US 20160299955A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2016/0299955 A1**

Jain et al. (43) **Pub. Date: Oct. 13, 2016**

(54) **TEXT MINING SYSTEM AND TOOL**

(71) Applicant: **MuSigma Business Solutions Pvt. Ltd.**, Bangalore (IN)

(72) Inventors: **Gaurav Jain**, Bangalore (IN); **Deepinder Dhingra**, Bangalore (IN); **Zubin Dowlaty**, Georgetown, TN (US); **Bharat Upadrasta**, Bangalore (IN)

(21) Appl. No.: **14/828,390**

(22) Filed: **Aug. 17, 2015**

(30) **Foreign Application Priority Data**

Apr. 10, 2015 (IN) 1879/CHE/2015

Publication Classification

(51) **Int. Cl.**

G06F 17/30 (2006.01)

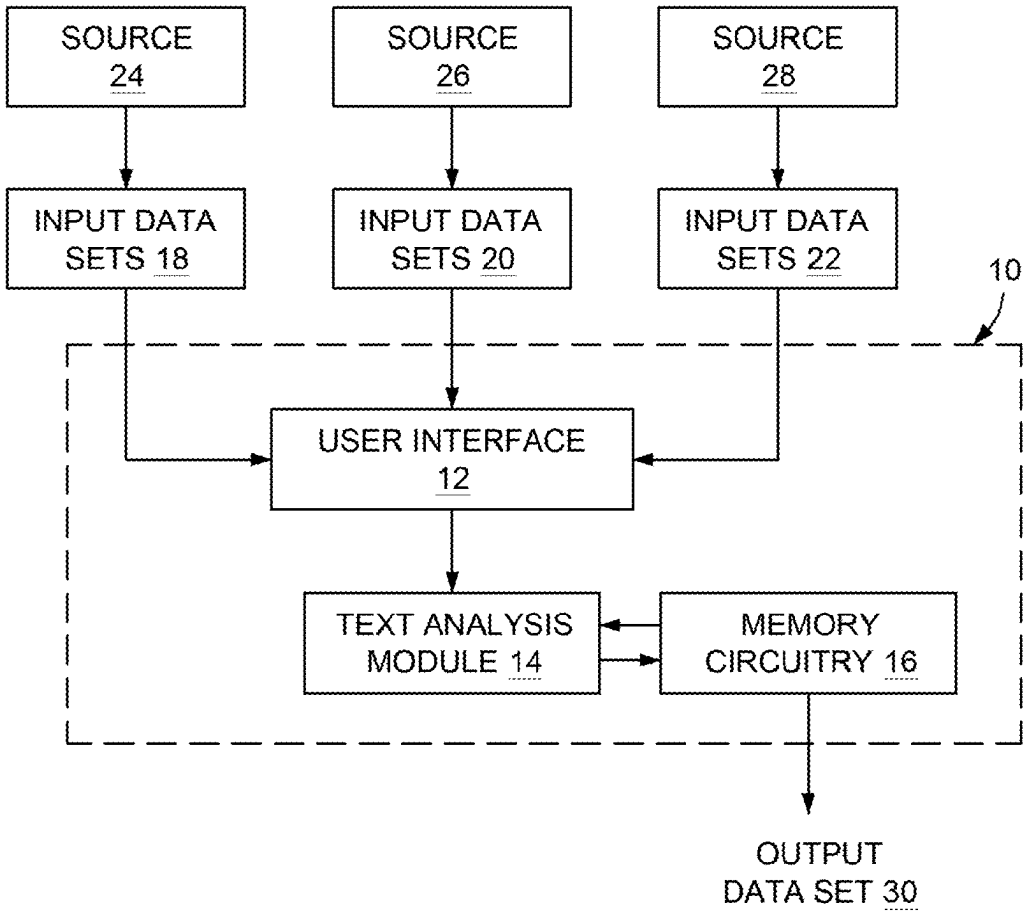
G06F 3/0484 (2006.01)

(52) **U.S. Cl.**

CPC ***G06F 17/30539*** (2013.01); ***G06F 3/0484*** (2013.01); ***G06F 17/30312*** (2013.01); ***G06F 17/30554*** (2013.01)

(57) **ABSTRACT**

A text mining system for extracting relevant text from a plurality of input data sets is provided. The text mining system includes an input interface module configured to enable one or more users to select a plurality of sources for a plurality of input data sets. The text mining system also includes a text analysis module configured to receive the plurality of input data sets and to generate an output data set by analyzing the plurality of input data sets. The text analysis module includes a data handling module configured to convert the plurality of input data sets to an analytics text set. The text analysis module also includes an exploratory analysis module configured to determine a plurality of correlations within the analytics text set. The text analysis module further includes a topic modeling module configured to identify a plurality of topics repeatedly occurring in the analytics text set and a reporting module configured to generate a plurality of reports for the text analysis module. The text mining system further includes memory circuitry configured to store the plurality of input data sets, the analytics text set and the output data set.



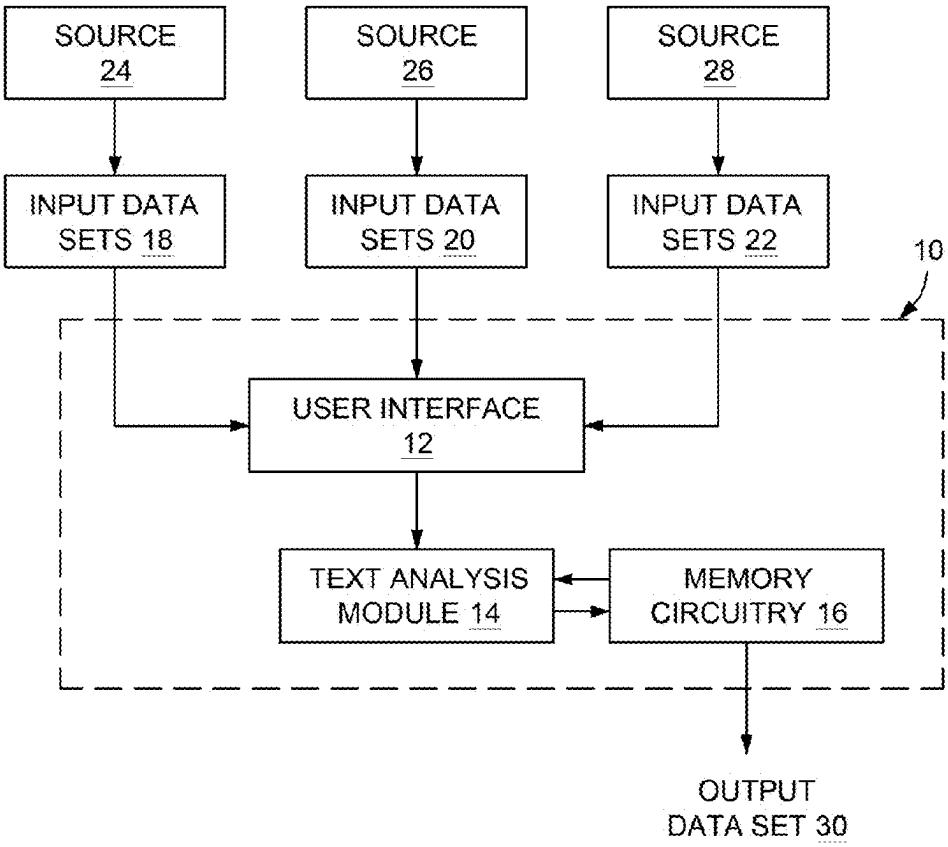

SOURCE 24
SOURCE 26
SOURCE 28
INPUT DATA SETS 18
INPUT DATA SETS 20
INPUT DATA SETS 22
10
USER INTERFACE 12
TEXT ANALYSIS MODULE 14
MEMORY CIRCUITRY 16
OUTPUT DATA SET 30

FIG. 1

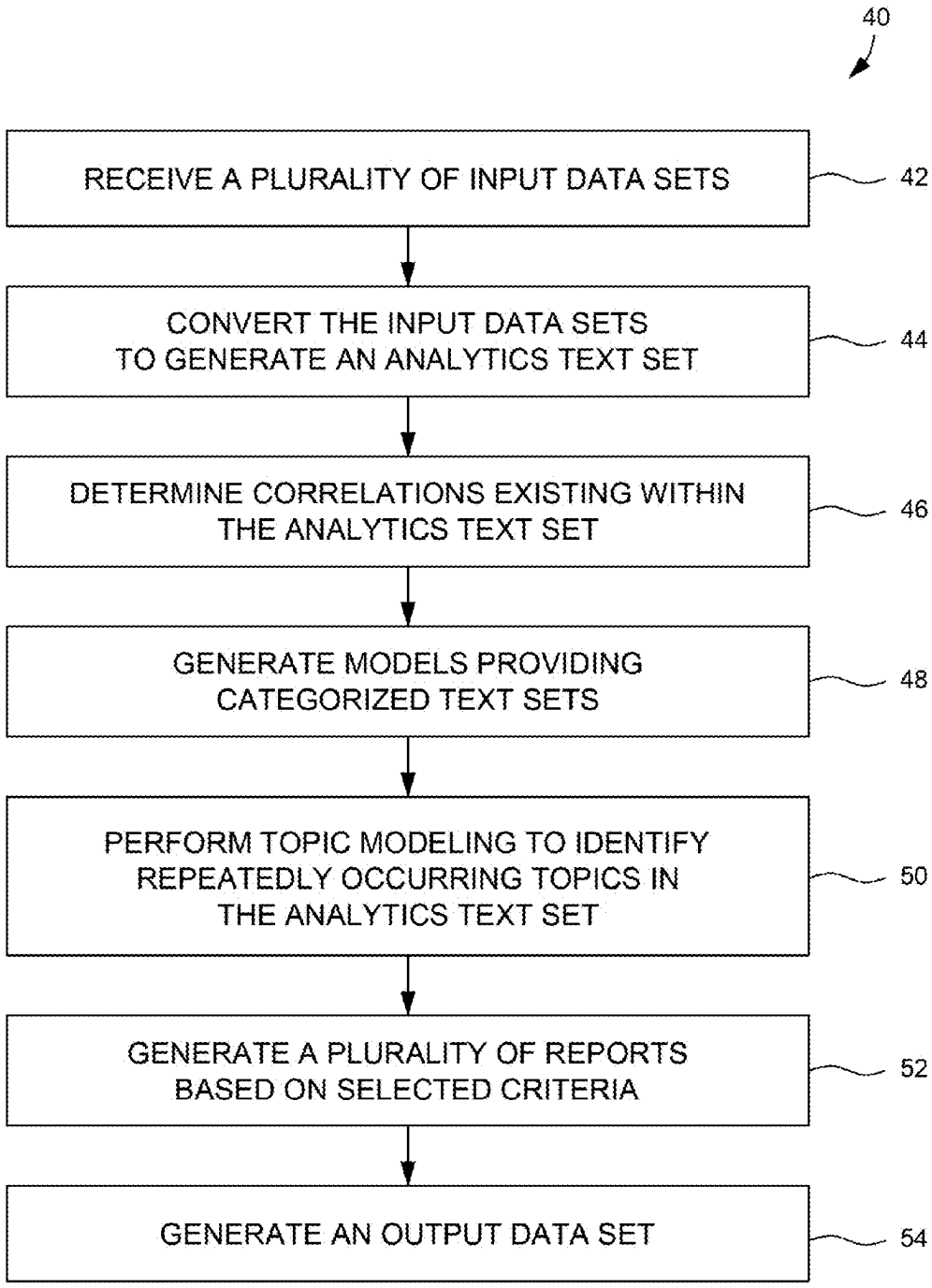

40
RECEIVE A PLURALITY OF INPUT DATA SETS
42
CONVERT THE INPUT DATA SETS TO GENERATE AN ANALYTICS TEXT SET
44
DETERMINE CORRELATIONS EXISTING WITHIN THE ANALYTICS TEXT SET
46
GENERATE MODELS PROVIDING CATEGORIZED TEXT SETS
48
PERFORM TOPIC MODELING TO IDENTIFY REPEATEDLY OCCURRING TOPICS IN THE ANALYTICS TEXT SET
50
GENERATE A PLURALITY OF REPORTS BASED ON SELECTED CRITERIA
52
GENERATE AN OUTPUT DATA SET
54

FIG. 2

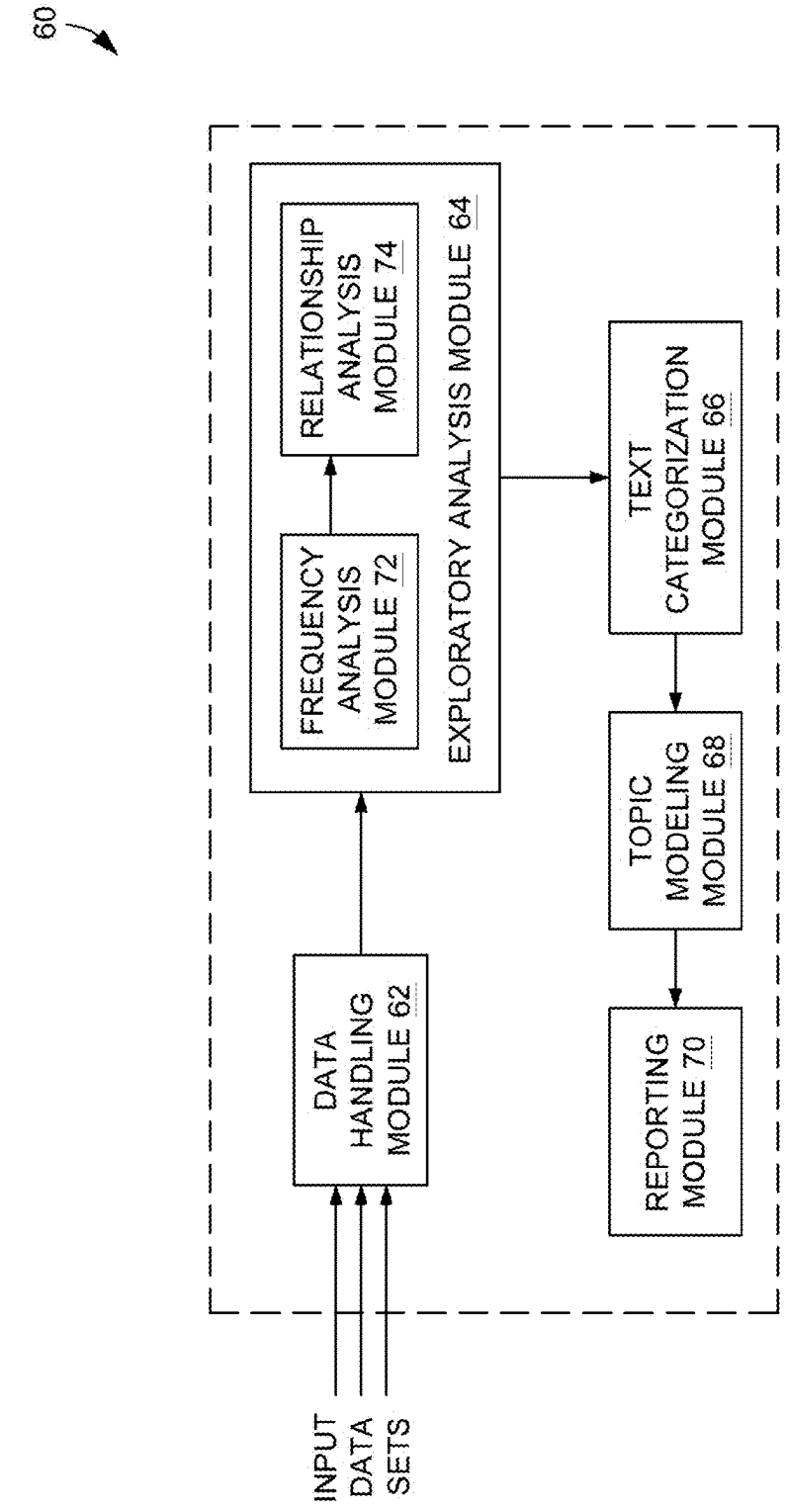

60
INPUT DATA SETS
DATA HANDLING MODULE 62
EXPLORATORY ANALYSIS MODULE 64
FREQUENCY ANALYSIS MODULE 72
RELATIONSHIP ANALYSIS MODULE 74
TEXT CATEGORIZATION MODULE 66
TOPIC MODELING MODULE 68
REPORTING MODULE 70

