Identifying Drug Related Events from Social Media

Final Project Report

April 27, 2017

Client: Long Xia (Ph.D. student, email: longxia1@vt.edu)
Weiguo Fan (Professor, email: wfan@vt.edu)

Students: Sung Ho Kim, Woo Jin Kye, Yoonju Lee, Jeongho Noh, Jisu You

Class: CS4624 Multimedia, Hypertext, and information Access
Instructor: Edward A. Fox

Department of Computer Science
Virginia Tech
Blacksburg, VA 24061

Table of Contents

Table of Tables

Table of Figures

Abstract

The overall goal of the project was to establish an innovative information system, which can automatically detect and extract content related to side effect of drugs from user reviews, determine whether they are talking about effectiveness or adverse drug events, extract keywords or phrases related to effectiveness or adverse drug events, and visualize the resulting information to doctors and patients. Our group was provided with crawled Twitter reviews and social network forum reviews on drugs that are used to treat diabetes. The raw data were manually labeled in four different label for named entity recognition in order to create training, testing, and validation sets. Using the training data set, a side effect dictionary was created using PamTAT. Side effect dictionary was then refined by removing neutral words to increase accuracy. To validate the accuracy of the generated side effect dictionary, the results of side effect analysis based on the generated dictionary and two other general negative word dictionaries were compared. The generated side effect dictionary performed better in recognizing side effect entities. After validation, the generated dictionary was further tested with a set of user reviews on a drug that is used to treat stroke. Using generated dictionary, the project accomplished to accurately determine if any reviews relates to the mention of side effect of specific drugs. The project successfully delivered to accurately detect mention of side effect from the reviews in > 90% accuracy. Resulting algorithm can be used to create innovative information system to detect and extract content related to side effect of drugs for any other drugs with creation of problem specific dictionary. The project should be further developed to incorporate automatic extraction of user reviews, analysis of data, and visualization of results.

1. Product/Service Description

The project aims to provide effectiveness and safety surveillance after a drug's release on the market. The information will be gathered from social media and visualized to be distributed to both doctors and patients for better understanding of drugs.

1.1 Product Context

Besides effectiveness, a side effect is another important factor that doctors will consider before prescribing a medication. For now, the information about adverse drug events is collected during clinical trials. However, many other side effects can take time to surface. Thus, it is not possible to establish a complete safety profile just from clinical trials. An effective and efficient way to monitor the safety of drugs after release becomes a significant goal for public health information systems. Social media and patient forums become more and more popular, and patients share their disease conditions and medications, including discussions about effectiveness and adverse drug events. This offers a great potential data source for creating a comprehensive information system to better monitor drugs. However, existing research often focuses on drugs that are used to treat a specific medical condition (e.g., diabetes) and data from only one platform (e.g., Twitter). As a result, they can only yield a small data set and moderate performance.

The project aims to build a generic model, which can be used to identify the effectiveness and adverse drug events of drugs for a wide range of diseases.

1.2 User Characteristics

- Patient
  - Desires effectiveness and side effect information of a certain drug.
  - Considers overall sentiment of a drug use.
- Doctor
  - Needs to consider patient's individual health condition and possible drug interactions.
  - Needs to inform a patient of possible side effects of a drug.

1.3 Assumptions

- The raw data have been retrieved from Twitter comments related to drugs for certain disease.
- The comments are written by Twitter users who may not be medical professionals.
- Reported side effects may not be consistent and accurate.
- Individual patient health condition is not considered.
- Manually labeled data should be unbiased.
- The project is dependent on user reviews.

1.4 Constraints

- Access to user reviews.
- Programming language.
- GPU availability.
- Data processing mechanisms.
- Amount of sample data.

2. Requirements

2.1 Functional Requirements

Req#	Requirement	Comments	Priority
RQ_01	Manually label raw data.	Label data based on named entity recognition and sentiment analysis.	1
RQ_02	Building machine learning model for named entity recognition and result analysis.	Build NLP model, machine learning model (SVM, Decision Tree, KNN, etc.), and deep learning model (CNN, RNN, LSTM, etc.). Implement current state-of-the-art model and test on different datasets.	1
RQ_03	Results analysis and model fine tuning.	Compare the results with benchmark, error analysis, and fine-tuning model.	2
RQ_04	Case studies.	Obtain small datasets to test generality of final model.	2
RQ_05	Visualization for side effect classification task and sentiment classification task.	Create visual representation of resulting data.	1

Table 1. Functional Requirements

2.2 Usability

- The system should be able to generate a visualized analysis upon taking a dataset.

2.3 Performance

- Machine learning process should take less than one week.
- After fine-tuning, the error margin should be less than five percent.

2.3.1 Capacity

- The system needs to handle at least one thousand comments at a time.

2.3.2 Consistency

- Manual labeling needs to be done consistently.

2.4 System Interface/Integration

- Text analytics will be conducted using PamTAT.
- PamTAT handles CSV file formats, on which manual labeling has been done.
- Raw data has been transformed into CSV file format.

2.5 Standards Compliance

- Existing research mainly focuses on a specific medical condition and data from one source.

3. User Manual

3.1 User Scenarios/Use Cases

3.1.1 General Characteristic

Scope	SocialMediaDrugEvents
Level	Client goal
Primary Actor	Client
Stakeholders and interests	Client: Wants Deep Learning Program that will analyze social media produced reviews on drug use and effects Dr. Fox: Wants to see well designed and implemented project Group: Need to label data

3.1.2 Main Success Scenario

1. Client asks to have reviews on social network forum analyzed
2. Group goes through portion of social network forum reviews and labels them manually
3. Program deep learns and analyzes rest of reviews
4. Client can now analyze any reviews on any drug with the resulting model

3.1.3 Special Requirements

1 Data labeling process should not be biased nor based on the person labeling the data

3.1.4 Technology and Data Variations List

1 Sentiment data labeling should be in two parts; first would be rating between 1 - 5, with 1 being negative and 5 being positive. Second label should be whether the drug has side effect (1) or not (0).

3.1.5 Other Issues

1 Biased data can result in biased results and will require re-do of labeling part.

2 More reviews should be labeled if results don’t match.

4. Developer Manual

4.1 Design

4.1.1 High-Level Design

The project follows a pipe-and-filter architectural design. Figure 1 illustrates the basic framework of processing the data. As shown, there are five major nodes. An output from each node goes into the next node as an input.

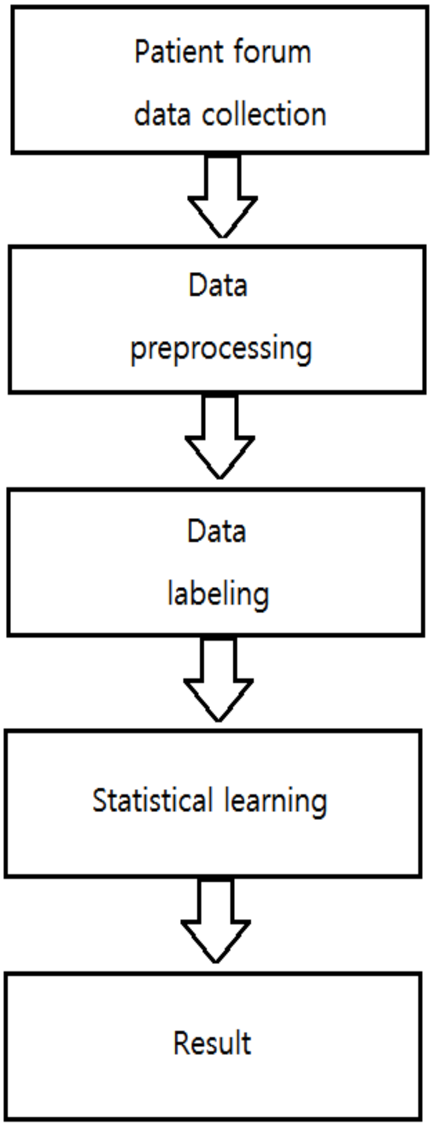


Figure 1. Pipe-and-filter architectural design. An output from each node is used

as the input for the next node.

4.1.1.1 Patient Forum Data Collection

An automated crawler was used to retrieve reviews from the WebMD, an online publisher of news and information pertaining to human health and well-being. The retrieved data were stored as comma-separated values (CSV) files. Each row in a file includes the content of the review, effectiveness score, ease of use score, and the satisfaction score written by a reviewer.

4.1.1.2 Data Preprocessing

During the data preprocessing, two additional types of CSV files were created for human labeling. The first one was used for Named Entity Recognition (NER). Each word in a review had to be classified into one of the following entities: D for drug name, S for side-effects, M for medical terms, and O for others. Each row in a file includes a word in a review and a column for a human to classify its entity. In total, there were more than three hundred thousand words to be classified.

