MWSUG 2019 – Paper IN64

Detecting Side Effects and Evaluating Effectiveness of Drugs from Customers' Online Reviews using Text Analytics and Data Mining Models

Thu Dinh, Oklahoma State University, Stillwater, Oklahoma

ABSTRACT

Drug reviews play a very important role in providing crucial medical care information for both healthcare professionals and consumers. Customers are increasingly utilizing online review sites, discussion boards and forums to voice their opinions and express their sentiments about experienced drugs. However, a potential buyer typically finds it very hard to review all of these online comments before making a purchase decision. Another big challenge would be the unstructured, qualitative, and textual nature of the reviews, which makes it difficult for readers to classify the comments into meaningful insights. The aim of this research is to create a data-mining model to evaluate the effectiveness and detect potential side effects from online customer reviews on specific prescriptive drugs. The study utilizes text parsing, text filtering, text topic, and text clustering within SAS® Enterprise Miner™ 14.3 for feature engineering and supervised learning algorithm for building multiple predictive models (logistic regression, decision tree, neural network, text rule builder) to identify the optimal model for reviews classification. The study's results show that the best predictive model for side effect classification is the text rule builder model with a validation average square error of 5.79% and a misclassification rate of 31.57%. Regarding effectiveness classification, text rule builder model also works best with 5.10% validation average square error and 29.08% misclassification rate. These models are further validated using transfer learning algorithm to evaluate model performance and generalization. The results can help as practical guidelines and useful references for 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 in making 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 are capable of providing more useful insights into the effectiveness and particular side effects of the drug, which overall ratings cannot. However, with a daily increasing number of textual comments from users, it has become more and more challenging for potential users to go through all of the reviews before making decisions. Therefore, an efficient structured algorithm is needed to explore the reviews and classify them into meaningful attributes which can serve as helpful recommendation to potential buyers. In light of that, the primary goal of this study is to construct an optimal data-mining model to evaluate the effectiveness and detect potential side effects of prescribed drugs from online customer reviews. The training data are collected from *druglib.com* to build predictive models which are then validated on the data gathered from *drugs.com* using transfer learning. The results of the study expect to provide some useful references and practical guidelines on drug effectiveness and side effects for prospective patients in making their informed purchase decisions.

DATA PREPARATION

DATA SOURCE

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 first converted to excel format and later imported to SAS® Enterprise Miner 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

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

ID	urlDrugName	rating	effectiveness	sideEffects	condition	benefitsReview	sideEffectsReview	commentsReview
1366	biaxin	9	Considerably Effective	Mild Side Effects	sinus infection	The antibiotic may have de	Some back pain, some	Took the antibiotics for 1
3724	lamictal	9	Highly Effective	Mild Side Effects	bipolar disorder	Lamictal stabilized my serio	Drowsiness, a bit of me	Severe mood swings betw
3824	depakene	4	Moderately Effective	Severe Side Effects	bipolar disorder	Initial benefits were compa	Depakene has a very th	Depakote was prescribed
969	sarafem	10	Highly Effective	No Side Effects	bi-polar / anxiety	It controlls my mood swing	I didnt really notice any	This drug may not be for
696	accutane	10	Highly Effective	Mild Side Effects	nodular acne	Within one week of treatm	Side effects included m	Drug was taken in gelatin
1380	biaxin	2	Marginally Effective	No Side Effects	sinus infection	By the end of the 10-day tr	I felt no significant side	Basically the treatment di
45	carbamazepine	8	Considerably Effective	Moderate Side Effects	seizure	reduction in seizures reduct	tired/sleepy very tired	took it for seizure took pil
1939	ultram-er	10	Highly Effective	Mild Side Effects	cervical disk degenerati	Ive been taking Tramadol f	I have had no side effec	Treating for neck, shoulde
2576	klonopin	10	Highly Effective	No Side Effects	panic disorder	I immediately stopped havi	I experienced no side e	I started both klonopin an
1093	effexor	1	Marginally Effective	Extremely Severe Side Effects	depression	the presumed benefits wer	here we go.the initial e	family doctor initially pres

Figure 1 - Partial Data of the *Druglib.com* Data Set

The second data set from *Drugs.com* consists of 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 2 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 d	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 fo	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 2 - Partial Data of the *Drugs.com* Data Set

METHODOLOGY

APPROACH

With a primary aim to detect side effects and evaluate effectiveness of prescription drugs from online customers' reviews by employing text analytics and data mining models, this study treats these tasks as classification problems. The text reviews are transformed into textual units which are then consolidated to new variables to form feature representations for classifiers. Next, we train the classifiers using supervised learning on the *Druglib.com* data set to build several predictive models in order to classify side effect levels and effectiveness levels. Then we use transfer learning algorithm to score the best performing model on *Drugs.com* data set to evaluate model validation and generalization.

This study approach can be visually illustrated by the following figure.

