Sentiment Analysis on Drug Reviews

ISM 6136 Fall 2019

Data Mining Project- Final Paper

Information System and Decision Sciences

**Text Miners**

Sai Suraj Argula

Vinay Kumar Reddy Baradi

Harika Yadav Gillela

Aitemir Yeskenov

Table of Contents

Introduction…………………………………………………………………………………...…...3

Dataset Overview………………………………………………………………………………….3

Data Exploration…………………………………………………………………………………..4

Process Flow…….………………………………………………………………………………...6

Data Cleaning and Pre-Processing steps…………………………………………………..6

Design and Develop Models………………………………………………………………7

Comparing Models………………………………………………………………………...9

Predictions with Test Data……….………………………………………………………………..9

Acknowledgement..……………………………………………………………………………...11

References………………………………………………………………………………………..11

**Introduction**

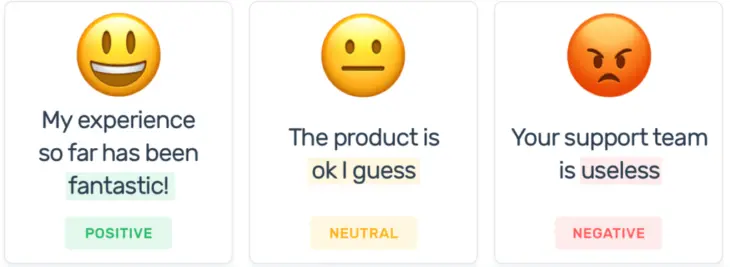
**Text Analytics**

Text analytics is an artificial intelligence (AI) technology that uses natural language processing (NLP) to transform the free, unstructured text in documents and databases into normalized, structured data suitable for analysis and drive machine learning (ML) algorithms to find Patterns and Trends.

**Sentiment Analysis**

Sentiment analysis is the automated process that uses AI to identify positive, negative and neutral opinions from text. Sentiment analysis is widely used for getting insights from social media comments, survey responses, and product reviews, and making data-driven decisions.

There are many types and flavors of sentiment analysis and tools of sentiment analysis range from systems that focus on polarity to system that detect the emotions, feelings and intentions.



Using sentiment analysis, we are analyzing the drug reviews and classify them as positive or negative. When a drug is used by a patient, the review provided by patient will have various aspects of a drug depending on their interest and expertise. We analyze different reviews provided by the patients and predict if the review is positive or negative.

**Dataset Overview**

The drug dataset is taken from the UCI repository. The dataset provides patient reviews on specific drugs along with the condition of the patient. The patient rates the drug from range of 1 to 10 based on their satisfaction.

**Attribute Information:**

|  |  |  |
| --- | --- | --- |
| Attribute Name | Type | Description |
| UniqueID | ID | Unique ID for each patient review |
| Drug Name | categorical | Name of the drug |
| Condition | categorical | Condition of the patient |
| Review | text | Patient review |
| Date | date | Date of review entry |
| Useful Count | numerical | Number of users who found the review as useful |
| Sentiment | text | Defined column, Rating > 5 as positive; rating<5 as negative |

For the sentiment analysis we have only considered the attributes review and sentiment as those are the only relevant text variables for sentiment analysis. The data is split into training data (70%) and test data (30%).

**Data Exploration**

The total number of entries including the training and test data are 161298. The dataset seems to be skewed as it has more positive reviews and less negative reviews.

A screenshot of a cell phone

Description automatically generated

Top 10 Drug Names based on review count:

|  |  |
| --- | --- |
| Drug Name | Count |
| Levonorgestrel | 3657 |
| Etonogestrel | 3336 |
| Ethinyl estradiol / norethindrone | 2850 |
| Nexplanon | 2156 |
| Ethinyl estradiol / norgestimate | 2117 |
| Ethinyl estradiol / levonorgestrel | 1888 |
| Phentermine | 1543 |
| Sertraline | 1360 |
| Escitalopram | 1292 |
| Mirena | 1242 |

Top 10 reviews based on condition:

|  |  |
| --- | --- |
| Drug Name | Count |
| Birth Control | 28788 |
| Depression | 9069 |
| Pain | 6145 |
| Anxiety | 5904 |
| Acne | 5588 |
| Bipolar disorder | 4224 |
| Insomnia | 3673 |
| Weight loss | 3609 |
| Obesity | 3568 |
| ADHD | 3383 |

As RapidMiner was very slow when we tried to use whole data to build the model. So, we filtered down the data by adding a single condition. We have used condition Acne that has 5588 in the training dataset and 1848 cases in test dataset.

Below is the plot of the negative and positive cases of the Acne condition.

A screenshot of text

Description automatically generated

Top 10 drug names of the condition Acne

|  |  |
| --- | --- |
| Drug Name | Count |
| Isotretinoin | 517 |
| Adapalene/benzoylperoxide | 397 |
| Epiduo | 394 |
| Doxycycline | 361 |
| Accutane | 321 |
| Benzoylperoxide/clindamycin | 262 |
| Minocycline | 232 |
| Tretinoin | 228 |
| Ethinylestradiol/norgestimate | 215 |
| Drospirenone/ethinylestradiol | 199 |

**Process Flow**

Using RapidMiner different models have been used for the sentiment prediction. The data was split in 70-30 splits for training using the models: Decision Tree, Logistic Regression-SVM and Naïve Bayes.