FIG. 3

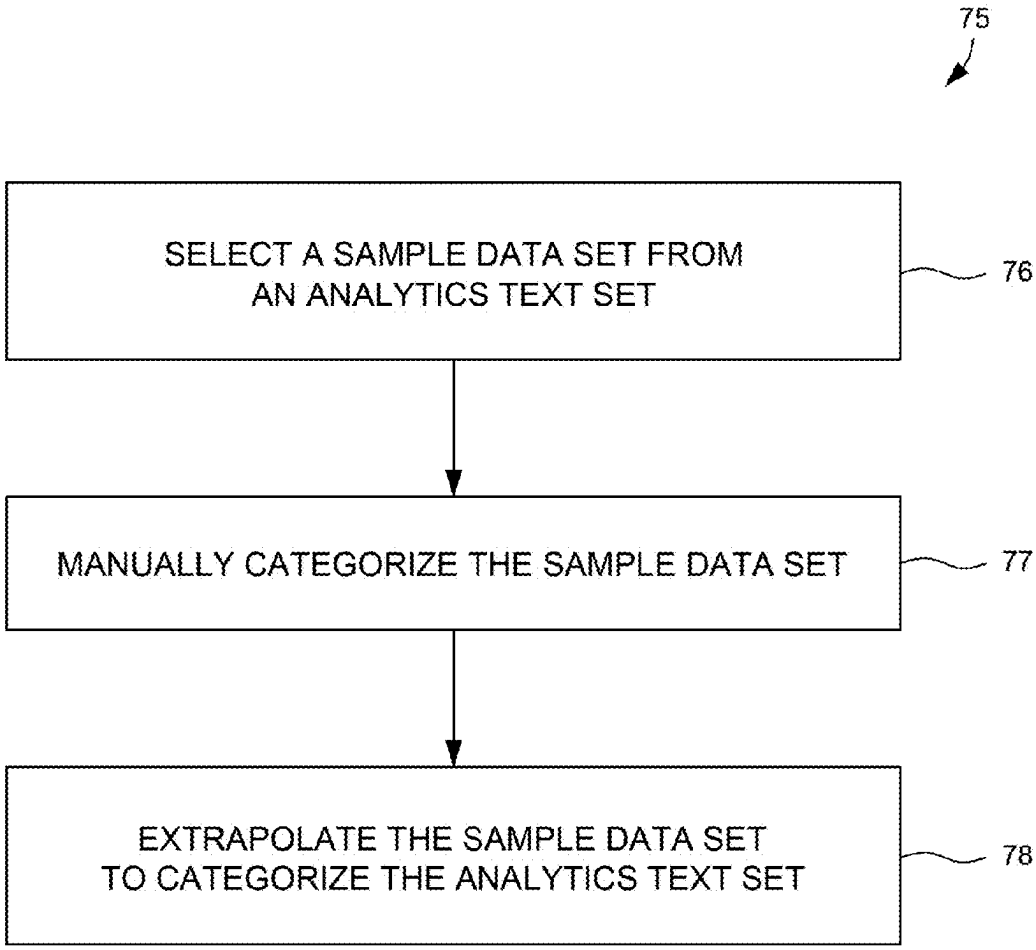

75
SELECT A SAMPLE DATA SET FROM AN ANALYTICS TEXT SET
76
MANUALLY CATEGORIZE THE SAMPLE DATA SET
77
EXTRAPOLATE THE SAMPLE DATA SET TO CATEGORIZE THE ANALYTICS TEXT SET
78

FIG. 4

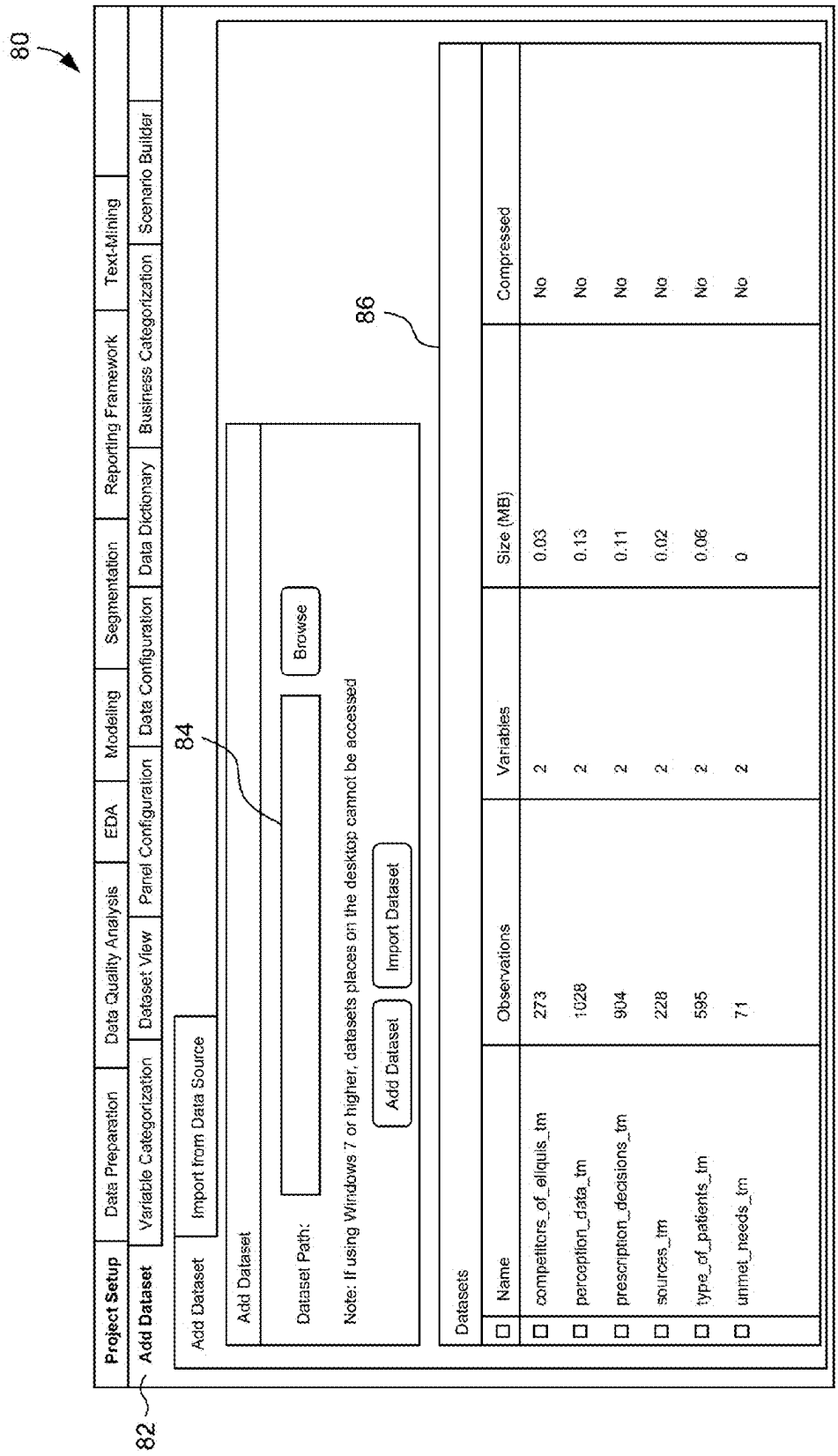

80
82
84
86
Project Setup
Data Preparation
Data Quality Analysis
EDA
Modeling
Segmentation
Reporting Framework
Text-Mining
Add Dataset
Variable Categorization
Dataset View
Panel Configuration
Data Configuration
Data Dictionary
Business Categorization
Scenario Builder
Add Dataset
Import from Data Source
Add Dataset
Dataset Path:
Browse
Note: If using Windows 7 or higher, datasets places on the desktop cannot be accessed
Add Dataset
Import Dataset
Datasets
Name
Observations
Variables
Size (MB)
Compressed
competitors_of_eliquis_tm
273
2
0.03
No
perception_data_tm
1028
2
0.13
No
prescription_decisions_tm
904
2
0.11
No
sources_tm
228
2
0.02
No
type_of_patients_tm
595
2
0.06
No
unmet_needs_tm
71
2
0
No

FIG. 5

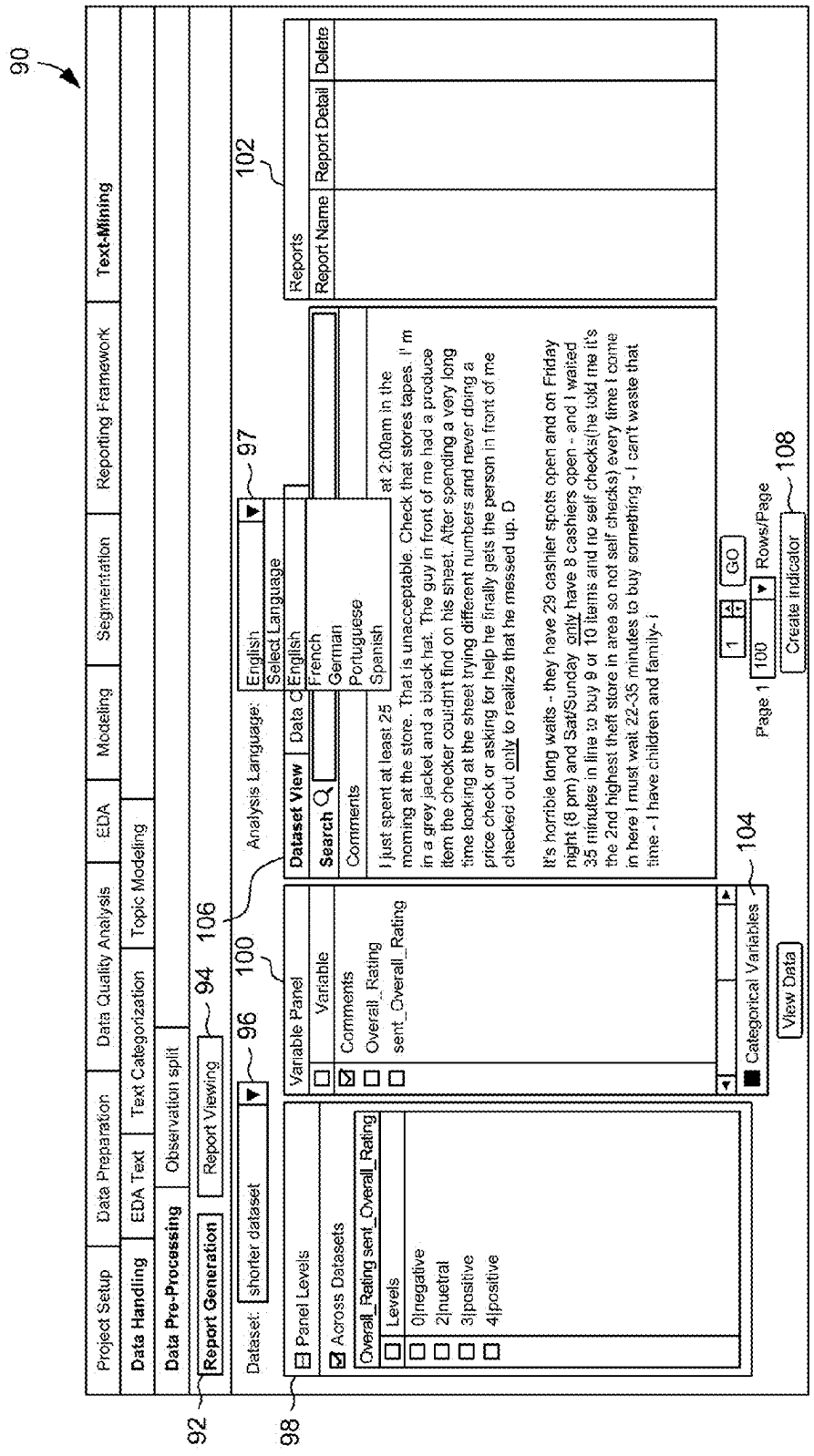

90
Project Setup
Data Preparation
Data Quality Analysis
EDA
Modeling
Segmentation
Reporting Framework
Text-Mining
Data Handling
EDA Text
Text Categorization
Topic Modeling
Data Pre-Processing
Observation split
Report Generation
Report Viewing
92
94
96
98
100
102
104
106
108
Dataset: shorter dataset
Analysis Language:
English
Select Language
97
Dataset View
Data C
English
French
German
Portuguese
Spanish
Search
Comments
I just spent at least 25 morning at the store. That is unacceptable. Check that stores tapes. I' m in a grey jacket and a black hat. The guy in front of me had a produce item the checker couldn't find on his sheet. After spending a very long time looking at the sheet trying different numbers and never doing a price check or asking for help he finally gets the person in front of me checked out only to realize that he messed up. D
at 2:00am in the
It's horrible long waits - they have 29 cashier spots open and on Friday night (8 pm) and Sat/Sunday only have 8 cashiers open - and I waited 35 minutes in line to buy 9 or 10 items and no self checks(the told me it's the 2nd highest theft store in area so not self checks) every time I come in here I must wait 22-35 minutes to buy something - I can't waste that time - I have children and family- (
Reports
Report Name
Report Detail
Delete
Panel Levels
Across Datasets
Overall_Rating sent_Overall_Rating
Levels
0|negative
2|nuetral
3|positive
4|positive
Variable Panel
Variable
Comments
Overall_Rating
sent_Overall_Rating
Categorical Variables
View Data
1
GO
Page 1
100
Rows/Page
Create indicator