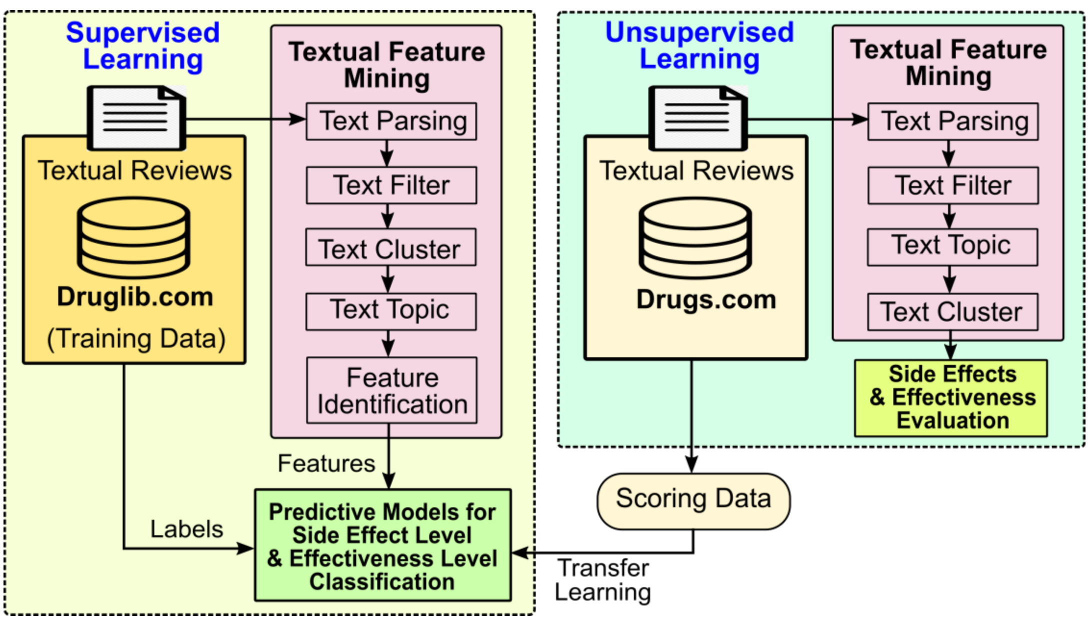


Figure 3 – Approach for side effect and effectiveness classification

TARGET VARIABLES

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

Target Variables	Values	Level	Frequency (Percentage)
DrugSideEffectLevel	No Side Effects	0	131 (20.00%)
	Mild / Moderate Side Effects	1	420 (64.12%)
	Severe / Extremely Severe Side Effects	2	104 (15.88%)
DrugEffectivenessLevel	Ineffective	0	61 (9.31 %)
	Marginally / Moderately Effective	1	128 (19.54%)
	Considerably / Highly Effective	2	466 (71.15%)

Table 3 – Models target variables

STATISTICAL TESTS

The study first performs cross tabulation and Chi-Square significant tests to determine whether there is any significant association:

- between the 10-star overall satisfaction rating (‘*rating*’ variable) and the three-level side effect rating (‘*DrugSideEffectLevel*’ variable), or
- between the 10-star overall satisfaction rating (‘*rating*’ variable) and the three-level effectiveness rating (‘*DrugEffectivenessLevel*’ variable).

The results of the above preliminary tests are summarized as below.

Statistics for Table of rating by DrugSideEffectLevel

Statistic	DF	Value	Prob
Chi-Square	18	2462.8184	<.0001
Likelihood Ratio Chi-Square	18	2187.6139	<.0001
Mantel-Haenszel Chi-Square	1	1543.5509	<.0001
Phi Coefficient		0.7714	
Contingency Coefficient		0.6108	
Cramer's V		0.5454	

Sample Size = 4139

Statistics for Table of rating by DrugEffectivenessLevel

Statistic	DF	Value	Prob
Chi-Square	18	3011.7882	<.0001
Likelihood Ratio Chi-Square	18	2657.2643	<.0001
Mantel-Haenszel Chi-Square	1	2059.6952	<.0001
Phi Coefficient		0.8530	
Contingency Coefficient		0.6490	
Cramer's V		0.6032	

Sample Size = 4139

Figure 4 - Statistical tests of rating and DrugSideEffectLevel/ DrugEffectivenessLevel

Figure 4 indicates that the p-values for both Chi-Square tests are less than the 5% level of significance (Prob < .0001). Hence there exists a statistically significant association between the overall rating and the side effect rating (the strength of the association is medium, as shown by the Cramer’s V value of 0.5454). Similarly, there is also a statistically medium strong association between the overall rating and the effectiveness rating (with Cramer’s V value of 0.6032). Overall, there is a significant relationship between each individual rating and the overall rating of prescribed drugs.

SIDE EFFECT CLASSIFICATION

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

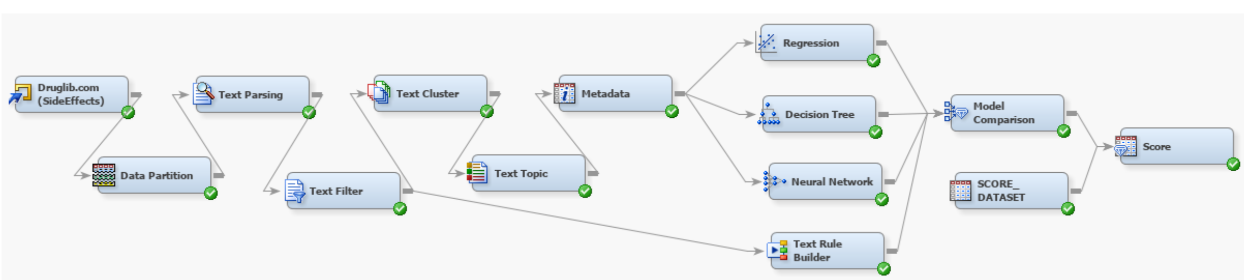


Figure 5 - Modeling diagram for side effect classification

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 “DrugSideEffectLevel” 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 effects 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

The Text Parsing node is connected to the Data Partition node with customized settings as below:

- The “Detect Different Parts of Speech” option is set to ‘yes’ to be able to treat the same words of different parts of speech as different.
- The “Detect Find Entities” option is set to ‘Standard’.
- The “Ignore Parts of Speech” list is set to include the following choices: ‘Abbr’, ‘Aux’, ‘Conj’, ‘Det’, ‘Interj’, ‘Num’, ‘Part’, ‘Prep’, ‘Pron’, ‘Prop’.
- The “Ignore Types of Attributes” is set to: ‘Num’, ‘Punct’.

As a result, the Text Parsing node generates a Term by Document matrix which helps identify the most frequently occurring words and the number of comments in which each word occurs. Figure 6 below displays partial Term by Document matrix for comments on side effects.

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
+ month ...	Noun	Alpha	202	170	Y	+	2649	38
+ notice ...	Verb	Alpha	193	169	Y	+	8409	39
x000d x000d ...	Noun	Mixed	397	167	Y		2274	40
+ problem ...	Noun	Alpha	186	165	Y	+	3331	41
+ increase ...	Verb	Alpha	181	160	Y	+	7586	42
+ sleep ...	Verb	Alpha	181	158	Y	+	3243	43
+ cause ...	Verb	Alpha	184	157	Y	+	2353	44
first ...	Noun	Alpha	169	154	Y		135	45
+ hour ...	Noun	Alpha	166	147	Y	+	8235	46
+ seem ...	Verb	Alpha	160	141	N	+	13023	47
+ stop ...	Verb	Alpha	165	141	Y	+	8207	47
+ bad ...	Adj	Alpha	154	140	Y	+	1639	49
+ treatment ...	Noun	Alpha	166	140	Y	+	1545	49

Figure 6 - Text Parsing results for reviews on side effects

Some of the most commonly used words by reviewers in the comments are “effect”, “dry”, “skin”, “nausea”, “pain”, “headache”, “stomach”, etc., which is expected as these words relate to some common side effects of prescription drugs.