**Data Cleaning and Pre-processing:**

A screenshot of a social media post

Description automatically generated

* **Tokenization**: Tokenization breaks a sequence of strings into pieces such as words, keywords, phrases and other elements called tokens.
* **Transform Cases:** Transform cases converts the text into either uppercase or lowercase letters.
* **Filter stop words**: This step involves filtering the most commonly used words such as A, IS, THIS, THE etc. We have also included a customized dictionary of stop words like I, Ha-ha, Ahhh and removed those from the dataset.
* **Stem (Porter):** In this step we reduce the words to their base forms. In this example worked is changed to work and looking is changing to look.
* **Filter token by lengths:** This step eliminates the words that have length less than a minimum specified number of letters or exceed a maximum number of characters set. For example here we have set minimum length as 2 so the words with less than 2 characters will be eliminated.
* **Generate N-gram:** An n-Gram is a combination of ‘n’ consecutive terms in a sentence. It creates an association between the words. We have used bigram as criteria to generate n-grams. RapidMiner separates n-gramswith an underscore (\_).

The preprocessing steps are explained below with an example sentence:

A screenshot of a cell phone

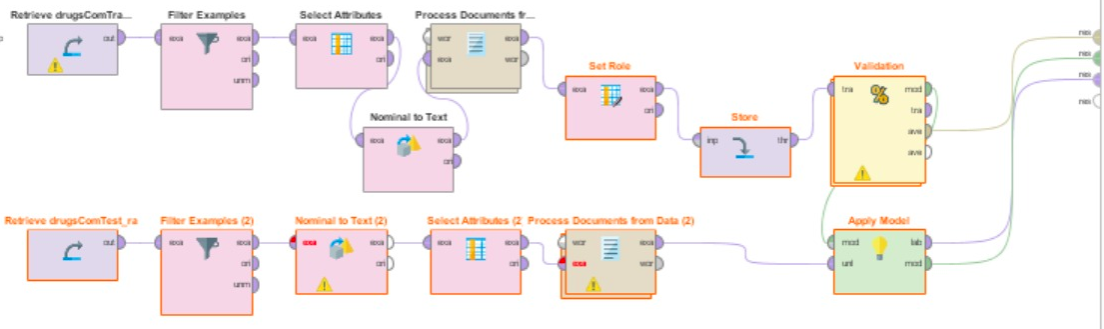
Description automatically generated

The result set after performing the above processing steps is as shown below:

A close up of a white wall

Description automatically generated

**Design and develop models to predict sentiment**

****

**A screenshot of a cell phone

Description automatically generated**

Naïve-Bayes model evaluation:

* Model accuracy and Confusion Matrix

A screenshot of a cell phone

Description automatically generated

Logistic Regression (SVM) model evaluation

* Model accuracy and Confusion Matrix

A screenshot of a cell phone

Description automatically generated

Decision Tree model evaluation

* Model accuracy and Confusion Matrix

A screenshot of a cell phone

Description automatically generated

**Compare results of different Models built**

Comparing results from the three different models we see that the Naïve Bayes has highest performance with highest accuracy and low classification error.

|  |  |  |
| --- | --- | --- |
| Model | Accuracy | Classification Error |
| Decision Tree | 76.57 | 24.43 |
| Naïve Bayes | 84.79 | 15.21 |
| SVM | 84.14 | 15.86 |

**Predictions on Test Data**

The predictions are done using the models Naïve Bayes and SVM as they have high accuracy.

Using the model Naïve Bayes the review with unique ID 179764, “If you can live with skin as is do not go on epiduo. It may eventually get better, but you will go through hell first.”, was predicted as negative based on the word ‘hell’.

The review with Unique ID 6760, “I have always had really bad skin and It has made it flawless! It is amazing! I love Beyaz!” is predicted as positive as it has words such as love, amaze, flawless etc.

Below are predictions done on the test data set using the Naïve Bayes model.

A screenshot of a computer

Description automatically generated

With the SVM model the review “I would be moody and extremely irritable all the time using it. It’s time to switch” with unique ID 33173 is predicted as negative. The prediction is done as negative as the review has word irritation.

“My skin looks amazing!  I now love going into public and my confidence is through the roof, I would definitely recommend for everyone!” with unique ID is predicted as positive as it has words such as amazing, love etc.

Below are predictions done on the test data set using the SVM model.

A screenshot of a social media post

Description automatically generated

**Acknowledgement**

We would like to thank Dr Balaji Padmanabhan for his guidance in every step of the project and explanation of concepts which we were able to incorporate in the project.

We would like to thank Surya Kallumadi of Kansas State University for providing this dataset public through UCI website.

**References**

<https://monkeylearn.com/sentiment-analysis/>

<https://www.paralleldots.com/sentiment-analysis>

<https://en.wikipedia.org/wiki/Text_mining>

<https://www.dataversity.net/roadmap-text-analytics-initiative/>