THE UNIVERSITY OF TEXAS AT DALLAS

OPINION MINING ON YELP ACADEMIC DATASET

FINAL PROJECT REPORT

PRASAD SUNIL PANDE

4/29/2015

ABSTRACT

Opinion mining is an area of Natural Language Processing where a system is developed that can identify and classify opinion or sentiment as represented in an electronic dataset. Opinion Mining is simple words is defined as identifying the subjective sentences from an input document and successively identify the polar phrase among the sentences as positive and negative. Previous attempts has defined opinion mining as a sequence of few standard processes namely subjectivity detection, polarity detection and degree of polarity identification. The scope of this project includes identification of subjective sentences and identify the polar phrases from those subjective statements based on the algorithm mentioned in A Review on Natural Language Processing in Opinion Mining [1].

INTRODUCTION

Today many human decision relies on the opinions. We always consider the experience of the other people while taking any decision in everyday life. If you want to visit a new restaurant in your area, you take opinions from the people who already visited that restaurant. Based on their opinion you decide whether you are going for the restaurant or not. Opinions are basically the user experience about particular product or service. In today's world of digital media people are sharing their experiences in the form of reviews. Reviews are available for all kind of products of most of the retail giants like Amazon, eBay. While purchasing any product from these retailers, people always read the reviews and comments from the others. Also, there are websites such as Yelp, Zomato, Rotten Tomatoes where people share their experiences about different places likes restaurants, clinics, clubs etc. These websites gives us wider views about the particular thing. Today our decision is not only dependent on the reviews from the one or two trusted people we know but on the all the reviews that people had mentioned through these websites. A survey showed 88% people trust the online reviews as personal recommendations [2]. Seems like is life is becoming more and more simple.

For this project I have considered the reviews about the businesses available on Yelp. Yelp has the list of reviews for most of the businesses in your area. According to survey, Yelp has 135 million average monthly unique visitors. There are around 67 million reviews available for different businesses as of October 2014 [3]. My project is considering reviews for different categories of businesses and applying the subjectivity detection and polarity identification algorithm as mentioned in algorithm mentioned in the paper [1].

PRIOR WORK

A lot of work has been previously done in the area of opinion mining. Lot of projects have been implemented on different customer reviews to find the opinions. M.Hu and B.Liu has designed a famous framework [4] on Mining and Summarizing the Customer Reviews. Extracting the opinion from the different product reviews is comparatively an easy task since the aspects of particular products are always limited and known. Suppose I am applying any opinion mining algorithm on the mobile reviews then the different features we need to consider for a particular mobile device are limited i.e. hardware configuration, operating system, available memory etc. These features are rarely changed and they are almost used in the same context. For testing my algorithm I have considered the reviews for Dr. Goldberg's clinic. For a subject like clinic or food restaurant people express their experience in different ways and these aspects are not the same always. The vocabulary for describing such businesses' reviews is also vast. There are more ambiguities in the sentences constructed for writing reviews. I have used different papers and frameworks designed for performing different task of opinion mining and try to collaborate those to implement my algorithm.

I referred to the subjectivity lexicons which contains the clues collected from different resources to recognize the subjectivity of the sentences [5]. I have used these clues to check the subjectivity of the main verb and check if the main verb of the sentence is strongly subjective or not.

I collaborated the SentiWordnet 3.0 framework – An Enhanced lexical resource for Sentiment Analysis and Opinion Mining [6]. To identify the polarity and object score of the lexicon, I implemented SentiWordnet methods to get the precise score.

There is another interesting paper is referred to get the overview of the orientation strength of a particular lexicon by M.J.M. Vermeji. It helped me to understand the concept of the orientation strength of the verb as well as the other for the other parts of speech [7].

SYSTEM DESIGN

DATA EXTRACTION

Yelp provides academic dataset which contains 1.6M reviews of about 61K business units. Data provided by the Yelp is in the form of JSON object. For example, review object in the Yelp data is represented as follows:

```
{
    'type': 'review',
    'business_id': (encrypted business id),
    'user_id': (encrypted user id),
    'stars': (star rating, rounded to half-stars),
    'text': (review text),
    'date': (date, formatted like '2012-03-14'),
    'votes': {(vote type): (count)},
}
```

The first step from the implementation perspective is the extraction of the data from the JSON format. Extraction of the data in the *text* field of a review object for particular *business_id* is the first step performed. Python json package is used to load the JSON data and stored the review sentences in the list format.