TEXT FILTER

Further, the Text Parsing node is connected to the Text Filter node which helps figure out the words that occur most/ least number of times as specified in the properties panel. Specifically, the settings are customized as below:

- The “Check Spelling” option is set to ‘yes’, which enables SAS to create correctly spelled synonyms for misspelled words.
- The “Term Weight” option is set to “Mutual Information” (with a categorical target variable, mutual information weighting technique can be used to derive meaningful weights to the terms).
- The “Minimum Number of Documents” option is set to 3 (any terms that occur in fewer than three documents will be excluded).

As an illustration of how the Text Filter node works, the below Term table from the Interactive Filter Viewer result shows various forms of some commonly used words in reviewers’ comments on side effect, such as “severe”, “nausea”, “pain”, “stomach”, “headache”. Each of these words are grouped together with its misspelled derivations into one general term by SAS Enterprise Miner.

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
headache	196	179	✓	0.104	Noun	Alpha
headache	99	91			Noun	Alpha
headahe	1	1			Noun	Alpha
headaces	1	1			Noun	Alpha
headeaches	1	1			Noun	Alpha
headachy	1	1			Noun	Alpha
headaches	93	88			Noun	Alpha
loss	227	179	✓	0.193	Noun	Alpha

Figure 7 - Text Filter results for reviews on side effects

Concept links

Concept links, which can be accessed under the Interactive Filter Viewer from the properties panel of the Text Filter node, help understand the association between terms based on their co-occurrence in the documents. The focal term of analysis is placed at the center of the concept link diagram whereas the terms that are associated with the centered term are connected to it using links. The hub and spoke structure of the link represents the association between those terms and the thickness of the link explains the strength of association. Below are the concept links for some of the most frequent terms:

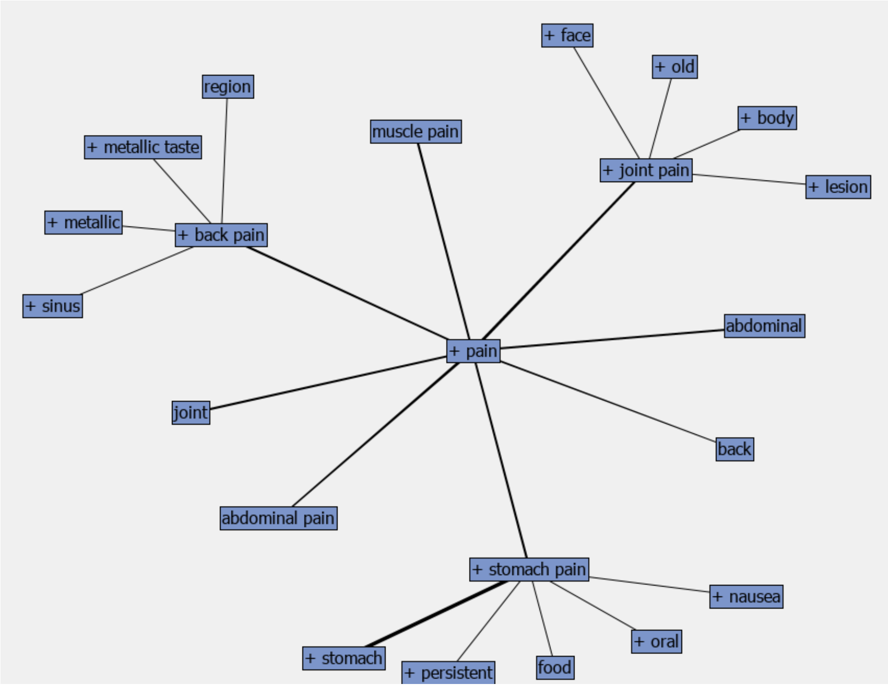


Figure 8 - Concept links for the term "pain"

The concept link diagram in Figure 8 shows that the term "pain" is associated with such terms as "muscle pain", "back pain", "abdominal pain", "stomach pain", "joint pain". Hence, it can be inferred that these are some commonly found "pain" side effects of prescription drugs.

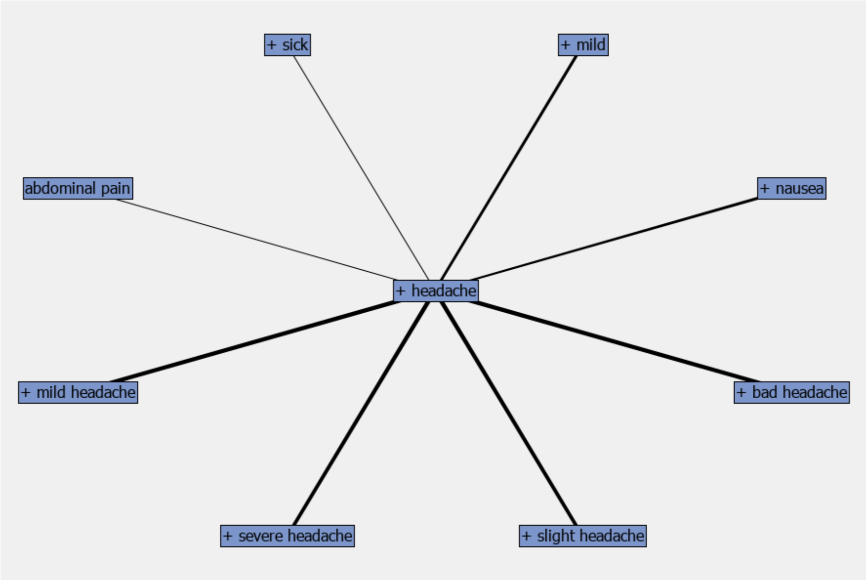


Figure 9 - Concept links for the term "headache"

Similarly, the concept link diagram in Figure 9 indicates that the term “headache” is strongly associated with “bad headache”, “slight headache”, “severe headache”, and “mild headache”.

TEXT CLUSTERING

The Text Cluster node is connected to the Text Filter node to group terms that closely relate to each other into separate clusters of related terms. Using a trial and error method, the properties settings for the Text Cluster node are customized as below to generate well-separated clusters in the cluster space.