The next file was created for sentiment and side effect scoring. Each review was given a score of 1-5 based on a sentiment of a review and a binary score that indicates a mention of the drug’s side effects. In total, there were about 5,600 reviews to be scored.

4.1.1.3 Data Labeling

Each CSV file created from the data processing was labeled by three humans for enhanced accuracy. Each member labeled 1862 reviews, or 117,500 entities. People manually label the data to train the machine. Each review gets scored manually by humans.

4.1.1.4 Statistical Learning

The code used for the statistical learning model is going to be strictly procedural since there is no need for any abstractions. The model is written in Python, which has rich built-in functions for text manipulation.

Labeled data were fed into the model for training purpose. Approximately 80% of the data was used as a training set and the remaining 20% was used as a test set.

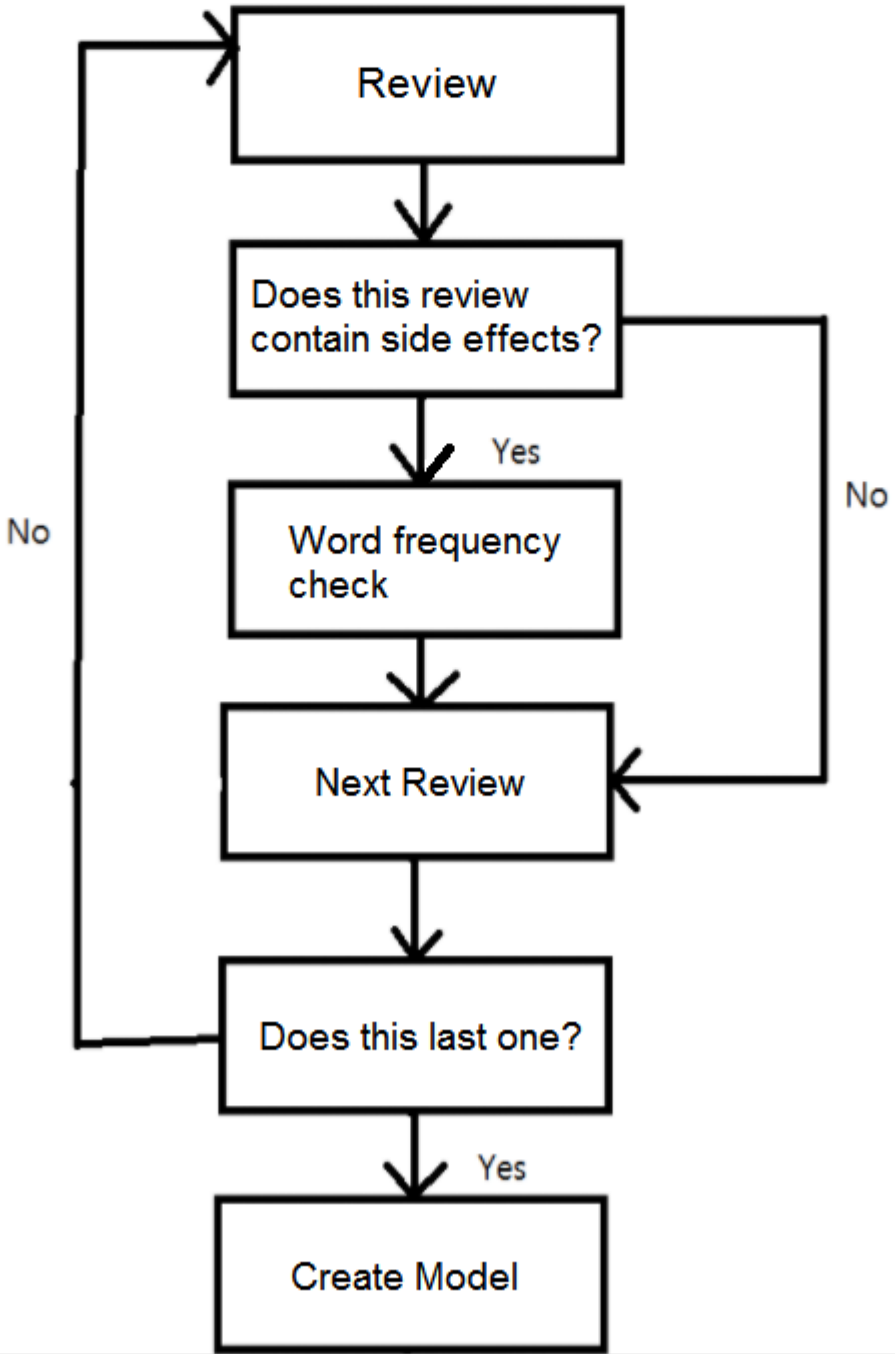


Figure 2. The process of building the statistical model

4.1.1.5 Result

One result of this project is the set of reviews scored by the trained model. After that result has been processed, it will be compared with the human labeled data for error analysis. We use standard machine-learning and text analysis evaluation metrics, precision, recall, and F-measure--to evaluate the performance of dictionary.

4.2 Implementation

4.2.1 Data Retrieval

For the data retrieval, a crawler to crawl reviews on a social media drug side effect sharing forum had to be created. A Python program is developed to reach social media page to retrieve reviews from side effect sharing pages using the “Beautiful Soup” library. Contents such as review ID, drug index, and review are retrieved from HTML contents using “Beautiful Soup”. Then those contents are stored in a CSV file (in this case for later use through Microsoft Excel) for further manual labeling for rating and the existence of side effects, by humans.

4.2.2 Data Labeling

For data labeling, two different types of data have been labeled: Named Entity Recognition and Sentiment.

4.2.2.1 Named Entity Recognition

4.2.2.1.1 Purpose

Manual labeling is necessary to build a problem specific dictionary. The built in dictionaries do not have enough information to correctly categorize the entities needed in this study. For more information on built in dictionary, please refer to section 5.3.3.3.

4.2.2.1.2 Categories

- Files - three different Excel files to give label:
    - NER_Training.csv – contains 235,000 words, and will be used for building dictionary, which is a list of drug entity and side effects entity.
    - NER_Validation.csv – contains 47,060 words, and will be used for making confusion matrix.
    - NER_Testing.csv – contains 45,029 words, and will be used for machine learning.
- Label - four different labels for different entities:
    - D – drug entity
    - S – side effect entity

- ○ M – miscellaneous medical terms that are not a drug entity or a side effect entity
- ○ O – others

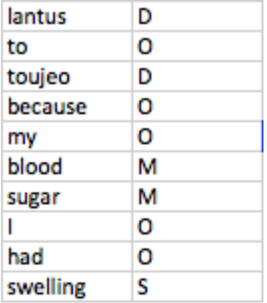
4.2.2.1.3 Personnel

Two different group members, Yoonju Lee and Woo Jin Kye gave one label for each row:

- NER_Training.csv:
  - Lee: row 1 – 117,500
  - Kye: row 117,501 - end
- NER_Validation.csv :
  - Lee: row 20,001 – 47,060
  - Kye: row 1 – 20,000
- NER_testing.csv:
  - Lee: row 1 – 20,000,
  - Kye: row 20,001 – 45,029

4.2.2.1.4 Method/Procedure

To identify the drug name and side effect entities from customer reviews, participants gave one label for each word as shown in Table 2. To keep consistency, two different members first worked together for the first three thousand rows and separately for the rest of the rows. Since this project is focused on the possible side effects of drugs, labels M and O do not affect the result of this project. They were distinguished for possible future use of the data.



lantus	D
to	O
toujeo	D
because	O
my	O
blood	M
sugar	M
I	O
had	O
swelling	S

Table 2. Named Entity Recognition Data

4.2.2.1.5 Result/Outcome

There are a total 2242 of different side effect entities and 412 different drug entities with approximately:

- 85 percent labeled as O,
- 11 percent labeled as S,
- 2 percent labeled as D,
- and 2 percent labeled as M.

4.2.2.2 Sentiment Labels

4.2.2.2.1 Purpose

Name Entity Recognition and Sentiment labels have to be analyzed separately to ensure the accuracy of deep learning. Name Entity Recognition labels will be used to create a confusion matrix(refers to 5.3.3) and smoke list(refers to 6.1), which are the important tools for creation and analysis of dictionary generated by PamTAT. Those tools will be further explained in the next section. Sentiment labels will be used to evaluate if reviews contain a side effect mention or not with a dictionary produced by either NER labels or PamTAT.

