UNIVERSITY OF GENOA

FACULTY OF COMPUTER SCIENCE

NATURAL LANGUAGE PROCESSING

FINAL PROJECT

by

Emre DEMİRCAN

Instructor

Assoc. Prof. Viviana Mascardi

June 15, 2020

GENOA

Testing SentiWordNet by applying Drug Review data set

Emre DEMİRCAN

4711588@studenti.unige.it

1. Introduction

1.1 Background Information

Medicines are compounds that treat or prevent disease or illness to provide better lifetime for creatures. Despite this specification, medicines can also cause unwanted, even sometimes deathful reactions. Recent researches show that medicine usage and medicine diversification are increasing over the years. Therewithal, it brings quite a few questions and problems to the patients about therapeutic and side effects. Adverse drug reactions (ADR) and side effects are unlikable, annoying, and harmful reactions caused by usage of medicinal products. The most stunning statistic is that the total number of deaths caused by drug reactions is nearly 110,000 per year [6]. Currently, ADRs are among the leading causes of death in many countries.

1.2 Problem Definition

The discovery of medicine treatments with less side effects are very difficult due to the lack of understanding of the mechanisms of drug interactions and uniqueness of individuals. Risk of developing side effects vary according to age, gender, genetic factors, lifestyle of person and so forth.

We are living in the age of data. It is exponentially growing and getting bigger each day. This gives us a huge capability to make research and look for some improvements in treatments, but it also creates other problems such as understanding and processing the data, cleaning low quality and inaccurate data.

In this project, Drug Review data set by UCI repository has been chosen to apply sentiment analysis that uses various of Natural Language Processing (NLP) methods and algorithms. The user's ratings and reviews about drug that can be found in the dataset will play the main role and SentiWordNet [3,4] will be used as lexical resource.

1.3 Goal / Contribution

Despite the problems mentioned above, the aim of this project is to experience and test the SentiWordNet on medical data reviews. Eventually, I want to recommend medicines that is more effective and has less side effects to people while providing healthier treatment process by reducing the number of individuals who have reactions.

2. Methodologies / Tools / Libraries

2.1 WordNet

WordNet is one of the largest and most used lexical databases that has semantic relations between words [5]. Synset is an interface to look up words in WordNet. They are basically the groupings of synonymous words that express the same concept. Each of WordNet have 117.000

synsets and they are all linked to each other by means of conceptual relations. Moreover, each synset has brief explanation with couple of examples of their usages in sentences. WordNet can be used to get information about the following for a given word or phrase:

- Synonyms - Words that have the same meaning (dog = pup)
- Hypernyms - The generic term (i.e. dog is a kind of animal)
- Hyponyms - A member of a class of terms (i.e. dalmatian is a kind of dog)
- Holonyms – Constituents that the item is contained in (i.e. canis)

There are a lot of implemented semantic networks and web interfaces based on WordNet. Additionally, it has remarkable impact for some fields of NLP such as Word-sense disambiguation, question answering and sentiment analysis.

2.2 SentiWordNet

SentiWordNet is a lexical resource that supports sentiment classification and opinion mining applications [3]. It is publicly available for research purposes and it can be easily imported by the help of natural language toolkit (nltk) which is a platform that makes easier to work with NLP [1]. SentiWordNet has all three numerical scores of each synset of WORDNET as Pos(s), Neg(s) and Obj(s). The closed interval of each of these three scores is [0, 1] and their sum is 1 for each synset. More details will be given during the implementation part of the project (Section 3).

2.3 Sentiment Analysis

Natural language processing (NLP) is a domain of artificial intelligence that help machines understand and process the human language. Sentiment analysis or also known as opinion mining is one of the NLP fields that helps to understand sentiments behind the text and classify polarity of statements such as positive, negative, or neutral. There are some main approaches and methods while applying for sentiment analysis (Figure 2.1).