- Max SVD Dimensions: 40
- Number of clusters: 15
- Cluster Algorithm: Expectation-Maximization
- Number of Descriptive Terms: 15

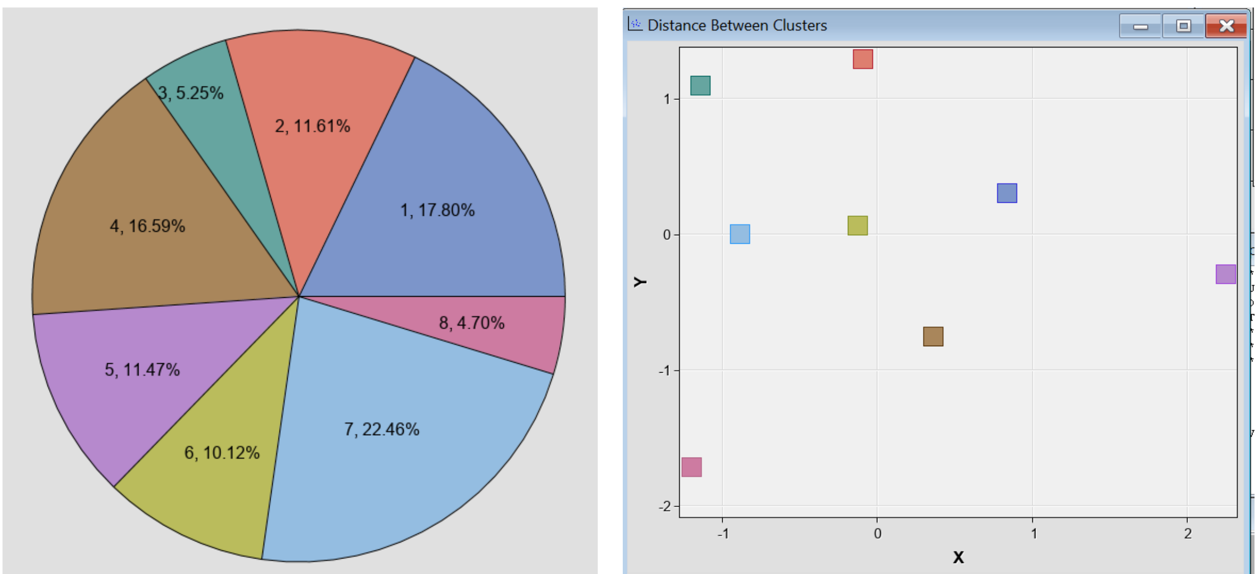


Figure 10 - Text Cluster node results for reviews on side effects

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 11 - Text Cluster descriptive terms for reviews on side effects

Text Cluster node generates eight well-separated clusters as shown in Figure 10 and Figure 11. Cluster 7 has the highest frequency (22%) with such descriptive terms as “week”, “start”, “feel”, “first”, “morning”, “hour”, “feeling”, etc., which often occur together. It can be interpreted that some side effects from the above cluster could be related to bad feeling, or not feeling like to eat in the morning, or hard to sleep at night which often happen on the first few hours/ days/ weeks using the drugs.

TEXT TOPIC

Text Topic node is connected to the Text Cluster node, which enables SAS to combine terms into topics for obtaining further valuable insights from data. The number of Multi-Term Topics has been set to 15 (through trial and error) to examine the features that reviewers are more interested to comment about the drugs.

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 12 - Text Topic results for reviews on side effects

Figure 12 shows 15 different topics with corresponding number of terms in each topic and also number of documents that contain the topic terms. For example, topic 2 indicates that drug users may experience side effects like severe nausea or diarrhea, whereas topic 5 addresses some 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, the 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 (DrugSideEffectLevel).

Target Value	Rule # ▲	Rule	Valid Precision	Precision	Valid Recall	Recall	Valid F1 score	F1 score	True Positive/Total	Valid True Positive/Total
1	1	mild & ~effect	83.64%	87.67%	6.97%	8.33%	12.87%	15.21%	129/150	47/56
1	2	peel	86.84%	88.14%	10.00%	11.13%	17.93%	19.76%	43/49	20/21
1	3	dry	84.25%	86.88%	18.64%	21.54%	30.52%	34.52%	195/239	66/87
1	4	decrease	83.22%	87.56%	18.79%	22.90%	30.66%	36.31%	24/27	4/6
1	5	occasional	82.39%	87.91%	19.85%	24.59%	31.99%	38.43%	41/46	15/19
1	6	slightly	82.74%	87.61%	21.06%	26.22%	33.57%	40.36%	30/36	13/15
1	7	reduce	81.97%	87.14%	22.73%	27.78%	35.59%	42.13%	37/46	16/22
1	8	a bit	81.44%	86.65%	23.94%	29.99%	37.00%	44.56%	46/59	17/23
1	9	decrease	81.86%	86.44%	25.30%	31.95%	38.66%	46.65%	46/56	18/21
1	10	slight	82.27%	85.67%	27.42%	35.39%	41.14%	50.09%	91/119	22/29
1	11	first	82.64%	85.57%	30.30%	39.36%	44.35%	53.92%	120/157	39/49
1	12	drowsiness	82.06%	85.43%	32.58%	41.18%	46.64%	55.58%	58/80	34/47
1	13	taste	81.68%	85.51%	33.79%	42.23%	47.80%	56.53%	32/43	12/19
1	14	dryness	80.84%	85.35%	35.15%	43.20%	49.00%	57.37%	43/52	20/26
1	15	appetite	80.46%	85.29%	36.82%	44.89%	50.52%	58.82%	76/101	22/34
1	16	constipation	80.19%	85.48%	38.03%	46.71%	51.59%	60.41%	56/76	19/32
1	17	beginning	79.87%	85.48%	38.48%	47.50%	51.94%	61.06%	31/39	12/15
1	18	tinnitus	80.00%	85.60%	38.79%	47.95%	52.24%	61.47%	15/17	3/4
1	19	gain	79.64%	85.33%	39.70%	49.58%	52.98%	62.72%	52/72	18/29
1	20	drowsy	80.00%	85.34%	40.61%	50.36%	53.87%	63.34%	21/26	11/12
1	21	increase	80.17%	84.98%	43.48%	53.03%	56.39%	65.30%	111/160	47/65
1	22	tired	79.67%	84.99%	44.55%	53.81%	57.14%	65.90%	27/34	13/22
1	23	little	79.21%	84.30%	45.61%	55.56%	57.88%	66.98%	86/116	28/44
1	24	stool	79.32%	84.41%	45.91%	56.02%	58.16%	67.34%	13/15	3/5
1	25	dream	79.18%	84.47%	46.67%	56.60%	58.72%	67.78%	34/43	13/20
1	26	sensation	79.13%	84.49%	47.12%	57.06%	59.07%	68.12%	26/34	7/14
1	27	tire	79.46%	84.17%	48.64%	58.10%	60.34%	68.75%	47/64	22/28
1	28	sensitivity	79.27%	83.87%	49.24%	58.88%	60.75%	69.19%	35/48	12/18
1	29	headache & ~effect	78.40%	83.41%	50.61%	60.18%	61.51%	69.92%	93/128	32/57
1	30	skin	78.75%	82.73%	53.33%	62.00%	63.60%	70.88%	164/229	73/92

Figure 13 - Text Rule Builder results for reviews on side effects

The above Rules Obtained table displays rules for predicting the target variable. These rules are presented as the conjunction of terms and their negations. For example, Rule 1 "mild & ~effect" says that for a document to satisfy this rule, it must contain the term “mild” and should not contain the term “effect”. This term has a valid precision of 83.64% which implies that the precision for validation data for all rules up to this point in the table for the target value for matching documents that are actually assigned to that target value is 83.64%.

