PharmaSUG 2020 - Paper EP-177

Detecting Side Effects and Evaluating the Effectiveness of Drugs from Customers' Online Reviews using Text Analytics, Sentiment Analysis, and Machine Learning Models

Thu Dinh, Oklahoma State University;
Goutam Chakraborty, Oklahoma State University

ABSTRACT

Drug reviews play a very significant role in providing crucial medical care information for both healthcare professionals and consumers. Customers are utilizing online review sites to voice opinions and express sentiments about experienced drugs. However, a potential buyer typically finds it very hard to go through all comments before making a purchase decision. Another big challenge is the unstructured and textual nature of the reviews, which makes it difficult for readers to classify comments into meaningful insights. For these reasons, this paper primarily aims to classify the side effect level and effectiveness level of prescribed drugs by using text analytics and predictive models in SAS® Enterprise Miner™. Additionally, the paper explores specific effectiveness and potential side effects of each prescription drug through sentiment analysis and text mining within SAS® Visual Text Analytics. The study's results show that the best performing model for side effect level classification is the rule-based model with a validation misclassification rate at 27.1%. Regarding effectiveness level classification, the text rule builder model also works best with a 22.4% validation misclassification rate. These models are further validated by using a transfer learning algorithm to evaluate performance and generalization. The results can be used to develop practical guidelines and useful references to facilitate prospective patients in making better informed purchase decisions.

INTRODUCTION

With the rapid growth in the number of available online reviews sites and discussion boards, today's consumers are increasingly relying on online resources to aid in purchase decisions. Review sites provide existing customers the opportunity to share objective feedback about products and services they have personal experience with, which in turn facilitates prospective consumers purchase decisions. According to recent customer behavior surveys, nearly 95% of shoppers read online reviews before making a purchase (Spiegel Research Center, 2017) and 97% of buyers consider online reviews as a major useful source of information when making a purchase decision (Fan and Fuel, 2016). Typically, online drug reviews consist of two parts - ratings and textual comments. While ratings indicate the overall evaluation of customer using a numeric scale, textual comments can often provide more useful insights into the effectiveness and side effects of the drug, which overall ratings cannot. However, with the increasing number of textual comments from users, it has become more and more challenging for potential users to go through all reviews before making decisions. Therefore, an efficient structured algorithm is needed to explore the reviews and classify them into meaningful features which can serve as helpful recommendation to potential buyers. In view of that, the primary goal of this study is to construct a data-mining model to classify the side effect level and effectiveness level of prescription drugs. Additionally, the study also attempts to detect the potential side effects and explore specific effectiveness of each prescribed drug to facilitate prospective patients in selecting the best drug for treatment. The training data are collected from *druglib.com* to build predictive models which are then validated using additional data gathered from *drugs.com* using transfer learning. The results of the study are expected to help develop some useful references and practical guidelines for prospective drug users in making informed purchase decisions.

DATA PREPARATION

DATA SOURCE

This paragraph uses the PaperBody style. This paragraph uses the PaperBody style. The data for this research paper are retrieved from two independent websites, *Druglib.com* and *Drugs.com*, which are

among the largest and most widely visited pharmaceutical information resources for both consumers and healthcare professionals. These data sets are stored in '.tsv' (tab separated values) files and originally compiled by Felix Gräßer *et al.,* 2018. The data are available for download within the UC Irvine Machine Learning Repository (UC-Irvine, 2018). The downloaded data sets are converted to excel format and then imported to SAS® Enterprise Miner and SAS® Visual Text Analytics for further analysis.

DATA DICTIONARY

The first data set from *Druglib.com* consists of patient reviews on 541 drugs along with 1,808 related conditions. Reviews are provided on three aspects including benefits, side effects and overall comment. Similarly, ratings are also available for three aspects: 5-level side effect rating, 5-level effectiveness rating, and 10-star overall satisfaction rating. There are a total of 4,143 observations with nine attributes as shown in Table 1 below:

Variable	Description	Datatype
ID	Index of review entry	Numerical
UrlDrugName	Name of drug	Categorical
Condition	Patient condition (reason for using drug)	Text
BenefitsReview	Patient review on benefits	Text
Effectiveness	*5-level effectiveness rating* (Ineffective, Marginally Effective, Moderately Effective, Considerably Effective, Highly Effective)	Categorical
SideEffectsReview	Patient review on side effects	Text
SideEffects	*5-level side effect rating* (No Side Effects, Mild Side Effects, Moderate Side Effects, Severe Side Effects, Extremely Severe Side Effects)	Categorical
CommentsReview	Patient overall comment	Text
Rating	10-star overall satisfaction rating	Numerical

Table 1. Variables in the Druglib.com data set

The second data set from *Drugs.com* provides patient reviews on 3,654 drugs along with 836 related conditions and a 10-star patient rating which reflects overall patient satisfaction. There are a total of 215,063 observations in the data set with seven attributes as presented in Table 2 below:

Variable	Description	Datatype
ID	Index of review entry	Numerical
DrugName	Name of drug	Categorical
Condition	Patient condition (reason for using drug)	Categorical
Review	Patient review	Text
Date	Date of review entry	Date
Rating	10-star overall satisfaction rating	Numerical
UsefulCount	Number of users who found the review useful	Numerical

Table 2 - Variables in the Drugs.com data set

A screenshot of the data retrieved from *Drugs.com* is provided in Figure 1 below:

ID	drugName	condition	review	rating	date	usefulCoun
163740	Mirtazapine	Depression	"I've tried a few antidepressants over	10	February 28, 201	22
206473	Mesalamine	Crohn's Disease, Maintenance	"My son has Crohn's disease and has c	8	May 17, 2009	17
159672	Bactrim	Urinary Tract Infection	"Quick reduction of symptoms"	9	September 29, 2(	3
39293	Contrave	Weight Loss	"Contrave combines drugs that were used foı	9	March 5, 2017	35
97768	Cyclafem 1 / 35	Birth Control	"I have been on this birth control for one cycl	9	October 22, 2015	4
208087	Zyclara	Keratosis	"4 days in on first 2 weeks. Using on arms ar	4	July 3, 2014	13
215892	Copper	Birth Control	"I've had the copper coil for about 3 m	6	June 6, 2016	1
169852	Amitriptyline	Migraine Prevention	"This has been great for me. I've been	9	April 21, 2009	32
23295	Methadone	Opiate Withdrawal	"Ive been on Methadone for over ten years a	7	October 18, 2016	21
71428	Levora	Birth Control	"I was on this pill for almost two years. It doe	2	April 16, 2011	3
196802	Paroxetine	Hot Flashes	"Holy Hell is exactly how I feel. I had been tal	1	February 22, 201	17

Figure 1 - Partial data

METHODOLOGY

APPROACH

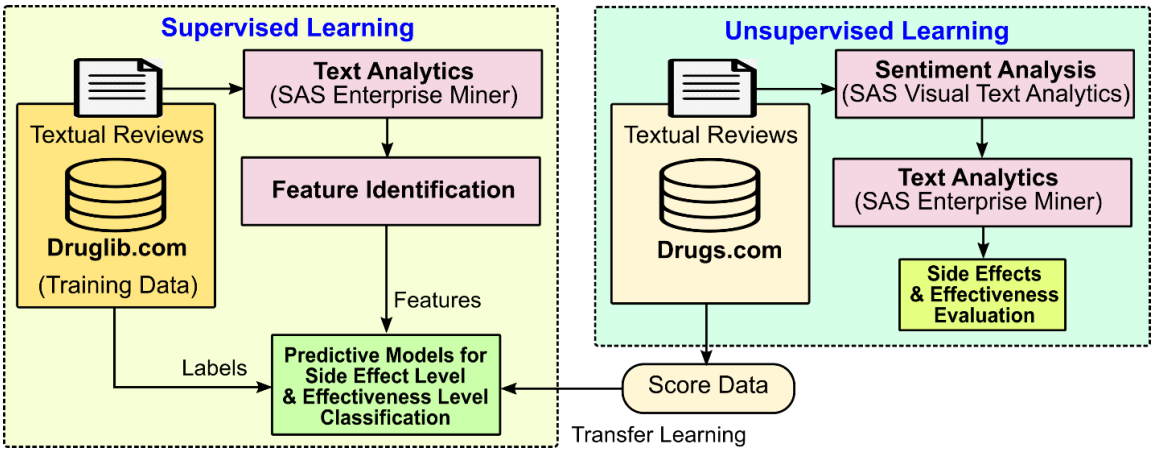


Figure 2 – Analysis approach

TARGET VARIABLES

The severity of side effects and the level of effectiveness in the *Druglib.com* data set are rated by reviewers using the 5-point Likert scale, while those in the *Drugs.com* are not rated. We randomly pick a subsample from the *Drugs.com* data set and manually annotate labels of side effect levels and effectiveness levels. In order to reduce the workload and the confusion of labeling, we create new target variables for the *Druglib.com* data set as following:

SideEffectLevel	No Side Effects	0	131 (20.00%)
	Mild / Moderate Side Effects	1	420 (64.12%)
	Severe / Extremely Severe Side Effects	2	104 (15.88%)
EffectivenessLevel	Ineffective	0	61 (9.31 %)
	Marginally / Moderately Effective	1	128 (19.54%)
	Considerably / Highly Effective	2	466 (71.15%)

Table 3 – Model target variables

SIDE EFFECT CLASSIFICATION

To classify the side effect levels of drugs from users' reviews, the following text mining and predictive modeling process is implemented.

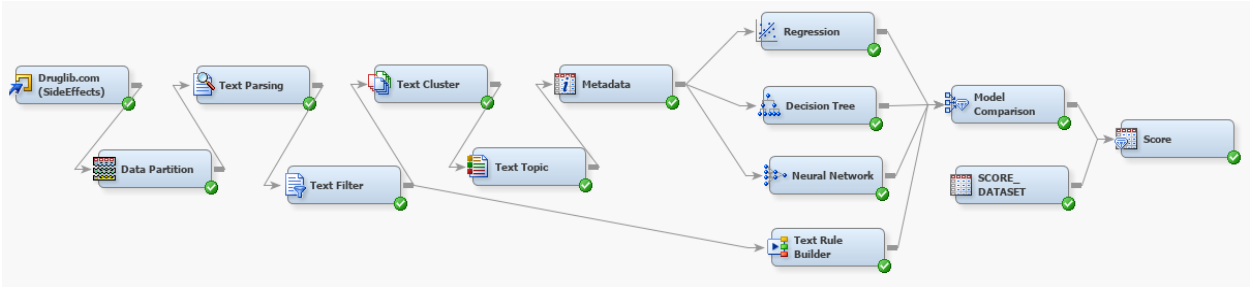


Figure 3 - Modeling diagram for side effect classification

The process flow and certain settings for individual nodes are customized based on best recommended practices in text analytics (Chakraborty, Pagolu, & Garla, 2014).

In this process flow, the "SideEffectLevel" variable is set as the categorical target variable and the "SideEffectsReview" variable is set as the text input variable to build predictive models for side effect level

classification. These models are implemented by employing text mining for features identification and machine learning techniques for building classification models.

DATA PARTITION

The *druglib*.com data set is imported to SAS® Enterprise Miner™ 14.3 via the Import File node and then partitioned in to 70% training data and 30% validation data via the Data Partition node.