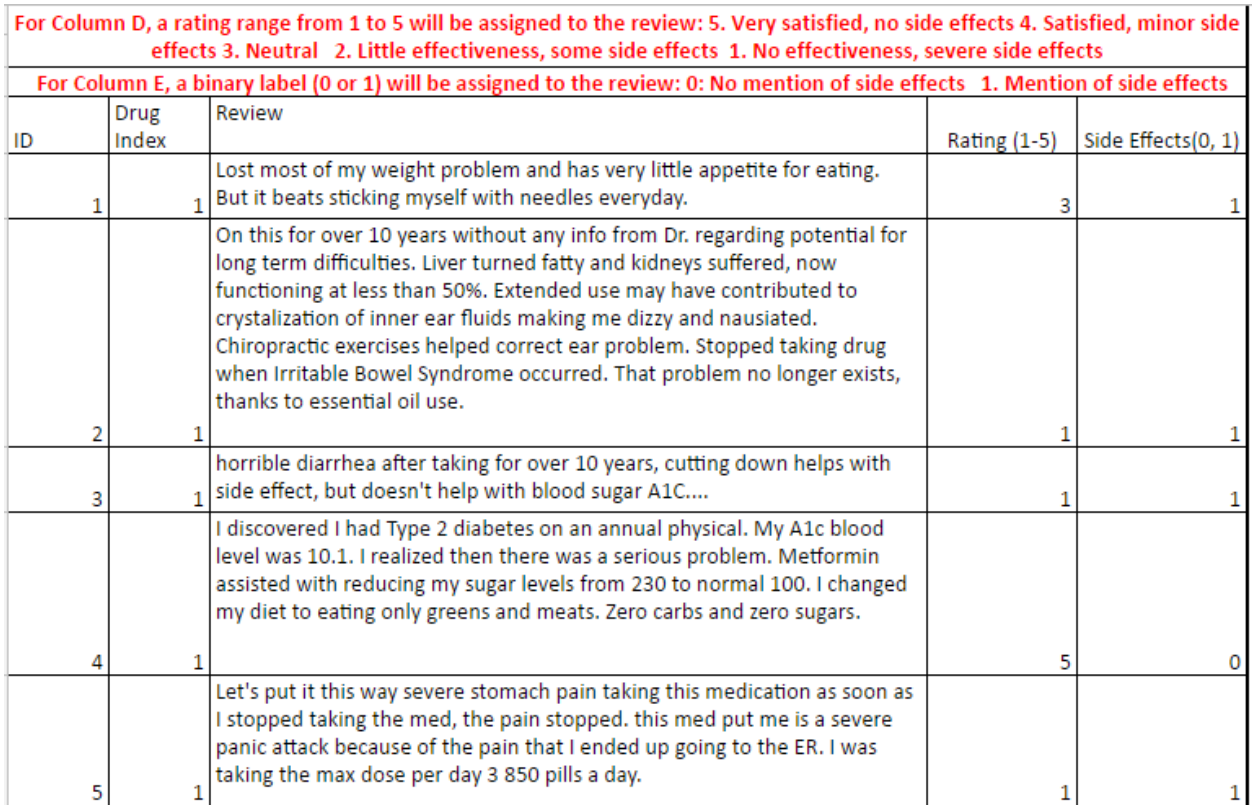
4.2.2.2.2 Personnel

Three different group members, Sung Ho Kim, Jisu You, and Jeongho Noh gave one label for each row.

- Label_Diabetes_Sentiment.xlsx:
  - Kim: row 1 – 1,862
  - You: row 1,863 - 3,724
  - Noh: row 3,725 - 5,586

4.2.2.2.3 Method/Procedure

Each review was given a sentiment score and a binary label. A sentiment score ranges from 1 to 5 to indicate a level of satisfaction, with 5 being most satisfied. A binary label can either be 0 or 1. 1 indicates that the review contains a mention of the drug’s side effects, and 0 indicates the opposite.



For Column D, a rating range from 1 to 5 will be assigned to the review: 5. Very satisfied, no side effects 4. Satisfied, minor side effects 3. Neutral 2. Little effectiveness, some side effects 1. No effectiveness, severe side effects

For Column E, a binary label (0 or 1) will be assigned to the review: 0: No mention of side effects 1. Mention of side effects

ID	Drug Index	Review	Rating (1-5)	Side Effects(0, 1)
1	1	Lost most of my weight problem and has very little appetite for eating. But it beats sticking myself with needles everyday.	3	1
2	1	On this for over 10 years without any info from Dr. regarding potential for long term difficulties. Liver turned fatty and kidneys suffered, now functioning at less than 50%. Extended use may have contributed to crystalization of inner ear fluids making me dizzy and nausiated. Chiropractic exercises helped correct ear problem. Stopped taking drug when Irritable Bowel Syndrome occurred. That problem no longer exists, thanks to essential oil use.	1	1
3	1	horrible diarrhea after taking for over 10 years, cutting down helps with side effect, but doesn't help with blood sugar A1C....	1	1
4	1	I discovered I had Type 2 diabetes on an annual physical. My A1c blood level was 10.1. I realized then there was a serious problem. Metformin assisted with reducing my sugar levels from 230 to normal 100. I changed my diet to eating only greens and meats. Zero carbs and zero sugars.	5	0
5	1	Let's put it this way severe stomach pain taking this medication as soon as I stopped taking the med, the pain stopped. this med put me is a severe panic attack because of the pain that I ended up going to the ER. I was taking the max dose per day 3 850 pills a day.	1	1

Table 3. Label_Diabetes_Sentiment.xlsx

4.2.3 Statistical Analysis of Labeled Data

Based on the manually labeled data, a side effect scoring dictionary was formed. The dictionary will be used for identifying side effects from new user reviews. To analyze the effectiveness of the dictionary, a confusion matrix was formed.

4.2.3.1 Software

PamTAT software was utilized for generating a smoke list, side effect scoring dictionary, and confusion matrix.

4.2.3.2 Smoke List

A smoke list is created by removing non-alphabetic characters that do not necessarily contribute to side effect recognition. It is a list of meaningful words separated in each row.

4.2.3.2.1 Purpose

- To create a problem specific entity recognition dictionary
- To validate labels before formatting the dictionary

4.2.3.2.2 Procedure

- Import labeling data to PamTAT worksheet
- Remove non-alphabetic characters
- Run significant values analysis with correlation coefficient algorithm
- Sort the list based on side effect scoring and filter out words not related to side effect
- Create a side effect entity dictionary with the filtered words

4.2.3.3 Confusion Matrix

The side effect scores generated by the side effect entity dictionary needs to be validated for effectiveness. Multiple scores generated by different algorithms were compared in a confusion matrix to show the difference between the effectiveness of built-in dictionaries and custom dictionary on distinguishing side effect entities.

4.2.3.3.1 Purpose

- To analyze the effectiveness of the side effect scoring dictionary
- To visualize the analyzed data

4.2.3.3.2 Complications

- Each scoring algorithm has different scoring method
- Requires careful data selection
- Requires some normalization of scores

4.2.3.3.3 Scoring Methods

- Side Effect Scoring Dictionary - custom built dictionary for identifying side effect entities
- AFFIN Negative Word Dictionary - built-in dictionary
- General Inquirer Negative Sentiment Dictionary - built-in dictionary

4.2.3.3.4 Score Normalization

Each method has a different scoring policy:

- Side Effect Scoring Dictionary - gives higher score based on frequency of the appearance of each word in side effect entity recognition in NER_Training data set
- AFINN Negative Word Dictionary - gives negative scores based on the sentiment expression of the word: bastard (-5) and bitter (-1)
- General Inquirer Negative Sentiment Dictionary - Any word with negative sentiment gets a score of 1 to determine word is relate with side effect or not