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. After trial and error, the customized setting with high Generalization Error, very low Purity of Rules and low Exhaustiveness produced the best results with lowest Average Square Error and Misclassification Rate.

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 5, in all these models, the categorical variable "DrugSideEffectLevel" is set to be 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 be 'Stepwise'
- Model selection criterion is set to be 'Validation Error'

DECISION TREE

The Decision Tree node is set up with below settings:

- Subtree selection method is set to be 'Assessment' (i.e., the smallest subtree with the best assessment value is chosen)
- Subtree assessment measure is set to be 'Average Square Error'

NEURAL NETWORK

The Neural Network node is set up with below setting:

- Model selection criterion is set to be 'Average Error'

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 the Model Comparison node are set up as following:

- Model selection statistic: Average Square Error
- Model selection table: Validation

The Model Comparison results are provided in the below table.

Selected Model	Model Description	Target Variable	Selection Criterion: Valid: Average Squared Error
Y	Text Rule Builder	DrugSideEffectLevel	0.057913
	Regression	DrugSideEffectLevel	0.135656
	Neural Network	DrugSideEffectLevel	0.138367
	Decision Tree	DrugSideEffectLevel	0.144103

Figure 14 – Comparison between models for side effect classification.

Figure 14 indicates that among the four interested 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 Average Squared Error (ASE) at 5.79% as compared to the other three models.

EFFECTIVENESS LEVEL CLASSIFICATION

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

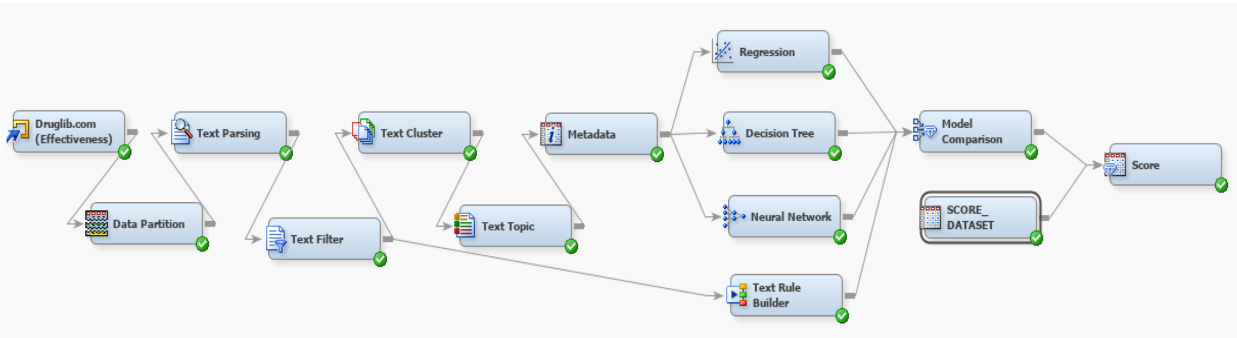


Figure 15 – Modeling diagram for effectiveness classification

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 "DrugEffectivenessLevel" 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

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
blood	...Noun	Alpha	179	138	Y		1838	50
+ night	...Noun	Alpha	174	137	Y	+	6641	51

Figure 16 - Text Parsing results for reviews on effectiveness

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.

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 17 - Text Filter results for reviews on effectiveness

Concept Links

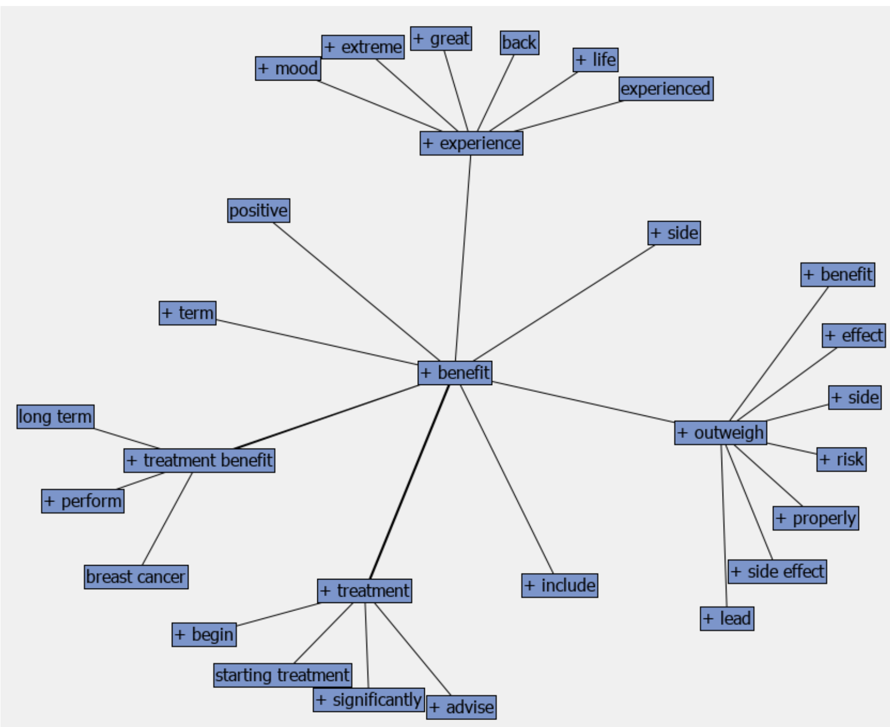


Figure 18 - Concept links for the term "benefit"

The concept link diagram in Figure 18 shows that the term “benefit” is associated with such terms as “experience”, “great”, “extreme”, “treatment benefit”, “significantly”, “long term”, “outweigh”, “benefit”. Hence, it can be inferred that some effectiveness of prescribed drugs can be illustrated by great experience (change in mood, life), treatment benefit in the long term, significantly benefit, or that benefits outweigh side effects.

