SAS® GLOBAL FORUM 2018

USERS PROGRAM

Analysis of Unstructured Data: Topic Mining & Predictive Modeling using Text

Ravi Teja Allaparthi

April 8 – 11 | Denver, CO

#SASGF

Analysis of Unstructured Data: Topic Mining & Predictive Modeling using Text

Ravi Teja Allaparthi

Oklahoma State University – Stillwater

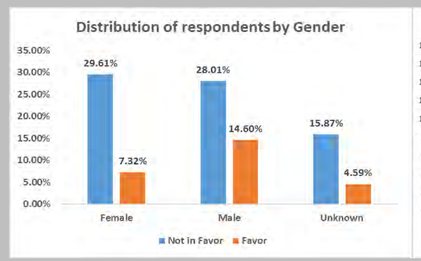
ABSTRACT

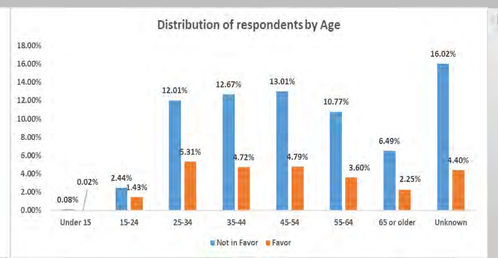
Eighty percent of the data generated in digital space is unstructured. While the amount of textual data is increasing rapidly, the ability to summarize and make sense of such data for making better business decisions remains challenging. This paper provides insights into how to analyze textual survey data for extracting public opinion from a huge collection of feedback forms and also to come up with rules in predicting the opinion of a user (support or oppose) . The dataset to be analyzed is collected based on the Toronto Casino Feedback Form which contains 17,000 records with information about open ended questions such as why do you not support the establishment of a casino as well as closed ended questions such as age group and gender

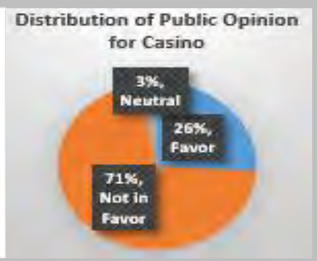
The primary objective is to understand and predict the opinion of a user towards the establishment of a casino by considering the survey filled by the user using unstructured data. Topics are extracted using the Text Topic node in SAS Enterprise Miner in identifying the public opinion. Text Rule Builder node is used to build rules that can differentiate the opinion of a user. From the analysis, we are able to identify that the majority of the public worried about gambling leading to addiction, increase in crime rate, traffic congestion and family and social relationships . People also opined positively about new jobs being created, tourism, additional revenue generated through taxes and tourism and increase in entertainment options.

DATA DISCOVERY

- City Council of Toronto wants to seek the public's input on the establishment of a casino as public consultation is an important part of it's decision making process and the data is collected based on the Toronto Casino Feedback Form. Following facts are discovered during the Exploratory Data Analysis
- 17,000 people responded to the survey. Out of them, 26% were in support to establish a Casino in Toronto, 71% were opposing the idea and the rest 3% were neutral. People with neutral feelings were excluded from the analysis
- 43% of the total respondents were males where as 37% were females.
- More than half of the total respondents were between age of 25-54

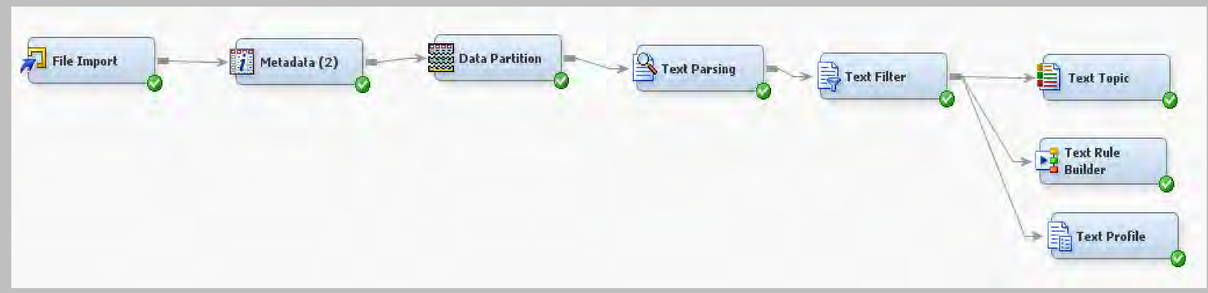






METHODOLOGY

- The data in the form of excel file is imported in SAS enterprise miner using file import node and analyzed using text parsing, text filter, text topic, text rule builder and text profile nodes



- Stratified Sampling is performed with Target as the Strata variable to segregate the dataset into Training (70%) and Validation (30%) datasets using the Data Partition node
- In order to process the text data, Text Parsing and Text Filter nodes are used. Text Filter Node applies filters to the text data and creates a transaction dataset that details which observations contain which words.