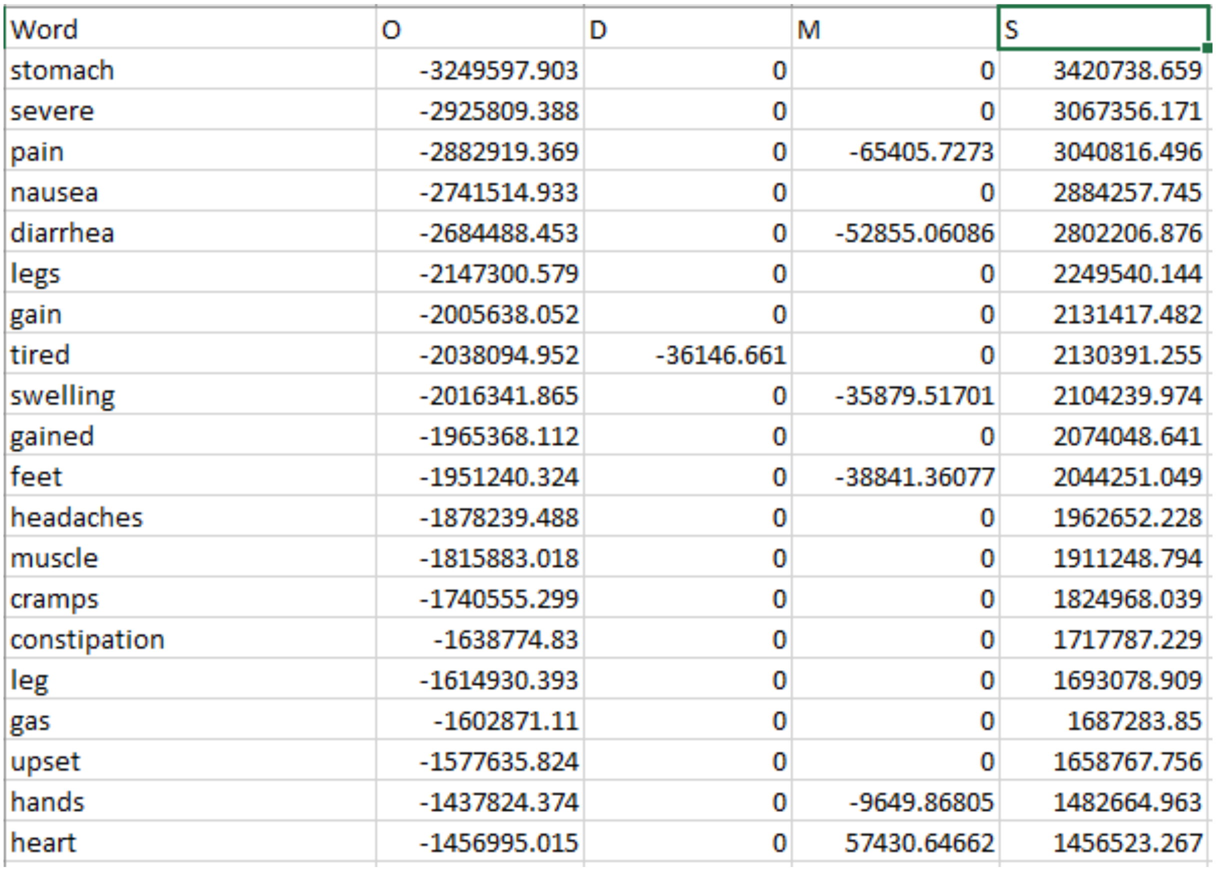
The magnitude of a score can be safely ignored since the analysis only requires identification of side effect entities. All scores are normalized to 1 that confirms side effect entity.

4.3 Prototype

Using NER data and 5,600 labeled reviews, three data sets were generated: a smoke list, a dictionary, and a confusion matrix.

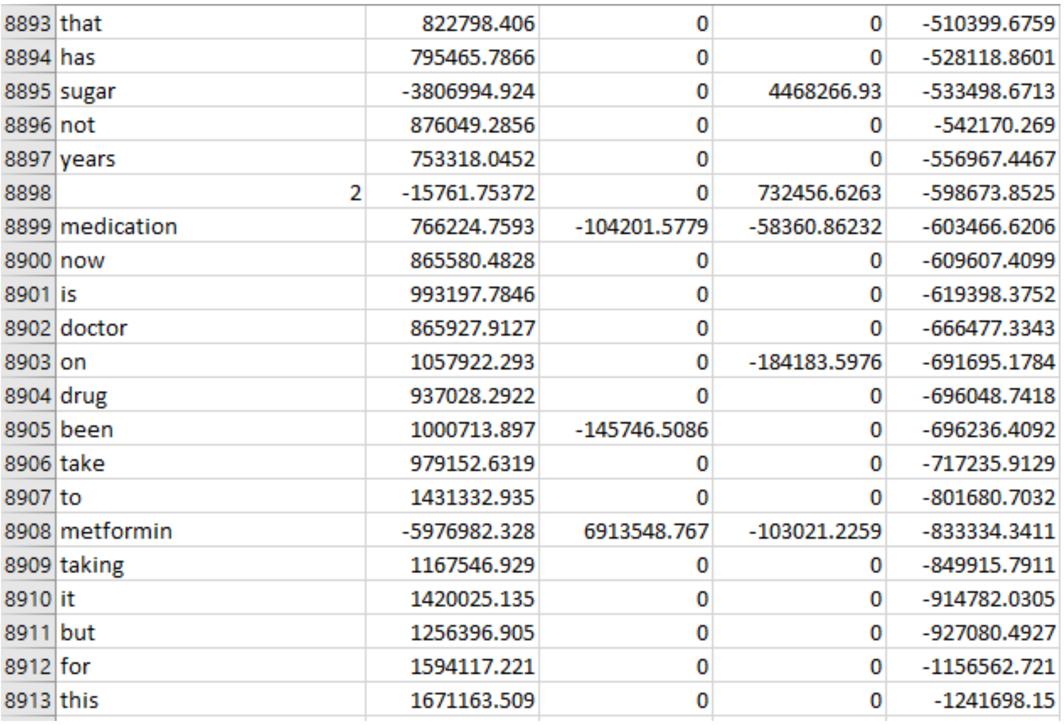
4.3.1 Smoke List

The smoke list created initially had five columns, including Word, O, M, S, and D. The Word column simply contains a word. O, M, S, and D each represents four different entities as mentioned earlier, D for drug entity, S for side effect entity, M for other medical term entity, and O for others. Each of those columns holds a score associated with a word within its row. The scores are calculated based on the number of occurrence of each word in regard to the entities. For instance, the word “stomach” has high score because it frequently appears in reviews with a mention of side effects.



Word	O	D	M	S
stomach	-3249597.903	0	0	3420738.659
severe	-2925809.388	0	0	3067356.171
pain	-2882919.369	0	-65405.7273	3040816.496
nausea	-2741514.933	0	0	2884257.745
diarrhea	-2684488.453	0	-52855.06086	2802206.876
legs	-2147300.579	0	0	2249540.144
gain	-2005638.052	0	0	2131417.482
tired	-2038094.952	-36146.661	0	2130391.255
swelling	-2016341.865	0	-35879.51701	2104239.974
gained	-1965368.112	0	0	2074048.641
feet	-1951240.324	0	-38841.36077	2044251.049
headaches	-1878239.488	0	0	1962652.228
muscle	-1815883.018	0	0	1911248.794
cramps	-1740555.299	0	0	1824968.039
constipation	-1638774.83	0	0	1717787.229
leg	-1614930.393	0	0	1693078.909
gas	-1602871.11	0	0	1687283.85
upset	-1577635.824	0	0	1658767.756
hands	-1437824.374	0	-9649.86805	1482664.963
heart	-1456995.015	0	57430.64662	1456523.267

Table 4. First twenty words in the smoke list



8893	that	822798.406	0	0	-510399.6759
8894	has	795465.7866	0	0	-528118.8601
8895	sugar	-3806994.924	0	4468266.93	-533498.6713
8896	not	876049.2856	0	0	-542170.269
8897	years	753318.0452	0	0	-556967.4467
8898	2	-15761.75372	0	732456.6263	-598673.8525
8899	medication	766224.7593	-104201.5779	-58360.86232	-603466.6206
8900	now	865580.4828	0	0	-609607.4099
8901	is	993197.7846	0	0	-619398.3752
8902	doctor	865927.9127	0	0	-666477.3343
8903	on	1057922.293	0	-184183.5976	-691695.1784
8904	drug	937028.2922	0	0	-696048.7418
8905	been	1000713.897	-145746.5086	0	-696236.4092
8906	take	979152.6319	0	0	-717235.9129
8907	to	1431332.935	0	0	-801680.7032
8908	metformin	-5976982.328	6913548.767	-103021.2259	-833334.3411
8909	taking	1167546.929	0	0	-849915.7911
8910	it	1420025.135	0	0	-914782.0305
8911	but	1256396.905	0	0	-927080.4927
8912	for	1594117.221	0	0	-1156562.721
8913	this	1671163.509	0	0	-1241698.15

