Understanding Human Sentiments with the help of Sentiment Analysis

[1]Pooja Bhatia, [2]Pratik P. Watwani

Students
Department of Computer Engineering
Vivekanand Education Society's Institute of Technology, Chembur, India

***Abstract*—Sentiment Analysis is a prominent application in the domain of Natural Language Processing.In most basic words, Sentiment Analysis is utilized to comprehend the sentiments from a single word to thousands of pages of the document, which is conveyed by either one or a huge mass of people. This paper concentrates on sharing in the first place an introduction to Sentiment Analysis; Understanding the need of utilizing sentiment analysis and why it requires a huge amount of consideration. We here attempt to examine about how sentiment analysis can be performed and utilized by any organization to comprehend the opinions of a single individual or a vast mass of individuals. Without a doubt, with the progressing time new systems, tools, APIs are being produced and overseen. Additionally, we've attempted to exhibit the working of sentiment analysis with a case of a hotel survey, legitimizing the sentiments of the guests about the hotel. With the development in Internet and users, sentiment analysis can be utilized by the organization to decide the nature of their administration being offered, be that as it may, not just this in the up and coming sections of this paper different examples have been talked about which are altogether long far from the business point of view, where it is used.**

***IndexTerms*—sentiment, sentiment analysis, natural language, natural language processing, NLTK.**

I. Introduction

Sentiment Analysis is a small chunk of Natural Language Processing field. Proposing as it seems to be, Natural Language Processing (NLP for short) is the undertaking of utilizing computational advances to comprehend and control the natural language otherwise known as human spoken language. Natural Language Processing utilizes computational methodologies to first comprehend the language and after that assistance, the client can perform distinctive demands. For instance, making a robot play out your guidelines. NLP is one of the huge fields of Artificial Intelligence. There can be two data sources that can be fed to the NLP system: Text or Speech.

NLP is developing each day, it is utilized in a variety of applications, which incorporate, Named Entity Recognition, Optical Character Recognition, Automatic Summarization, Lexical Semantics, Speech Recognition, Speech Processing, Question Answering, Sentiment Analysis, Intent Analysis, Text-to-Speech and so on.

II. Sentiment Analysis

A. Introduction

Human correspondence is impractical without the supplication of one of the three vital elements viz. sentiments, emotions or opinions. It is not hard to state but rather this correspondence is not, in the least, conceivable without these elements. Now, simply put, sentiments or opinions are a reaction to a specific external activity. Say, between a correspondence of a supervisor and a worker who just got fired from his job, the worker will undoubtedly certainly demonstrate a few sentiments and feelings or even express his opinions as far as offering a change and development in his work.

The approach of SA is very systematized. The primary segment is the input, which can be as either text or speech. Utilizing any of the accessible differentiating strategies, the information is dissected for general sentiments of the opinion holder towards an entity. For instance, a product provided by the organization. So, a user utilizing that specific product is the opinion holder, holding the opinion about the product i.e. the entity.

Most usually, SA is utilized by an organization who sells a product or service to an individual or a mass of the public. In the developing world of internet, users tend to share their reviews, opinions on a wide and distinct assortment of sites. Manual analysis is in this manner impractical, you can't just look through many sites, perusing about reviews of the product or service you offer. Hence, automating this task is crucial.

Users give their reviews and opinions about the product or service, this data fills in as the contribution to the input of the system, which at last decides the polarity of the feedback.

B. The Need

Consider the accompanying examples of genuine cases where Sentiment Analysis has been used, these themselves explain the need for sentiment analysis:

a) For 2012 Presidential race, Obama organization used sentiment analysis with public opinion for their campaign policies.

b) Do you see, when news channels get some information about conclusions to open? This is one of the utilizations of Sentiment Analysis, presumably been done since years.

c) IBM's Watson can be utilized to investigate an organization's tweets – and tweets about the organization – to find issues, openings, and experiences into client estimations.

d) Audio sentiment analysis is being utilized to quantify stress levels in call centers with the goal that client representative agents can gauge how steamed the client is and intercede prior before things raise. Clients frequently talk into the recipient while they are on hold or tuning into the relieving music, and they additionally can likewise make different sounds, for example, overwhelming sighing, which can show that they are becoming progressively baffled. [1]

e) Even Wimbledon started utilizing opinion mining this year to help foresee which features and news points rising out of the competition would most interest its fans. Their frameworks could investigate existing Tweets, redesigns, and remarks and make prescient recommendations about the sorts of stories that fans would be well on the way to respond to decidedly. [1]

f) Expedia in Canada utilized opinion analysis to discover that the music going with one of their advertisements was accepting an overwhelmingly negative reaction on the web, and they could react to that feeling suitably: by discharging another form of the advertisement in which the culpable violin was suddenly smashed.