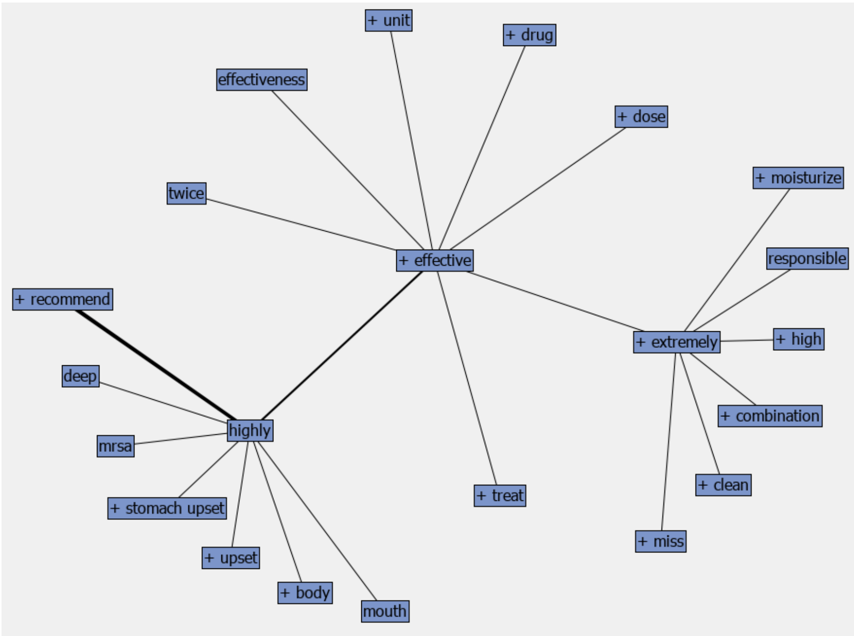


Figure 19 - Concept links for the term “effective”

Similarly, the concept link diagram in Figure 19 indicates that the term “effective” is associated with “highly”, “twice”, “effectiveness”, “extremely”, “treat”, etc., of which the association between “effective” and “highly recommend” is the strongest one.

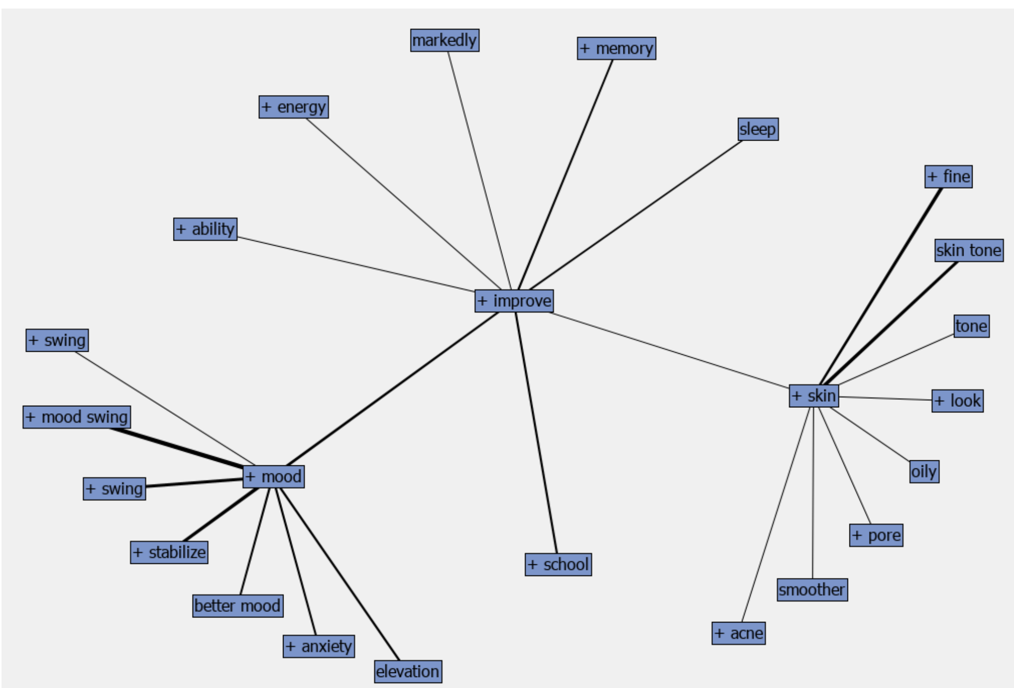
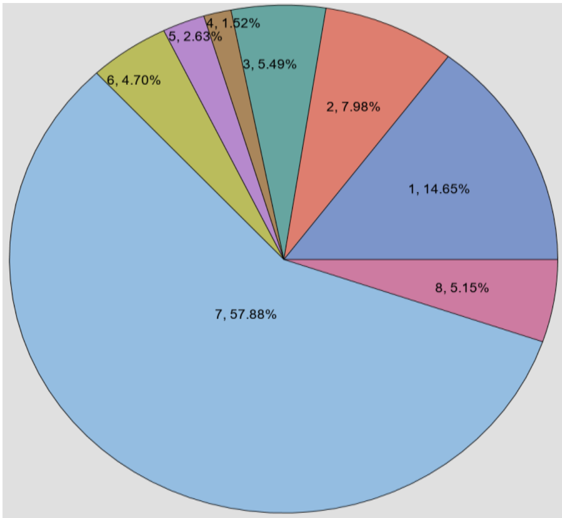


Figure 20 - Concept links for the term “improve”

The concept links in Figure 20 show that improvement in mood, skin, energy, memory, sleep, ability are also possible effects of analyzed drugs.

TEXT CLUSTERING



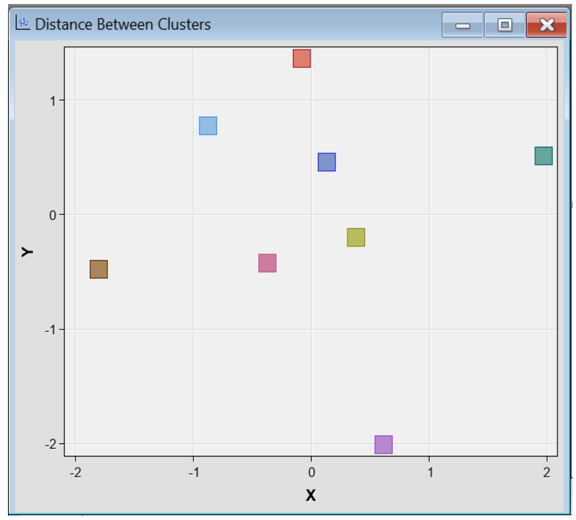


Figure 21 – Text Cluster node results for reviews on effectiveness

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 22 - Text Cluster descriptive terms for reviews on effectiveness

The Text Cluster node generates eight well-separated clusters as shown in Figure 21 and Figure 22. 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. It can be inferred 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 23 - Text Topic results for reviews on effectiveness

Figure 23 shows 15 different topics with corresponding number of terms in each topic and also 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 (DrugEffectivenessLevel). After trial and error, the customized setting with very low Generalization Error, very low Purity of Rules and very low Exhaustiveness produced the best results with lowest Average Squared Error and Misclassification Rate.