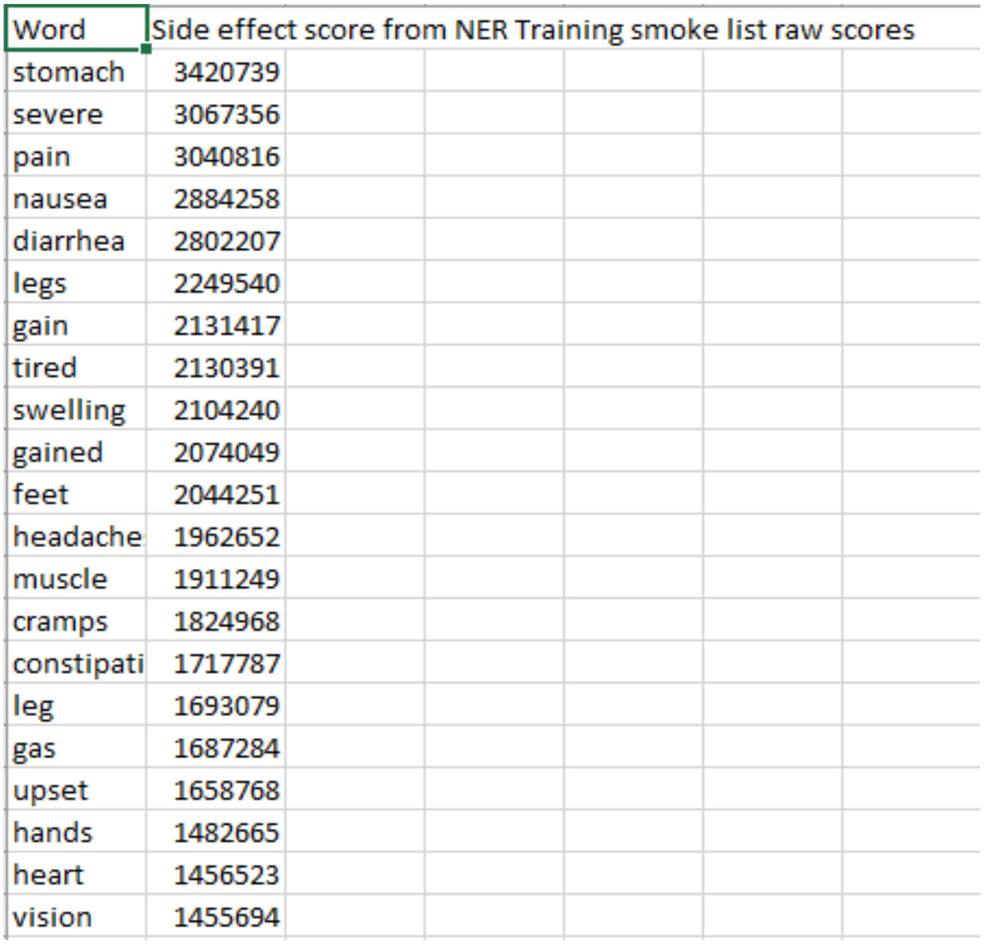
Table 5. Last twenty words in the smoke list

Since side effect entity is the most important factor in this experiment, the smoke list is sorted by scores in the S column, which range from -1241698.15 to 3420738.659. The Table 4 and Table 5 show general overviews of the smoke list.

As shown in Table 4, the words that are highly relevant to the side effects of a drug--such as “severe”, “stomach”, “nausea”, “pain”, and etc.--scored the highest. On the other hand, Table 5 shows the bottom of the list, which contains the lowest scored words that are not relevant to the side effects of the drug. The problem with this list, however, is that there are some meaningless words--such as “this”, “for”, “now”, “on”, and etc.--that can be associated with reviews with and without side effects.

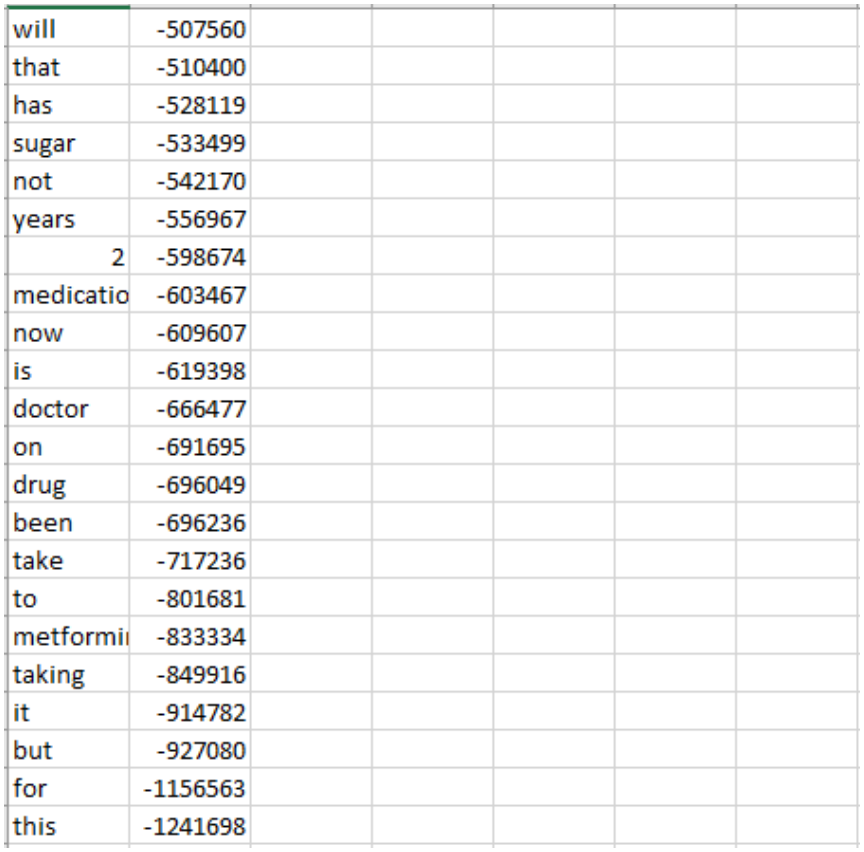
4.3.2 Dictionary

Using the smoke list, a dictionary was created. The dictionary is used to score a review based on words that appear in the review. The dictionary created is shown in Table 6 and Table 7. Again, tables show the top and bottom of the list.



Word	Side effect score from NER Training smoke list raw scores
stomach	3420739
severe	3067356
pain	3040816
nausea	2884258
diarrhea	2802207
legs	2249540
gain	2131417
tired	2130391
swelling	2104240
gained	2074049
feet	2044251
headache	1962652
muscle	1911249
cramps	1824968
constipati	1717787
leg	1693079
gas	1687284
upset	1658768
hands	1482665
heart	1456523
vision	1455694

Table 6. First twenty words in the dictionary



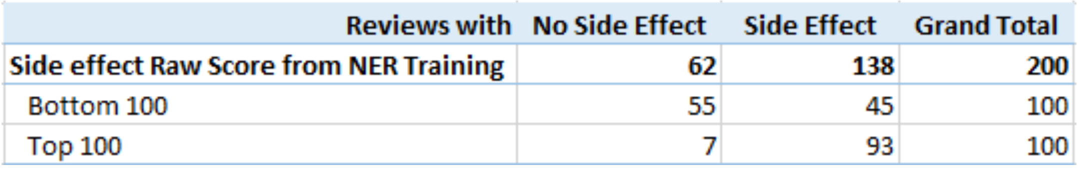
will	-507560
that	-510400
has	-528119
sugar	-533499
not	-542170
years	-556967
2	-598674
medicatio	-603467
now	-609607
is	-619398
doctor	-666477
on	-691695
drug	-696049
been	-696236
take	-717236
to	-801681
metformi	-833334
taking	-849916
it	-914782
but	-927080
for	-1156563
this	-1241698

Table 7. Last twenty words in the dictionary

The dictionary only contains the S column since that is the only column with relevant scores. Similar to the smoke list, the words in the top of the dictionary seem reasonable. Some meaningless words, however, are still present in the dictionary. The neutral words like “this”, “for”, “but”, and “is” have potential to defect the accuracy of the result.

4.3.3 Confusion Matrix

The confusion matrix in Table 8 compares the top 100 reviews relevant to side effects and the bottom 100 reviews not relevant to side effects. This confusion matrix was generated with raw scoring dictionary.



Reviews with	No Side Effect	Side Effect	Grand Total
Side effect Raw Score from NER Training	**62**	**138**	**200**
Bottom 100	55	45	100
Top 100	7	93	100

Table 8. Confusion Matrix with Raw Data Dictionary

As shown in Table 8, the top 100 reviews have 93 reviews with side effect and 7 reviews with no side effect. However, the bottom 100 reviews contain 45 reviews with side effect. The custom dictionary with raw score data has 93% accuracy on the top 100 reviews and 45% accuracy on the bottom 100 reviews. Normalization of data is required to obtain better results and get higher accuracy.

