unigrams or combination of two words i.e. bigrams
4. Finds the top terms

Frequency Analysis

Words satisfying the given user inputs are searched throughout the text document and can be done on the extracted parts of speech as well. The analysis can be performed on any of the POS or the entire text data.

EDA Text – Relationship Analysis

Description

Depending upon the Variable, Parts of speech and no. of Top Keywords selected from the navigation component, those exact no. of most frequently occurring keywords are generated by R and displayed. On selecting any top keyword we find the associated words in the text document. The Associated Score tells us the strength of association that existed between other words with the selected one. With the association result generated from R a Ravis graph is also created that depicts the same information in a graphical way.

Association Scores

Word Association scores are used to find those terms (words in this case) which correlate with the analyzed keyword. Association Score is the correlation between words in the text document. Example below illustrates the calculation.

Document1 - pen is good

Document2 - pen is bad

Document	Pen	Is	Good	Bad
D1	1	1	1	0
D2	1	1	0	1

Term Frequency

TF= No. of occurrences of a term in a document

Association score between "pen" and "is" is 1 in the above example. Words having association score less than 0 are not shown.

Text Categorization

Description

Text Categorization uses supervised machine learning algorithms that examine texts to provide new methods of navigating digitized information. Models can be created by training the data using algorithms like SVM, Random Forest, GLMNET, and Maximum Entropy.

Training Data

A training set needs to be created in order to build the predictive model. It contains an exhaustive set of all the variables (unique words in this case) along with the final result variable containing required categories (sentiment class in this case – positive, negative, neutral etc.).

Matrix Creation

The data is represented in the form of a matrix with rows corresponding to actual text documents and unique words as columns. The weights used to evaluate how important a word is to a document in a collection or corpus are –

1. **TfIdf** – Multiplicative product of Term Frequency and Inverse Document Frequency.

The term count in the given document is simply the number of times a given term appears in that document. This count is usually normalized to prevent a bias towards longer documents (which may have a higher term count regardless of the actual importance of that term in the document) to give a measure of the importance of the term t within the particular document d. Thus we have the term frequency tf (t, d), defined in the simplest case as the occurrence count of a term in a document. The inverse document frequency is a measure of the general importance of the term (obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient).

$$\mathrm{idf}(t) = \log \frac{|D|}{|\{d : t \in d\}|}$$ with

| D |: cardinality of D, or the total number of documents in the corpus

$|\{d : t \in d\}|$: Number of documents where the term t appears (i.e. $\mathrm{tf}(t, d) \neq 0$). If the term is not in the corpus, this will lead to a division-by-zero. It is therefore common to adjust the formula to

$1 + |\{d : t \in d\}|$.

Then

$\text{tf-idf}(t, d) = \mathrm{tf}(t, d) \times \mathrm{idf}(t)$

A high weight in tf–idf is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents; the weights hence tend to filter out common terms.

2. **Tf** – Simple term frequency as described above
3. **Binary** – 1 or 0 signifying presence or absence of a word in a document
   Test Dataset : It is used for checking the stability or accuracy of the model.

Heuristics : These are the various machine learning algorithms that can be used to build the model. Some of the popular ones are Naïve Bayesian, SVM, Decision Trees, etc.

Workflow

Raw Data/Training Corpus

Supervised Learning

ML Rules/Manually coded Rules

Classify Virgin Text

Machine Learning
SVM
Random Forest
GLMNET
MAXENT

SVM (Support Vector Machines)

In machine learning, **support vector machines** are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

Random Forest

Random forest (or **random forests**) is an ensembleclassifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees.

Each tree is constructed using the following algorithm

1. Let the number of training cases be *N*, and the number of variables in the classifier be *M*.

2. We are told the number *m* of input variables to be used to determine the decision at a node of the tree; *m* should be much less than *M*.