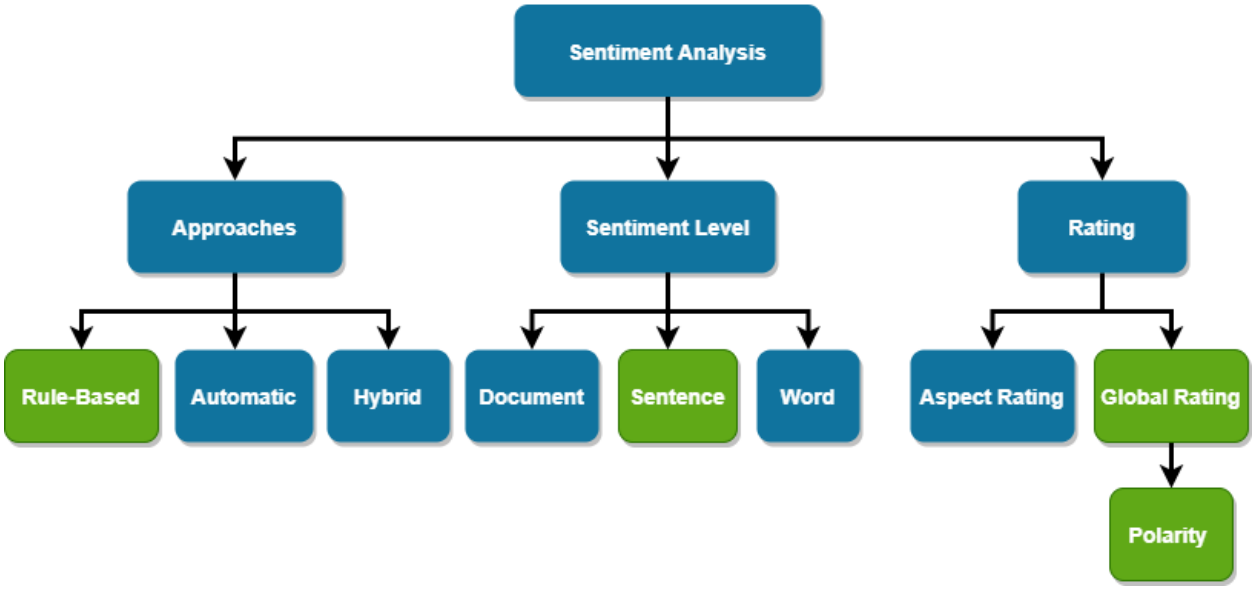

Figure 2.1: Main approaches and methods of sentiment analysis

2.3.1 Rule-Based Approach

Rule-based approaches uses a set of rules which is also known as lexicon. The aim of this approach is to identify the words according to their polarity as positivity and negativity. While positive refers to words which have desired states, negative refers to words which have undesired states. There are three classification level in sentiment analysis as Aspect-level, Sentence-level, and Document-level. In Aspect-level, we aim to classify all the parts which are related to a specific aspect in the text or document. In Sentence-level, we focus on each sentence to understand whether they express positive or negative opinion about the main subject of the text. It is mostly used to do sentiment analysis of reviews and comments. In Document-level, we consider complete document as an information unit to understand the positiveness or negativeness of the main opinion. I can say that there is no remarkable difference between sentence and document level because the sentences are just small pieces that form documents. While applying this approach, the following steps are needed to be performed:

- Extracting the text to be applied for sentiment analysis.
- Defining the list of positive and negative words or using libraries and resources.
- Using some NLP algorithms such as tokenization, stop words removal, Part-of-speech (POS) tagging, stemming, lemmatization.
- Sentiment analysis.
- Specifying the polarity according to sentiment score of words which is between 1 (totally positive) and -1 (totally negative).
- Understanding the results and have an inference about the data.

2.4 Programming Language (PL) and Integrated Development Environment (IDE)

Python programming language will be used to implement this project. It incorporates all the standard statistical tests, models, and analyses, as well as providing a comprehensive language for managing and manipulating data.

PyCharm is an integrated development environment (IDE) produced by JetBrains. It gives easy debugging opportunity and it helps to write high-quality codes while providing high readability and comprehensibleness of the code. Moreover, it supports scientific and NLP libraries of python that facilitate to build projects related to data science and machine learning.

3. Implementation

3.1 Introduction

In this chapter, the implementation part of the planned project will be discussed based on the Section 2. Implemented project will be shown with I/O file formats, pseudo codes, libraries, mathematical formulas, and descriptions of each applied part. It will refer some additions and new methods with the difference of proposed study.

3.2 Data

Nowadays, it is quite easy to find a data set which is related to medicine and drug with the review of individual’s experiences. In this study, the data set has been obtained by the UCI Machine Learning Repository. It helps people to have empirical analysis of machine learning

algorithms with its collection of databases, domain theories, and data generators [2]. Drug review data set provides patient reviews on the experience of specific drugs along with related conditions. Reviews are linked to the aspects: benefits, side effects and overall comment. Moreover, ratings are available to make inference about overall satisfaction of user about the used drug and each rating has interval [0, 10].