STANFORD PARSING

Stanford dependency parser is used in order to find the main verb of the sentence. Stanford dependency parser, parse function is used to find the dependencies of the sentences. Dependencies returned are parsed to find ROOT of the sentence structure which in most of the cases is the verb. To re-confirm if the ROOT is the main verb of the sentence, POS tag is evaluated and verified for the ROOT word. Sample Dependency tree.

[[u'root', u'ROOT', u'offers'], [u'nn', u'goldberg', u'dr.'], [u'nsubj', u'offers', u'goldberg'], [u'iobj', u'offers', u'everything'], [u'dobj', u'offers', u'i'], [u'dep', u'offers', u'look'], [u'prep', u'look', u'for'], [u'pcomp', u'for', u'in'], [u'det', u'practitioner', u'a'], [u'amod', u'practitioner', u'general'], [u'pobj', u'in', u'practitioner']]

As we can see for most of the sentences, verb comes as the ROOT of the sentence which help me to detect the main verb of particular sentence.

For this project, I used to the python wrapper designed for Stanford dependency parser [8].

Dependency parser using python wrapper is implemented as a client-server based application where server keeps on running throughout the execution of the code and user program acts as a client.

SUBJECTIVITY DETECTION

Subjectivity Lexicon list obtained from the Multi-perspective Question Answering Project is loaded in the MySQL database. Subjectivity lexicon list has the subjectivity defined for 8215 different words. Subjectivity lexicon list has the following attributes:

Each line in the file contains one subjectivity clue. Below is an example

{type=strongsubj len=1 word1=abuse pos1=verb stemmed1=y priorpolarity=negative}

- **type** - Either strongsubj or weaksubj. A clue that is subjective in most context is considered strongly subjective (strongsubj), and those that may only have certain subjective usages are considered weakly subjective (weaksubj).
- **len** - Length of the clue in words. All clues in this file are single words.
- **word1** - Token or stem of the clue
- **pos1** - part of speech of the clue, may be any pos (any part of speech)
- **stemmed1** - y (yes) or n (no). Is the clue word1 stemmed? If stemmed1=y, this means that the clue should match all unstemmed variants of the word with the corresponding part of speech. For example, "abuse", above, will match "abuses" (verb), "abused" (verb), "abusing" (verb), but not "abuse" (noun) or "abuses" (noun)
- **priorpolarity** - positive, negative, both, neutral. The prior polarity of the clue. Out of context, does the clue seem to evoke something positive or something negative?

Subjectivity for the main verb of the sentence is identified using the dependency parsing and subjectivity of the verb is retrieved from MySQL database.

POLARITY DETECTION

Once the subjective sentences are detected and stored in the file using subjectivity detection algorithm. Next step is to find the polarity of the sentence. Polarity detection algorithm extracts the subjective sentences and based on subjective phrases assign the polarity to the sentences as positive, negative and neutral. If there are more than one subjective phrases in the statements then based on the cumulative score of the subjective phrases polarity is decided for the statement.

ALGORITHM

Algorithm of Subjectivity Detection

1. Input Document
2. Segregate Sentences
3. Search for any Strong Subjectivity Word List from Subjectivity Word List
4. Either search for more than one weak subjectivity words
5. Parsing sentence using Dependency parser.
6. Look for finite verb chunk in present system
7. Take any decision depending upon the orientation strength of the main verb
8. If the orientation strength is high then the sentence is subjective itself
9. If it is low then search for any other POS categories like Noun, Adjective, Adverb, Verb, and accumulate their polarity values from SentiWordNet. If the summation value is more than threshold value (0.5 here) then it is assumed that the sentence is subjective itself.