FIG. 6A

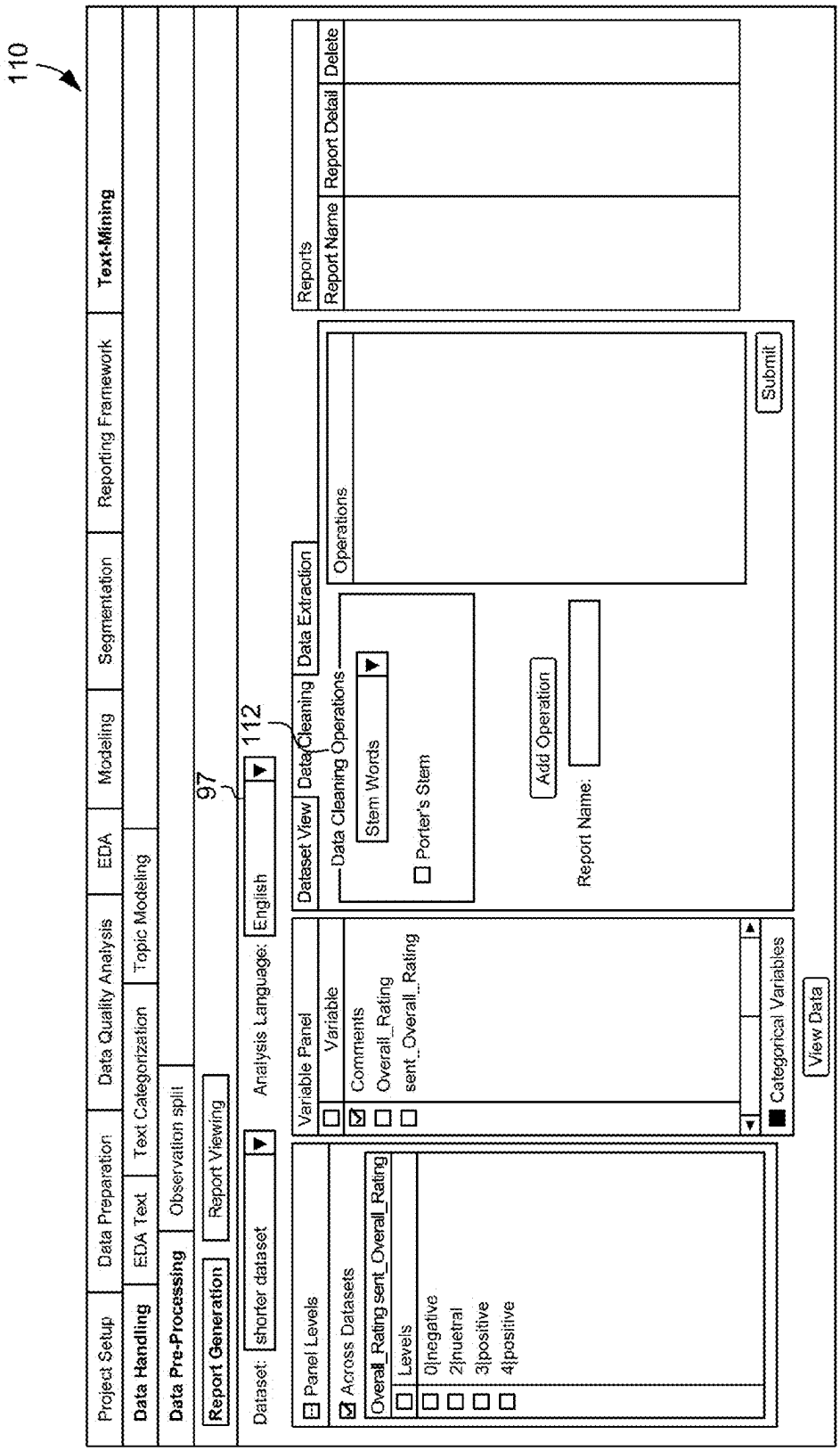

110
Project Setup
Data Preparation
Data Quality Analysis
EDA
Modeling
Segmentation
Reporting Framework
Text-Mining
Data Handling
EDA Text
Text Categorization
Topic Modeling
Data Pre-Processing
Observation split
97
Report Generation
Report Viewing
Dataset:
shorter dataset
Analysis Language:
English
112
Panel Levels
Across Datasets
Overall_Rating sent_Overall_Rating
Levels
0|negative
2|nuetral
3|positive
4|positive
Variable Panel
Variable
Comments
Overall_Rating
sent_Overall_Rating
Categorical Variables
View Data
Dataset View
Data Cleaning
Data Extraction
Data Cleaning Operations
Stem Words
Porter's Stem
Add Operation
Report Name:
Operations
Submit
Reports
Report Name
Report Detail
Delete

FIG. 6B

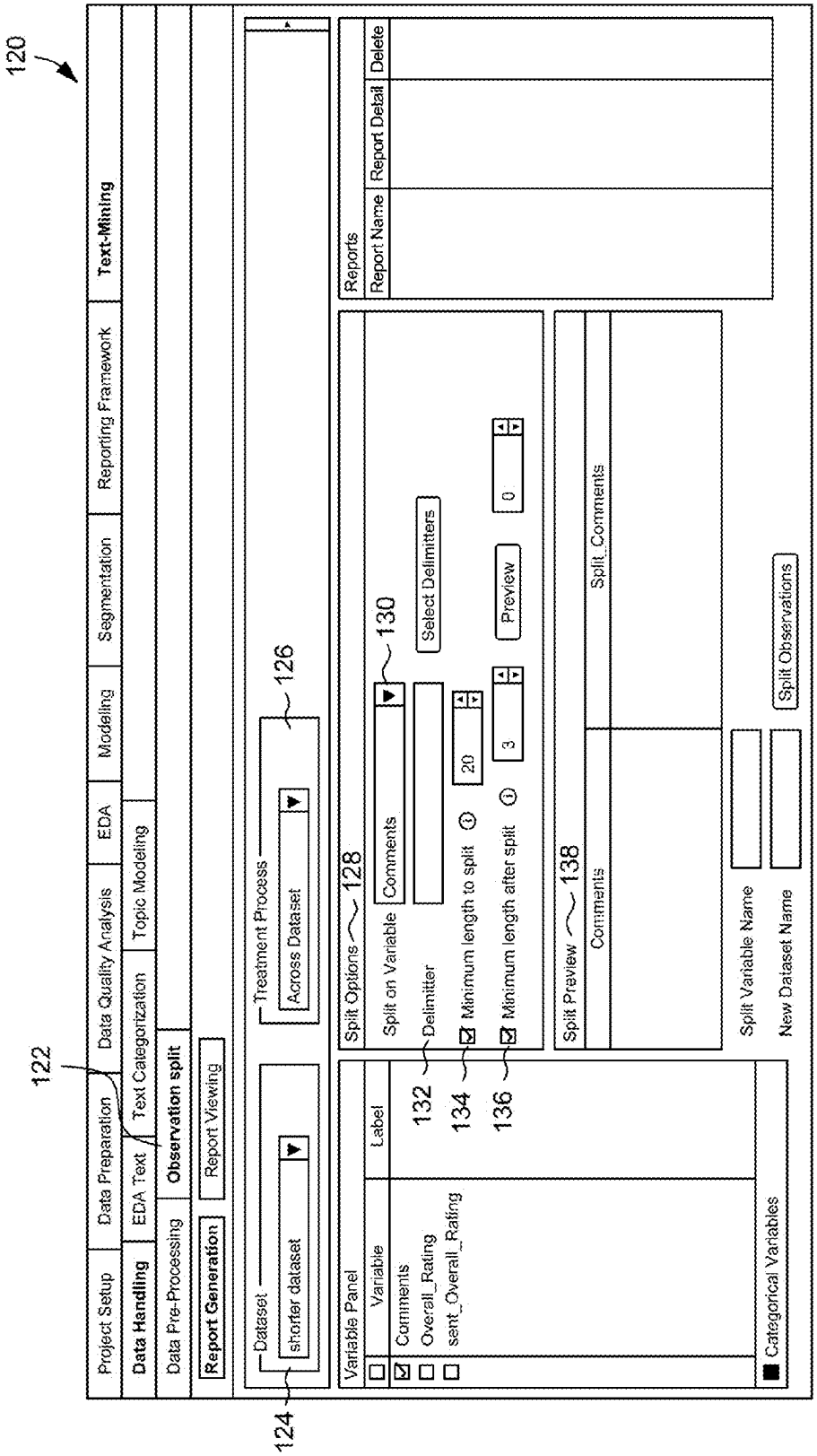

120
122
124
126
128
130
132
134
136
138
Project Setup
Data Preparation
Data Quality Analysis
EDA
Modeling
Segmentation
Reporting Framework
Text-Mining
Data Handling
EDA Text
Text Categorization
Topic Modeling
Data Pre-Processing
Observation split
Report Generation
Report Viewing
Dataset
shorter dataset
Treatment Process
Across Dataset
Variable Panel
Variable
Label
Comments
Overall_Rating
sent_Overall_Rating
Categorical Variables
Split Options
Split on Variable
Comments
Delimiter
Select Delimiters
Minimum length to split
20
Minimum length after split
3
Preview
0
Split Variable Name
New Dataset Name
Split Observations
Split Preview
Comments
Split_Comments
Reports
Report Name
Report Detail
Delete

FIG. 6C

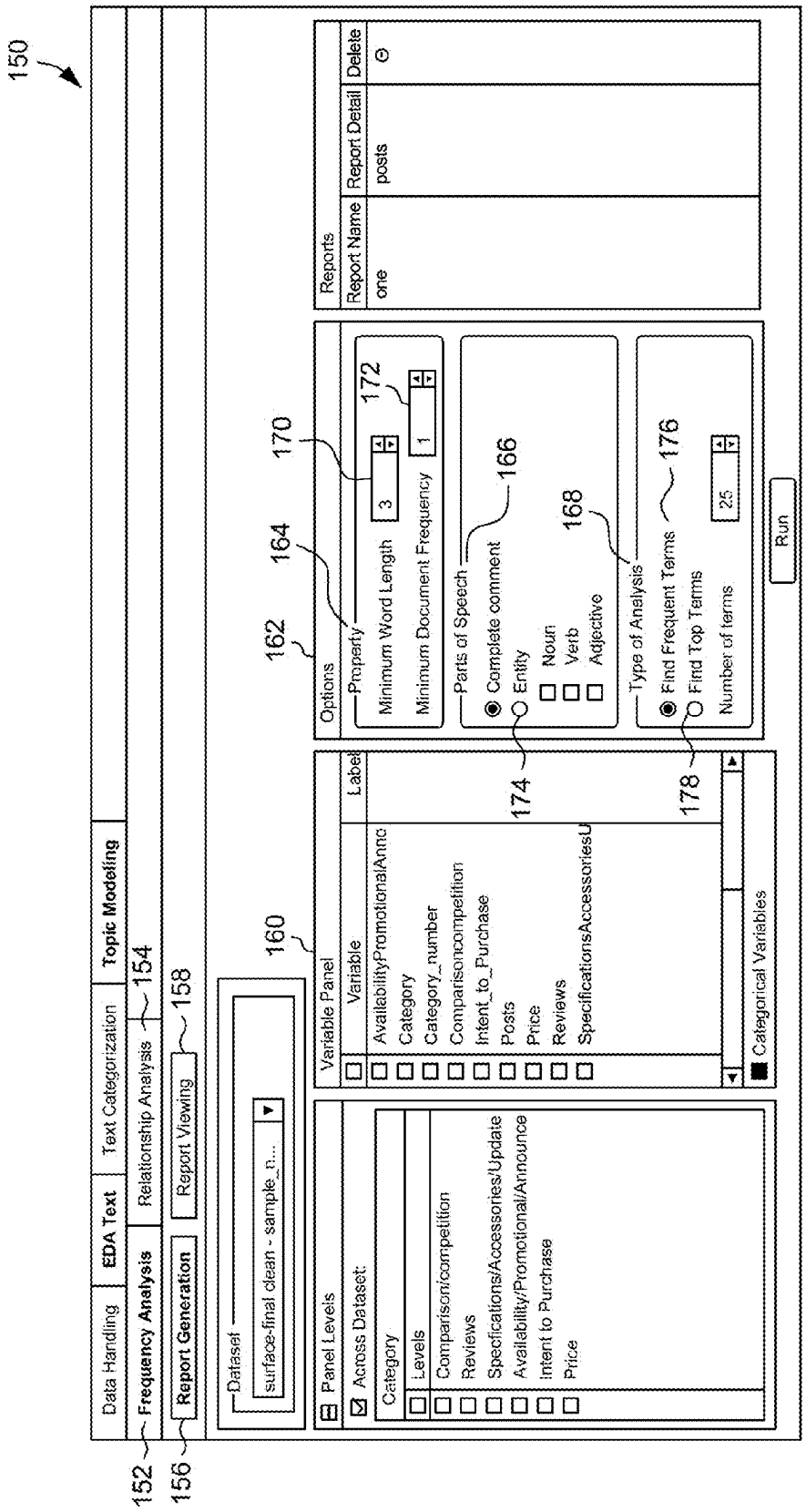

150
Data Handling
EDA Text
Text Categorization
Topic Modeling
Frequency Analysis
Relationship Analysis
Report Generation
Report Viewing
152
154
156
158
Dataset
surface-final clean - sample_n...
160
Panel Levels
Across Dataset:
Category
Levels
Comparisoncompetition
Reviews
Specifications/Accessories/Update
Availability/Promotional/Announce
Intent to Purchase
Price
Variable Panel
Variable
Label
AvailabilityPromotionalAnnc
Category
Category_number
Comparisoncompetition
Intent_to_Purchase
Posts
Price
Reviews
SpecificationsAccessoriesU
Categorical Variables
162
164
170
172
174
178
Options
Property
Minimum Word Length
3
Minimum Document Frequency
1
Parts of Speech
Complete comment
Entity
Noun
Verb
Adjective
166
168
Type of Analysis
Find Frequent Terms
Find Top Terms
Number of terms
25
176
Run
Reports
Report Name
Report Detail
Delete
one
posts