Table 1: Data set information

Attribute	Date Type	Information	Example
UniqueID	Numerical	Unique id	163740
drugName	Categorical	Name of drug	Mirtazapine
condition	Categorical	Name of condition	Depression
review	Text	Patient review	"Quick reduction of symptoms"
rating	Numerical	10-star patient rating	9
date	Continuous	Date of review entry	29-Seb-2017
usefulCount	Numerical	Number of users who found review useful	17

The review and rating attributes are the most important parts of the data set to apply and test SentiWordNet with sentiment analysis. However, reviews cannot be used directly as an input for the program due to it has some duplicated, inaccurate, or incorrect and we need to apply for the data cleansing phase of NLP.

3.3 Data Cleansing

Data cleansing is the process of increasing the quality of data by modifying and preparing the data that have the mentioned shortcomings of reviews. Some of data cleansing techniques that have been applied in this study will be shown step by step in the following:

1. Irrelevant observations that do not actually fit our problem have been removed.
2. Single and unique conditions have been eliminated due to there is no chance to have comparation with another one.
3. Since we want to apply for sentiment analysis, numerical content in the text is not useful so I kept only the text content.
4. Nltk corpus gives us an opportunity to detect meaningful words among all words in the text so all meaningless words are discarded.

3.4 SentiWordNet and WordNet

SentiWordNet has been theoretically explained in Section 2.2. In this part, I will explain how to use SentiWordNet and its relationship with WordNet. As a search param, I will consider the word ‘wonder’ to illustrate the results. The WordNet part gets all synsets against our param. SentiWordNet part gets the polarities for each synset. This process is the smallest piece of acquiring sentiment results for each review of patients in our data set. Lemmas have been shown for illustrative purposes to understand synsets in better way.

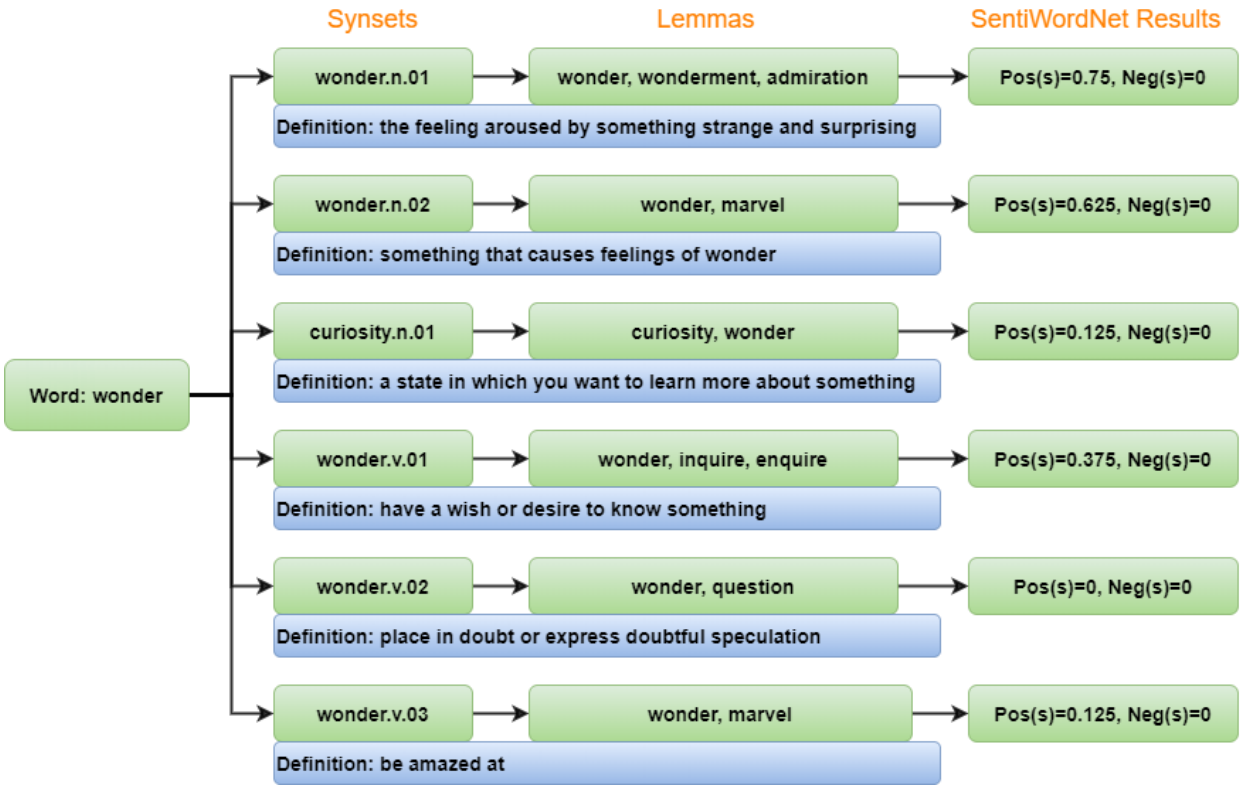

Figure 3.1: The relationship of SentiWordNet and WordNet

As you can see in the Figure 3.1, each synset has POS tag. If we consider the first synset 'wonder.n.01', its tags can be interpreted as the following:

- Wonder: the word we need to get polarity of
- n: Part-of-speech (n = noun)
- 01: the most common usage of wonder as a noun. It indicates the frequency of usage.

POS tagging is the process of reading the text in a corpus and assign corresponding part of a speech tag to each word such as noun (n), verb (v), adjective (a), adverb (r), etc.

3.5 Data Transformation

Firstly, I will show the steps that has been done during the data transformation process and explain them after.

1. Replacing negations with antonyms.
2. Tokenizing sentences.
3. Removing stop words and tokenizing words.
4. Finding POS tags of each word.
5. Lemmatizing and acquiring synset that has the most common usage of lemma.
6. Normalizing the ratings.

Since the words are the basic input of our program, we must handle negations such as 'not happy', 'not effective' etc. WordNet gives us a change to detect these words and replace them by supporting antonyms of lemmas. Antonyms are words that have opposite or contrasting meaning

of another word. In our example, while ‘not happy’ becomes ‘unhappy’, ‘not effective’ becomes ‘ineffective’ with the help of WordNet.

After replacing the negations with their antonyms, all sentences and words have been tokenized. It means that all review text has been split firstly into sentences, and then each sentence has been split into words. After this process, stop words have been removed to clean useless words. Stop words refer to words that does not add much meaning to a sentence such as the, she, is, a etc. They can be ignored without changing the meaning of sentence. After we remove stop words, we acquired the plain words. These words have been used to identify POS tags that will make easier to distinguish synsets and find the correct input for our sentiment analysis (see Figure 3.1). WordNetLemmatizer has been used with obtained POS tags for lemmatization. After this process, the most suitable synsets have been chosen with lemmas. The synset that has the most common usage is used to apply for sentiment analysis.

Lastly, that ratings which have interval [0, 10] has been normalized. If rating is higher than *X*, that has been considered as positive review (+1), if it is lower than *Y*, that has been consider as negative review (-1) and other values became neutral (0). X and Y thresholds have been changed during the implementation phase for testing purposes.

3.6 Application of Sentiment Analysis

Rule-based approach has been applied for sentiment analysis (see Section 2.3.1). Our rule has been defined as the following.

1. Acquiring the polarities for each word by using SentiWordNet (*Pos(s), Neg(s)*).
2. Calculating the sentiment result for each word with the acquired polarities by the formula:

   $$WSR(w) = Pos(s) - Neg(s) \quad \text{(Eq. 3.1)}$$

   where w represents the applied word whereas WSR is the sentiment result of w.
3. Calculating the average result of all review text by the formula:

   $$Avg(r) = \frac{Total\ sentiment\ result\ each\ word\ in\ review}{The\ number\ of\ total\ word\ in\ review} \quad \text{(Eq. 3.2)}$$

   where r represents all the review whereas Avg is the average sentiment result of all review. All steps that are created our rules have been shown. The result for each review is become our prediction while the ratings are actual ones.

3.6.1 Evaluation of Sentiment Analysis

Some evaluation methods have been used to measure the performance of sentiment analysis. All the performance metrics are based on confusion matrix whose content is given in Figure 3.2.

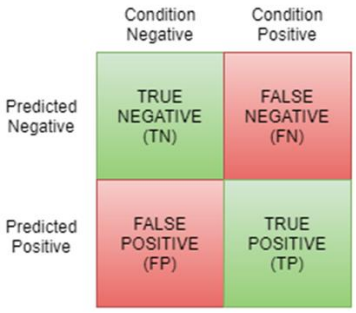

Figure 3.2: An illustration of confusion matrix

True positives (TP) are the cases when the actual label of review was positive, and the predicted label is also positive. True negatives (TN) are the cases when both the predicted label of review and actual one is negative. FP is false positive whose true label is negative, but the method predicted as positive. Similarly, FN is false negative, whose true label is positive, but the method assigned as negative.