TEXT PARSING

Term	Role	Attribute	Freq	# Docs	Keep	Parent/Child Status	Parent ID	Rank for Variable numdocs
+ drug	... Noun	Alpha	387	286	Y	+	7896	16
+ medication	... Noun	Alpha	341	271	Y	+	4412	17
+ experience	... Verb	Alpha	308	266	Y	+	3907	18
+ time	... Noun	Alpha	306	262	Y	+	5104	19
+ effect	... Verb	Alpha	273	258	Y	+	6419	20
+ go	... Verb	Alpha	303	254	N	+	13046	21
+ week	... Noun	Alpha	305	252	Y	+	6193	22
+ dry	... Adj	Alpha	276	239	Y	+	5436	23
+ skin	... Noun	Alpha	327	229	Y	+	9169	24
any	... Adv	Alpha	235	219	N		13221	25
+ make	... Verb	Alpha	231	202	N	+	13131	26
+ mild	... Adj	Alpha	224	200	Y	+	2619	27
severe	... Adj	Alpha	225	195	Y		9985	28
+ weight	... Noun	Alpha	237	192	Y	+	8923	29
+ start	... Verb	Alpha	246	191	Y	+	1698	30
+ mouth	... Noun	Alpha	205	188	Y	+	9213	31
+ nausea	... Noun	Alpha	197	188	Y	+	7104	31
+ pain	... Noun	Alpha	263	186	Y	+	6382	33
i	... Noun	Alpha	383	181	N		13262	34
loss	... Noun	Alpha	227	179	Y		7258	35
+ headache	... Noun	Alpha	192	175	Y	+	5280	36
stomach	... Noun	Alpha	201	174	Y		2344	37

Figure 4 - Text Parsing results for reviews on side effects

TEXT FILTER

TERM	FREQ	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
week	307	254	✓	0.166	Noun	Alpha
dry	276	239	✓	0.155	Adj	Alpha
dry	270	234			Adj	Alpha
drier	6	6			Adj	Alpha
skin	327	229	✓	0.098	Noun	Alpha
mild	224	200	✓	0.162	Adj	Alpha
severe	226	196	✓	0.486	Adj	Alpha
severe	225	195			Adj	Alpha
servere	1	1			Noun	Alpha
start	252	195	✓	0.187	Verb	Alpha
nausea	204	195	✓	0.179	Noun	Alpha
nauseas	2	2			Noun	Alpha
nausea	195	186			Noun	Alpha
nausiea	1	1			Noun	Alpha
nausa	1	1			Noun	Alpha
nasea	1	1			Noun	Alpha
nauseau	2	2			Noun	Alpha
nasuea	1	1			Noun	Alpha
nause	1	1			Noun	Alpha
mouth	213	195	✓	0.141	Noun	Alpha
weight	238	193	✓	0.1	Noun	Alpha
pain	264	187	✓	0.366	Noun	Alpha
pain	243	178			Noun	Alpha
plain	1	1			Adj	Alpha
pains	20	18			Noun	Alpha
stomach	214	183	✓	0.123	Noun	Alpha
stomache	6	6			Noun	Alpha
stomach	201	174			Noun	Alpha
stomach	5	5			Verb	Alpha
stomac	2	1			Noun	Alpha

Figure 5 - Text Filter results for reviews on side effects

Concept links

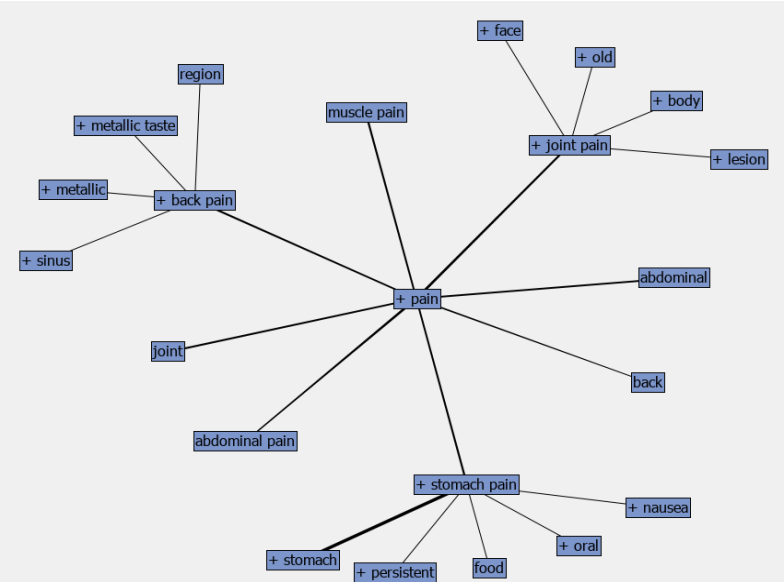
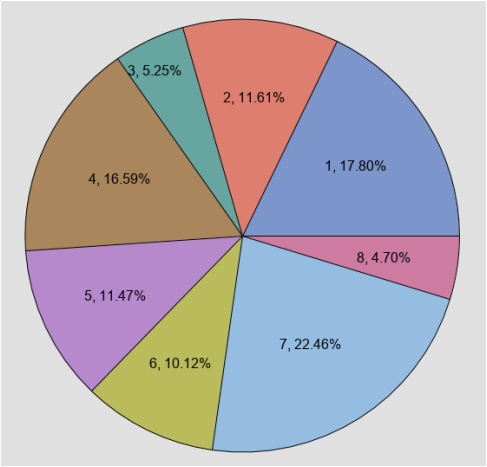


Figure 6 - Concept links for the term "pain"

The concept link diagram in Figure 6 shows that the term "pain" is associated with such terms as "muscle pain", "back pain", "abdominal pain", "stomach pain", "joint pain". This indicates that these are some commonly found "pain" side effects of prescription drugs.

TEXT CLUSTERING



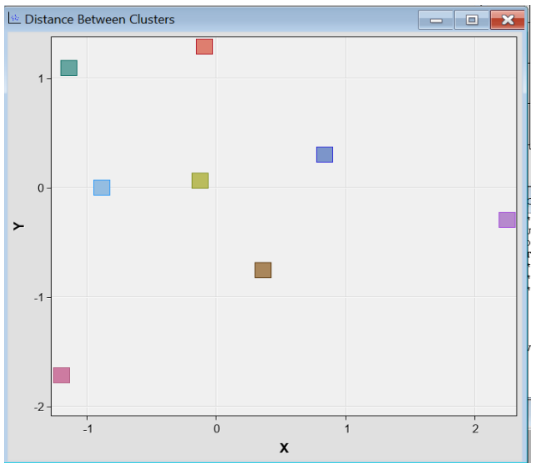


Figure 7 - Text Cluster node results for reviews on side effects

Cluster ID	Descriptive Terms	Frequency	Percentage
1	+effect +side +'side effect' +experience aware +negative +notice 'at all' +medication +bad +problem +blood sex +slight +year ...	515	18%
2	+dry +mouth loss +weight +'dry mouth' gain +depression 'weight gain' +mild +memory anxiety +fatigue 'dry skin' +appetite sexual ...	336	12%
3	+skin +rash +body +develop +peel +red +face +itchy +itch +sensitive +redness +dryness +area +flake +irritation ...	152	5%
4	+pain +severe +extreme +depression +day +cramp +ache +start +mood anxiety 'a day' +work +muscle +month +nausea ...	480	17%
5	+effect side +'side effect' +'no side effect' +experience +note +treatment +drug aware +medication +notice +drowsiness +sun +decrease 'weight gain'...	332	11%
6	+muscle +reaction chest +breath +pressure +cause +mood +ache +blood +extremely +note +cramp +stomach +swell +constipation ...	293	10%
7	+stop +week +little +start +feel +first +eat +morning +month +hour +bad +sleep first +feeling +night ...	650	22%
8	+day 'a day' +few +couple +tire first +late +feel +time +morning +sleep +appetite +first +eat +bad ...	136	5%

Figure 8 - Text Cluster descriptive terms for reviews on side effects

Text Cluster node generates eight well-separated clusters as shown in Figure 7 and Figure 8. Cluster 7 has the highest frequency (22%) with such descriptive terms as “week”, “start”, “feel”, “first”, “morning”, “hour”, “feeling”, etc., which often occur together. This implies that some side effects from the above cluster could be related to bad feeling, not feeling like to eat in the morning, or hard to sleep at night which often happen on the first few days/ weeks using the drugs.

TEXT TOPIC

Category	Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
Multiple	1	0.330	0.024	side,+side effect,+effect,+notice,+drug	10	501
Multiple	2	0.151	0.025	+severe,side,severe nausea,+nausea,+diarrhea	17	196
Multiple	3	0.133	0.026	+day,a day,+notice,+sleep,+feel	29	361
Multiple	4	0.139	0.025	+effect,+side effect,+side,+experience,+notice	19	221
Multiple	5	0.125	0.026	+pain,+muscle,chest,joint,abdominal	38	187
Multiple	6	0.120	0.026	+effect,+side,+notice,negative side,+far	31	295
Multiple	7	0.113	0.026	+dry,+mouth,+dry mouth,+skin,+mild	42	248
Multiple	8	0.106	0.027	+depression,anxiety,+mood, x000d x000d ,+swing	58	176
Multiple	9	0.104	0.029	+stop,anxiety,+feel,+week,+start	111	337
Multiple	10	0.124	0.026	+experience,+mild,+week,+nausea,+effect	47	272
Multiple	11	0.113	0.024	+no side effect,+effect,side,at all,+experience	18	71
Multiple	12	0.099	0.026	+extreme,+horrible,+mood,+nausea,anxiety	57	88
Multiple	13	0.103	0.028	+rash,+body,+develop,+skin,+cause	97	294
Multiple	14	0.095	0.026	aware,+experience,+night,+effect,+side	39	48
Multiple	15	0.098	0.027	loss,gain,+weight,+hair,weight gain	67	276

Figure 9 - Text Topic results for reviews on side effects

Figure 9 shows 15 different topics with corresponding number of terms in each topic and number of documents that contain the topic terms. For example, topic 2 indicates that drug users may experience some side effects like severe nausea or diarrhea, whereas topic 5 addresses other side effects related to pains in muscle, chest, join, or abdominal pains. Topic 7 mentions dry mouth or dry skin as possible side effects while from topic 12, other major concerns that reviewers express are regarding the extreme horrible mood or anxiety. Meanwhile, topic 11 indicates that some reviewers experience no side effect at all.

TEXT RULE BUILDER

The Text Rule Builder node is a Boolean rule-based categorizer that automatically generates an ordered set of rules that are useful in describing and predicting the target variable (SideEffectLevel).

The Text Rule Builder node is designed with five different settings (*Very High/ High/ Medium/ Low/ Very Low*) for Generalization Error, Purity of Rules and Exhaustiveness.

- ***Generalization Error*** determines the predicted probability for rules that use an untrained data set. Higher values do a better job of preventing overtraining at a cost of not finding potentially useful rules.
- ***Purity of Rules*** determines how selective each rule is by controlling the maximum p-value necessary to add a term to a rule. Highest value results in the fewest, purest rules while lowest value results in the most rules that handle the most terms.
- ***Exhaustiveness*** determines the exhaustiveness of the rule search process, or how many potential rules are considered at each step. As Exhaustiveness increases, the amount of time that the Text Rule Builder node requires and the probability of overtraining the model are also increased.