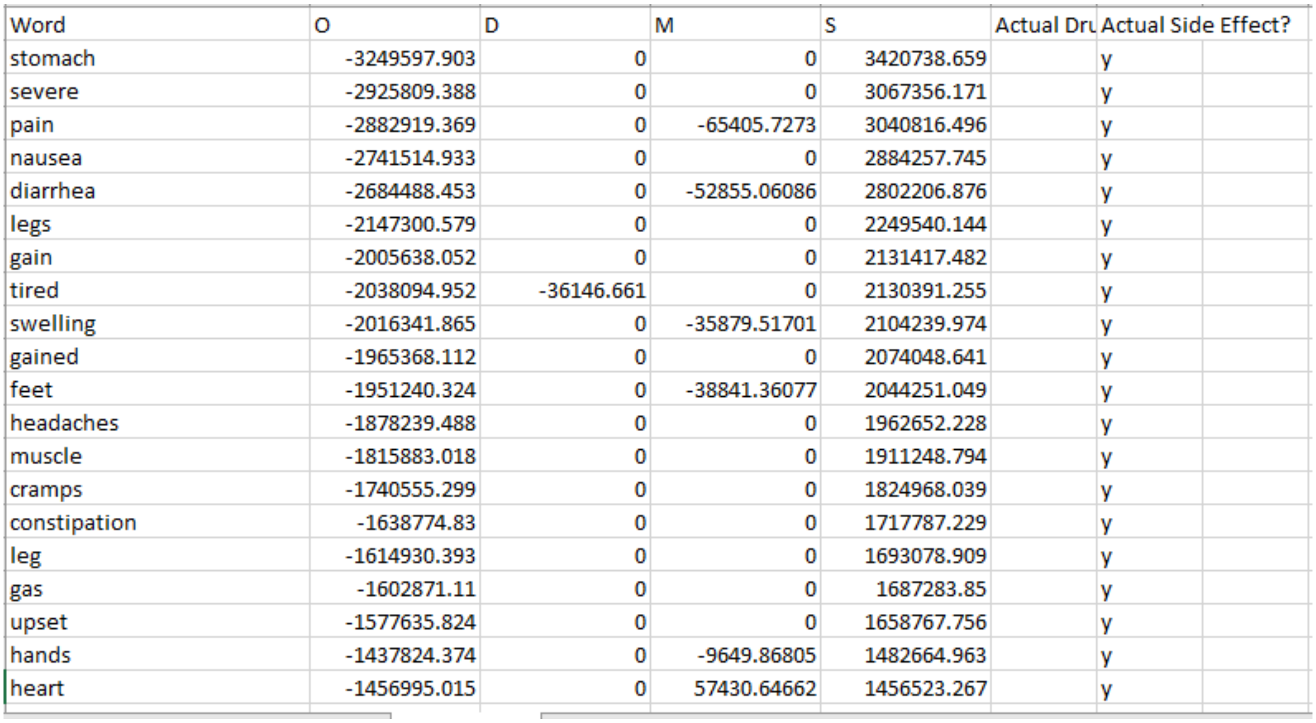
5. Refinement

5.1 Smoke List

The initial smoke list had five columns, including Word, O, M, S, and D. However, two more columns – actual drug and actual side effect – were added to increase the accuracy of the smoke list. The actual side effect column in the smoke list was manually labeled, using y or n: y representing that word relates to side effect and n representing that word does not relate to side effect. Then the smoke list has been normalized to have two labels that show the relation of word to side effect. Both labels show the relation of word to side effect but they are different that one is automated by pamTAT whereas the other one was manually labeled by people to create a more accurate dictionary. Then a new dictionary with the addition of actual side effect column of smoke list was created (see Table 9). Each of those columns holds a score associated with a word within its row.

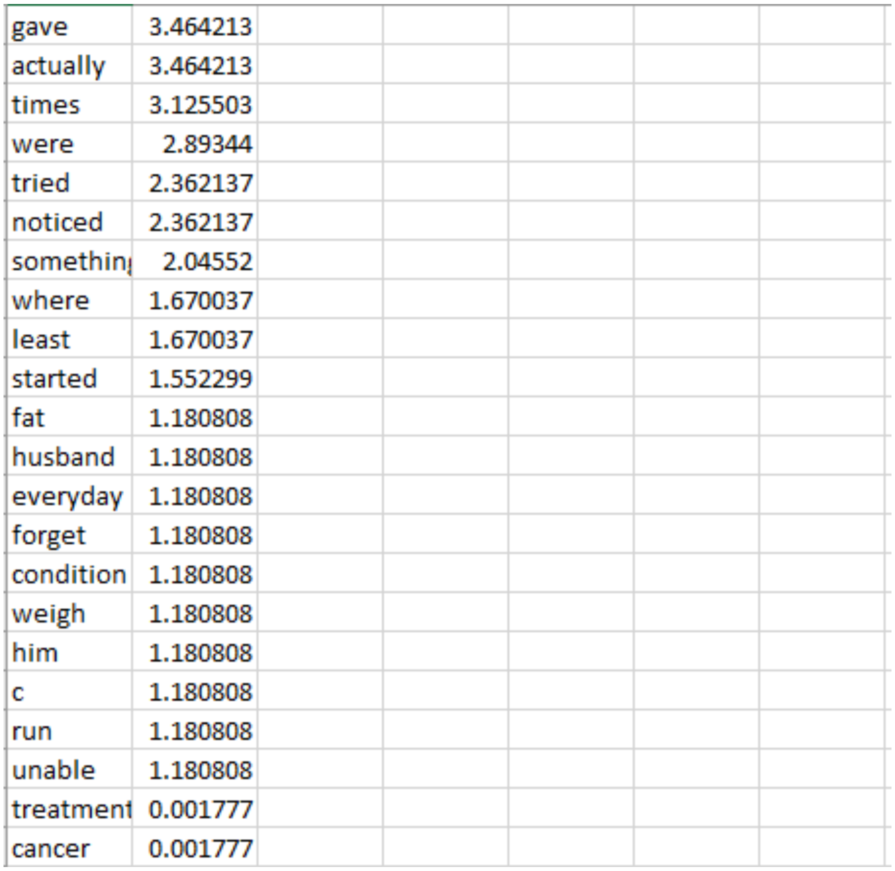
5.2 Dictionary

Using newly refined smoke list, a dictionary was recreated. Along with recreating a dictionary with the refined smoke list, the dictionary itself was also refined. The dictionary was first refined by filtering out neutral words and prepositions such as “this”, “for”, “but”, “where”, “when”, and etc. (Table 10). Then the dictionary was normalized by dividing the score created by smoke list by number of occurrences of the word to get better scores of words since the first dictionary, regardless of the relation to side effect, gave higher scores for the words that occurred more often (Table 11).



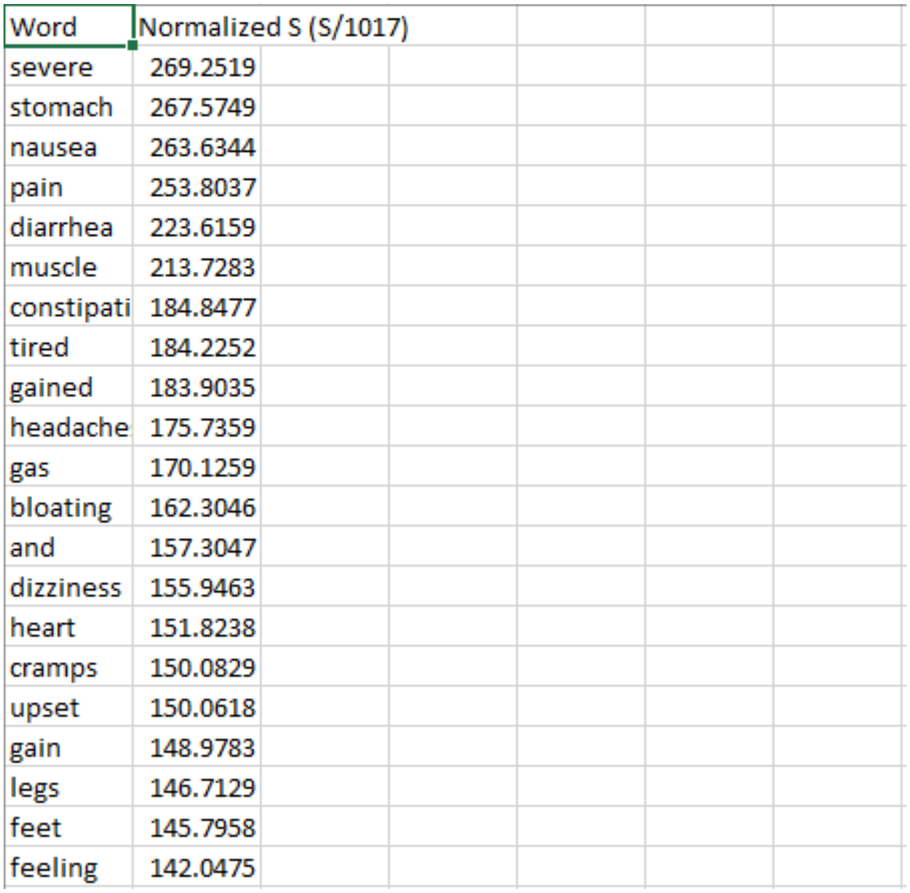
Word	O	D	M	S	Actual Dru	Actual Side Effect?
stomach	-3249597.903	0	0	3420738.659		y
severe	-2925809.388	0	0	3067356.171		y
pain	-2882919.369	0	-65405.7273	3040816.496		y
nausea	-2741514.933	0	0	2884257.745		y
diarrhea	-2684488.453	0	-52855.06086	2802206.876		y
legs	-2147300.579	0	0	2249540.144		y
gain	-2005638.052	0	0	2131417.482		y
tired	-2038094.952	-36146.661	0	2130391.255		y
swelling	-2016341.865	0	-35879.51701	2104239.974		y
gained	-1965368.112	0	0	2074048.641		y
feet	-1951240.324	0	-38841.36077	2044251.049		y
headaches	-1878239.488	0	0	1962652.228		y
muscle	-1815883.018	0	0	1911248.794		y
cramps	-1740555.299	0	0	1824968.039		y
constipation	-1638774.83	0	0	1717787.229		y
leg	-1614930.393	0	0	1693078.909		y
gas	-1602871.11	0	0	1687283.85		y
upset	-1577635.824	0	0	1658767.756		y
hands	-1437824.374	0	-9649.86805	1482664.963		y
heart	-1456995.015	0	57430.64662	1456523.267		y