Text Topic Analysis

- The Text Topic Node uses the transaction data created by the Text Filter Node and creates topics, which are groups of words that are automatically determined to be related. Each topic is characterized by several key words, though the total number of terms in a topic is indicated in the Topics table. In the picture below you can see that Topic 3 is characterized by the keywords “gamble, encourage, addictive, poor, people” but contains a total of 35 different words.

Topic ID	Document Cutoff	Topic	Term Cutoff	Number of Terms
1	0.331	+job,+create,+existing job,union,+sustain	0.018	12
2	0.269	+revenue,+tax revenue,+city,+source,+additional reve...	0.018	15
3	0.168	+gamble,+encourage,addictive,+poor,+people	0.019	35
4	0.241	tourism,+job,+revenue,+help,+convention	0.018	6
5	0.160	+entertainment,+entertainment option,+option,+busi...	0.018	18
6	0.151	+city,+casino,+help,image,+culture	0.019	40

Analysis of Unstructured Data: Topic Mining & Predictive Modeling using Text

Ravi Teja Allaparthi

Oklahoma State University – Stillwater

METHODOLOGY (FOLLOWED)

Text Profiling

- The Text Profile node was used to profile a target variable using terms found in the documents. For each level of a target variable, the node outputs a list of terms from the collection that describe the level. Profiled Variables table in the results window of the Text Profile node indicates that the users who does not support the idea of establishing the casino in Toronto used the terms "gamble, crime, addiction, problem, traffic" as shown below.

Target Value	Term 1	Term 2	Term 3	Term 4	Term 5
0	Gamble	Crime	Addiction	Problem	Traffic
1	Job	Tourism	Revenue	Employment	Creation

Text Rule Builder

- The Text Rule Builder node creates an ordered set of rules from small subsets of terms that together are useful in describing and predicting a target variable. Each rule in the set is associated with a specific target category that consists of a conjunction that indicates the presence or absence of one or a small subset of terms as show below. Complex set of rules can be built by altering the Generalization error, Purity of rules and Exhaustiveness options but we need to be careful as complex set of rules leads to overfitting.

Target Value	Rule	Precision
0	addiction & ~tourism &~favor	99.5%
0	gamble & ~job & ~tourism & ~keep & ~crime	99.6%
0	crime & ~help & ~entertainment	99.4%

- 13.2 % Validation misclassification rate of the predictive model built by Text Builder node seems reasonable and further the accuracy of the model can be improved by working through the Change Target Values option in the Text Rule Builder node

Fit statistics	Train	Validation
Average square error	2.4%	2.5%
Misclasssification rate	11.2%	13.2%

RESULTS

- Like two sides of a coin, public always had two diverse opinions on a topic. In this survey, majority of the public opposed the idea of establishing a casino in Toronto. Prominent topics that worry the public were mentioned below.
  - Gambling leads to addiction which can degrade the economic conditions of the people
  - Increase in crime rate
  - Traffic Congestion
  - Destroys the family and social relationships
- Public also supported the idea on the basis of following topics.
  - New jobs will be created.
  - Tourism
  - Additional Revenue will be generated through taxes and tourism
  - Increase in entertainment options
- In addition to identifying the public opinion the rules formed from mining the textual data can be easily understood to segregate the user's perception towards the outcome.

DISCUSSION

- By analyzing the textual data, we were able to answer questions such as why public was opposing the establishment of Casino in Toronto and how the comments provided by users helped in predicting an user's perception on establishing a Casino. The Text mining analysis done using SAS Enterprise Miner was able to give us insights about these questions. As a general rule of thumb, we recommend using Text Topic node to identify the public opinion in similar contexts.

REFERENCES

- https://communities.sas.com/t5/SAS-Communities-Library/Tip-How-to-Create-Models-with-Text-Data-Using-SAS-Enterprise/ta-p/221216
- http://www1.toronto.ca/wps/portal/contentonly?vgnextoid=26136f951f6dd310VgnVCM10000071d60f89RCRD
- Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS by Goutam Chakraborty, Murali Pagolu, Satish Garla
- https://support.sas.com/documentation/cdl/en/tmgs/67510/PDF/default/tmgs.pdf
- http://www.wellesleyinstitute.com/wp-content/uploads/2013/01/Real-Cost-of-a-Casino_Wellesley-Institute_2013.pdf

SAS® GLOBAL FORUM 2018

April 8 – 11 | Denver, CO
Colorado Convention Center

#SASGF