FIG. 7

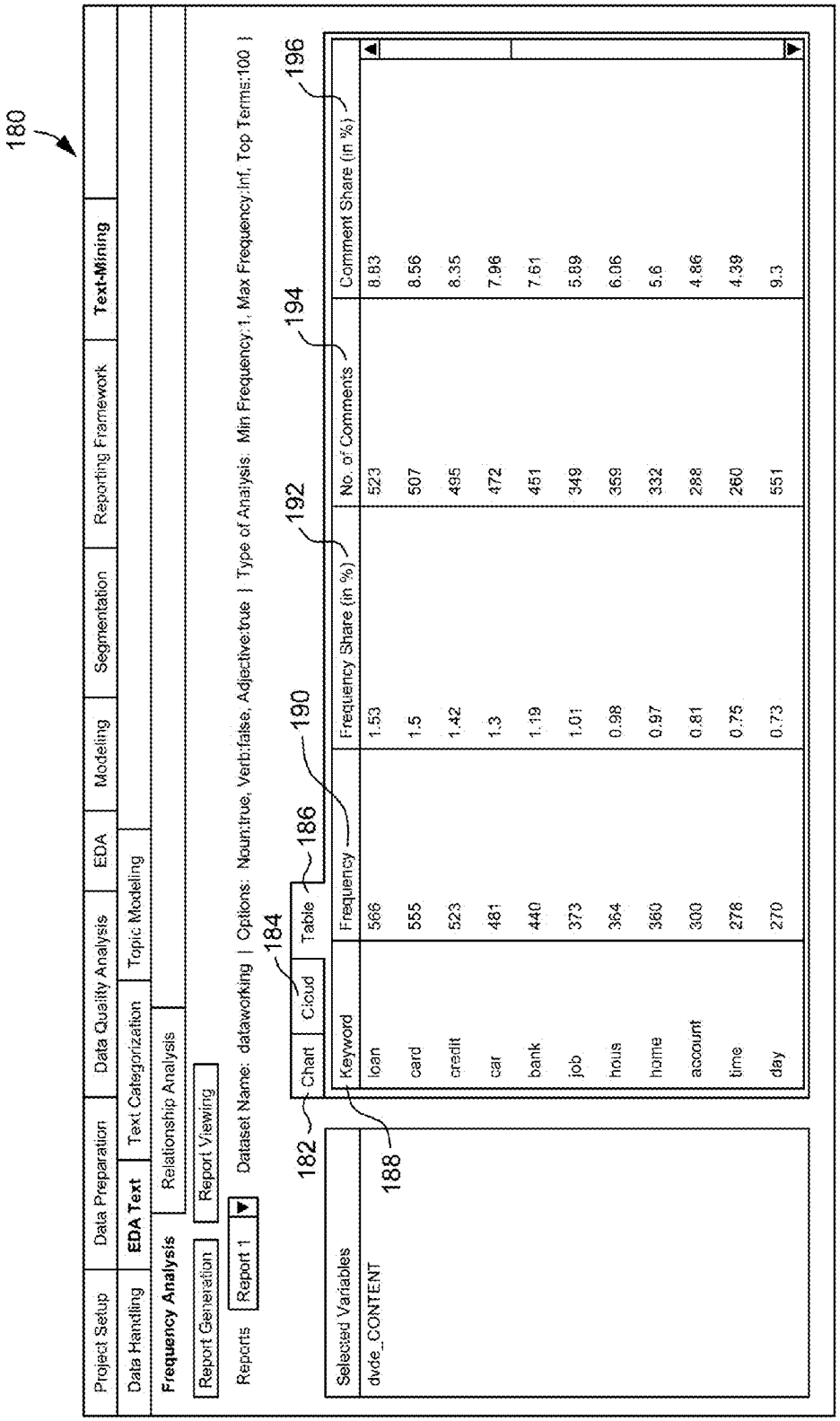

180
Project Setup
Data Preparation
Data Quality Analysis
EDA
Modeling
Segmentation
Reporting Framework
Text-Mining
Data Handling
EDA Text
Text Categorization
Topic Modeling
Frequency Analysis
Relationship Analysis
Report Generation
Report Viewing
Reports
Report 1
Dataset Name: dataworking | Options: Noun:true, Verb:false, Adjective:true } Type of Analysis: Min Frequency:1, Max Frequency:Inf, Top Terms:100 }
182
184
186
190
192
194
196
Chart
Cloud
Table
Selected Variables
dvde_CONTENT
188
Keyword
Frequency
Frequency Share (in %)
No. of Comments
Comment Share (in %)
loan 566 1.53 523 8.83
card 555 1.5 507 8.56
credit 523 1.42 495 8.35
car 481 1.3 472 7.96
bank 440 1.19 451 7.61
job 373 1.01 349 5.89
hous 364 0.98 369 6.06
home 360 0.97 332 5.6
account 300 0.81 288 4.86
time 278 0.75 260 4.39
day 270 0.73 551 9.3

FIG. 8A

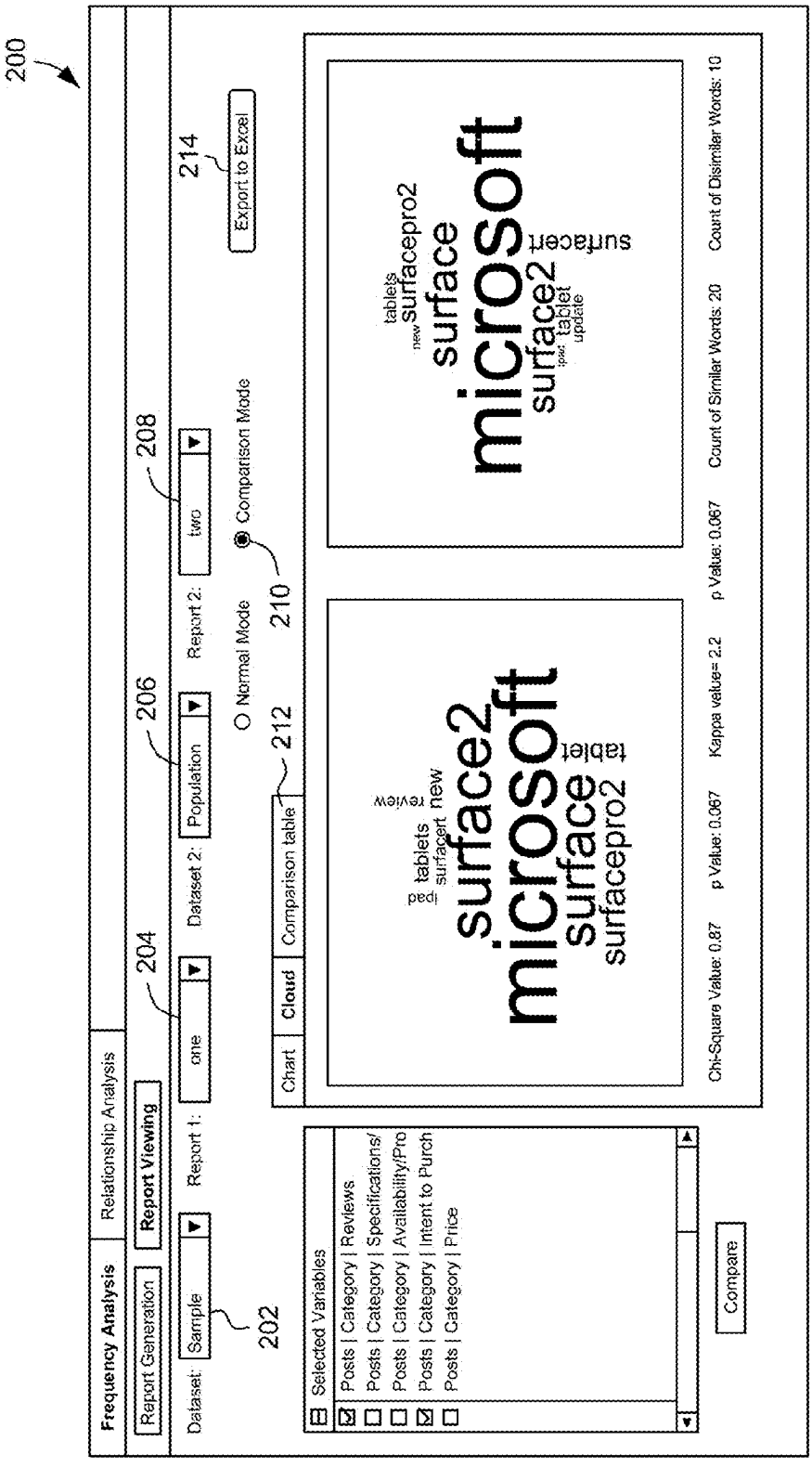

200
Frequency Analysis
Relationship Analysis
Report Generation
Report Viewing
204
206
208
214
Dataset: Sample
Report 1: one
Dataset 2: Population
Report 2: two
Export to Excel
Comparison Mode
Normal Mode
202
210
212
Chart
Cloud
Comparison table
Selected Variables
Posts | Category | Reviews
Posts | Category | Specifications/
Posts | Category | Availability/Pro
Posts | Category | Intent to Purch
Posts | Category | Price
Compare
Chi-Square Value: 0.87
p Value: 0.067
Kappa value= 2.2
p Value: 0.067
Count of Similar Words: 20
Count of Dissimilar Words: 10

FIG. 8B

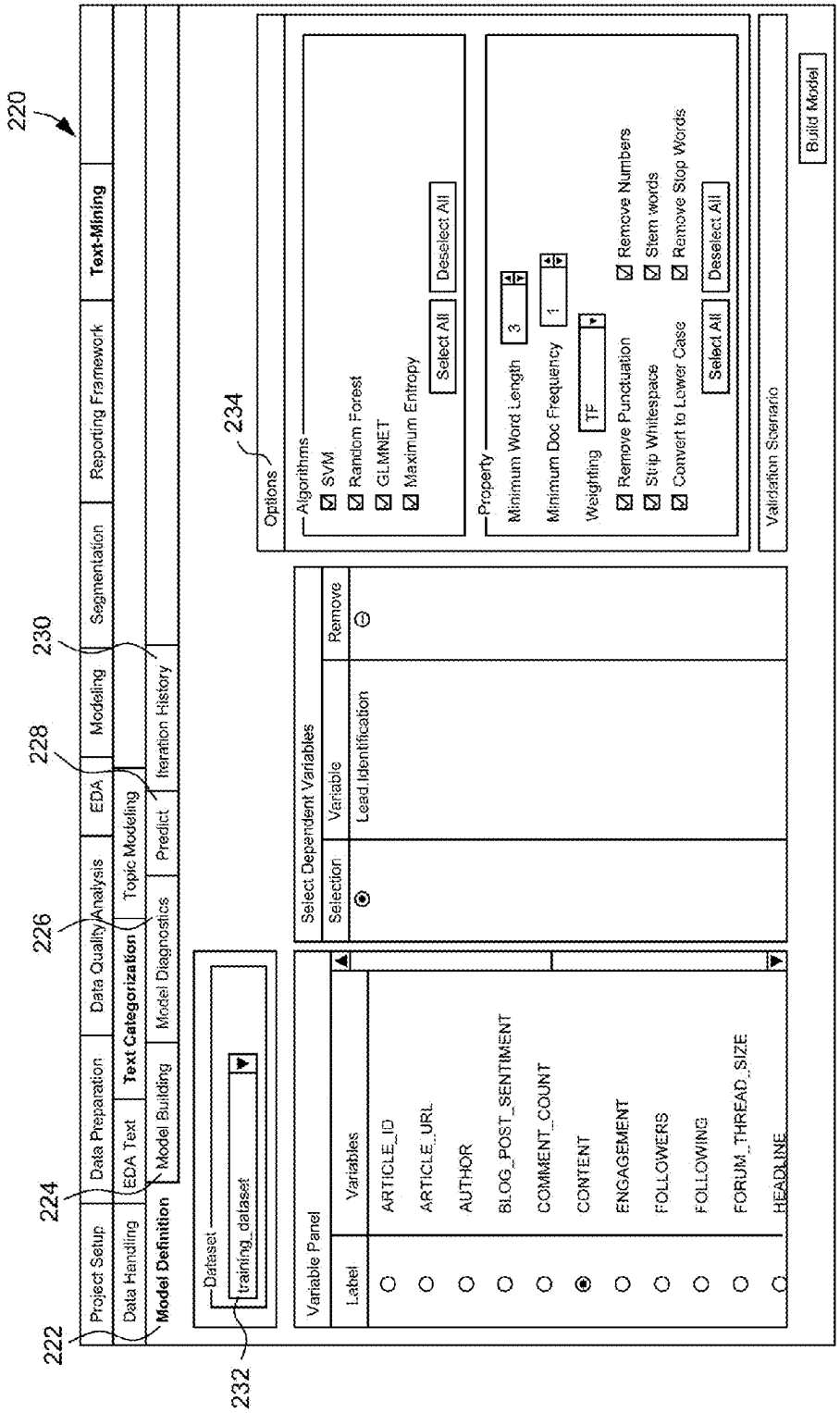

220
222
224
226
228
230
232
234
Project Setup
Data Preparation
Data Quality
Analysis
EDA
Modeling
Segmentation
Reporting Framework
Text-Mining
Data Handling
EDA Text
Text Categorization
Topic Modeling
Model Definition
Model Building
Model Diagnostics
Predict
Iteration History
Dataset
training_dataset
Variable Panel
Label
Variables
ARTICLE_ID
ARTICLE_URL
AUTHOR
BLOG_POST_SENTIMENT
COMMENT_COUNT
CONTENT
ENGAGEMENT
FOLLOWERS
FOLLOWING
FORUM_THREAD_SIZE
HEADLINE
Select Dependent Variables
Selection
Variable
Remove
Lead Identification
Options
Algorithms
SVM
Random Forest
GLMNET
Maximum Entropy
Select All
Deselect All
Property
Minimum Word Length
3
Minimum Doc Frequency
1
Weighting
TF
Remove Punctuation
Strip Whitespace
Convert to Lower Case
Remove Numbers
Stem words
Remove Stop Words
Select All
Deselect All
Validation Scenario
Build Model

FIG. 9

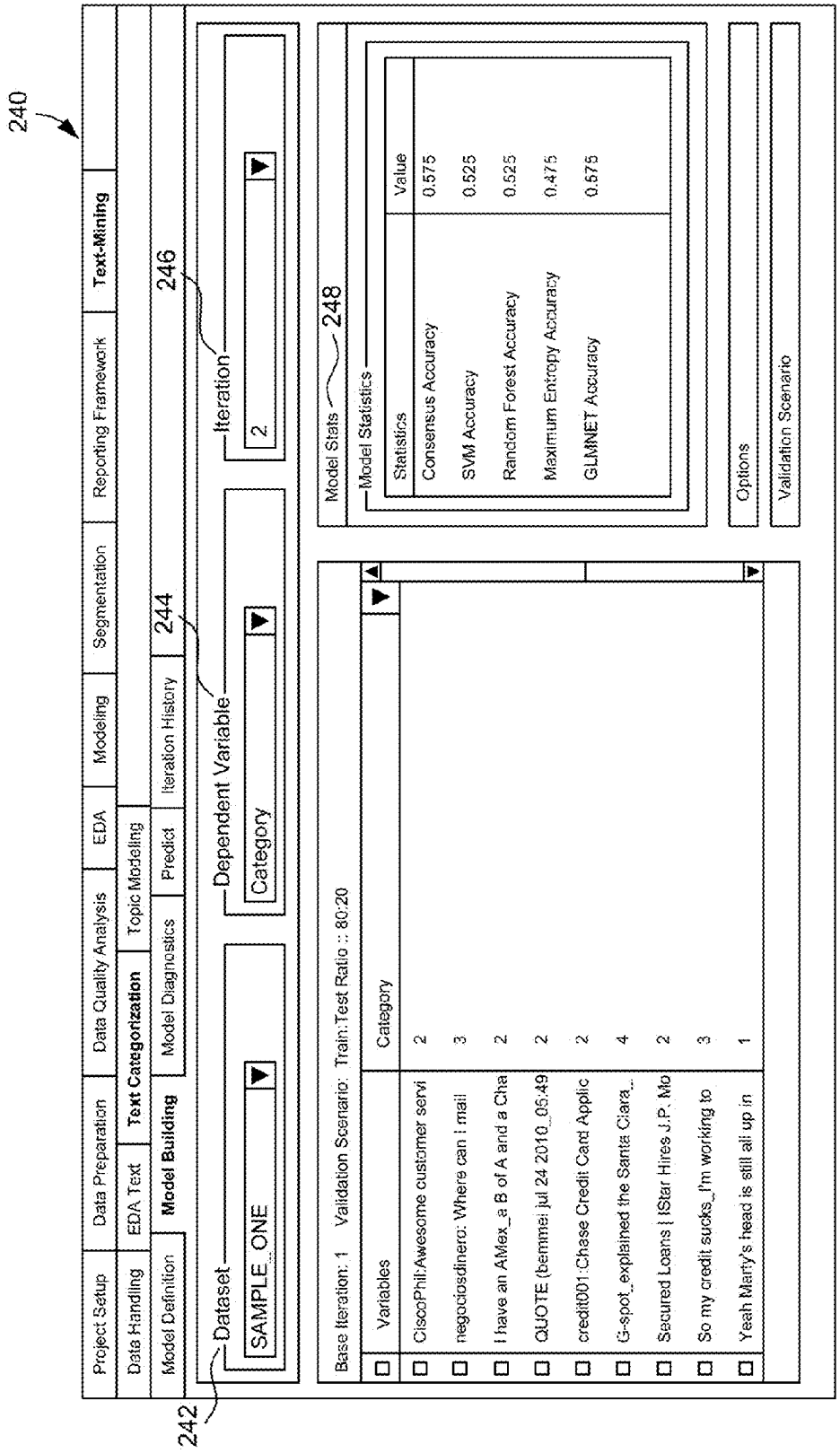

240
Project Setup
Data Preparation
Data Quality Analysis
EDA
Modeling
Segmentation
Reporting Framework
Text-Mining
Data Handling
EDA Text
Text Categorization
Topic Modeling
Model Definition
Model Building
Model Diagnostics
Predict
Iteration History
244
246
242
Dataset
SAMPLE_ONE
Dependent Variable
Category
Iteration
2
Base Iteration: 1 Validation Scenario: Train:Test Ratio :: 80:20
Variables
Category
CiscoPhil:Awesome customer servi 2
negociosdinero: Where can I mail 3
I have an AMex_a B of A and a Cha 2
QUOTE (bemma: jul 24 2010_05:49 2
credit001:Chase Credit Card Applic 2
G-spot_explained the Santa Clara_ 4
Secured Loans | iStar Hires J.P. Mo 2
So my credit sucks_I'm working to 3
Yeah Marty's head is still all up in 1
Model Stats
248
Model Statistics
Statistics
Value
Consensus Accuracy 0.575
SVM Accuracy 0.525
Random Forest Accuracy 0.525
Maximum Entropy Accuracy 0.475
GLMNET Accuracy 0.575
Options
Validation Scenario