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 24 - Text Rule Builder results for reviews on effectiveness

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 15Figure 5, in all these models, the categorical variable "DrugEffectivenessLevel" is set to be the target variable and the text variable "benefitsReview" is set as the input variable.

Other key settings for the Model Comparison node are:

- Model selection statistic: Average Squared Error
- Model selection table: Validation

The Model Comparison results are provided as below.

Selected Model	Model Description	Target Variable	Selection Criterion: Valid: Average Squared Error
Y	Text Rule Builder	DrugEffectivenessLevel	0.051049
	Regression	DrugEffectivenessLevel	0.135394
	Neural Network	DrugEffectivenessLevel	0.136451
	Decision Tree	DrugEffectivenessLevel	0.137548

Figure 25 - Comparison between models for effectiveness classification

Figure 25 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 Average Squared Error (ASE) at 5.10% as compared to the other three models.

TRANSFER LEARNING

With Text Rule Builder model being the best predictive model in both side effect levels classification and effectiveness levels classification, transfer learning algorithm is used to apply this selected model on a new independent score data set to evaluate model performance and validation. 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 CLASSIFICATION

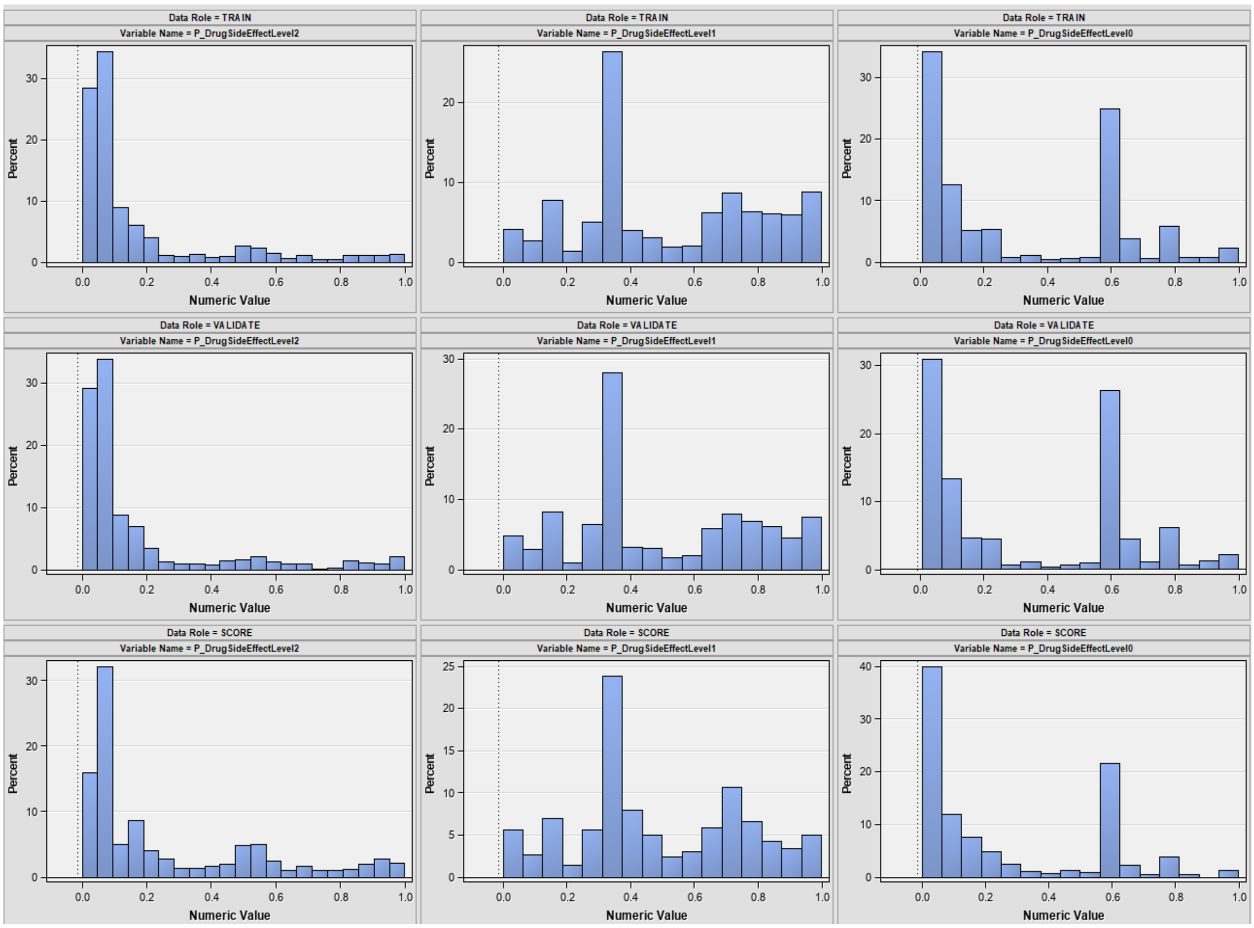


Figure 26 - Comparison of probability distribution of side effect classification across train, validate, and score data sets

Figure 26 illustrates the probability distribution of each side effect level's categorization across train, validate, and score data sets. For example, the three histograms vertically 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). Overall, they all 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 side effect rating.