Given the above setting properties, after trial and error, the customized setting with *very high* Generalization Error, *very low* Purity of Rules and *medium* Exhaustiveness produces the best results with lowest validation Misclassification Rate and Average Square Error in classifying side effect levels.

The Text Rule Builder node also has one special feature – the ***Change Target Values*** window that allows active learning so that a user can interact with the algorithm to iteratively build a better predictive model. Specifically, the Change Target Values window enables the flexibility to view and reassign target values, then rerun the Text Rule Builder node and iteratively refine the model.

Change Target Values-WORK.TRCHANGE

Text	Data Partition	Target Variable	Original	Predicted	Assigned Target
Mild tiredness, sensitivity to bright sunlight, and mild joint aches. Some slight difficulty sleeping and tenderness particularly in finger joints.	Validate	DrugSideEffectLevel	0	1	0
I experienced no side effects at all from this nasal spray. This was great because I had significant side effects from all the pills I had ever tried for my hay fever/nasal allergies (dry mouth and d	Validate	DrugSideEffectLevel	0	1	0
While the drug did give the desired effect, I was groggy and tired. The worst side effect would be that I had short-term memory loss during when I took it. I would see notes that I took at the time	Validate	DrugSideEffectLevel	2	1	2
There were two side effects to the drug. The first side effect was that I was very drowsy. I knew that this would be a side effect and so that is why I took the medication about a half hour to an hour	Validate	DrugSideEffectLevel	0	1	0
Sleep-eating, and night-time snacking. Definitely stimulates appetite shortly (w/in 30 mins) after taking. I do not have this side-effect with other sleep preparations.	Validate	DrugSideEffectLevel	2	1	2
my lips, mouth, and face began going numb and tingling after only first dose. i thought it was all in my head until i continued the treatment and it kept getting worse. i had muscle spasms all over m	Validate	DrugSideEffectLevel	1	2	1

Figure 10 – The Change Target Values Window from Text Rule Builder node

Figure 10 shows some observations when a rule predicted a target value that is different from the assigned target value. For example, the first row mentions about mild tiredness, mild joint aches, and slight difficult sleeping, all of which suggest that the level of side effect should be 1 (Mild/ Moderate Side Effect). Hence, the assigned Target value should be changed from 0 to 1 which matches the predicted target value. Similarly, for the fourth observations, the comment addresses about two possible side effects of the drug, whereas the original target value is 0 (No Side Effect). Therefore, the assigned target should be changed from 0 to 1. For the last observation in Figure 10, given the severe side effects described in the comment, the assigned target value should be changed from 1 to 2 to accurately reflect the true level/ degree of the side effects. Overall, by utilizing the adaptive learning feature of the Text Rule Builder node, we can improve the resulting text categorization accuracy.

Some of the final rules generated by the Text Rule Builder model node are provided as below.

Rules Obtained

Target Value	Rule #	Rule	Valid Precision	Precision	Valid Recall	Recall	Valid F1 score	F1 score	Valid True Positive/Total	True Positive/Total
0	52	no side effect	100.0%	100.0%	5.57%	3.96%	10.56%	7.62%	17/17	30/31
0	53	aware	100.0%	95.08%	8.20%	7.66%	15.15%	14.18%	8/12	28/36
1	6	peel	95.24%	93.52%	33.61%	38.09%	49.69%	54.13%	21/21	46/49
1	5	tire	94.98%	93.35%	31.79%	36.20%	47.64%	52.17%	25/28	53/64
1	10	occasional	94.84%	93.73%	41.18%	47.01%	57.42%	62.62%	17/19	43/46
1	7	a bit	94.80%	93.40%	35.71%	40.04%	51.88%	56.05%	20/23	53/59
1	9	first	94.68%	93.66%	39.92%	45.51%	56.16%	61.25%	40/49	131/157
1	8	slightly	94.64%	93.60%	37.11%	41.36%	53.32%	57.37%	14/15	36/36
1	3	slight	94.63%	94.67%	19.75%	24.58%	32.68%	39.02%	26/29	106/119
1	4	dry	94.50%	93.71%	28.85%	33.69%	44.21%	49.56%	73/87	202/239
1	2	little	94.44%	94.57%	16.67%	18.60%	28.33%	31.09%	53/59	121/131
1	1	mild	94.37%	95.34%	9.38%	11.57%	17.07%	20.63%	70/76	187/200
2	43	horrible	94.19%	80.97%	33.61%	32.62%	49.54%	46.51%	14/15	43/48
1	11	reduce	94.12%	93.33%	42.58%	48.40%	58.63%	63.74%	18/22	41/46
1	17	constipation	94.12%	91.36%	53.78%	58.45%	68.45%	71.29%	20/32	54/76
1	12	decrease	94.01%	93.32%	43.98%	50.03%	59.92%	65.14%	19/21	48/56
1	13	increase	93.96%	92.00%	47.90%	53.49%	63.45%	67.65%	54/65	115/160
1	16	beginning	93.95%	91.62%	52.24%	57.07%	67.15%	70.33%	13/15	34/39
1	15	dryness	93.86%	91.79%	51.40%	56.25%	66.43%	69.76%	24/26	45/52
1	14	drowsiness	93.63%	91.84%	49.44%	55.19%	64.71%	68.94%	36/47	64/80
2	42	severe	93.42%	79.35%	29.46%	26.02%	44.79%	39.19%	77/91	150/196
1	18	weight	93.18%	90.63%	57.42%	61.97%	71.06%	73.61%	58/83	130/193
1	26	sleepiness	93.04%	89.13%	65.55%	68.57%	76.91%	77.51%	16/18	29/37

Figure 11 - Text Rule Builder results for classifying reviews on side effects

The above Rules Obtained table displays some rules for predicting the target variable. For example, Rule 52 says that for a document to satisfy this rule, it must contain the term “no side effect” so the target variable is assigned value “0”. Similarly, Rule 10, 7, 8, and 1 mention that if the comment contains any one of those terms “occasional”, “a bit”, “slightly”, or “mild”, then it is classified as level 1 side effect (target value =1). Meanwhile, based on rules 42 and 43, if a document has such terms as “horrible”, “severe”, “extreme”, the target variable should be assigned value “2”.

The order of the rules for each category in the table is important. The rule in the first row for each category is discovered by considering all documents and is the first rule that is added into the rule set. The rule in the second row of the table for each category is learned by analyzing all documents that were not covered by the first rule, and so on. The remaining columns in the table indicate the accuracy of the rules. Take rule 43 for example, this rule has a valid precision of 94.19% which implies that the precision (True Positive/ Total Category) in validation data for all rules up to this point in the table for the target value in matching documents that are actually assigned to that target value is 94.19%. The number of correctly matching documents in the validation data for this rule is 43 out of total 48 documents that match this rule, as indicated by the Valid True Positive/ Total column.

The Text Rule Builder model is then compared with other data mining models including Regression, Decision Tree, and Neural Network to find out the optimal model in classifying side effects reviews into three respective levels of rating. As previously mentioned in Figure 3, in all these models, the categorical variable “SideEffectLevel” is set as the target variable and the text variable “SideEffectsReview” is set as the input variable. Other key settings are specified as below.

REGRESSION

The Regression node is set up with below settings:

- Model selection method is set to *Stepwise*
- Model selection criterion is set to *Validation Misclassification*

DECISION TREE

The Decision Tree node is set up with below settings:

- Subtree selection method is set to *Assessment* (i.e., the smallest subtree with the best assessment value is chosen)
- Subtree assessment measure is set to *Misclassification*

NEURAL NETWORK

After trial and error, the Neural Network node is set up with below setting:

- Network architecture: *Normalized Radial Basis Function with equal width and height*
- Number of hidden units: *3*
- Target Layer Combination Function: *Linear*
- Target Layer Activation Function: *Softmax*
- Target Layer Error Function: *MBernoulli (Multiple Bernoulli)*
- Model selection criterion is set to *Misclassification*

MODEL COMPARISON

The Model Comparison node is connected to all four predictive model nodes including Text Rule Builder, Regression, Decision Tree, and Neural Network to find out the optimal model in classifying side effects reviews into three respective levels of rating. The settings for Model comparison nodes are set as below.

- Model selection statistic: Misclassification Rate
- Model selection table: *Validation*

The Model Comparison results are provided in Figure 12.

Fit Statistics

Selected Model	Model Description	Target Variable	Selection Criterion: Valid: Misclassification Rate ▲	Valid: Average Squared Error	Valid: Kolmogorov-Smirnov Statistic	Valid: Roc Index	Valid: Gini Coefficient
Y	Text Rule Builder	DrugSideEffectLevel	0.270635	0.057946	0.615	0.847	0.693
	Regression	DrugSideEffectLevel	0.278715	0.137069	0.493	0.822	0.643
	Neural Network	DrugSideEffectLevel	0.281124	0.138462	0.484	0.819	0.637
	Decision Tree	DrugSideEffectLevel	0.301205	0.14676	0.417	0.778	0.556

Figure 12 – Fit statistics comparison between models for side effect level classification.

Given the study's decision prediction goal, the most relevant selection criteria to rate model performance should be based on the Validation Misclassification Rate. Figure 12 indicates that among the four competing models, the Text Rule Builder appears to be the best performing model in classifying side effect reviews into the three respective levels (No Side Effects – Mild/ Moderate Side Effects - Severe / Extremely Severe Side Effects) since it has the lowest Validation Misclassification Rate at 27.06% as compared to the other three models. Additionally, this rule-based model also performed better than three remaining models in terms of lowest Validation Average Square Error and highest values for all other three metrics (Kolmogorov-Smirnov Statistics, Roc Index, and Gini Coefficient).

EFFECTIVENESS LEVEL CLASSIFICATION

To classify the effectiveness level of drugs from patients' benefits comments, the following text mining and predictive modeling process is implemented.

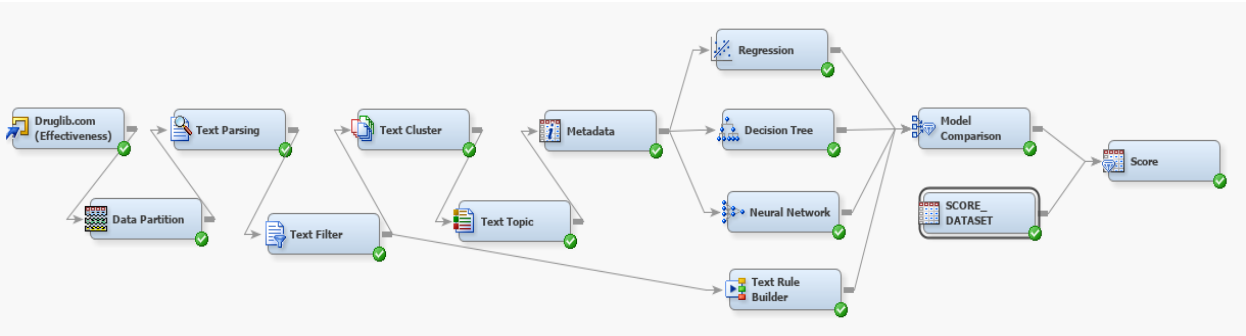


Figure 13 – Modeling diagram for effectiveness classification

The process flow is basically similar to that of side effect level classification, apart from the difference that the categorical target variable is now set to be "EffectivenessLevel" and the text input variable is "benefitsReview".

DATA PARTITION

The *druglib*.com data set is imported to SAS® Enterprise Miner™ 14.3 via the Import File node and then partitioned into 70% training data and 30% validation data via the Data Partition node.