FIG. 10

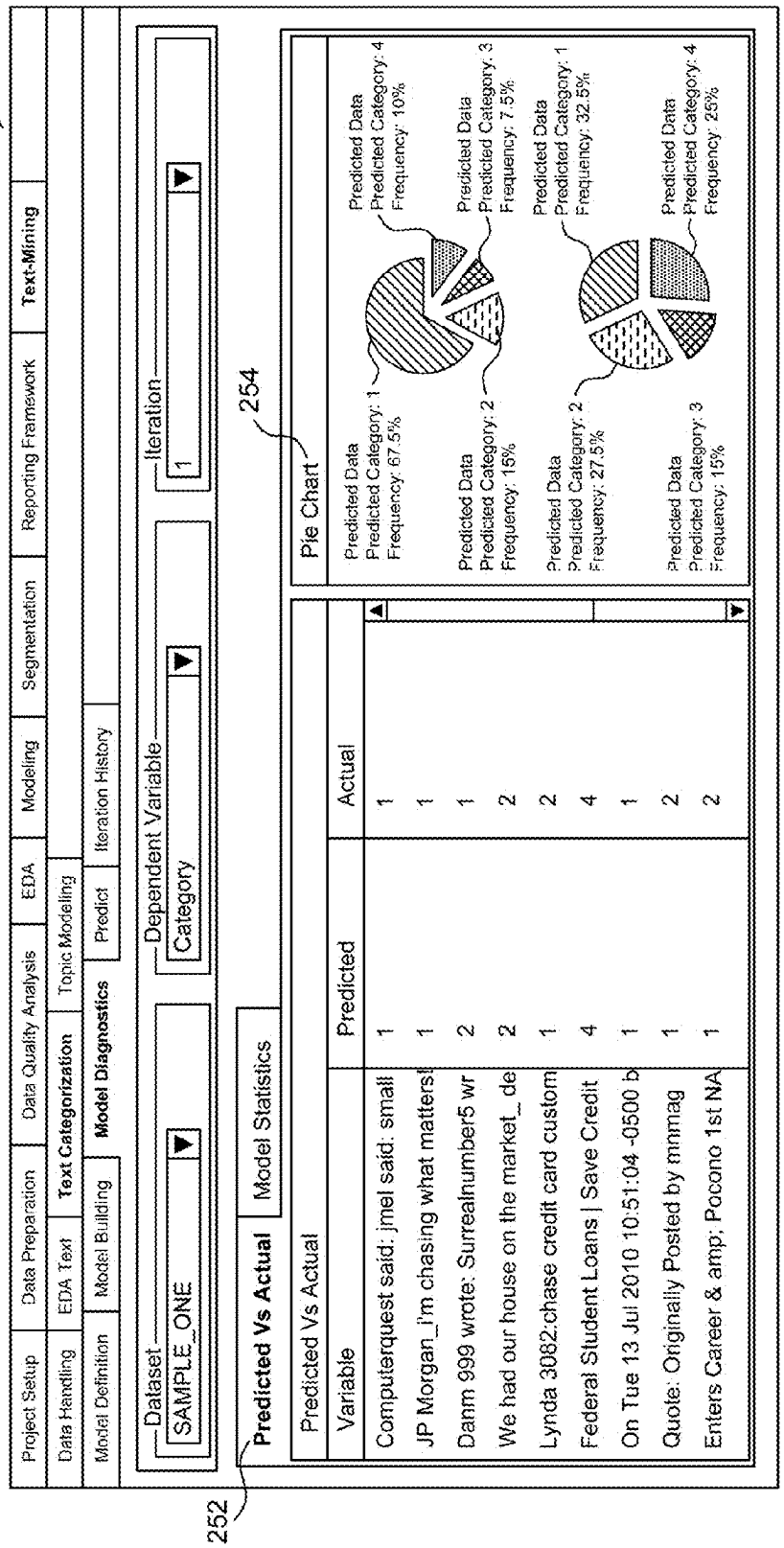

250
Project Setup
Data Preparation
Data Quality Analysis
EDA
Modeling
Segmentation
Reporting Framework
Text-Mining
Data Handling
EDA Text
Text Categorization
Topic Modeling
Model Definition
Model Building
Model Diagnostics
Predict
Iteration History
Dataset
SAMPLE_ONE
Dependent Variable
Category
Iteration
1
252
Predicted Vs Actual
Model Statistics
254
Predicted Vs Actual
Variable
Predicted
Actual
Computerquest said: jmel said: small
JP Morgan_I'm chasing what matters!
Danm 999 wrote: Surrealnumber5 wr
We had our house on the market_ de
Lynda 3082:chase credit card custom
Federal Student Loans | Save Credit
On Tue 13 Jul 2010 10:51:04 -0500 b
Quote: Originally Posted by mnmag
Enters Career & Pocono 1st NA
Pie Chart
Predicted Data
Predicted Category: 1
Frequency: 67.5%
Predicted Data
Predicted Category: 4
Frequency: 10%
Predicted Data
Predicted Category: 2
Frequency: 15%
Predicted Data
Predicted Category: 3
Frequency: 7.5%
Predicted Data
Predicted Category: 2
Frequency: 27.5%
Predicted Data
Predicted Category: 1
Frequency: 32.5%
Predicted Data
Predicted Category: 3
Frequency: 15%
Predicted Data
Predicted Category: 4
Frequency: 25%

FIG. 11

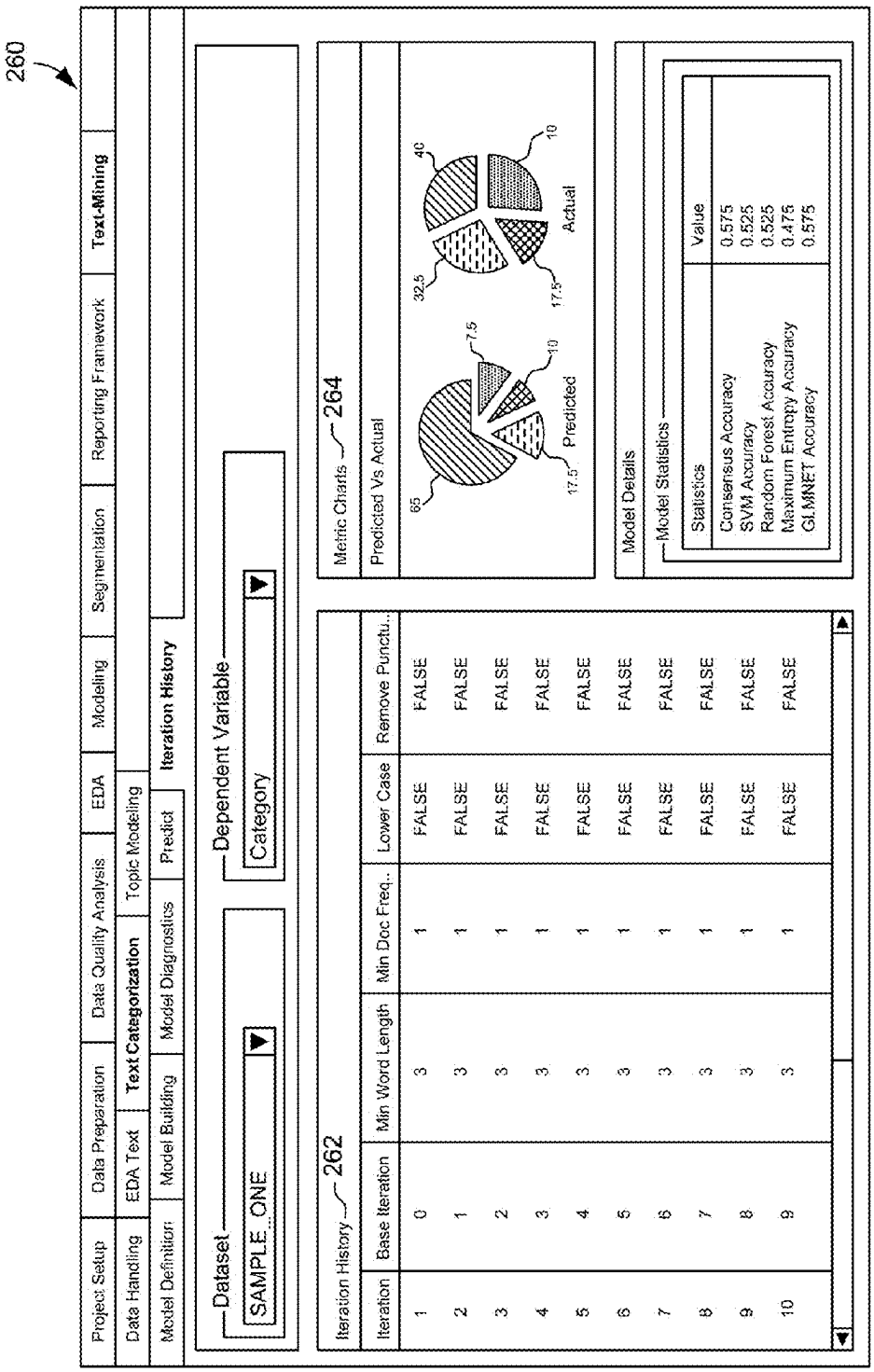

260
Project Setup
Data Preparation
Data Quality Analysis
EDA
Modeling
Segmentation
Reporting Framework
Text-Mining
Data Handling
EDA Text
Text Categorization
Topic Modeling
Model Definition
Model Building
Model Diagnostics
Predict
Iteration History
Dataset
SAMPLE_ONE
Dependent Variable
Category
Iteration History
262
Iteration
Base Iteration
Min Word Length
Min Doc Freq..
Lower Case
Remove Punctu..
1 0 3 1 FALSE FALSE
2 1 3 1 FALSE FALSE
3 2 3 1 FALSE FALSE
4 3 3 1 FALSE FALSE
5 4 3 1 FALSE FALSE
6 5 3 1 FALSE FALSE
7 6 3 1 FALSE FALSE
8 7 3 1 FALSE FALSE
9 8 3 1 FALSE FALSE
10 9 3 1 FALSE FALSE
Metric Charts
264
Predicted Vs Actual
65
7.5
10
17.5
Predicted
32.5
40
17.5
10
Actual
Model Details
Model Statistics
Statistics
Value
Consensus Accuracy 0.575
SVM Accuracy 0.525
Random Forest Accuracy 0.525
Maximum Entropy Accuracy 0.475
GLMNET Accuracy 0.575

FIG. 12

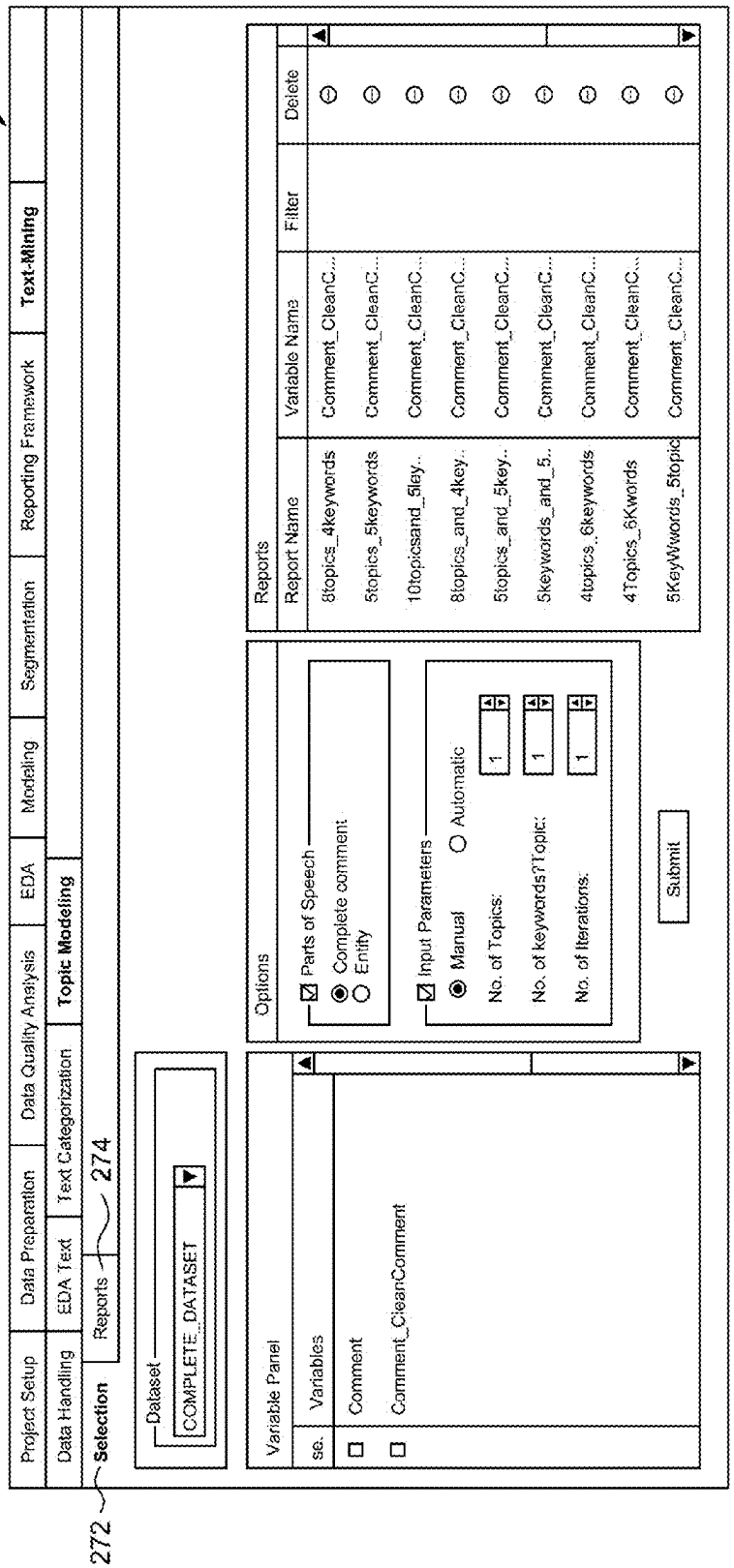

270
Project Setup
Data Preparation
Data Quality Analysis
EDA
Modeling
Segmentation
Reporting Framework
Text-Mining
Data Handling
EDA Text
Text Categorization
Topic Modeling
Selection
Reports
274
272
Dataset
COMPLETE_DATASET
Variable Panel
se.
Variables
Comment
Comment_CleanComment
Options
Parts of Speech
Complete comment
Entity
Input Parameters
Manual
Automatic
No. of Topics:
No. of keywords/Topic:
No. of Iterations:
Submit
Reports
Report Name
Variable Name
Filter
Delete
8topics_4keywords
5topics_5keywords
10topicsand_5key..
8topics_and_4key..
5topics_and_5key..
5keywords_and_5..
4topics_6keywords
4Topics_6Kwords
5KeyWwords_5topic
Comment_CleanC...