SCORING RESULTS FOR EFFECTIVENESS CLASSIFICATION

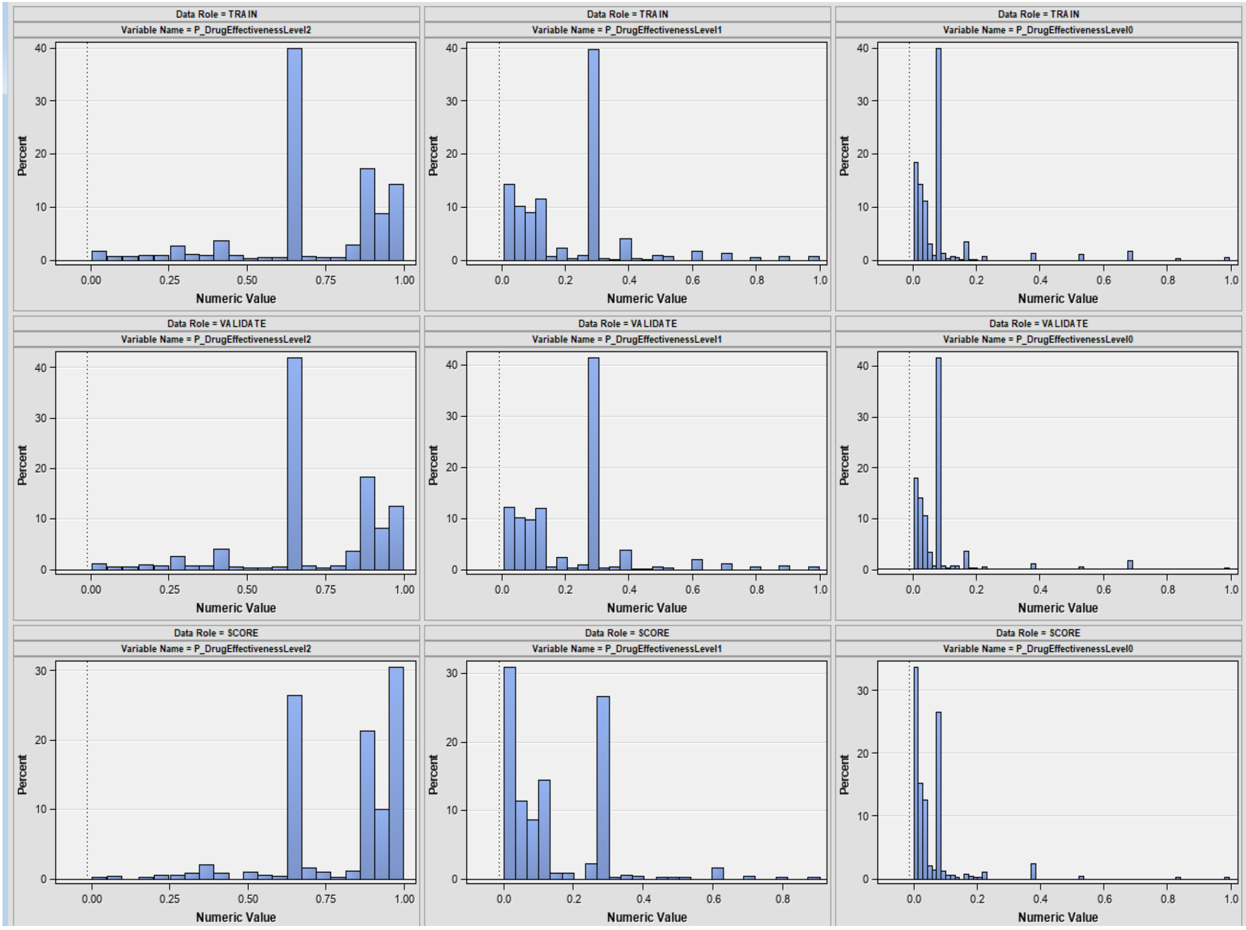


Figure 27 - Comparison of probability distribution of effectiveness classification across train, validate, and score data sets

Figure 27 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 better generalization in drug reviews classification.

DRUG EFFECTIVENESS EVALUATION

For the purpose of evaluating the effectiveness of a given specific drug, all users' overall reviews for five common prescription drugs to treat depression have been chosen for analysis. Reviews for these drugs are obtained from *Drugs.com* which are later used for text analytics with SAS® Enterprise Miner.

Accordingly, the drugs which are selected for analysis in this part are:

- Wellbutrin XL
- Lexapro
- Prozac
- Cymbalta
- Effexor

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 in order to evaluate the effectiveness of each drug.
The following diagram illustrates the process flow for the analysis:

Figure 28 - Unsupervised learning diagram for drug effectiveness evaluation

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

Figure 29 shows eight clusters generated for Wellbutrin XL 1-7 rating data. Clusters 8 and 1 have highest frequency percentages, indicating some common effects of Wellbutrin XL could be 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 30 - Text Cluster node output for Wellbutrin XL rating 8-10 data

Figure 30 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 could be 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 31 - Text Cluster node output for Lexapro rating 1-7 data

Fourteen clusters are generated for Lexapro 1-7 rating data as shown in Figure 31. The top frequency percentage clusters depict that some common effects of Lexapro could be 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 32 - Text Cluster node output for Lexapro rating 8-10 data

Figure 32 identifies nine clusters for Lexapro 8-10 rating data. Clusters 2, 6, and 1 have highest frequency percentage, indicating some effectiveness of Lexapro could be 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 33 - Text Cluster node output for Prozac rating 1-7 data

Ten clusters are generated for Prozac 1-7 rating data as shown in Figure 33. 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 34 - Text Cluster node output for Prozac rating 8-10 data

Figure 34 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 35 - 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 35. 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 36 - Text Cluster node output for Cymbalta rating 8-10 data

Figure 36 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, hepful, life saving anti-depressant treatment, whereas some claim several 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 37 - Text Cluster node output for Effexor rating 1-7 data

Seven clusters are generated for Effexor 1-7 rating data as shown in Figure 37. Cluster 2 has highest frequency percentages, which implies that some side effects of Effexor are 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 38 - Text Cluster node output for Effexor rating 8-10 data

Figure 38 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. Some side effects are sweating, crying, and weight gain.

COMPARISON OF EFFECTIVENESS OF FIVE DRUGS

Drug	Low rating evaluation	High rating evaluation	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, often happening in the morning	better feeling, happy mood	7.29
Cymbalta	nausea, back pain, sweating, weight gain, dizziness, and anxiety	best, hepful, 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

Figure 39 – Comparison of effectiveness of five anti-depressant drugs

Figure 39 helps understand the specific benefits and side effects of each of the five selected prescribed drugs, which can serve as practical guidelines to prospective clients in making their 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, 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 own 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 purchased 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, for the purpose of illustration, text analytics with unsupervised learning algorithm are also employed to detect the specific side effects and effectiveness of several selected anti-depression drugs which can help as practical guidelines for potential users. Overall, the study expects to provide valuable insights to assist prospective patients in making their informed purchase decisions and improve monitoring public health by revealing collective experience. A future challenge would be to fully analyze the reviews at sentence and phrase level by employing more sophisticated aspect-based sentiment analysis and more powerful 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 Their Engagement with Consumers Through Online Reviews". Accessed March, 2019. https://spiegel.medill.northwestern.edu/_pdf/Spiegel_Online%20Review_eBook_Jun2017_FINAL.pdf

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

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

ACKNOWLEDGMENTS

I sincerely thank my professors Dr. Goutam Chakraborty and Dr. Miriam McGaugh for their constant support and valuable guidance through every phase of this paper.

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