TEXT PARSING

As shown by Figure 14 below, some of the most commonly used words by reviewers in the comments are "benefit", "effective", "better", "improve", etc., which is expected as these words generally relate to some benefits of prescription drugs.

Term	Role	Attribute	Freq	# Docs	Keep	Parent/Child Status	Parent ID	Rank for Variable numdocs
+ skin	...Noun	Alpha	347	248	Y	+	10329	24
+ time	...Noun	Alpha	291	244	Y	+	5838	25
+ work	...Verb	Alpha	268	235	Y	+	9858	26
+ benefit	...Noun	Alpha	253	234	Y	+	4229	27
+ start	...Verb	Alpha	286	234	Y	+	1949	27
more	...Adv	Alpha	264	227	N		14726	29
more	...Adj	Alpha	271	225	N		14721	30
+ symptom	...Noun	Alpha	270	216	Y	+	6682	31
+ make	...Verb	Alpha	234	212	N	+	14817	32
+ stop	...Verb	Alpha	232	209	Y	+	9268	33
+ depression	...Noun	Alpha	246	200	Y	+	6198	34
+ use	...Verb	Alpha	244	194	N	+	14769	35
+ anxiety	...Noun	Alpha	223	183	Y	+	4585	36
acne	...Noun	Alpha	257	181	Y		2620	37
+ sleep	...Verb	Alpha	222	181	Y	+	3742	37
now	...Adv	Alpha	199	172	N		14946	39
effective	...Adj	Alpha	184	163	Y		6196	40
i	...Noun	Alpha	311	163	N		14961	40
better	...Adj	Alpha	177	161	Y		9911	42
+ improve	...Verb	Alpha	186	154	Y	+	4347	43
+ life	...Noun	Alpha	183	153	Y	+	1324	44
+ mood	...Noun	Alpha	171	152	Y	+	10686	45
still	...Adv	Alpha	166	152	N		14954	45
+ seem	...Verb	Alpha	168	148	N	+	14711	47
+ increase	...Verb	Alpha	175	144	Y	+	8561	48
better	...Adv	Alpha	152	143	Y		10046	49

Figure 14 - Text Parsing results for reviews on effectiveness

TEXT FILTER

	TERM	FREQ	# DOCS	KEEP ▼	WEIGHT	ROLE	ATTRIBUTE
⊞	skin	356	251	✓	0.121	Noun	Alpha
⊞	month	302	250	✓	0.024	Noun	Alpha
⊞	week	286	250	✓	0.06	Noun	Alpha
⊟	benefit	264	245	✓	0.379	Noun	Alpha
	bendfits	1	1			Noun	Alpha
	benfits	1	1			Noun	Alpha
	benrfits	1	1			Miscellaneous Pr...	Entity
	bennefit	1	1			Noun	Alpha
	benefit	1	1			Miscellaneous Pr...	Entity
	benefit	72	67			Noun	Alpha
	benifit	2	2			Noun	Alpha
	benifits	4	4			Noun	Alpha
	benefits	181	172			Noun	Alpha
⊞	time	292	245	✓	0.15	Noun	Alpha
⊞	start	288	235	✓	0.064	Verb	Alpha
⊞	work	268	235	✓	0.057	Verb	Alpha
⊞	symptom	293	233	✓	0.04	Noun	Alpha
⊞	stop	232	209	✓	0.025	Verb	Alpha
⊞	depression	251	202	✓	0.043	Noun	Alpha
⊞	sleep	230	186	✓	0.031	Verb	Alpha
⊞	anxiety	227	186	✓	0.04	Noun	Alpha
⊞	acne	260	183	✓	0.023	Noun	Alpha
⊟	effective	186	165	✓	0.052	Adj	Alpha
	effectiv	1	1			Noun	Alpha
	effective	184	163			Adj	Alpha
	effecive	1	1			Noun	Alpha
	better	177	161	✓	0.043	Adj	Alpha
⊞	improve	188	156	✓	0.057	Verb	Alpha

Figure 15 - Text Filter results for reviews on effectiveness

Concept Links

The concept links in Figure 16 show that improvement in mood, skin, energy, memory, sleep, ability are possible effects of analyzed drugs.

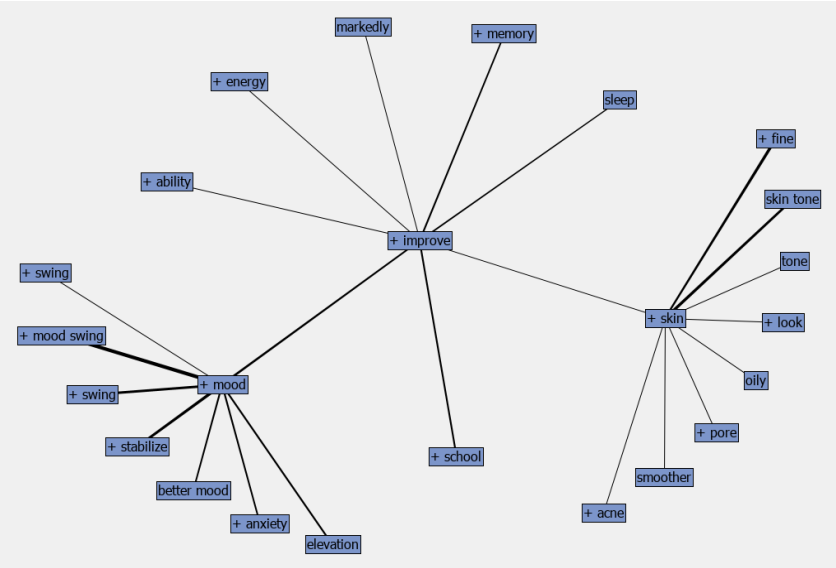
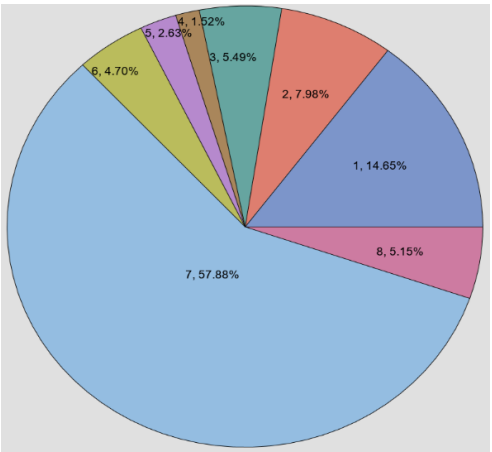


Figure 16 - Concept links for the term "improve"

TEXT CLUSTERING



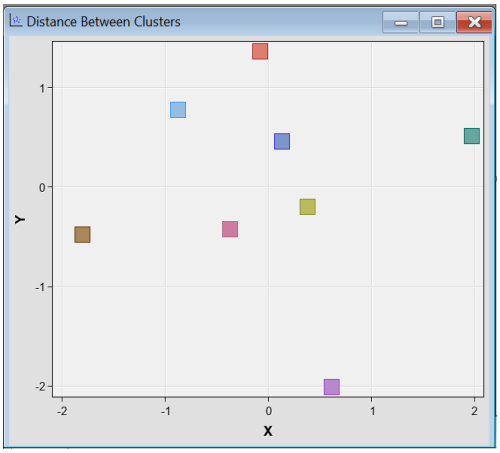


Figure 17 – Text Cluster node results for reviews on effectiveness

Cluster ID	Descriptive Terms	Frequency	Percentage
1	+effect +side +'side effect' +infection +antibiotic 'at all' +drug quickly +experience +treatment +mg +'treatment benefit' +medicine +long +benefit...	424	15%
2	+doctor +prescribe +lower +cholesterol +medicine +'blood pressure' +pressure +blood back +level +high +bad +year +osteoporosis +back ...	231	8%
3	+benefit +'treatment benefit' +treatment +include +little +relief +month +pregnancy +bad +stop +medication +good +experience +continue +ca...	159	5%
4	x000d x000d + x000d x000d x000d x000d +minute +find +look +little +attack +hour +symptom +back +relief +experience +first back...	44	2%
5	+benefit +treatment +advise +include +'treatment benefit' +bad +notice +significantly +overall +mood +continue 'at all' +headache clarity +side ...	76	3%
6	+reaction +discontinue long allergic adverse +severe quickly +pregnancy +area +'side effect' +hair +side +effect +minute +experience ...	136	5%
7	+help +skin +able +acne +clear +night +sleep +improve +attack +time +look +reduce +feel +anxiety better ...	1675	58%
8	+increase +bone 'at all' density 'bone density' +progression +mg +osteoporosis clarity +difference +depress +loss +notice +energy +mood ...	149	5%

Figure 18 - Text Cluster descriptive terms for reviews on effectiveness

The Text Cluster node generates eight well-separated clusters as shown in Figure 17 and Figure 18. Cluster 7 has the highest frequency (58%) with such descriptive terms as "help", "skin", "able", "clear", "improve", "look", "reduce", "feel", "better", etc., which often occur together. This indicates that some effectiveness from the above cluster could be regarding better sleep, acne cleared, improved skin/ look, reduced anxiety, and better feeling.

TEXT TOPIC

Category	Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
Multiple	1	0.200	0.022	+benefit,+treatment benefit,+treatment,+receive,+outweigh	8	244
Multiple	2	0.179	0.022	+benefit,+treatment,+short,+advise,+bad	4	88
Multiple	3	0.131	0.023	+side,+effect,+bad,at all,+side effect	23	121
Multiple	4	0.110	0.024	+doctor,+prescribe,+effect,+time,+know	40	118
Multiple	5	0.128	0.024	+side effect,+effect,+side,+side,+drug	21	173
Multiple	6	0.111	0.025	+drug,+help,at all,+effect,+know	52	288
Multiple	7	0.112	0.025	+skin,+line,+wrinkle,+improvement,+treatment	91	269
Multiple	8	0.096	0.024	+treat,+patient,+treatment,+add,+medicine	42	76
Multiple	9	0.102	0.026	+time,at all,+short,+severe,+able	93	265
Multiple	10	0.091	0.023	+bone,density,bone density,+increase,+side	42	44
Multiple	11	0.096	0.024	+lower,+blood,+blood pressure,+patient,+pressure	56	175
Multiple	12	0.097	0.023	long,at all,+medicine,+effect,+benefit	36	96
Multiple	13	0.095	0.024	+antibiotic,+effect,+medicine,+amoxicillin,+sinus infection	49	105
Multiple	14	0.098	0.025	+prescribe,+side,+discontinue,+bad,+severe	49	173
Multiple	15	0.098	0.025	+medicine,+help,+help,slightly,+effect	59	327

Figure 19 - Text Topic results for reviews on effectiveness

Figure 19 shows 15 different topics with corresponding number of terms in each topic and number of documents that contain the topic terms. Topic 1 shows that there are some drugs which benefits outweigh side effects. Topic 7 identifies some improvement in skin treatment like reducing lines and wrinkles, whereas, topic 11 addresses lower blood pressure. Topic 15 indicates that some medicines only show slightly effectiveness.

TEXT RULE BUILDER

The Text Rule Builder node generates an ordered set of rules that together are useful in describing and predicting the target variable (EffectivenessLevel). After trial and error, the customized setting with *very low* Generalization Error, *very low* Purity of Rules and *low* Exhaustiveness produce the best results with lowest Validation Misclassification Rate and Average Squared Error.