FIG. 13

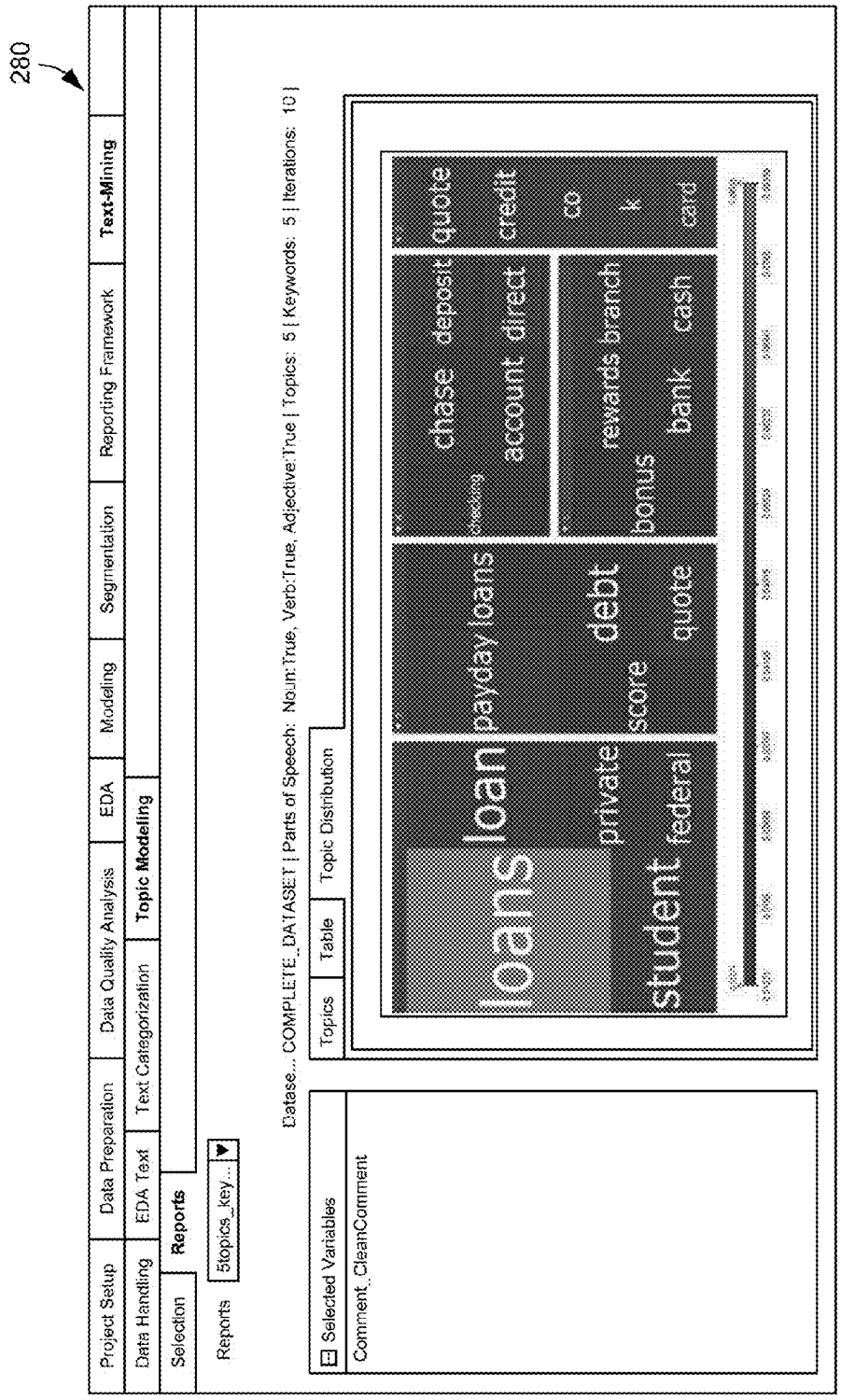

280
Project Setup
Data Preparation
Data Quality Analysis
EDA
Modeling
Segmentation
Reporting Framework
Text-Mining
Data Handling
EDA Text
Text Categorization
Topic Modeling
Selection
Reports
Reports
5topics_key...
Datase... COMPLETE_DATASET | Parts of Speech: Noun:True, Verb:True, Adjective:True | Topics: 5 | Keywords: 5 | Iterations: 10 |
Topics
Table
Topic Distribution
Selected Variables
Comment_CleanComment
loans
loan
private
student
federal
payday loans
debt
score
quote
chase
deposit
account
direct
rewards
branch
bonus
bank
cash
quote
credit
co
k
card

FIG. 14

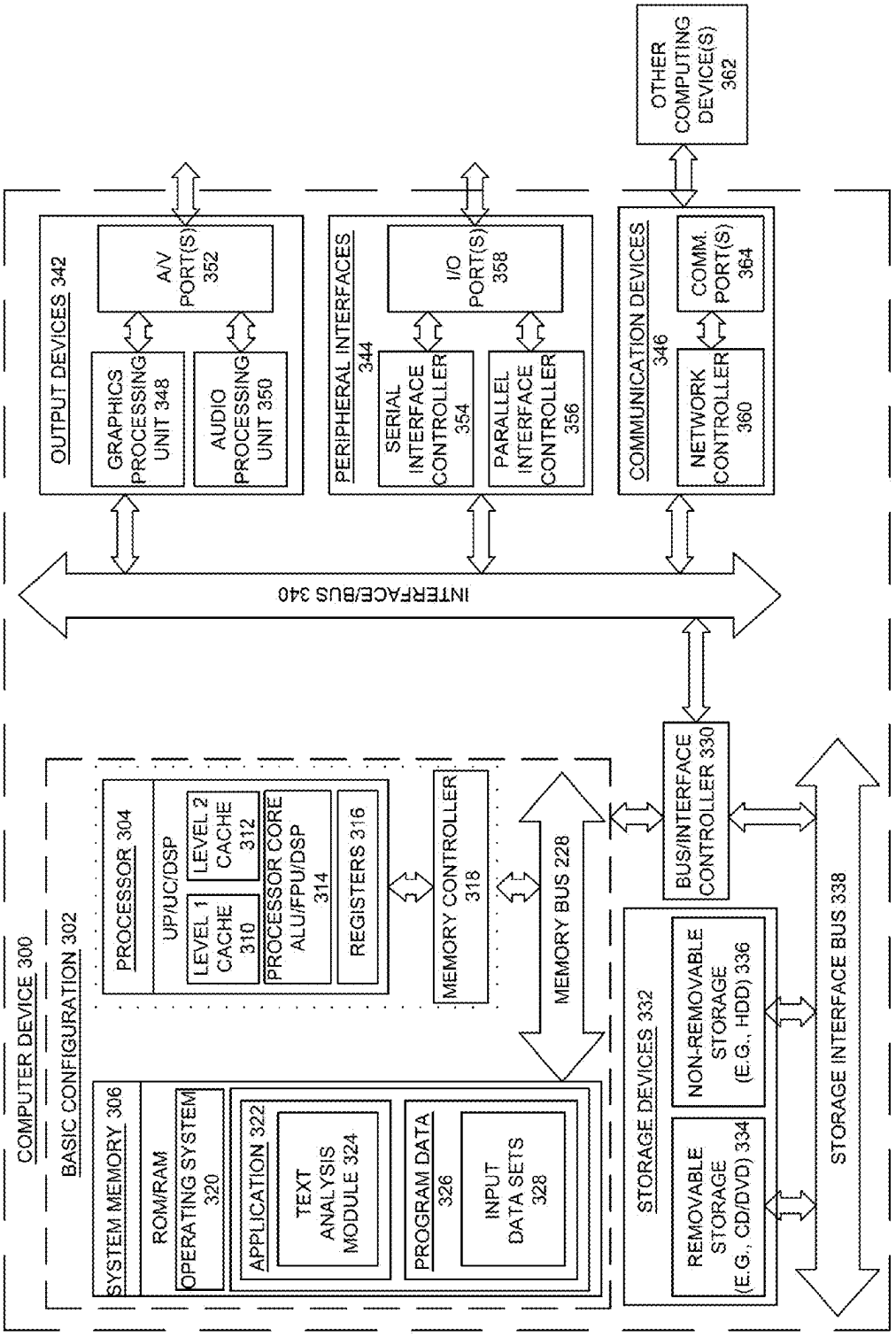

COMPUTER DEVICE 300
BASIC CONFIGURATION 302
PROCESSOR 304
UP/UC/DSP
LEVEL 1 CACHE 310
LEVEL 2 CACHE 312
PROCESSOR CORE ALU/FPU/DSP 314
REGISTERS 316
MEMORY CONTROLLER 318
SYSTEM MEMORY 306
ROM/RAM
OPERATING SYSTEM 320
APPLICATION 322
TEXT ANALYSIS MODULE 324
PROGRAM DATA 326
INPUT DATA SETS 328
MEMORY BUS 228
BUS/INTERFACE CONTROLLER 330
STORAGE DEVICES 332
REMOVABLE STORAGE (E.G., CD/DVD) 334
NON-REMOVABLE STORAGE (E.G., HDD) 336
STORAGE INTERFACE BUS 338
INTERFACE/BUS 340
OUTPUT DEVICES 342
GRAPHICS PROCESSING UNIT 348
AUDIO PROCESSING UNIT 350
A/V PORT(S) 352
PERIPHERAL INTERFACES 344
SERIAL INTERFACE CONTROLLER 354
PARALLEL INTERFACE CONTROLLER 356
I/O PORT(S) 358
COMMUNICATION DEVICES 346
NETWORK CONTROLLER 360
COMM. PORT(S) 364
OTHER COMPUTING DEVICE(S) 362

FIG. 15

TEXT MINING SYSTEM AND TOOL

BACKGROUND

[0001] The invention relates generally to text mining systems, and more particularly to a system and tool for deriving relevant information from text derived from several sources.

[0002] Text mining, sometimes alternately referred to as text data mining, or text analytics, refers to the operation of deriving relevant information from text received from several sources. Typical text mining tasks include text categorization, text clustering, concept or entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling among others.

[0003] Text mining systems can be used to build large dossiers of information about specific events. Text mining can be broadly applied to fulfill a wide variety of research and business needs in various fields such as security, biomedical, online media, marketing sentiment analysis, academics and software, etc. Moreover, text mining can also be used in certain email spam filters as a way of determining the characteristics of messages that are likely to be advertisements or other unwanted material.

[0004] However, with the current text mining systems the end user of an analytic application must be sufficiently skilled to accomplish all the tasks, some of which require substantial expertise and hence turns out to be expensive affair. Also, the huge amount of data collected in text mining is mostly semi-structured, unstructured and ill-organized that contains lexical, syntactic and semantic ambiguities. The available text mining tools use text-based searches, which can only find documents containing specific user-defined words or phrases and requires human intervention to interpret the information and to turn it actionable.

[0005] Therefore, it is desirable to automate text mining, thus reducing the need for the users to have special expertise in the field.

SUMMARY

[0006] Briefly, according to one aspect of the invention, a text mining system for extracting relevant text from a plurality of input data sets is provided. The text mining system includes an input interface module configured to enable one or more users to select a plurality of sources for a plurality of input data sets. The text mining system also includes a text analysis module configured to receive the plurality of input data sets and to generate an output data set by analyzing the plurality of input data sets. The text analysis module includes a data handling module configured to convert the plurality of input data sets to an analytics text set. The text analysis module also includes an exploratory analysis module configured to determine a plurality of correlations within the analytics text set. The text analysis module further includes a topic modeling module configured to identify a plurality of topics repeatedly occurring in the analytics text set and a reporting module configured to generate a plurality of reports for the text analysis module. The text mining system further includes memory circuitry configured to store the plurality of input data sets, the analytics text set and the output data set.

[0007] In accordance with another aspect, a text mining tool for extracting relevant text from a plurality of input data sets is provided. The text mining tool includes an input interface module configured to enable a user to select a plurality of sources for a plurality of input data sets and a data handling interface configured to enable the user to select one or more variables to trigger a data handling task. The data handling task converts the plurality of input data sets to an analytics text set. The text mining tool also includes an exploratory analysis interface configured to enable the user to select one or more types of analysis to trigger exploratory analysis task. The exploratory analysis task determines a plurality of correlations within the analytics text set. The text mining tool further includes a topic modeling interface configured to enable the user to select one or more input parameters to trigger topic modeling task. The topic modeling task identifies a plurality of topics repeatedly occurring in the analytics text set and a reporting interface configured to generate a plurality of reports based on selected criteria.

[0008] In accordance with yet another aspect, a method for extracting relevant text from a plurality of input data sets is provided. The method includes selecting a plurality of input data sets from a plurality of sources and converting the plurality of input data sets to generate an analytics text set. The method also includes determining correlations existing within the analytics text set by performing exploratory analysis and generating one or more models based on the results of the exploratory analysis. The method further includes performing topic modeling to identify repeatedly occurring topics in the analytics text set, generating a plurality of reports based on selected criteria and generating an output data set.

BRIEF DESCRIPTION OF THE DRAWINGS

[0009] These and other features, aspects, and advantages of the present invention will become better understood when the following detailed description is read with reference to the accompanying drawings in which like characters represent like parts throughout the drawings, wherein:

[0010] FIG. **1** is a block diagram of a text mining system implemented according to aspects of the present technique;

[0011] FIG. **2** is a flow diagram of one method of extracting relevant text from the input data sets using a text mining system implemented according to aspects of the present technique;

[0012] FIG. **3** is a block diagram of an example text analysis module implemented according to aspects of the present technique;

[0013] FIG. **4** is a flow diagram of one method of categorizing an analytics text set implemented according to aspects of the present technique;

[0014] FIG. **5** is an example home screen of a text mining tool implemented according to aspects of the present technique;

[0015] FIG. **6**A through **6**C are example data handling screens of a text mining tool implemented according to aspects of the present technique;

[0016] FIG. **7** is an example exploratory analysis screen of a text mining tool implemented according to aspects of the present technique;

[0017] FIGS. **8**A and **8**B are example report generation screens of a text mining tool implemented according to aspects of the present technique;

[0018] FIG. **9** is an example text categorization screen illustrating model definition of a text mining tool implemented according to aspects of the present technique;

[0019] FIG. **10** is an example model building screen of a text mining tool implemented according to aspects of the present technique;

[0020] FIG. **11** is an example model diagnostic screen of a text mining tool implemented according to aspects of the present technique;

[0021] FIG. **12** is an example iteration history viewing screen of a text mining tool implemented according to aspects of the present technique;

[0022] FIG. **13** is an example topic modeling screen of a text mining tool implemented according to aspects of the present technique;

[0023] FIG. **14** is an example topic distribution chart viewing screen of a text mining tool implemented according to aspects of the present technique; and

[0024] FIG. **15** is a block diagram of a general purpose computer arranged for extracting relevant text from a plurality of input data sets implemented according to aspects of the present technique.