3. Choose a training set for this tree by choosing *n* times with replacement from all *N* available training cases (i.e. take a bootstrap sample). Use the rest of the cases to estimate the error of the tree, by predicting their classes.

4. For each node of the tree, randomly choose *m* variables on which to base the decision at that node. Calculate the best split based on these *m* variables in the training set.

5. Each tree is fully grown and not pruned (as may be done in constructing a normal tree classifier).

For prediction a new sample is pushed down the tree. It is assigned the label of the training sample in the terminal node it ends up in. This procedure is iterated over all trees in the ensemble, and the mode vote of all trees is reported as random forest prediction.

Maximum Entropy

Maximum entropy is a probability distribution estimation technique widely used for a variety of natural language tasks, such as language modeling, part-of-speech tagging, and text segmentation. The underlying principle of maximum entropy is that without external knowledge, one should prefer distributions that are uniform. Constraints on the distribution, derived from labeled training data, inform the technique where to be minimally non-uniform. The maximum entropy formulation has a unique solution which can be found by the improved iterative scaling algorithm. In this paper, maximum entropy is used for text classification by estimating the conditional distribution of the class variable given the document. In experiments on several text datasets we compare accuracy to naive Bayes and show that maximum entropy is sometimes significantly better, but also sometimes worse. Maximum entropy is a general technique for estimating probability distributions from data. The over-riding principle in maximum entropy is that when nothing is known, the distribution should be as uniform as possible, that is, have maximal entropy. Labeled training data is used to derive a set of constraints for the model that characterize the class-specific expectations for the distribution. Constraints are represented as expected values of features," any real-valued function of an example.

GLMNET

The models include linear regression, two class logistic regression, and multinomial regression problems while the penalties include the lasso, ridge regression and mixtures of the two (the elastic net). The algorithms use cyclical coordinate descent, computed along a regularization path.

Topic Modeling – Selection

Description

Topic modeling uses a suite of unsupervised new machine learning algorithms that examine texts to provide new methods of navigating digitized information. With topic models, we can search and explore a collection of documents based on the themes that run through it. We can zoom in and zoom out to find specific or broader themes; we can look at how those themes changed through time; we can see how themes are connected to each other.

Notation and Terminology

A word is the basic unit of discrete data, defined to be an item from a vocabulary indexed by {1… V}. We represent words using unit-basis vectors that have a single component equal to one and all other components equal to zero.

A document is a sequence of N words denoted by $w = (w_1, w2...w_n)$, where w_n is the nth word in the sequence.

A corpus is a collection of M documents denoted by $D = \{w_1, w2... wn\}$

Latent Dirichlet Allocation

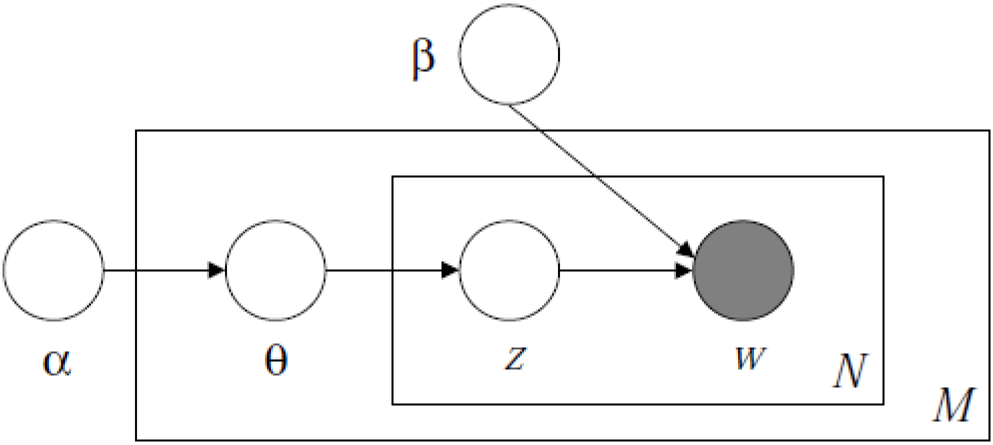
Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

LDA assumes the following generative process for each document w in a corpus D:

1. Choose $N \sim \text{Poisson}\ (\xi)$
2. Choose $\theta \sim \text{Dir}\ (\alpha)$
3. For each of the N words w_n,

(a) Choose a topic $z_n \sim \text{Multinomial}\ (\theta)$.