Target Value	Rule # ▲	Rule	Precision	Valid Precision	Recall	Valid Recall	F1 score	Valid F1 score	True Positive/ Total	Valid True Positive/ Total
2	1	life & work	100.0%	87.50%	0.91%	0.78%	1.81%	1.55%	19/19	7/8
2	2	able & ~chip & normal	100.0%	81.25%	1.63%	1.45%	3.21%	2.86%	16/16	6/8
2	3	able & ~chip & ~stop & ~observe & ~resistant & start	100.0%	87.50%	2.78%	3.13%	5.42%	6.05%	29/29	15/17
2	4	life & suffer	100.0%	89.19%	3.46%	3.69%	6.68%	7.09%	14/14	6/6
2	5	greatly & ~brown spot	99.18%	90.70%	5.81%	4.36%	10.98%	8.32%	53/54	7/7
2	6	year & ~bone density & ~worse & ~stop & week & ~b...	99.33%	92.98%	7.11%	5.93%	13.26%	11.15%	30/30	20/20
2	7	cold sore	99.40%	91.80%	7.92%	6.26%	14.67%	11.73%	18/18	3/4
2	8	dryness	99.46%	91.30%	8.79%	7.05%	16.14%	13.08%	19/19	8/9
2	9	prior	99.50%	91.78%	9.60%	7.49%	17.51%	13.86%	19/19	5/5
2	10	lexapro	99.53%	88.10%	10.27%	8.28%	18.62%	15.13%	18/18	10/14
2	11	control & ~moderate & ~bp & ~theory & birth	99.58%	89.01%	11.38%	9.06%	20.42%	16.45%	26/26	8/8
2	12	wake & able	99.60%	89.00%	12.00%	9.96%	21.42%	17.91%	19/20	9/11
2	13	calm	99.62%	89.42%	12.63%	10.40%	22.41%	18.64%	13/13	5/5
2	14	normal & week	99.64%	89.72%	13.20%	10.74%	23.31%	19.18%	12/12	4/4
2	15	release	99.65%	89.09%	13.78%	10.96%	24.21%	19.52%	13/13	2/3
2	16	basis	99.67%	89.34%	14.35%	12.19%	25.09%	21.46%	15/15	11/12
2	17	able & ~chip & ~stop & ~observe & ~improvement & ...	99.68%	88.71%	14.98%	12.30%	26.04%	21.61%	21/21	3/4
2	18	lift	99.42%	88.15%	16.42%	13.31%	28.18%	23.13%	31/32	10/13
2	19	all the time	99.44%	87.41%	16.95%	13.98%	28.96%	24.11%	14/14	8/11
2	20	drug & ~benefit	99.47%	87.16%	17.91%	14.43%	30.35%	24.76%	27/27	6/7

Figure 20 - Text Rule Builder results for reviews on effectiveness

These rules are presented as the conjunction of terms and their negations. For example, Rule 5 "greatly & ~brown spot" says that for a document to satisfy this rule, it must contain the term “greatly” and should not contain the term “brown spot”.

MODEL COMPARISON

The Model Comparison node is connected to all four predictive model nodes including Text Rule Builder, Regression, Decision Tree, and Neural Network to find out the optimal model in classifying benefits reviews into three respective levels of rating. As previously mentioned in Figure 13, in all these models, the categorical variable “EffectivenessLevel” is set as the target variable and the text variable

“benefitsReview” is set as the input variable. The Model selection statistic is set to be the Validation Misclassification Rate. The Model Comparison results are provided as below.

Fit Statistics

Selected Model	Model Description	Target Variable	Selection Criterion: Valid: Misclassification Rate	Valid: Average Squared Error	Valid: Kolmogorov-Smirnov Statistic	Valid: Roc Index	Valid: Gini Coefficient
Y	Text Rule Builder	DrugEffectivenessLevel	0.224441	0.044305	0.412	0.771	0.542
	Regression	DrugEffectivenessLevel	0.26506	0.135422	0.226	0.636	0.273
	Decision Tree	DrugEffectivenessLevel	0.265863	0.138375	0.166	0.592	0.184
	Neural Network	DrugEffectivenessLevel	0.266667	0.136348	0.213	0.641	0.283

Figure 21 – Fit statistics comparison between models for effectiveness level classification

Figure 21 indicates that Text Rule Builder is still the best performing model in classifying benefits reviews into three effectiveness levels (Ineffective – Marginally / Moderately Effective - Considerably / Highly Effective) since it has the lowest Validation Misclassification rate at 22.44% as compared to the other three models. This rule-based model also performs better than three remaining models in terms of lowest Validation Average Square Error and highest values for all other three metrics (Kolmogorov-Smirnov Statistics, Roc Index, and Gini Coefficient).

GENERALIZATION

The best performing model in side effect levels classification is scored on a new independent score data set to evaluate model validation and generalization. The score data set is created by randomly picking a sample of 500 observations from the second original data set retrieved from *Drugs.com* with manually annotated labels. The results from scoring are provided as below.

SCORING RESULTS FOR SIDE EFFECT LEVEL CLASSIFICATION

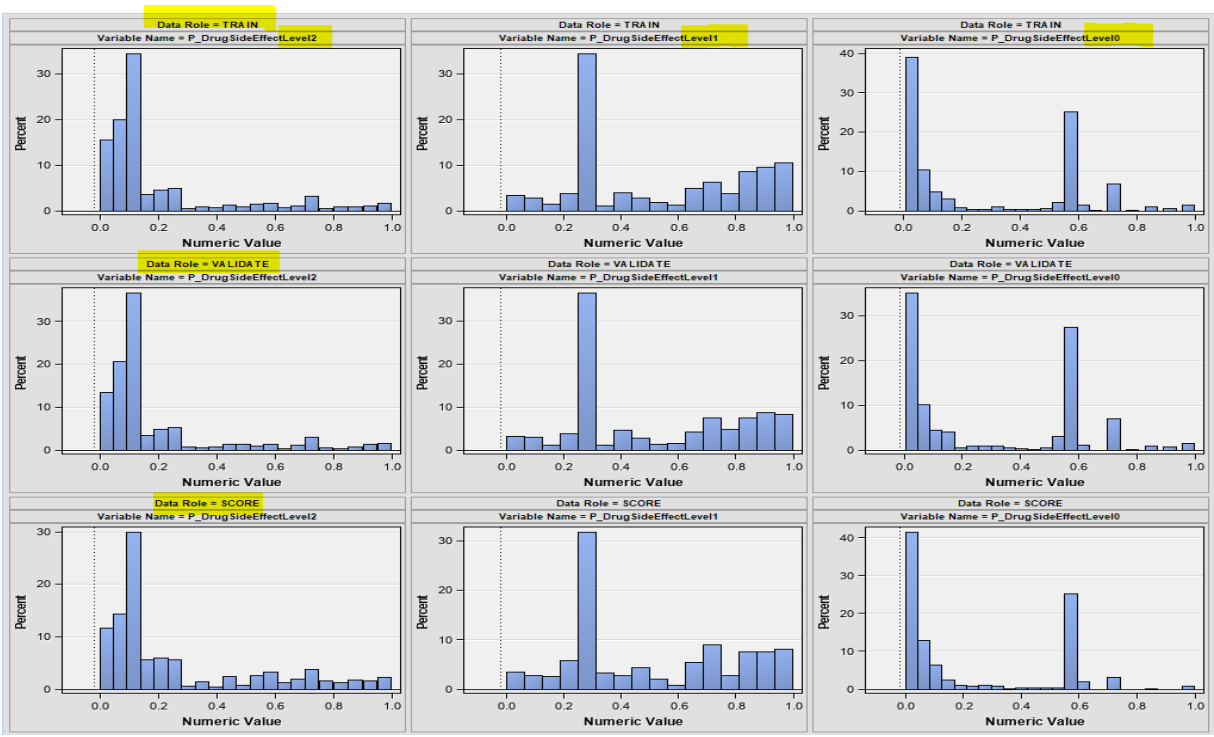


Figure 22 - Comparison of probability distribution of side effect level classification across train, validate, and score data sets

Figure 22 illustrates the probability distribution of each side effect level's categorization across train, validate, and score data sets. For example, the three vertical histograms on the far left depict the probability distribution of classifying users' comments into level 2 rating (Severe / Extremely Severe Side Effects) across three independent data sets. These three histograms have similar patterns (gradually decreasing) either in the train data set (first row), in the validate data set (second row), or in the score data set (third row). The same rules can be observed in the distribution of the probability of categorizing drug users' comments into level 1 rating - Mild / Moderate Side Effects (evidenced by the three vertical histograms in the middle) or into level 0 rating - No Side Effects (shown by the three vertical histograms on the far right). Also, looking from a horizontal dimension, the histograms depict different patterns across three levels, thus the model seems to work well in classification among three respective levels.

Overall, the histograms show consistent patterns for each rating level across various data sets, and different patterns across three levels. This implies that the selected text rule builder model is stable and robust, therefore can be used for further generalization in classifying drug side effect levels.

SCORING RESULTS FOR EFFECTIVENESS LEVEL CLASSIFICATION

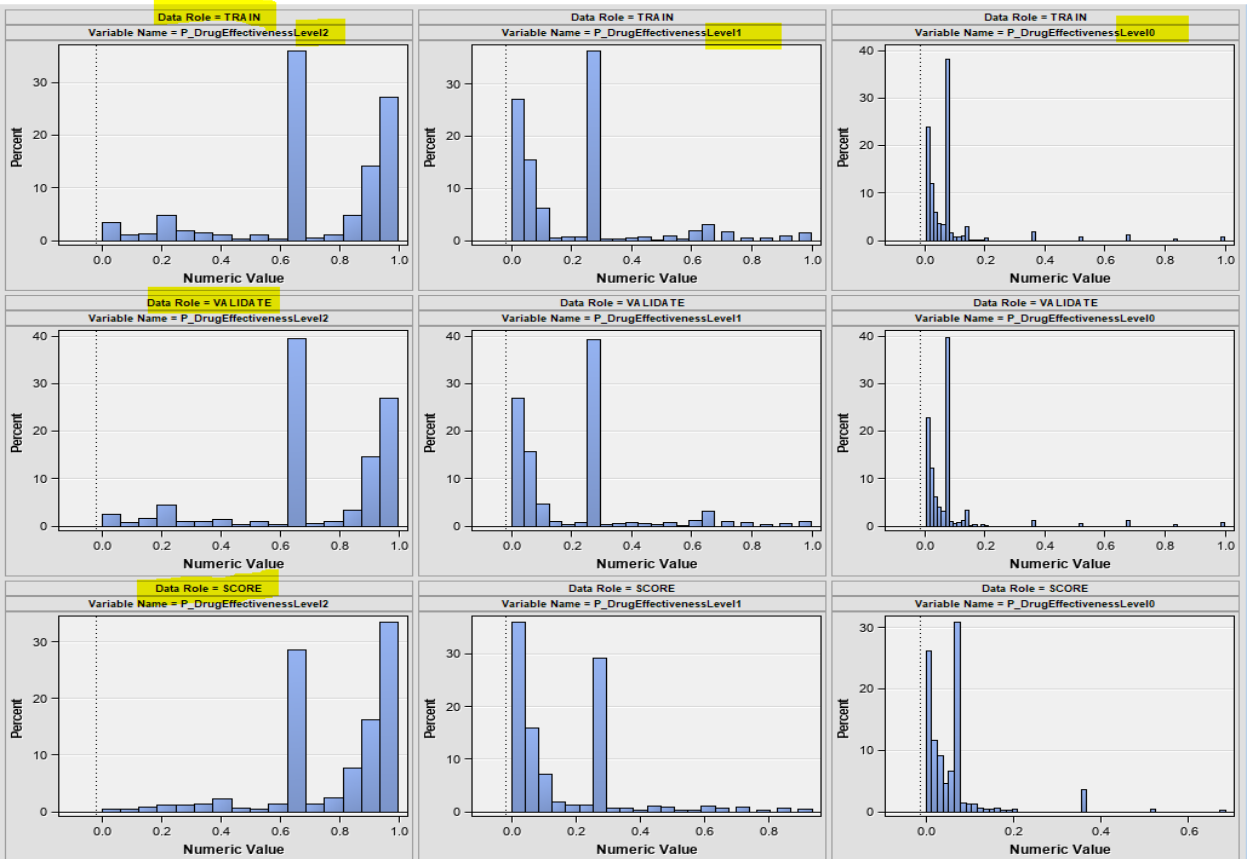


Figure 23 - Comparison of probability distribution of effectiveness level classification across train, validate, and score data sets

Figure 23 illustrates the probability distribution of categorizing each effectiveness level across train, validate, and score data sets. Similar to the scoring results of side effect classification, the histograms for effectiveness classification have consistent patterns for each rating level across train, validate and score data sets. This implies that the selected text rule builder model is working well in classifying the reviews in the score data set into three respective levels of drug benefits rating.