Algorithm of Polarity Identification

1. Consider extracted subjective sentences
2. If there is only one subjective expression in the sentence, then simply assign exact polarity value as in the subjectivity word list
3. In case of multiple subjective phrases in the sentence; system look for any dependency relationship between them.
4. In case of dependency relation the overall polarity of a sentence is calculated by modifier's polarity

\

SYSTEM DESIGN DIAGRAM

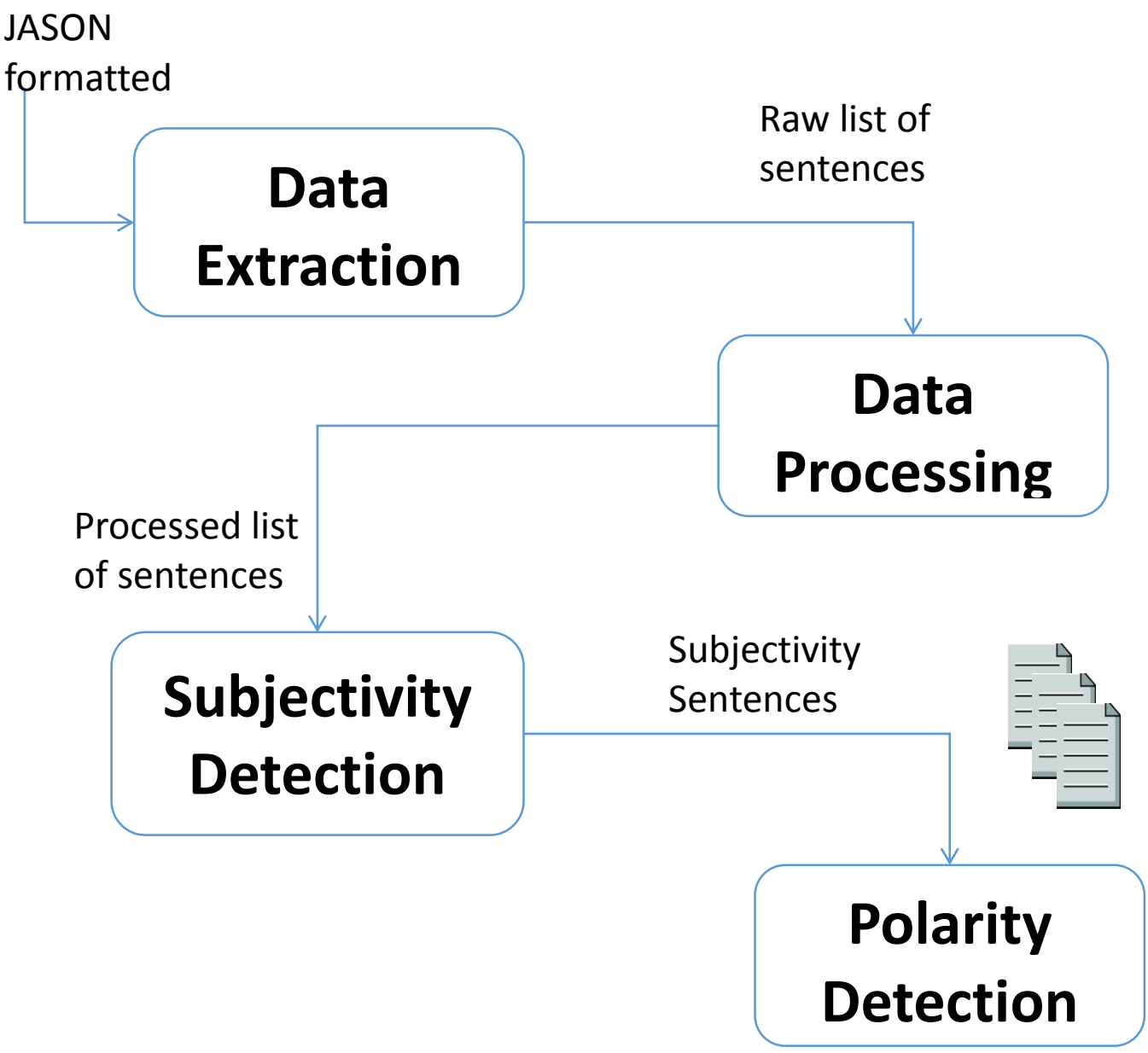

Diagram: System design flow

INPUT AND OUTPUT

INPUT

Input for the program is the file containing the review JSON object. Sample review object is as follows:

{

"votes": {

"funny": 0,

"useful": 2,

"cool": 1

},

"user_id": "Xqd0DzHaiyRqVH3WRG7hzg",

"review_id": "15SdjuK7DmYqUAj6rjGowg",

"stars": 5,

"date": "2007-05-17",

"text": "dr. goldberg offers everything i look for in a general practitioner. he's nice andeasy to talk to without being patronizing; he's always on time in seeing his patients; he's affiliated with a top-notch hospital (nyu) which my parents have explained to me is very important in case something happens and you need surgery; and you can get referrals to see specialists without having to see him first. really, what more do you need? i'm sitting here trying to think of any complaints i have about him, but i'm really drawing a blank.",

"type": "review",

"business_id": "vcNAWiLM4dR7D2nwwJ7nCA"

}

JSON formatted review is parsed in the program and text is extracted for particular *business_id*. Review text is preprocessed to segregate and store the each sentence in the list. Sentences from the list is then undergone a subjectivity detection and polarity detection algorithms.

OUTPUT

Output of the program are following text files:

- Text File for Subjective Sentences – Output of the Subjectivity Detection Step
- Text File for Positive Opinions – Output of the Polarity detection algorithm.
- Text File for Negative Opinions - Output of the Polarity detection algorithm.
- Text File for Neutral Opinions - Output of the Polarity detection algorithm

Sample outputs from Positive and Negative opinions:

Sample Positive opinions extracted:

He will be missed very much.I think finding a new doctor in NYC that you actually like might almost be as awful as trying to find a date!
I love Dr. Goldberg.
Today I actually got to see the doctor a few minutes early!
You deserve better and they will not be there when you really need them.
i am sitting here trying to think of any complaints i have about him, but i am really drawing a blank.
Can not say I am surprised when I was referred to him by another doctor who I think is wonderful and because he went to one of the best medical schools in the country.It is really easy to get an appointment.

Sample Negative Opinions extracted:

And to make matters even worse, his office staff is incompetent.I do not know what Dr. Goldberg was like before moving to Arizona, but let me tell you, STAY AWAY from this doctor and this office.
I think I was one of his 1st patients when he started at MHMG.The entire office has an attitude like they are doing you a favor.
Unfortunately, the frustration of being Dr. Goldberg's patient is a repeat of the experience I have had with so many other doctors in NYC -- good doctor, terrible staff.
Got a letter in the mail last week that said Dr. Goldberg is moving to Arizona to take a new position there in June.It is with regret that I feel that I have to give Dr. Goldberg 2

CONCLUSION AND FUTURE WORK

Opinion mining on wide range of review dataset is still an active research in progress. Accuracy of any particular system depends of many factors considered during the system design. In my project, I have taken the subjectivity of the part of speech categories of the sentences and SentiWordnet classifier to detect the subjectivity and polarity of the sentences. There are many other factors that can be considered to improve the accuracy of this algorithm. Subjectivity of the sentence is limited to the subjectivity wordlist that I have considered. Extending the subjectivity lexicon list and their subjectivities will give us more accuracy in evaluating subjectivity of the sentences. Semantics of the negated sentences are not considered in the current system. Considering the semantics of these type of sentences will help to improve the results of the polarity detection. Some of the complex emotions such as sarcasm are hard to detect which are not covered as a part of this experiment.

Future work in this project involves sarcasm and negativity detection to handle the complex emotions of the people. Also, project can be extended to include the subjectivity and semantics of the pronouns which are ignored in the current system.

ACKNOWLEDGEMENT

I want to thank Prof. Dr. Dan Moldovan for giving me this opportunity to work on such a vast topic and guiding me throughout the course of the project.

REFERENCES

[1] A Review on Natural Language Processing in Opinion Mining by Debnath Bhattacharyya, Susmita Biswas and Tai-hoon Kim-International Journal of International Journal of Smart Home, Vol.4, No.2, April, 2010 Vol.4, No.2, April, 2010

[2] Search Engine Land Reviews – About local consumer review survey conducted in 2014.

[3] By the Numbers: 40 Amazing Yelp Statistics by DMR

[4] Mining and Summarizing Customer Reviews by M.Hu and B.Liu

[5] Subjectivity Lexicon - Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis - Theresa Wilson, Janyce Wiebe and Paul Hoffmann

[6] SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining - Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani

[7] THE ORIENTATION OF USER OPINIONS THROUGH ADVERBS, VERBS AND NOUNS by M.J.M. Vermeij

[8] A python wrapper for the Stanford CoreNLP java library. https://github.com/kedz/corenlp