(b) Choose a word wn from $p\ (w_{nj}, z_n|\ \beta)$, a multinomial probability conditioned on the topic z_n



Graphical model representation of LDA

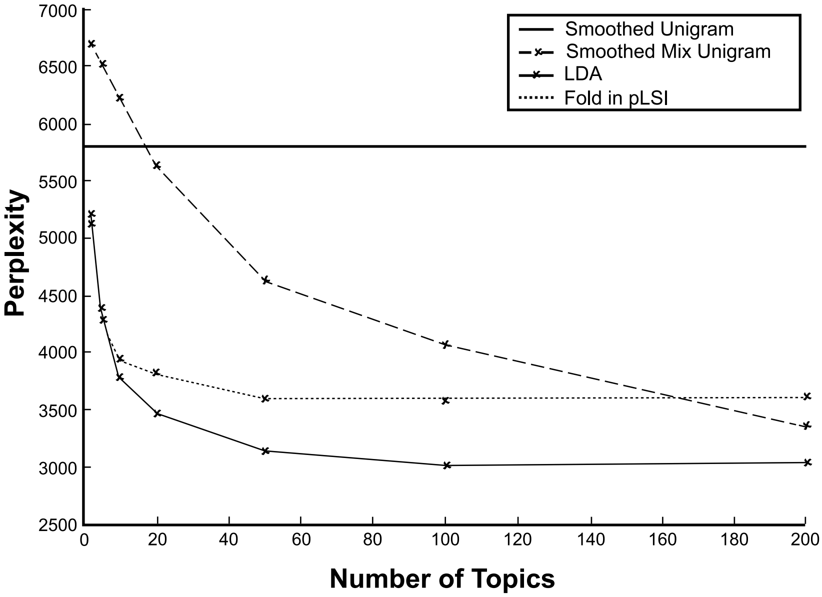
The boxes are “plates” representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

Model Selection

For fitting the LDA model or the CTM to a given document-term matrix the number of topics needs to be fixed a-priori. Additionally, estimation using Gibbs sampling requires specification of values for the parameters of the prior distributions. Griffths and Steyvers (2004) suggest a value of 50=k for alpha and 0.1 for beta. Because the number of topics is in general not known, models with several different numbers of topics are fitted and the optimal number is determined in a data-driven way. Model selection with respect to the number of topics is possible by splitting the data into training and test data sets.The perplexity is often used to evaluate the models on held-out data and is equivalent to the geometric mean per-word likelihood.