To sum up, the scoring results for both side effect classification and effectiveness classification indicate that the probability distribution of classifying users' comments into three respective levels of either side effects or effectiveness in the score data set looks considerably similar to those in the training and

validation data sets. This essentially implies that the selected Text Rule Builder models are validated and likely to work well for the score data, hence, they can be further improved for generalization in drug reviews classification.

DRUG EFFECTIVENESS EVALUATION

SENTIMENT ANALYSIS BY SAS VISUAL TEXT ANALYTICS

In order to detect or evaluate the specific effectiveness of a given drug, users' overall reviews for anti-depression drugs from *Drugs.com* have been subset to a new data set (approximately 14,425 reviews). This subset is then imported to SAS Visual Text Analytics for Natural Language Processing (using Concepts and Text Parsing), Sentiment Analysis, Feature Extraction (via Text Topic), and Text Modeling (i.e., Document Categorization).

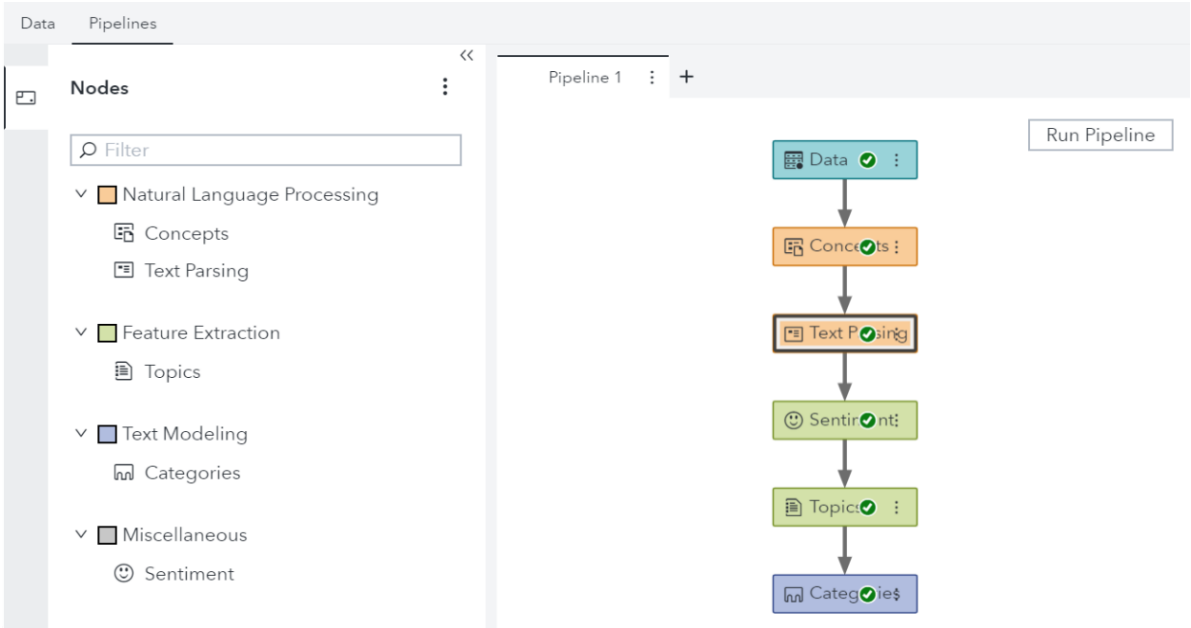


Figure 24 – Sentiment analysis flow

Concepts

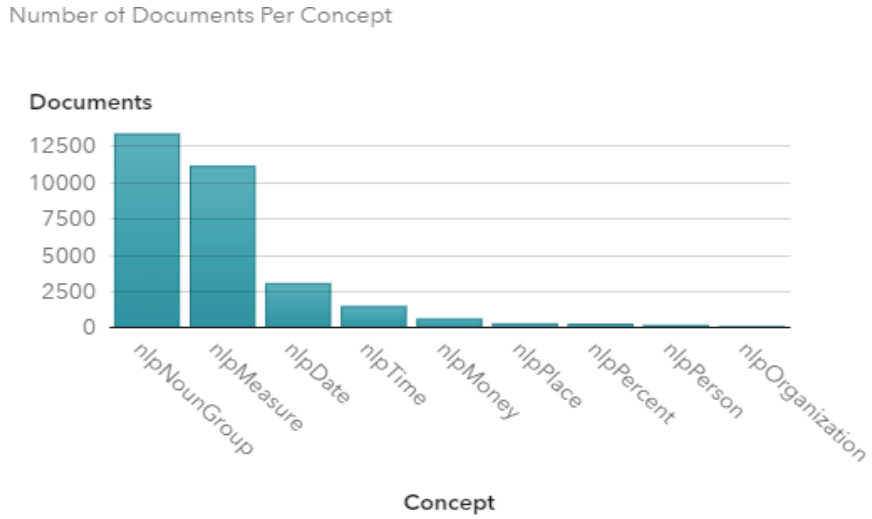


Figure 25 – Number of documents per concept

As shown in Figure 25, predefined concepts such as nlpDate, nlpMeasure, nlpMoney, nlpOrganization, nlpPerson, and nlpPlace are built by the Concept node to extract useful facts that could be helpful for indexing and searching, as well as additional analysis (for example, automatic concept extraction in future narratives). Most of the documents fall into nlpNounGroup, nlpMeasure, and nlpDate concepts.

Text Parsing

The Text Parsing node automatically extracts terms and noun groups from text by associating different parts of speech and understanding the context. Figure 26 below displays the proportion of Kept versus Drop terms across different groups.

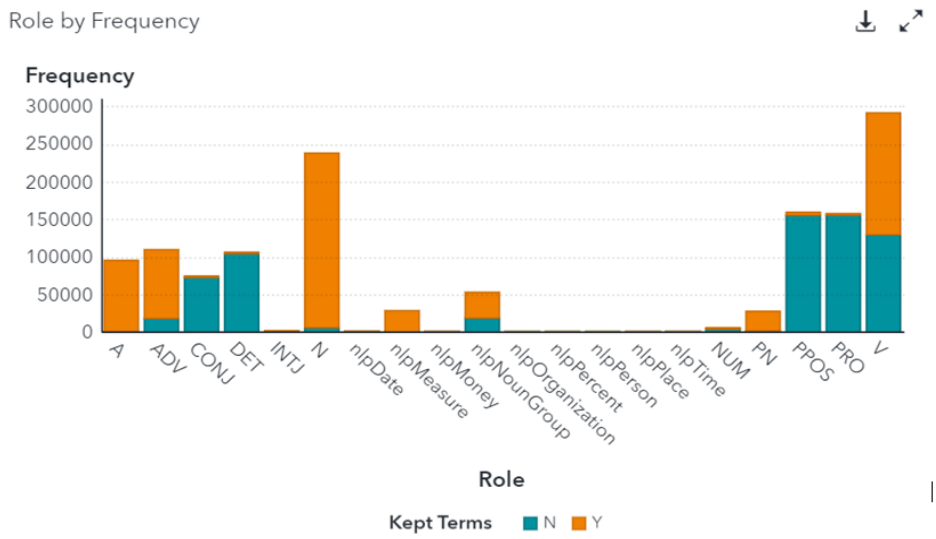


Figure 26 – Proportion of Kept versus Drop terms across different groups

Sentiment

The Sentiment node uses a domain-independent model that is included with SAS Visual Text Analytics. This rules-based analytic model computes sentiment relevancy for each post and classifies the emotion in unstructured text as *positive*, *negative*, or *neutral*. The results of Sentiment node are embedded in the Topic node's results as shown in Figure 27.

Text Topic

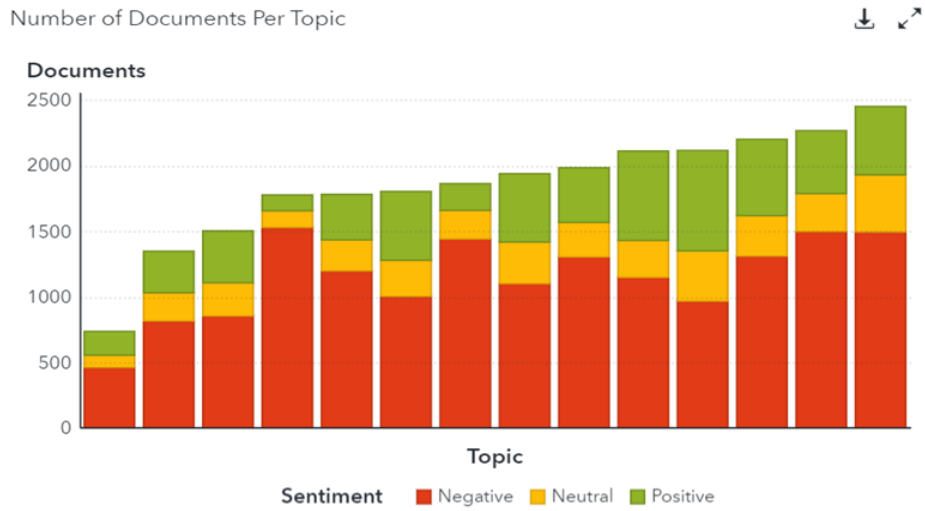


Figure 27 – Sentiment across different topics

Figure 27 demonstrates the sentiment classification across different topics. Overall, the documents that are categorized as negative sentiment accounts for large proportion in each topic. This implies that the overall reviews in the anti-depression data subset from drugs.com are mostly regarding side effects of the drugs. Some examples of observed common effects (both positive and negative) are provided in Figure 28 below.