DETAILED DESCRIPTION

[0025] The present invention provides a text mining system configured to extract relevant text from input data sets to enable accurate data analysis. The text mining system derives relevant information from text by structuring the input text, deriving patterns within the structured text, evaluation and interpretation of the structured text. In the example embodiment, text mining technique includes various tasks like data handling, exploratory analysis, text categorization, topic modeling and report generation. These tasks can be performed separately as per requirement and need not follow the sequence as specified.

[0026] References in the specification to "one embodiment", "an embodiment", "an exemplary embodiment", indicate that the embodiment described may include a particular feature, structure, or characteristic, but every embodiment may not necessarily include the particular feature, structure, or characteristic. Moreover, such phrases are not necessarily referring to the same embodiment. Further, when a particular feature, structure, or characteristic is described in connection with an embodiment, it is submitted that it is within the knowledge of one skilled in the art to affect such feature, structure, or characteristic in connection with other embodiments whether or not explicitly described.

[0027] FIG. **1** is a block diagram of a text mining system implemented according to aspects of the present technique that is arranged for extracting relevant text from input data sets in accordance with the present technique. The text mining system **10** typically includes a user interface **12**, a text analysis module **14** and memory circuitry **16**. Each component is described in further detail below.

[0028] The text mining system **10** is configured to receive input data sets **18**, **20**, **22** from several sources **24**, **26** and **28**. Examples of input data sets include substantially large amount of text, alphanumeric data etc. obtained from several sources like social media platforms, sales and marketing channels, financial reports and the like. For the purposes of this specification and claims, the term "social media platform" may relate to any type of computerized mechanism through which persons may connect or communicate with each other. Some social media platforms may be applications that facilitate end-to-end communications between users in a formal manner. Other social networks may be less formal, and may consist of a user's email contact list, phone list, mailing list, or other database from which a user may initiate or receive communication. Also, it may be noted that, the term "user" may refer to both natural people and other entities that operate as a "user". Examples include corporations, organizations, enterprises, teams, or other group of peoples.

[0029] The user interface **12** is configured to enable a user to provide a set of keywords for a pre-defined operation. Input data sets related to the keywords are obtained from several sources generally referred by reference numerals **24**, **26**, **28**. Examples of sources are social media networks such as Twitter, Facebook, etc., business reports from various business units, trends and predictions from specific stock markets, and the like.

[0030] Text analysis module **14** is coupled to the user interface **12** and is configured to receive the input data sets **18**, **20**, **22** derived from the keywords specified by the user and generates an output data set **30** by perusing the input data sets. The output data set **30** refers to the relevant text extracted from the input data sets. The text analysis module **14** performs various operations like data handling, exploratory analysis, text categorization, topic modeling and report generation related to the selected keywords to extract relevant text from the input data sets **18**, **20**, **22**. The text analysis module **14** is further configured to provide language compatibility by allowing the user to select the input data sets from a plurality of languages.

[0031] Memory circuitry **16** is coupled to the text analysis module **14** and is configured to store the input data sets **18**, **20**, **22** and the output data set **30**. The manner in which relevant text is extracted from the input data sets **18**, **20**, **22** is described in further detail below.

[0032] FIG. **2** is a flow diagram of one method for extracting relevant text from the input data sets using a text mining system implemented according to aspects of the present technique. The input data sets may be derived from various social media platforms as described above. Each step of the process is described below.

[0033] At block **42**, the input data sets derived from the keywords specified by the user are received. The keywords are provided by the user via user interface **12**. In general, the input data sets may include keywords for a certain product, the product name, a name of a business or an organization, and the like. In one embodiment, the input data sets can be in any language based on the language preference specified by the user. Examples of the languages include, but are not limited to, English, German, Spanish, Portuguese, French, and the like.

[0034] At block **44**, the input data sets are converted to an analytics text set. In one embodiment, the input data sets are pre-processed to filter non-relevant text by performing a data handling task. For example, stop words, special characters, phone numbers, URL's, white spaces, email addresses etc. are some of the example non-relevant text that is removed from the input data sets. In another example, non-relevant text such as nouns, verbs, adjectives, etc. are either removed or grouped together to form the analytics text set.

[0035] At block **46**, exploratory analysis is performed to determine correlations existing within the analytics text set. Exploratory analysis establishes the intricacies relationships

existing amongst the input data sets. Examples of exploratory analysis include frequency analysis and relationship analysis.

[0036] At block **48**, one or more models providing one or more categorized text sets are generated based on the results of the exploratory analysis. Each model provides one or more categorized text sets to achieve a pre-defined goal determined by the user. The process of text categorization includes recognizing inherent structure in the analytics text set and grouping variables together by similarity into one or more categories.

[0037] At block **50**, topic modeling is performed to identify frequently appearing topics in the analytics text set. The analytics text set can either be a categorized text set or a non-categorized text set. The topics are identified based on several themes present in the analytics text sets. The process captures the identification of repeatedly occurring text in a mathematical framework, to allow examining the analytics text set based on the statistics of the words, identifying the topic and determining the balance of topics in each analytics text set. Further, a relative importance of each word within a topic is determined.

[0038] At block **52**, several reports are generated based on desired criteria provided by the user. Multiple reports can be generated at various stages of the process flow. Different reports can be viewed at one place in reporting framework and results can be compared across reports with ease.

[0039] At block **54**, an output data set is generated based on the results of exploratory analysis, categorization and topic modelling steps described above. The generated output data set is then used for various analytic operations. The manner in which the text analysis module operates is described in further detail below.

[0040] FIG. **3** is a block diagram of an example text analysis module implemented according to aspects of the present technique. The text analysis module **60** includes a data handling module **62**, an exploratory analysis module **64**, a text categorization module **66**, a topic modeling module **68** and a reporting module **70**. Each component is described in further detail below.

[0041] Data handling module **62** is configured to convert the input data sets to an analytics text set. The data handling module **62** performs this operation by cleaning up the input data sets. In one embodiment, the data handling module **62** is configured to perform a pre-processing task by filtering non-relevant elements from the input data sets. The input data sets provided by the user can be in any language based on the language preference specified by the user. Examples of the languages include, but are not limited to, English, German, Spanish, Portuguese, French, and the like. The cleaning of input data sets involves detecting, correcting or removing non-relevant text. The data handling module **62** further performs various tasks including tokenization, sentence segmentation, speech tagging, extraction of named entity, chunking, parsing, co-reference resolution and the like.

[0042] Exploratory analysis module **64** operates on the analytics text set generated by the data handling module **62** and is configured to determine a various correlations that are present within the analytics text set. In one embodiment, the exploratory analysis module **64** further includes a frequency analysis module **72** and a relationship analysis module **74** which is described in further detail below.

[0043] Frequency analysis module **72** is configured to perform detailed analysis of the analytics text set. The detailed analysis includes operations such as the removal of sparse terms, identification of words with minimum threshold frequency for analysis, identification of most frequently occurring unigrams or bigrams (combination of two words) and identification of top terms in the analytics text set.

[0044] Relationship analysis module **74** is configured to determine a frequency of occurring keywords depending on the variables, parts of speech and number of top keywords. In one example embodiment, on selection of any top keyword by the user, the associated words in the analytics text set are searched. For each of the associated word in the analytics text set an associated score is calculated. The associated score indicates the strength of association that exists between other words with the selected one. Further, parameters like term frequency indicating the number of occurrences of a particular term in the analytics text set is also calculated.

[0045] Text categorization module **66** is configured to generate a plurality of models of the analytics text set based on the results of the exploratory analysis module **64**. As mentioned earlier, the analytics text set can either be a categorized text set or a non-categorized text set. The text categorization module **66** performs several operations like model building, model diagnostics, predict and iteration history using machine learning models.

[0046] In one embodiment, the text categorization is performed by first manually categorizing a subset (e. g. a sample data set) of the analytics text set. The text categorization module **66** categorizes the analytics text set by creating an actual categorization module by identifying a plurality of categories for sample data set and then creates a predictive categorization module by applying the identified categories on the analytics text set. The text categorization module **66** further compares the actual categorization module and the predictive categorization module in an iterative manner.

[0047] The parameters used for manual categorization is then extrapolated to the remainder of the analytics text set. In one embodiment, supervised machine learning algorithms are applied to the analytics text set. The supervised machine learning can be customized using machine learning rules or manually coded rules. For example, models can be created during model building by using training data and algorithms like support vector machine (SVM), random forest, GLMNET, and maximum entropy etc.

[0048] Topic modeling module **68** is configured to identify a plurality of topics repeatedly occurring in the analytics text set. Topic modeling module **68** provides a simple way to analyze the substantially large volumes of unlabeled text. Typically, the analytics text set includes a cluster of words that frequently occur together. The topic modeling module **68** connects words with similar meanings and distinguishes between uses of words with multiple meanings using contextual clues. Further, the topic modeling module **68** identifies the hidden topical patterns that pervade the collection through statistical regularities and annotate texts with these topics. The topic annotations are further used to organize, summarize and search texts.

[0049] Topic modeling module **68** makes use of a suite of unsupervised machine learning algorithms to examine texts. In one example embodiment, Latent Dirichlet Allocation (LDA) is used. The LDA algorithm generates probabilistic

mode of a corpus that allows sets of observations to be explained by unobserved groups to explain why some parts of the text are similar.

[0050] Reporting module **70** is configured to enable the user to access several reports generated by the text analysis module **60**. The reports are generated in such a way so as to allow viewing topics and keywords per topic as word cloud as well as to provide possibility to view topic distribution charts. The reporting module **70** further facilitates storing the reports to enable the user to access several reports from a single location. The manner in which the analytics text set is categorized manually is described in further detail below.

[0051] FIG. **4** is a flow diagram of one method of categorizing an analytics text set implemented according to aspects of the present technique. Each step of the process is described below.

[0052] At block **76**, a sample data set is selected from analytics text set. As mentioned earlier, the sample data set is a subset of the analytics text set. At block **77**, the sample data set is manually categorized using multiple parameters that are defined by the user to create an actual categorization module. The process of text categorization includes recognizing inherent structure in the input data sets and grouping variables together by similarity into one or more categories. Further, a predictive categorization module is created by applying the identified categories on the analytics text set. The actual categorization module and the predictive categorization module are compared in an iterative manner.

[0053] At block **78**, the sample data set is extrapolated to categorize the remainder of the analytics text set. The extrapolation is done by performing operations like model building, model diagnostics, predict and iteration history using machine learning models. For example, models can be created during model building by using training data and algorithms like support vector machine (SVM), random forest, GLMNET, and maximum entropy etc.

[0054] The above described text mining system may be implemented as a text mining tool that is configured to execute on a computing device. The text mining tool is configured to extract relevant text from the input data sets and includes several interfaces. Some of the relevant interfaces are described in further detail below.

[0055] FIG. **5** is an example home screen of a text mining tool implemented according to aspects of the present technique. The home screen **80** enables the users to add an input data set by using the "ADD DATASET" tab **82**. A path for the input data sets to be added can be specified through the "DATASET PATH" tab **84**. Further, the various existing input data sets can be viewed using the pane **86**.

[0056] FIG. **6**A through **6**C are example data handling screens of a text mining tool implemented according to aspects of the present technique. The data handling screens **6**A through **6**C enable the user to perform several data handling operations on the input data sets to generate an analytics text set. In the illustrated embodiment, the data pre-processing screen **90** enables the user to perform operations mainly related to report generation (cell **92**) and report viewing (cell **94**). During report generation operations user can select input data sets using dataset field (cell **96**) provided in the data pre-processing screen **90**. The data handling screens **6**A and **6**B further enable the user to perform the operations related to data handling in a plurality of languages like English, German, Spanish, Portuguese and French based on the language preference specified by the user. The user can specify the language preference using analysis language field (cell **97**). In the illustrated embodiment, the language preference specified by the user is English.

[0057] The data pre-processing screen **90** further includes panes pertaining to panel levels **98**, variable panel **100**, and reports **102**. The variable panel **100** allows the user to select a plurality of variables including categorical variables (cell **104**). Additionally, a dataset view panel (cell **106**) is provided for a quick view of the data to the user for the selected variable. The dataset view panel (cell **106**) also allows the user to search for a specific term in the selected variables. The user can further create an indicator variable using tab "Create Indicator" (cell **108**) for the searched data that can later be used to perform analysis.

[0058] FIG. **6**B illustrates a data cleaning screen **110** that enables the user to perform several data cleaning operations (cell **112**). The data cleaning screen **110** facilitates the user to select new variables or manipulate existing ones. The data cleaning operations (cell **112**) remove noise from the input data sets. Examples of the data cleaning operations performed include removal of phone numbers, removal of special characters, removal of stop words, removal of URLs, removal of white spaces, removal of email addresses and the like. The data cleaning screen **110** also allows the user to order a sequence of data cleaning operations and the sequence can be changed by the user as per requirement. Further, the user is allowed to create a variable at any stage/step of the ordered sequence of data cleaning operations.

[0059] FIG. **6**C illustrates an observation split screen **120** that enables the user to perform observation split (cell **122**) by splitting the input data sets on the basis of certain delimiters provided by the user. The input data sets after split can further be used for performing analysis. The observation split (cell **122**) allows better understanding of the sentiments/categories present in the input data sets. An input data set and a treatment process is chosen using dataset (cell **124**) and treatment process (cell **126**) fields respectively. The several split options (cell **128**) are specified using fields pertaining to split on variable (cell **130**), delimiter (cell **132**), minimum length to split (cell **134**), and minimum length after split (cell **136**). The split preview pane (cell **138**) provided in the observation split screen **120** facilitates the user to preview the comments related to the selected split options.

[0060] FIG. **7** is an example exploratory analysis screen of a text mining tool implemented according to aspects of the present technique. In the illustrated embodiment, the exploratory analysis screen **150** includes frequency analysis (cell **152**) and relationship analysis (**154**). Each of the frequency analysis (cell **152**) and relationship analysis (**154**) further includes fields pertaining to report generation (cell **156**) and report viewing (cell **158**).

[0061] The frequency analysis (cell **152**) does a detailed analysis of the analytics text set and performs some of the actions like removal of sparse terms, identification of words with minimum threshold frequency for analysis, identification of most frequently occurring unigrams or bigrams (combination of two words) and identification of top terms. In the example embodiment, user can select a variable using variable panel **160** along with several options from options pane **162**. The several options provided in the options pane **162** include property (cell **164**), parts of speech (cell **166**)

and type of analysis (cell **168**). The user can specify parameters like minimum word length (cell **170**), minimum document frequency (cell **172**), type of entity (cell **174**), frequent terms (cell **176**) and top terms (cell **178**).