Table 9. Refined smoke list with two labels (actual drug and actual side effect) added



gave	3.464213
actually	3.464213
times	3.125503
were	2.89344
tried	2.362137
noticed	2.362137
somethin	2.04552
where	1.670037
least	1.670037
started	1.552299
fat	1.180808
husband	1.180808
everyday	1.180808
forget	1.180808
condition	1.180808
weigh	1.180808
him	1.180808
c	1.180808
run	1.180808
unable	1.180808
treatment	0.001777
cancer	0.001777

Table 10. Normalized dictionary contains no neutral words

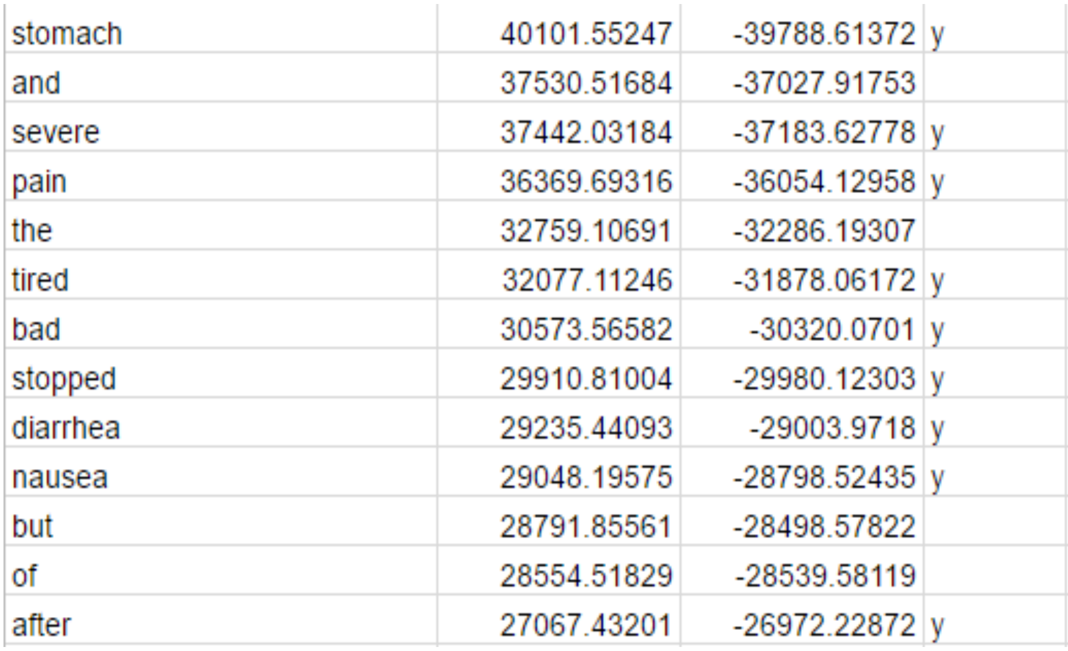


Word	Normalized S (S/1017)
severe	269.2519
stomach	267.5749
nausea	263.6344
pain	253.8037
diarrhea	223.6159
muscle	213.7283
constipati	184.8477
tired	184.2252
gained	183.9035
headache	175.7359
gas	170.1259
bloating	162.3046
and	157.3047
dizziness	155.9463
heart	151.8238
cramps	150.0829
upset	150.0618
gain	148.9783
legs	146.7129
feet	145.7958
feeling	142.0475

Table 11. Normalized dictionary with updated score

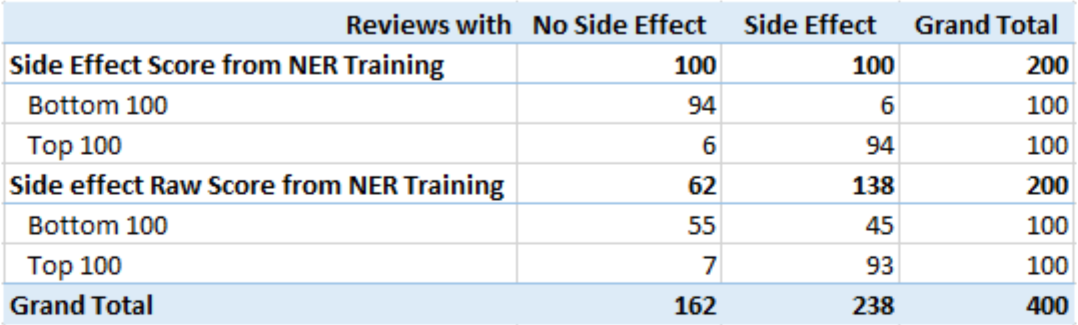
5.3 Confusion Matrix

In the confusion matrix, the bottom 100 reviews need to have higher accuracy. The normalized data is required to create an effective custom side effect dictionary. In Table 12. Words related to side effects were manually labeled with y, since magnitude of score is ignored in normalized data.



stomach	40101.55247	-39788.61372	y
and	37530.51684	-37027.91753	
severe	37442.03184	-37183.62778	y
pain	36369.69316	-36054.12958	y
the	32759.10691	-32286.19307	
tired	32077.11246	-31878.06172	y
bad	30573.56582	-30320.0701	y
stopped	29910.81004	-29980.12303	y
diarrhea	29235.44093	-29003.9718	y
nausea	29048.19575	-28798.52435	y
but	28791.85561	-28498.57822	
of	28554.51829	-28539.58119	
after	27067.43201	-26972.22872	y

Table 12. Manually labeled data



Reviews with	No Side Effect	Side Effect	Grand Total
Side Effect Score from NER Training	**100**	**100**	**200**
Bottom 100	94	6	100
Top 100	6	94	100
Side effect Raw Score from NER Training	**62**	**138**	**200**
Bottom 100	55	45	100
Top 100	7	93	100
Grand Total	**162**	**238**	**400**

Table 13. Confusion Matrix with raw data dictionary and normalized dictionary

Created custom side effect dictionary from normalized data to get higher accuracy to determine side effect related reviews. As shown in Table 13 the confusion matrix data achieved 94% of accuracy on both top and bottom.

6. Testing

The accuracy of the custom side effect scoring dictionary was tested by comparing the results from manual labeling and the scoring results from custom dictionary.

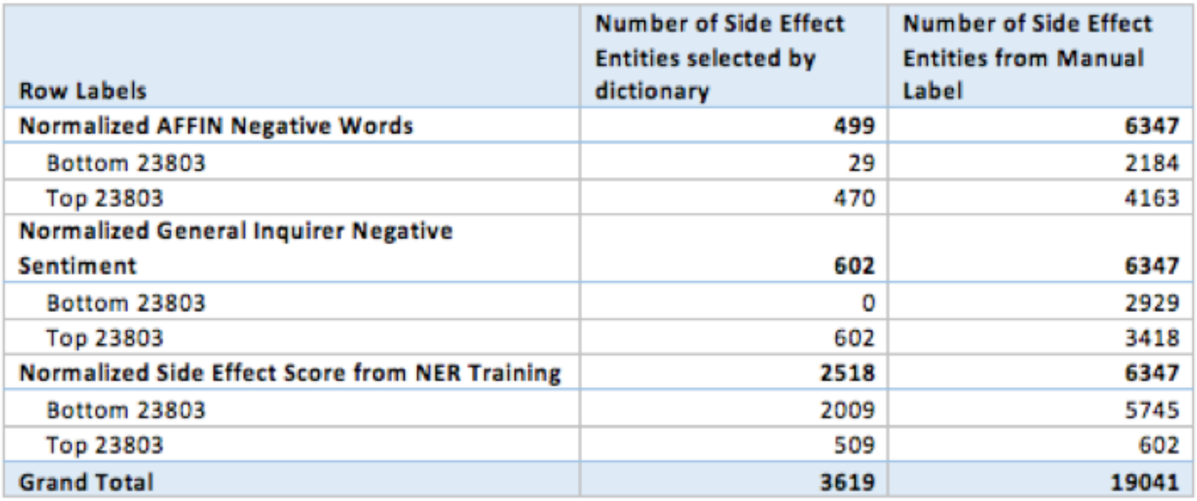
6.1 Testing Process with NER_Validation file

6.1.1: Purpose

The custom side effect scoring dictionary should detect a good amount of side effect entities. This testing will compare the number of side entities from manual labeling and those selected by the custom dictionary and built-in dictionaries.