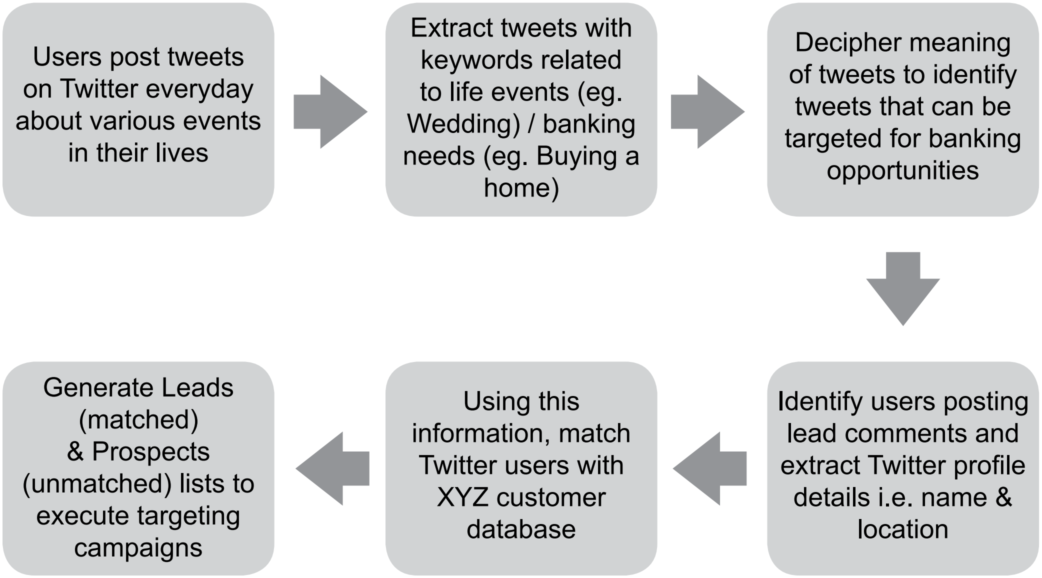
$$\text{Perplexity}(w) = \exp\left\{-\frac{\log(p(w))}{\sum_{d=1}^{D}\sum_{j=1}^{V} n^{(jd)}}\right\}$$

$n^{(jd)}$ denotes how often the j^{th} term occurred in the j^{th} document.



Dig: Perplexity results on the nematode (Top) and AP (Bottom) corpora for LDA< the unigram model and mixture of unigrams and pLSI

The overall approach for the execution for the execution of the project is as follows:



Users post tweets on Twitter everyday about various events in their lives

Extract tweets with keywords related to life events (eg. Wedding) / banking needs (eg. Buying a home)

Decipher meaning of tweets to identify tweets that can be targeted for banking opportunities

Identify users posting lead comments and extract Twitter profile details i.e. name & location

Using this information, match Twitter users with XYZ customer database

Generate Leads (matched) & Prospects (unmatched) lists to execute targeting campaigns

The tweets are extracted using Radian 6 based on keywords in the following 18 categories

Category No.	Category Name	Triggers
1	Indirect Need	Competition/Dissatisfied with Other Banks
2,3,4,5	Direct Need - Divided into 4 - CC, Mortgage, D & T, Personal Loan	Directly expressing a banking need
6	Child	Pregnant
6	Child	Child Birth
6	Child	Adoption
6	Child	Child care
7	Education	Graduation/Postgrad/getting admitted
7	Education	HECS
8	Marriage	Getting engaged
8	Marriage	Marriage
8	Marriage	Children getting engaged/married
9	Job	First job/salary
9	Job	Promotion/pay hike
9	Job	Job shift
9	Job	Apprenticeship
9	Job	Part time Job
10	Travel	Going Overseas/holiday
10	Travel	Honeymoon
11	Home	General home move
11	Home	Relocation
11	Home	Looking/Buying house
11	Home	Renovation
12	Health	Hospital/Health Problems
13	Loss	Loss events 1 - Assaulted/Burglary/looted/mugged
13	Loss	Loss events 2 - Natural disaster/earthquake/flood/tornado/fire/accident

14	Business	Getting an ABN
14	Business	Growing/expanding business
14	Business	Start a company/business
15	Selling Asset	Selling car/home
16	New vehicle	Vehicle buy/need
17	Divorce	Getting divorced
18	Education Overseas	Overseas study
18	Education Overseas	Exchange students

Once the tweets are extracted, it is important to decipher meaning out of these tweets. The process for this using text mining has been explained in detailed above.

Machine learning algorithms such as SVM, GLMNET, Maximum Entropy, Random forest are used to train a sample data set to identify leads classified as 1.

After the algorithms are trained, the master data set can be uploaded to predict and identify leads for a bank.Once the tweets are extracted, it is important to decipher meaning out of these tweets. Text mining will be used for the same.