[0062] The relationship analysis (cell **154**) generates and displays frequency of occurring keywords depending upon the variable, parts of speech and number of top keywords selected by the user.

[0063] FIG. **8**A is an example report generation screen **180** of a text mining tool implemented according to aspects of the present technique. As illustrated, the report generated on performing the frequency analysis can be viewed in form of several visualizations like bar chart (cell **182**), text tag cloud (cell **184**) or tables (cell **186**). Several parameters related to the frequency analysis are viewed in the tabular form like keywords (cell **188**), frequency (cell **190**), frequency share (cell **192**), number of comments (cell **194**) and comment share (cell **196**).

[0064] FIG. **8**B illustrates a comparison screen **200** that enables the user to compare between two frequency analysis operations performed on two different input data sets. The input data sets and respective reports for comparison can be selected through selection fields provided in the screen **200** represented by reference numerals **202** thorough **208**. The comparison mode is selected using radio button **210** and viewed using comparison table (cell **212**). The comparison results highlight the key comparison attributes like count of similar words, count of dissimilar words, kappa value, chi-square value and the like. The comparison screen **200** provides option to the user to export the comparison results in various user friendly formats (tab **214**).

[0065] FIG. **9** is an example text categorization screen illustrating model definition of a text mining tool implemented according to aspects of the present technique. The text categorization screen **220** includes several fields pertaining to model definition (cell **222**), model building (cell **224**), model diagnostics (cell **226**), predict (cell **228**) and iteration history (cell **230**). On invoking the model definition (cell **222**) tab, several machine learning models can be created using a training data set (cell **232**) and various algorithms available in the "options" field **234** like support vector machines (SVM), random forest, GLMNET, and maximum entropy and the like. The training data set **232** includes an exhaustive set of all the variables along with the final result variable containing specified categories. For example, the variables may depict the unique words of the document while required categories may depict the sentiment class like positive, negative and neutral.

[0066] FIG. **10** is an example model building screen of a text mining tool implemented according to aspects of the present technique. The model building screen **240** includes several fields pertaining to the selection of input data sets (cell **242**), dependent variables (cell **244**) and number of iterations (cell **246**). The model building screen **240** further includes a pane **248** to indicate the statistics related to the selected model.

[0067] FIG. **11** is an example model diagnostic screen of a text mining tool implemented according to aspects of the present technique. As illustrated, once the model is build it is further evaluated based on the model statistics as part of the model diagnostics using the model diagnostic screen **250**. The model is evaluated using the predicted versus the actual data related to a particular model as shown using pane **252**. The same evaluation can also be viewed using several visualizations like pie chart (cell **254**).

[0068] FIG. **12** is an example iteration history viewing screen of a text mining tool implemented according to aspects of the present technique. Once the model diagnostics are performed as described above, it is followed by the predict step which requires scoring on the larger input data set involving model section to categorize the text. The outcome of the predict step leads to the iteration history which facilitates the comparison of various iterations (cell **262**) with the help of tables and charts (cell **264**).

[0069] FIG. **13** is an example topic modeling screen of a text mining tool implemented according to aspects of the present technique. The topic modeling screen **270** includes selection (cell **272**) and reports (cell **274**) field that allows model selection with respect to the number of topics and generates reports based on the one or more criteria selected by the user. In addition, the topic modeling screen **270** also allows searching and exploring a collection of documents based on pre-defined themes. Reports can be generated as an outcome of the topic modeling which allows viewing topic and keywords per topic as word cloud as well as provides possibility to view topic distribution chart as shown in FIG. **14** (topic distribution screen **280**).

[0070] The above described systems provide several advantages including handling of data sets in multiple languages. In addition, the technique described herein provides for categorization of data into specified categories using actual categorization techniques and predictive techniques. Further, the techniques described herein also include modelling of words repeatedly occurring in the text under different themes, etc.

[0071] The technique described above can be performed by the text mining system described in FIG. **1** and FIG. **3**. The technique described above may be embodied as devices, systems, methods, and/or computer program products. Accordingly, some or all of the subject matter described above may be embodied in hardware and/or in software (including firmware, resident software, micro-code, state machines, gate arrays, etc.) Furthermore, the subject matter may take the form of a computer program product such as an analytical tool, on a computer-usable or computer-readable storage medium having computer-usable or computer-readable program code embodied in the medium for use by or in connection with an instruction execution system. In the context of this description, a computer-usable or computer-readable medium may be any medium that can contain, store, communicate, propagate, or transport the program for use by or in connection with the instruction execution system, apparatus, or device.

[0072] The computer-usable or computer-readable medium may be, for example but not limited to, an electronic, magnetic, optical, electromagnetic, infrared, or semiconductor system, apparatus, device, or propagation medium. By way of example, and not limitation, computer readable media may comprise computer storage media and communication media.

[0073] When the subject matter is embodied in the general context of computer-executable instructions, the embodiment may comprise program modules, executed by one or more systems, computers, or other devices. Generally, program modules include routines, programs, objects, components, data structures, etc. that performs particular tasks or implement particular abstract data types. Typically, the func-

tionality of the program modules may be combined or distributed as desired in various embodiments.

[0074] FIG. **15** is a block diagram of an example computing system **300** that is arranged for extracting relevant text from a plurality of input data sets in accordance with the present technique is shown. In a very basic configuration **302**, computing system **300** typically includes one or more processors **304** and a system memory **306**. A memory bus **308** may be used for communicating between processor **304** and system memory **306**.

[0075] Depending on the desired configuration, processor **304** may be of any type including but not limited to a microprocessor (μP), a microcontroller (μC), a digital signal processor (DSP), or any combination thereof. Processor **304** may include one or more levels of caching, such as a level one cache **310** and a level two cache **312**, a processor core **314**, and registers **316**. An example processor core **314** may include an arithmetic logic unit (ALU), a floating point unit (FPU), a digital signal processing core (DSP Core), or any combination thereof. An example memory controller **318** may also be used with processor **304**, or in some implementations memory controller **318** may be an internal part of processor **304**.

[0076] Depending on the desired configuration, system memory **306** may be of any type including but not limited to volatile memory (such as RAM), non-volatile memory (such as ROM, flash memory, etc.) or any combination thereof. System memory **306** may include an operating system **320**, a text analysis module **324** as an application **322** and a plurality of input data sets **328** as a program data **326**.

[0077] Text analysis module **324** is configured to receive the input data sets **328** and to generate an output data set by analyzing the input data sets **328**. This described basic configuration **302** is illustrated in FIG. **15** by those components within the inner dashed line.

[0078] Computing system **300** may have additional features or functionality, and additional interfaces to facilitate communications between basic configuration **302** and any required devices and interfaces. For example, a bus/interface controller **330** may be used to facilitate communications between basic configuration **302** and one or more data storage devices **332** via a storage interface bus **338**. Data storage devices **332** may be removable storage devices **334**, non-removable storage devices **336**, or a combination thereof.

[0079] Examples of removable storage and non-removable storage devices include magnetic disk devices such as flexible disk drives and hard-disk drives (HDD), optical disk drives such as compact disk (CD) drives or digital versatile disk (DVD) drives, solid state drives (SSD), and tape drives to name a few. Example computer storage media may include volatile and nonvolatile, removable and non-removable media implemented in any method or technology for storage of information, such as computer readable instructions, data structures, program modules, or other data.

[0080] System memory **306**, removable storage devices **334** and non-removable storage devices **336** are examples of computer storage media. Computer storage media includes, but is not limited to, RAM, ROM, EEPROM, flash memory or other memory technology, CD-ROM, digital versatile disks (DVD) or other optical storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, or any other medium which may be used to store the desired information and which may be accessed by the computing system **300**. Any such computer storage media may be part of the computing system **300**.

[0081] Computing system **300** may also include an interface bus **340** for facilitating communication from various interface devices (e.g., output devices **342**, peripheral interfaces **344**, and communication devices **346**) to basic configuration **302** via bus/interface controller **330**. Example output devices **342** include a graphics processing unit **348** and an audio processing unit **350**, which may be configured to communicate to various external devices such as a display or speakers via one or more A/V ports **352**.

[0082] Example peripheral interfaces **344** include a serial interface controller **354** or a parallel interface controller **356**, which may be configured to communicate with external devices such as input devices (e.g., keyboard, mouse, pen, voice input device, touch input device, etc.) or other peripheral devices (e.g., printer, scanner, etc.) via one or more I/O ports **358**. An example communication device **346** includes a network controller **360**, which may be arranged to facilitate communications with one or more other computing device(s) **362** over a network communication link via one or more communication ports **364**.

[0083] The network communication link may be one example of a communication media. Communication media may typically be embodied by computer readable instructions, data structures, program modules, or other data in a modulated data signal, such as a carrier wave or other transport mechanism, and may include any information delivery media. A "modulated data signal" may be a signal that has one or more of its characteristics set or changed in such a manner as to encode information in the signal. By way of example, and not limitation, communication media may include wired media such as a wired network or direct-wired connection, and wireless media such as acoustic, radio frequency (RF), microwave, infrared (IR) and other wireless media. The term computer readable media as used herein may include both storage media and communication media.

[0084] Computing system **300** may be implemented as a portion of a small-form factor portable (or mobile) electronic device such as a cell phone, a personal data assistant (PDA), a personal media player device, a wireless web-watch device, a personal headset device, an application specific device, or a hybrid device that include any of the above functions. It may be noted that computing system **300** may also be implemented as a personal computer including both laptop computer and non-laptop computer configurations.

[0085] It will be understood by those within the art that, in general, terms used herein, and especially in the appended claims (e.g., bodies of the appended claims) are generally intended as "open" terms (e.g., the term "including" should be interpreted as "including but not limited to," the term "having" should be interpreted as "having at least," the term "includes" should be interpreted as "includes but is not limited to," etc.). It will be further understood by those within the art that if a specific number of an introduced claim recitation is intended, such an intent will be explicitly recited in the claim, and in the absence of such recitation no such intent is present.

[0086] For example, as an aid to understanding, the following appended claims may contain usage of the introductory phrases "at least one" and "one or more" to introduce claim recitations. However, the use of such phrases should

not be construed to imply that the introduction of a claim recitation by the indefinite articles "a" or "an" limits any particular claim containing such introduced claim recitation to embodiments containing only one such recitation, even when the same claim includes the introductory phrases "one or more" or "at least one" and indefinite articles such as "a" or "an" (e.g., "a" and/or "an" should be interpreted to mean "at least one" or "one or more"); the same holds true for the use of definite articles used to introduce claim recitations. In addition, even if a specific number of an introduced claim recitation is explicitly recited, those skilled in the art will recognize that such recitation should be interpreted to mean at least the recited number (e.g., the bare recitation of "two recitations," without other modifiers, means at least two recitations, or two or more recitations).

[0087] While only certain features of several embodiments have been illustrated and described herein, many modifications and changes will occur to those skilled in the art. It is, therefore, to be understood that the appended claims are intended to cover all such modifications and changes as fall within the true spirit of the invention.

1. A text mining system for extracting relevant text from a plurality of input data sets, the system comprising:

an input interface module configured to enable one or more users to select a plurality of sources for a plurality of input data sets;

a text analysis module configured to receive the plurality of input data sets and to generate an output data set by analyzing the plurality of input data sets, the text analysis module comprising:

a data handling module configured to convert the plurality of input data sets to an analytics text set;

an exploratory analysis module configured to determine a plurality of correlations within the analytics text set;

a topic modeling module configured to identify a plurality of topics repeatedly occurring in the analytics text set; and

a reporting module configured to generate a plurality of reports for the text analysis module; and

memory circuitry configured to store the plurality of input data sets, the analytics text set and the output data set.

2. The system of claim **1**, wherein the data handling module is further configured to perform a pre-processing task by filtering non-relevant elements from the plurality of input data sets.

3. The system of claim **1**, wherein the text analysis module further comprises a text categorization module configured to generate a plurality of models based on the results of the exploratory analysis module; wherein each model provides one or more categorized text sets to achieve a pre-defined goal determined by a user.

4. The system of claim **3**, wherein the text categorization module is further configured to categorize the analytics text set by:

creating an actual categorization module by identifying a plurality of categories for a sample data set; and

creating a predictive categorization module by applying the identified categories on the analytics text set; wherein the sample data set is a subset of the analytics text set.

5. The system of claim **3**, wherein the text categorization module is further configured to compare the actual categorization module and the predictive categorization module in an iterative manner.

6. The system of claim **1**, wherein the exploratory analysis module is configured to perform a frequency analysis on the analytics text set to determine frequently occurring unigrams, bigrams and texts with frequency in the specified range.

7. The system of claim **1**, wherein the exploratory analysis module is configured to perform a relationship analysis on the analytics text set to determine an association score representing correlation between words in the analytics text set.

8. The system of claim **1**, wherein the exploratory analysis module is further configured to generate visual representations corresponding to the frequency analysis and relationship analysis in form of bar charts, text tag cloud, tables or combinations thereof.

9. The system of claim **1**, wherein the topic modeling module identifies the plurality of topics repeatedly occurring in the analytics text set using a plurality of machine learning algorithms.

10. The system of claim **1**, wherein the reporting module is further configured to enable the users to access a plurality of reports generated by the text analysis module.

11. The system of claim **1**, wherein the text analysis module is configured to operate on a plurality of languages.

12. A text mining tool for extracting relevant text from a plurality of input data sets, the text mining tool comprising:

an input interface module configured to enable a user to select a plurality of sources for a plurality of input data sets;

a data handling interface configured to enable the user to select one or more variables to trigger a data handling task, wherein the data handling task converts the plurality of input data sets to an analytics text set;

an exploratory analysis interface configured to enable the user to select one or more types of analysis to trigger exploratory analysis task wherein the exploratory analysis task determines a plurality of correlations within the analytics text set;

a topic modeling interface configured to enable the user to select one or more input parameters to trigger topic modeling task wherein the topic modeling task identifies a plurality of topics repeatedly occurring in the analytics text set; and

a reporting interface configured to generate a plurality of reports based on selected criteria.

13. The text mining tool of claim **12**, wherein the data handling interface is further configured to enable the user to select between one or more data cleansing task.

14. The text mining tool of claim **12**, wherein the exploratory analysis interface is further configured to enable the user to select between a frequency analysis and a relationship analysis.

15. The text mining tool of claim **12**, wherein the text analysis module is configured to analyze input data sets in a plurality of languages.

16. A method for extracting relevant text from a plurality of input data sets, the method comprising:

selecting a plurality of input data sets from a plurality of sources;

converting the plurality of input data sets to generate an analytics text set;

determining correlations existing within the analytics text set by performing exploratory analysis;

generating one or more models based on the results of the exploratory analysis;

performing topic modeling to identify repeatedly occurring topics in the analytics text set;

generating a plurality of reports based on selected criteria; and

generating an output data set.

17. The method of claim **16**, further comprising performing a frequency analysis on the analytics text set to determine frequently occurring unigrams, bigrams and texts with frequency in the specified range.

18. The method of claim **16**, further comprising performing a relationship analysis on the analytics text set to determine an association score representing correlation between words in the analytics text set.

19. The method of claim **16**, further comprising storing the plurality of reports to enable the user to access the plurality of reports from a single location.

20. The method of claim **16**, wherein the plurality of input data sets is multi-lingual.

* * * * *