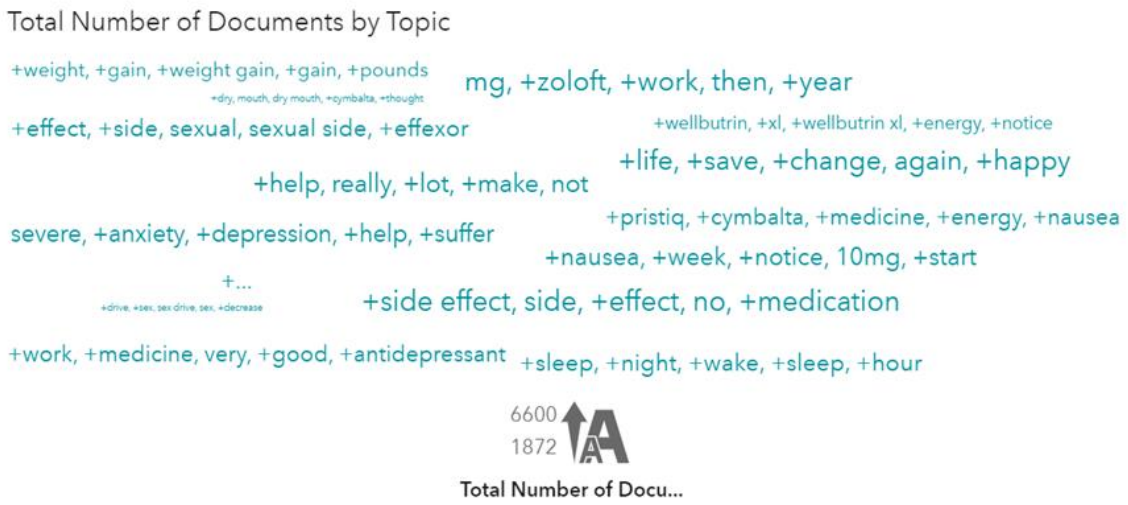


Figure 28 – Users' experienced effects from anti-depression drugs

TEXT ANALYTICS BY SAS ENTERPRISE MINER

Figure 29 - Unsupervised learning diagram for drug effectiveness evaluation

Figure 29 illustrates the process flow for unsupervised text analytics. For this part, only users' overall reviews for five common anti-depression drugs have been chosen for analysis. The SAS data set for each of these drugs is created and imported into SAS® Enterprise Miner 14.3, which is then partitioned into two data sets using the Filter node, one for low and medium ratings (from 1 to 7) and the other for high ratings (from 8 to 10). Next, text analytics with unsupervised learning algorithm is applied on these data sets, in which the overall 'reviews' variable is treated as the only text variable with no target variable to evaluate the effectiveness of each drug. The node properties settings for *Text Parsing*, *Text Topic*, and *Text Cluster* are customized the same as those in the Side Effect Classification part. Only the settings for "Term Weight" option and "Minimum Number of Documents" option in the *Text Filter* node are switched to default settings. The final results from the *Text Cluster* nodes for each drug are provided as below.

Wellbutrin XL

Cluster ID	Descriptive Terms	Percentage
8	dry mouth 'dry mouth' amp +deal +generic +add +headache +doctor +people +bupropion +totally xl better 'a lot of' ...	25%
1	'a lot' +hard +fall +help +first +week +know +long +hope appetite +depress +increase +feeling +'side effect' better ...	24%
2	+zoloft completely prozac +definitely +dose +diagnose +dream +lower +problem '150 mg' +lexapro +drug +people +big +stop ...	12%
5	+issue +meds 'a month' sad 300mg +life +gain +150mg +antidepressant +lose +definitely +suffer +month +mood +day ...	12%
3	last +cause +miserable +stand +experience +family +medication +subside '150 mg' +lexapro +prescribe +night +diagnose +lower angry ...	9%
4	+decide +several recently +right +mentally +totally few +back +suffer +life +down +first 'a month' +bed +diagnose ...	8%
7	+'side effect' side +effect +450mg +find +long sad +down +medicine +increase +symptom haven little +mentally +problem ...	8%
6	+bupropion +stand angry xl '2 weeks' 'at all' +family +gain irritable 'a lot of' +headache +medicine +antidepressant +people +know ...	1%

Figure 30 - Text Cluster node output for Wellbutrin XL rating 1-7 data

Figure 30 shows eight clusters generated for Wellbutrin XL 1-7 rating data. Clusters 8 and 1 have highest frequency percentages, indicating some potential common effects of Wellbutrin XL are dry mouth, headache, and loss of appetite.

Cluster ID	Descriptive Terms	Percentage ▼
3	+positive better +depress +medicine +medication +happy +thing +mood +recommend +notice energy +weight life +best +bad ...	25%
2	severe +stop +add +know +doctor +back +experience +want +month +side +attack +'side effect' anxiety life 300mg ...	23%
5	+brand insurance +generic +'300 mg' +mg difference +switch +great side +'side effect' +year +last +work +good +effect ...	20%
1	sexual +dream +negative +zoloft +side +sex +experience +increase +notice +quit +dose +bad +wellbutrin +lose +drive ...	16%
6	+insomnia sleep +night +major +happy +thing +dose +keep +recommend +sex +best +good +drive +150mg +great ...	13%
4	constipation always +wake +mind +slightly few +attack +drive +mood +amaze +love +sex +normal +increase +time ...	4%

Figure 31 - Text Cluster node output for Wellbutrin XL rating 8-10 data

Figure 31 shows six clusters generated for Wellbutrin XL 8-10 rating data. Cluster 3 has highest frequency percentage at 25%, indicating some effectiveness of Wellbutrin XL are positive effect, better feeling, happy mood, and more energy.

Lexapro

Cluster ID	Descriptive Terms	Percentage ▼
7	definitely +headache +'side effect' +emotion +long better +feel +night side +high +focus +notice +day +week +mood ...	15%
10	+weight gain +gain 'weight gain' +hard good +amaze 'a month' +lose +pound +working +antidepressant +exercise +know +cold ...	12%
5	+cold amp first +last next +eventually 5mg +med +20mg +prescribe +experience +finally nausea +'2 years' +head ...	10%
3	+keep dry +pain +nightmare +cause stomach +medicine insomnia +function few +feeling +'side effect' +hour +bed +episode ...	10%
14	suicide +drug reason +recommend +dream +doctor +constant +decide +hit +low +result +thought +cry insomnia +episode ...	8%
9	'10 mg' 'in the morning' 'sex drive' +difference +notice +mg panic +drive morning +attack +sex +prescribe +dose +experience anxious ...	8%
1	difficult effective better +antidepressant +well +effect +negative +high +long +normal +dose +work +feel +eventually reason ...	6%
2	+'work out' +eat +pill +pound different +frustrate +healthy +minute +daily +drive +half +upset +amaze +negative +year ...	6%
12	+back +normal 'all the time' lexapro +medication '20 mg' +positive +working anymore +finally anxious +tire 'a month' +episode control ...	6%
4	+fall +bed asleep sleep +eat +head morning +back +stop 'all the time' +dream +day +major +terrible 'in the morning' ...	5%
11	+medication depressive +major +tire help +bit +episode +eventually +friend +frustrate control +begin +exercise +focus +low ...	4%
6	+'suicidal thought' suicidal +thought control +people extreme lexapro +depress +begin +care +concentrate +decide +down +happen +attack...	3%
8	+result hospital severe +half +happen +right next +know lexapro +friend +healthy +minute +depress +care +late ...	3%
13	+antidepressant 'put on' +late +treat old +right '6 months' +10mg +bit +find +20mg +notice +function +back next ...	3%

Figure 32 - Text Cluster node output for Lexapro rating 1-7 data

Fourteen clusters are generated for Lexapro 1-7 rating data as shown in Figure 32. The top frequency percentage clusters depict that some possible common effects of Lexapro are headache, weight gain, nausea, nightmare, and insomnia.

Cluster ID	Descriptive Terms	Percentage
2	+friend people bed +life +mg life +lose +'10 mg' +back +bad able +medicine +day +month finally ...	18%
6	+zoloft back +difference +save +suffer +down +attack +drug +work +time +thought +first +depression side +experience ...	18%
1	'severe depression' severe first +prescribe +increase +notice +depression +anxiety +20mg +dose +attack +day +difference +experience +begin ...	15%
7	far +weight +10mg +mood insomnia +little +week +good gain side +find +notice 'weight gain' better couple ...	11%
3	+effexor +drug +escitalopram +well +year +begin +symptom +time great +dose +depression +anxiety +'20 mg' +work +20mg ...	9%
5	'weight gain' gain +'20 mg' +weight +mg couple +calm +'side effect' +effect +gain +well side +'10 mg' +mood able ...	8%
9	+'negative thought' +negative +thought +depress +long +function +cry +down morning +thing finally +feeling +life +little +sleep ...	8%
4	+drive +sex 'sex drive' +decrease appetite +weight side +feeling +effect +'side effect' +lose +good +medicine +gain +zoloft ...	7%
8	'haven't haven +night +swing 'at night' 'in the morning' morning +mood insomnia +decrease +notice +suffer +'side effect' +experience +help ...	5%

Figure 33 - Text Cluster node output for Lexapro rating 8-10 data

Figure 33 identifies nine clusters for Lexapro 8-10 rating data. Clusters 2, 6, and 1 have highest frequency percentage, indicating some effectiveness of Lexapro are life saving, able to help, finally work better.

Prozac

Cluster ID	Descriptive Terms	Percentage
9	+prescribe '10 mg' mg +feeling +'long time' severe +psychiatrist finally +dose +right first +happen anxiety +hope +stop ...	14%
2	'in the morning' +fluoxetine +night +morning +good few +well +love +little +last +10mg +daily +doctor +week +sleep ...	13%
7	+continue +high pill +wellbutrin definitely +dosage +notice +loss +different +add +issue +keep half +mood +medicine ...	12%
1	'a year' +help +year +wait +back always finally +numb half +weight +well +add +care +last +mood ...	11%
3	wish +happy body +begin 'to the point' completely +'suicidal thought' suicidal +lot +keep +psychiatrist +20mg +thought +low +switch ...	10%
8	+panic +attack +experience +symptom +cause +improvement 'a month' 'to the point' +lexapro +increase +mood +depression +10mg +major anxiety ...	10%
4	hospital +'suicidal thought' '3 weeks' +thought suicidal +life +depress +great +bad experience +loss +time +first +major always ...	9%
10	'sex drive' drive +sex +gain '3 months' +weight energy +care +function +low +issue +love +numb +happy +end ...	9%
5	+horrible +eat +hour +lose experience +end +medication +day +bad +cry +depress 'to the point' completely side +'side effect' ...	6%
6	+lexapro +major +disorder +right +switch +happen '3 weeks' +40mg +low +know +cry +issue +doctor 'a month' +notice ...	6%

Figure 34 - Text Cluster node output for Prozac rating 1-7 data

Ten clusters are generated for Prozac 1-7 rating data as shown in Figure 34. Most of the terms in high frequency clusters show negative side effects, examples being severe anxiety, trouble sleeping, often happening in the morning.