3.6.1.1 Accuracy

Accuracy refers to the closeness of a predicted value to an actual (real) value (Eq. 3.3).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad \text{(Eq. 3.3)}$$

3.6.1.2 Precision and Recall

Precision and recall are useful measures of success of prediction when the classes are very imbalanced. Precision is a measure of result relevancy (Eq. 3.4), while recall is a measure of how many truly relevant results are returned (Eq. 3.5).

$$Precision = \frac{TP}{TP + FP} \quad \text{(Eq. 3.4)}$$

$$Recall = \frac{TP}{TP + FN} \quad \text{(Eq. 3.5)}$$

A high precision shows that an algorithm returned substantially more relevant results than irrelevant ones, while high recall means that an algorithm returned most of the relevant results.

3.6.1.3 F1-measure

If a machine learning algorithm is good at recall, it does not mean that algorithm will provide a high precision as well. That is why we also need F1-measure which is the harmonic mean of recall and precision to evaluate an algorithm (Eq. 3.6).

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall} \qquad \textit{(Eq. 3.6)}$$

4. Test and Experiments

4.1 SentiWordNet and WordNet

These topics have been detailly mentioned in Section 3.4. The test classes have been created to discover and experience the usage of SentiWordNet and WordNet in NLP with their methods while understanding the relationship between them. WordNet's structure and easily integrability makes it very useful tool for NLP. All the topics that have been mentioned during Section 3 has been tested and experienced during the implementation part of the project. Despite SentiWordNet gives quite successful polarity results while applying its functions, since we focus on each word and we are not considering the word combinations it requires to clean and prepare the data in more detailed and careful way. However, both sources will keep going to play significant role in computational linguistics and natural language processing.

4.2 Rule-based Approach for Sentiment Analysis

Rule-based systems are very naive because the sequence of word combinations is not considering. More advanced approaches can be made but it quickly gets very complicated and it is hard to maintain. Accuracy and efficiency depend the defining rules. When we add a new rule to the current system, we must observe how it is interacting with older ones. As a result, these systems require an expert to create rules and have careful maintenance. Since we are talking about medical experiments, it would be better to work concurrently with a doctor.

4.3 Results

The combination of Rules-based approach with SentiWordNet did not give quite successful results due to the shortcomings that have been mentioned. Moreover, the changing of decision thresholds of user ratings played the main role because it was our target value while training phase of the program. We can see the results for different thresholds in the following figures.

	Precision	Recall	F1-score
-1	0.47	0.77	0.58
1	0.79	0.49	0.6
Total Accuracy: 0.60			

Figure 4.1: The result of sentiment analysis when rating thresholds (3,9)

	Precision	Recall	F1-score
-1	0.38	0.75	0.51
1	0.81	0.47	0.59
Total Accuracy: 0.55			

Figure 4.2: The result of sentiment analysis when rating thresholds (3,8)

	Precision	Recall	F1-score
-1	0.34	0.75	0.47
1	0.83	0.46	0.59
Total Accuracy: 0.54			

Figure 4.3:The result of sentiment analysis when rating thresholds (4,7)

	Precision	Recall	F1-score
-1	0.31	0.75	0.44
1	0.84	0.45	0.59
Total Accuracy: 0.52			

Figure 4.4:The result of sentiment analysis when rating thresholds (4,6)

We can see in the figures that our result can be changed according to rating thresholds (x,y) where x represents the threshold for positive label whereas y is the threshold for negative labels.

5. Conclusion and Future Works

The easy application and quite successful results of SentiWordNet are making it very useful for feature analysis and experiences of NLP applications. Since we were based on rule-based approach in this study, the results were not as expected. Other approaches can be more suitable to work with drug review data set for sentiment analysis purposes.

On the other hand, the data set and features based on users and their experiences. There are a lot of missing and unmatched data. Machine learning methods have better performance while it has cleaner data with more purposeful features. These points can play a significant role to improve performance.

REFERENCES

[1] Bird, Steven, Edward Loper and Ewan Klein (2009), *Natural Language Processing with Python*. O'Reilly Media Inc.

[2] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[3] Esuli and F. Sebastiani, SentiWordNet: A PubliclyAvailable Lexical Resource for Opinion Mining.Proceedings from International Conference onLanguage Resources and Evaluation (LREC), Genoa,2006.

[4] Felix Gräßer, Surya Kallumadi, Hagen Malberg, and Sebastian Zaunseder. 2018. Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. In Proceedings of the 2018 International Conference on Digital Health (DH '18). ACM, New York, NY, USA, 121-125.

[5] George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

[6] Lazarou J., Pomeranz B.H.,Corey P.N. Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies. *Journal of the American Medical Association*. 1998;279(15):1200-1205. PM:9555760.