6.1.2: Result

- Input file: NER_Validation.csv file is used. NER_Validation file contains numerous diabetes drug reviews. There are a total of 47,606 words.
- Manual Label: 6,347 words are labeled as side effect entities.
- AFINN Negative Words Dictionary: 499 side effect entities are detected.
- General Inquirer Negative Sentiment Dictionary: 602 side effect entities are detected.
- Custom Side Effect Scoring Dictionary: 2518 side effect entities are detected.



Row Labels	Number of Side Effect Entities selected by dictionary	Number of Side Effect Entities from Manual Label
Normalized AFFIN Negative Words	**499**	**6347**
Bottom 23803	29	2184
Top 23803	470	4163
Normalized General Inquirer Negative Sentiment	**602**	**6347**
Bottom 23803	0	2929
Top 23803	602	3418
Normalized Side Effect Score from NER Training	**2518**	**6347**
Bottom 23803	2009	5745
Top 23803	509	602
Grand Total	**3619**	**19041**

Table 14. Confusion Matrix from Validation Set

6.1.3: Percent Error

For determining the precision of the custom dictionary, the percent error formula is used as shown below. The formula is given by: $\%ERROR = |\frac{\#EXPERIMENTAL - \#THEORETICAL}{\#EXPERIMENTAL}| \times 100$

- AFINN Negative Words Dictionary
  $|\frac{499-6347}{6347}|\times 100 = 92.16\ \%$

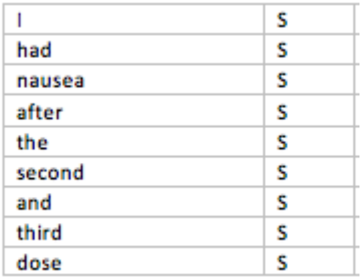
- General Inquirer Negative Sentiment Dictionary
  $|\frac{602-6347}{6347}|\times 100 = 90.52\ \%$

- Custom Side Effect Scoring Dictionary
  $|\frac{2518-6347}{6347}|\times 100 = 60.33\ \%$

As shown above, the error percentages of AFFIN Negative Words Dictionary and General Inquirer Negative Dictionary are over 90, while the error percentage of the custom side effect scoring dictionary is 60.33. The error percentages show that applying the custom side effect scoring dictionary to the data set significantly improves the result. However, 60.33 percent error is still enormous.

6.1.4: Error Analysis

- Major errors might occur in the labeling process. As shown below, we labeled “I”, “had”, “after”, “the”, “second”, “and”, “third”, and “dose” as side effect entities. It is because every word in a phrase containing a description of side effect is treated as a side effect entity. It was noticeable that those words were frequently occurring in neutral and positive reviews. The custom dictionary was refined by removing those words. See section 7 for more details.



I	S
had	S
nausea	S
after	S
the	S
second	S
and	S
third	S
dose	S

Table 15. Named Entity Recognition

- Minor errors might occur because of the lack of consistency. Even though two different members worked together on labeling the first three thousand rows for the name entity recognition part, labeling 162,045 words separately might cause in low consistency. During testing, we noticed that Kye labeled “insulin” as drug entity, while Lee did not.

6.2 Testing process with Aggrenox reviews

6.2.1: Purpose

Since we plan to build a generic model that can be used to identify side effects of drugs for wide range of diseases, the custom dictionary should detect side effects of drugs other than diabetes drug. This testing will show the effectiveness of the custom dictionary in detecting side effects of Aggrenox, which is a type of stroke medication.

6.2.2: Input Data

- 10 different user reviews for Aggrenox from Drugs.com.
- Ratings by the reviewer range from 0 to 10, where 10 is the highest.
- There are 3 10-star ratings, 1 8-star ratings, 1 5-star ratings, and 5 1-star ratings.
- There is no side effect stated in the reviews number 5, 6, and 10.
- Please see column 1, 2, and 5 of table 16 in section 8.2.3 for more information.

6.2.3: Result

- The specific rank was taken by using the custom side effects scoring dictionary as shown in column 3 of the table 16. The specific rank of each review ranges from 1 to 10, where 1 is the most negative and 10 is the most positive.
- Specific Rank: The specific ranks obtained by the custom dictionary were relatively accurate except for review number 8. Specific rank of the review number 8 is supposed to be less than 5. Other than the review number 8, the custom dictionary seems to be effective.

Review Number	**Review**	**Specific Rank** (1 - negative, 10 - positive)	**Side Effect** (1- yes, 0 -No)	**Ratings by the reviewer** (out of 10)
1	Took one Aggrenox in the morning and one at night and it caused extreme nausea, headache, diarrhea, stiff joints, and weakness all night and vomiting in the morning. This was prescribed after one cardiovascular event. Had tried Plavix 75 mg for 4 months with poor results. Then tried Plavix 159 mg for 6 weeks and still no results in platelet test.	1	1	1
2	I've had 4 bad Transient Ischemic Attacks. I tried Plavix once but had severe side effects. Just started on Aggrenox. Bad headache, nausea, stomach pain. But I'll do it and hopefully my body will adapt -- soon hopefully.	2	1	5
3	I had been taking Plavix without any side effects since experiencing transient ischemic stroke, but was advised by my doctor to try Aggrenox because a clotting test raised a question whether Plavix was effective. To start with I took one dose per day because I had had been advised to expect headaches until my body adjusted. I felt my head was going to explode	3	1	1

	and got extremely nauseated. I took it for three days. The headaches got worse, and I became nauseous, and my blood pressure went very high. I now list Aggrenox as a medication that I am allergic to because of the severe side effects. I continue to use Aspirin, Plavix and other medications to try to prevent strokes.			
4	I was put on aggrenox to keep from having strokes. I was fine with taking it until about 7 months later it caused me to have another. Before and after my stroke I had excruciating headaches. I took my own self off aggrenox and the headaches disappeared. I would suggest if anyone is taking it to STOP.	4	1	1
5	I had a few TIAs after bypass surgery, although the two may not be directly related; that is, the bypass surgery causing the TIAs. More likely it relates to my chemistry. The Drs. had me on Coumadin, which of cource is a real pain. After experiencing a TIA I was admitted and my attending physician was a stroke neurologist. He questioned me being on Coumadin and felt strongly that Aggrenox was the right medicine for my situation. I have used it for at least over 10 years with absolutely no side effects or occurrences.	5	0	10
6	I was told by my physician in January 2005, that I had several mini-stokes before the one that hospitalized me on January 31, 2005. It has been 7 years now and I had no noticeable side effects with Aggrenox. When I started the medication, I always took it as prescribed - 1 capsule in the morning and one capsule in the evening. About 1 year ago, I only started taking 1 capsule when I would take my blood pressure pill. My pharmacist begged me to start back taking the medication as it was originally prescribed to be taken.	6	0	10
7	I had a series of TIA's almost 10 years ago. After Numerous Drs and hospitals I went to the Mayo Clinic. After many tests they diagnosed the problem and prescribed Aggrenox. I had some mild headaches the first few weeks. I have been TIA free for almost a decade.	7	1	10
8	This medication was used to help prevent a stroke but it actually caused me to have a stroke. I wouldn't advise anyone to take this medication. Someone else I knew had several strokes before the dr figured out it was the aggrenox. I was on it 8 months before having my stroke.	8	1	1

9	Fourteen years i have been taking Aggrenox. Total life saver for me. Headaches when I first started, took Tylenol if absolutely necessary. ? Generic. Am I going to be forced to take generic? Pretty scary.	9	1	10
10	I have been taking Aggrenox now for 4 years for ITA--no side effects and no further attacks to date.	10	0	8

Table 16. User reviews for Aggrenox from Drugs.com

7. References

David Z. Adams, Richard Gruss, Alan S. Abrahams, Automated discovery of safety and efficacy concerns for joint & muscle pain relief treatments from online reviews, International Journal of Medical Informatics 100(2017) 108-120

Xiao Liu, Hsinchun Chen. Identifying Adverse Drug Events from patient social media, A case study for diabetes,IEEE Computer Society, May 2015, (2015)1541-1672

Alan Abrahams, George, Zach. Webex_PamTAT_Tutorial, Jan 27,2017, Virginia Tech, Blacksburg

"Imaging the Universe." Percent Error Formula | Imaging the Universe. 09 May 2017. <http://astro.physics.uiowa.edu/ITU/glossary/percent-error-formula/>.

"Prescription Drug Information, Interactions & Side Effects." *Drugs.com*. 09 May 2017. <https://www.drugs.com/>.