Cluster ID	Descriptive Terms	Percentage
2	'a year' 'a lot' world +fluoxetine +happy +people +bad +thing +few +good +old +thought +depress difference +little ...	26%
1	+'20 years' dosage +deal +mood +attack severe +antidepressant +help +'20 mg' +medicine +notice +anxiety +back +different mg...	18%
4	+live +night +exercise +low feel back +hope +time +lose +several prozac +month +first able +sleep ...	17%
6	+'side effect' side +effect appetite +haven +antidepressant +little +weight +switch +lose +several +cause +different +few +sleep ...	14%
5	mg +'10 mg' finally +begin difference anymore +day +notice +night +doctor +major +medication severe +'20 mg' +month ...	13%
3	'life saver' saver +event +decrease +sex life +save +medicine +depress +prescribe +20mg +life +great +low +major ...	13%

Figure 35 - Text Cluster node output for Prozac rating 8-10 data

Figure 35 shows that six clusters are generated for Prozac 8-10 rating data. Cluster 2 has highest frequency percentages, which indicates that Prozac receives some good reviews like a better feeling and happy mood.

Cymbalta

Cluster ID	Descriptive Terms	Percentage ▼
4	+leave +feeling +dose +night +wake +keep +mood +well +first +'30 mg' +nausea +doctor +day +mg +stop ...	28%
1	+side +'side effect' +effect 'a month' back +pain +bad first +notice +month +depression +tire +sweat +help +discontinue ...	19%
5	amp +'weight gain' gain 'a week' +gain +weight +difference +little +help +drug last +30mg anxiety +great +120mg ...	13%
6	+withdrawal +'withdrawal symptom' +symptom +dizziness +medication terrible +discontinue +120mg +awful +doctor +antidepressant +recommend...	13%
7	+sex +drug +drive +health +mind +reduce +recommend +problem horrible side +headache +medication +feel +life +awful ...	11%
2	+lose loss libido stomach +appetite +gain +back +find +tire +weight terrible +want +great +antidepressant +feeling ...	10%
3	suicidal +'suicidal thought' +thought +completely insomnia +daily +worsen +head +symptom +long +prescribe +120mg +side severe +drug ...	7%

Figure 36 - Text Cluster node output for Cymbalta rating 1-7 data

There are seven generated clusters for Cymbalta 1-7 rating data as shown in Figure 36. Clusters 4 and 1 have highest frequency percentages. Overall, Cymbalta is likely to have more side effects than benefits, some symptoms being nausea, back pain, sweating, weight gain, dizziness, and anxiety.

Cluster ID	Descriptive Terms	Percentage ▼
4	+save +love +life +antidepressant life +want best +drug finally +help +lose body different +depress +know ...	12%
5	+headache +begin +hour energy 'a day' +tire few +withdrawal +happy +stop back +60mg +cry horrible +month ...	11%
8	mg '30 mg' '60 mg' +doctor +lose +cry +start +want +thing +miss +suffer +zoloft +increase +prescribe energy ...	10%
12	paxil +effexor 'a year' prozac great +side +keep +little better +'side effect' +zoloft +wellbutrin +medication able +effect ...	10%
7	+pay +hard +know +attack +live +anxiety insurance +people brain +far +switch +full +long +dose +meds ...	9%
1	+night 'at night' sleep +sleep +down +wake working +morning +half +tire +medicine +stop +anti-depressant +side 'a lot' ...	9%
11	+first '3 weeks' +mood +week +day +long appetite few nausea +far +problem +effect +headache +morning +recommend ...	9%
10	amp +meds chronic +pain +find +diagnose +ptsd different +increase 'a day' +cripple +completely +feeling +last +look ...	8%
6	+real 'a lot of' +weight +depressive +gain dizziness +eat appetite +drug +major +lose nausea +well +'side effect' +decrease ...	7%
9	+thought +job +'suicidal thought' suicidal suicide +cripple +depress +thing +lift able +completely +life +morning +hard +long ...	7%
2	+pain +physician body +look life +ptsd +heart 'a lot' +nerve back +attack chronic +recommend +find +suffer ...	5%
3	+'withdrawal symptom' +symptom +withdrawal +miss 'at night' horrible +problem brain best +night +dose +want +60mg +begin +switch...	4%

Figure 37 - Text Cluster node output for Cymbalta rating 8-10 data

Figure 37 demonstrates 12 clusters that are generated for Cymbalta 8-10 rating data. Clusters 4, 5, 8, and 12 have highest frequency percentages, which show a blend of both benefits and side effects. Some reviewers compliment this drug as best, helpful, life saving anti-depressant treatment, whereas others claim a couple of negative effects including insomnia, nausea, headache, weight gain, loss of appetite, dizziness, and suicidal thought.

Effexor

Cluster ID	Descriptive Terms	Percentage ▼
2	+shake +wake +attack +night +drug +life +week horrible 'put on' +feel effexor prozac +numb +sweat +job ...	32%
7	+'withdrawal symptom' 'weight gain' gain +symptom side dosage +effect brain +'side effect' +miss weight +zap +withdrawal +few +150mg...	20%
6	'75 mg' +know +down +little +mg +mood +'6 months' +back +depress +day +good 75mg +sleep +year +pill ...	19%
3	+past +gain +'effexor xr' +feeling +effect +stop 75mg side +depression +'side effect' +medication +week +work +awful +extremely ...	13%
1	+read 'cold turkey' turkey +review meds +cold system med terrible +vomit +eye +cause +recommend 'one day' nauseous ...	9%
4	+blur +immediately vision +handle clearly 'one day' +job +absolutely +vomit +numb prozac +head +different +lose 75mg ...	4%
5	barely orgasm nauseous 'put on' nausea +body +advise +concentrate +awful +extremely +mood +pill +sweat dizzy +big ...	4%

Figure 38 - Text Cluster node output for Effexor rating 1-7 data

Seven clusters are generated for Effexor 1-7 rating data as shown in Figure 38. Cluster 2 has highest frequency percentages, which implies that some side effects of Effexor are that it takes long time for the drug to show effects, trouble sleeping, horrible feelings, numbness, and sweating.

Cluster ID	Descriptive Terms	Percentage
6	+weight +dream +far +major different 75mg +dose +'side effect' +thing +drug +meds +antidepressant +experience +happy +depress...	24%
1	first +well +cry +last little +high +long +sweat +good +prescribe +happy +start +work +few +stay ...	20%
4	+switch 'a day' +day +low +prozac +miss +'effexor xr' +year +'2 years' +extremely +help +begin +dosage +best +normal ...	18%
3	mg +'75 mg' +attack panic +want +honestly +late +normal +stay +anxiety better +know +dosage +medication +time ...	15%
5	+notice 'a week' +week difference +celexa side +effect +back +find +dosage +drug +thing +normal +completely +great ...	12%
2	+medicine +begin suicidal good +experience +save +honestly +meds +problem +forget +prescribe +feeling +best different +lose ...	11%

Figure 39 - Text Cluster node output for Effexor rating 8-10 data

Figure 39 depicts six clusters generated for Effexor 8-10 rating data. Clusters 6 and 1 have highest frequency percentages, which implies that some effectiveness of Effexor are happy mood, well working antidepressant. Other experienced side effects are sweating, crying, and weight gain.

COMPARISON OF EFFECTIVENESS OF FIVE DRUGS

Drug	Low-rating effects	High-rating effects	Average rating
Wellbutrin XL	dry mouth, headache, loss of appetite	better feeling, happy mood, more energy	7.59
Lexapro	headache, weight gain, nausea, nightmare, insomnia	life-saving, able to help, finally work better	7.58
Prozac	severe anxiety, trouble sleeping, mostly in the morning	better feeling, happy mood	7.29
Cymbalta	nausea, back pain, sweating, weight gain, dizziness, anxiety	best, helpful, life-saving anti-depressant treatment	6.47
Effexor	trouble sleeping, horrible feelings, numbness, sweating, take long time to show effects	happy mood, well working antidepressant	5.82

Table 4 – Comparison of effectiveness of five anti-depressant drugs

Table 4 helps understand the specific benefits and side effects of each of the five selected prescribed anti-depression drugs, which can serve as practical guidelines to prospective clients in making informed decisions of choosing the best and suitable drug for anti-depressant treatment. For example, they may take into thorough consideration the possible side effects of a given drug and determine if the benefits can outweigh the side effects and then compare these features with those of other similar drugs. Hence, overall, sentiment analysis and text analytics with unsupervised learning algorithm as analyzed above can facilitate patients in exploring experienced users' reviews and provide them with helpful recommendations in selecting the best drug for their treatment.

CONCLUSION

Increasingly, customers are using social media and other Internet-based applications (e.g., online review sites, discussion forums) to express their sentiments on experienced drugs. These reviews contain a wealth of useful information regarding user preferences and experiences over multiple prescription drugs which can be further leveraged to provide valuable insights to both health care professionals and drug users. However, given the unstructured, qualitative, and textual nature of the comments, potential customers would find it overwhelmingly challenging to go through all online reviews before making purchase decisions. The present paper utilizes best practices of text mining and supervised learning

algorithm within SAS® Enterprise Miner™ 14.3 to perform text analytics on online drugs reviews for feature engineering. Multiple predictive models are then optimized and trained on the extracted feature representations, among which the Text Rule Builder is found to be the best performing model for drug side effects classification as well as for effectiveness classification. In addition, the paper also examines the transferability of the selected trained classification models to ensure for better validation and generalization across independent data sources. Further, unsupervised sentiment analysis and text mining using SAS® Visual Text Analytics from SAS® Viya are also employed to detect the specific side effects and effectiveness of several selected anti-depression drugs which can serve as practical guidelines for potential users. Overall, the study expects to provide valuable insights to assist prospective drug users in making informed purchase decisions and improve monitoring public health by revealing collective experience. A future challenge would be fully analyzing the reviews at deeper level by employing more sophisticated aspect-based sentiment analysis and more powerful advanced machine learning models for improved results.

REFERENCES

Chakraborty, G., Pagolu, M., & Garla, S. (2014). *Text mining and analysis: practical methods, examples, and case studies using SAS*. SAS Institute.

Fan and Fuel, 2016. "No online customer reviews means BIG problems in 2017". Accessed March, 2019. https://fanandfuel.com/no-online-customer-reviews-means-big-problems-2017/

Gräßer, F., Kallumadi, S., Malberg, H., & Zaunseder, S. 2018. "Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning." *Proceedings of the 2018 International Conference on Digital Health*, 121-125. ACM.

Liu, J., Sarkar, M.K., & Chakraborty, G. (2013). "Feature-Based Sentiment Analysis on Android App Reviews Using SAS® Text Miner and SAS® Sentiment Analysis Studio'. *Proceedings of the SAS Global Forum 2013.*

Spiegel Research Center. 2017. "Data-Driven Insights on How Retailers Can Maximize the Value of Engagement with Consumers Through Online Reviews". Accessed February, 2020. https://spiegel.medill.northwestern.edu/_pdf/Spiegel_Online%20Review_eBook_Jun2017_FINAL.pdf

UC-Irvine. 2018. "Machine Learning Repository: Drug Review Data set (Druglib.com) Data set". Accessed March, 2019. https://archive.ics.uci.edu/ml/data sets/Drug+Review+Data set+%28Druglib.com%29

UC-Irvine. 2018. "Machine Learning Repository: Drug Review Data set (Drugs.com) Data set". Accessed March, 2019. https://archive.ics.uci.edu/ml/data sets/Drug+Review+Data set+%28Drugs.com%29

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Thu Dinh
Oklahoma State University
Stillwater, OK 74078
Phone: (405) 612-2129
Email: thu.dinh@okstate.edu

Dr. Goutam Chakraborty
Oklahoma State University
Stillwater, OK 74078
Phone: (405) 744-7644
Email: goutam.chakraborty@okstate.edu