A sample of tweets/textual data is manually classified by reading every statement. 1 meaning, the person is need of financial aid and 0 meaning the person in not in need of any financial aid and thus not a potential customer for the Bank.

Once the manual categorization is complete, the data is uploaded and various algorithms are used, such that these algorithms learn the basis on which the categorization is done. Post this, the algorithms predict and classify the manually categorized tweets as 0 and 1 based on their learning. The output displays the percentage of agreement.

For eg: if 90% is the agreement, the algorithm has classified 90% of the tweets correctly, i.e it matches the manual categorization.

Once a favorable number is obtained, usually above 80%, Machine learning algorithms such as SVM, GLMNET, Maximum Entropy, Random forest are used to predict and classify the master data set tweets on which there is no manual categorization.

All the four algorithms can be used for categorization. Thus a tweet can be classified as 1 if majority of the algorithms classify the tweet as 1 and 0 if majority classify it as 0.All the tweets which are classified also have the probability number with which they are classified as 1 or 0. Eg: If a tweet is classified as 1 with the probability 0.9, it means, that the algorithm predicts the category as 1 with 90% probability.

Hence the entire data set will be classified as 1 or 0, 1 implying the tweeters are leads for a bank.

Implications of the study

Majority of the population in developed countries in active on social media, sharing information, some of which can be triggers to banking needs. 50% of the Australians are active on Twitter and Facebook for example.

Through this study, the use of robust algorithms to decipher Social Media conversations & generate leads will be established. Companies can further deploy an end – end automated lead generation framework to generate leads on a daily basis

Social Media leads are recommended to be treated differently & learnt over a period of time. Companies can build a data mart capturing social media profiles of users

Text mining of social media data will increase the conversion rate of credit cards for banks as targeting will now by based on ends generated using social media data. The calls will not be random.

The cost of targeting is thus bound to decrease and the efficiently of targeting will increase coupled with an increase in the conversion rate.

Banks will be able to identify and interpret social media identities, conversations and events from Facebook and Twitter using Radian6 across the target population. Banks will have an integrated view of social and CRM data of its customers and prospects

References

Ghosh, S., and Reilly, D.L, 1994. Credit Card Fraud Detection with a Neural-Network, 27th Hawaii International l Conference on Information Systems, 3 (2003) : 621-630

Guangli at al,2011, Credit card churn forecasting by logistic regression and decision tree." Expert systems with applications: *An International Journal*. 38 (11) : 15273 – 15285.

Joseph Pun, Yuri Lawryshyn, 2012, Improving Credit Card Fraud detection. *International Journal of Computer Applications*. 56 (10)

Stolfo, S. J., Fan, D. W., Lee, W., Prodromidis, A., and Chan, P. K., 2000. Cost-Based Modeling for Fraud and Intrusion Detection: Results from the JAM Project, Proceedings of DARPA Information Survivability Conference and Exposition, 2 (2000) : 130-144

Syeda, M., Zhang, Y. Q., and Pan, Y., 2002 Parallel Granular Networks for Fast Credit Card Fraud Detection, Proceedings of IEEE International Conference on Fuzzy Systems : 572-577

Shailesh S Dhok and Bamnote.G.R 2012,Credit Card Fraud Detection Using Hidden Markov Model.

White Paper, Leveraging Data and Analytics for customer centricity and innovation, Asian Banker. http://www.theasianbanker.com/assets/media/dl/whitepaper/SAP_WP_2013_1.pdf

http://www.tcs.com/SiteCollectionDocuments/WhitePapers/BFS-Whitepaper-Unstructured-Text-Data-Banks-0613-2.pdf

http://www.infosys.com/FINsights/Documents/pdf/issue10/fraud-risk-social-media.pdf

http://www.ciosummits.com/media/solution_spotlight/EMC_Banking_on_Customer_Behavior.pdf (Article, Banking on customer behaviour)