g) Researchers at the Microsoft Research Labs in Washington found that it was conceivable to anticipate with content based sentiment examination which ladies were at danger of postnatal melancholy just by breaking down their Twitter posts. The research concentrated on verbal signs that the mother would utilize weeks before conceiving an offspring. The individuals who battle with parenthood tended to utilize words that indicated at a fundamental tension and despondency. There was greater antagonism in the language utilized, with an expansion in words, for example, baffled, hopeless, disappointed, miserable and abhor, and additionally an expansion in the utilization of "I" – showing a separation from the "we" of looming parenthood. [1]

h) Sentiments in electronic mails were utilized to discover how sexual orientations contrasted on enthusiastic emotional axes [2]

i) Emotions in books and novels were studied [3]

The above examples simply demonstrate the fact as to how sentiment analysis has been used in various fields by different organizations to understand human language

III. Levels of Analysis

There are three unmistakable levels at which analysis can be carried out: [4]

a) Document Level: This level fixations the whole record. The sentiment scores made at this level depict the limit at the report level. For an item analysis on this level, the outcomes would address either a beneficial outcome or a negative impact, i.e. positive or negative.

b) Sentence Level: In this level, the document is divided into different sentences. Each sentence thusly addresses positivity, negativity or neutrality. Not at all like document level, this level addresses the furthest point of extremity on each sentence.

c) Aspect Level: The most lessened level at which the analysis can be performed. Each sentence is additionally disembodied into words, on each of these words analysis is performed.

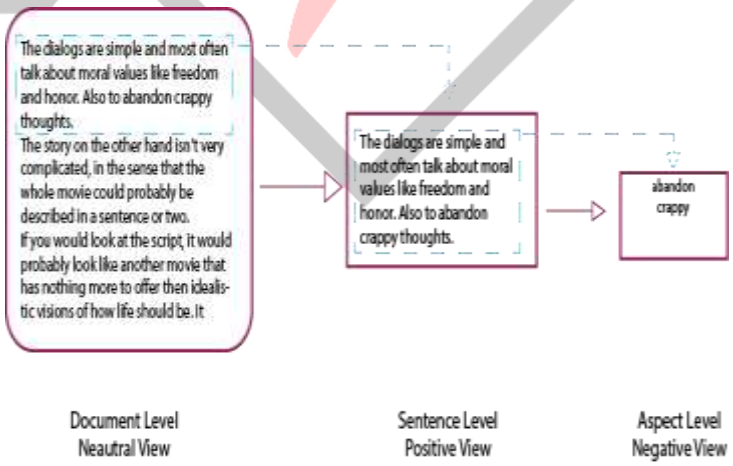

FIG 3.1. LEVELS OF SENTIMENT ANALYSIS.

It has always been a subject of argument whether which parts of speech manifestly describe the sentiments of the natural language.An examination coordinated describes whether the usage of adverbs or use of adjective-adverb yields better results in the analysis. Regardless, the escorting example depicts the representation showing that adverbs do influence the nature of a given sentiment.

- (S1) The match was good.
- (S2) The match was altogether good.
- (S3) The match was awesome.

Each of the three sentences is positive, however, we can concur that sentiments grow stronger from sentences (S1) to (S3)[5]

IV. Opinion

An opinion can be portrayed as the judgment confined by a person about an element. Sentiments can be either positive, neutral or negative which is known as polarity. Opinions are the fundamental bit of decision making, we usually tend to make, approach or look for assessments in each phase of life.

Opinions are categorized as:

a) Regular Opinions: These are a direct or an indirect target of sentiments on entities.

Ex: I love the phone. (or) The hotel service was pathetic.

b) Comparative Opinions: These opinions are generally comparison of two or more entities.

Ex: I find Android far better than iOS.

c) Expressed Opinions: These can be either implicit or explicit.

We can characterize opinion as: [6]

$(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$

This characterization is based on five terms, element name(e_{ij}), a part of this element(e_{ij}) is the aspect(a_{ij}), feeling(s_{ijkl}) on the aspect(a_{ij}) of the element(e_{ij}), the opinion holder(h_k) and the time(t_l) is when the sentiment is communicated by the holder(h_k).The $terms_{ijkl}$ can either be positive, neutral or negative.

V. Implementation Procedure

With its headway as of late, sentiment analysis can be performed with the assistance of different systems, tools, and procedures. There are first class tools and resources that have been developed and maintained. This paper will center to exhibit sentiment analysis utilizing the approach of the lexicon.

The 4 most imperative steps that must be conveyed to perform sentiment analysis on a set of data includes:

1. Text Normalization
2. Tokenization
3. Tagging-Identifying Words
4. Applying Lexicon

1. Text Normalization

Any piece of text we write contains anabundant amount of secondary information. This secondary information includes the numerical digits, punctuations, symbols, abbreviations, acronyms. This secondary information will not at all contribute towards the desired output. The aim of this step is to normalize the given text, this is done by eliminating all this secondary information from the raw string. It is always desired that the text should be normalized because it helps in generating a consistent and desired output.

Text Normalization provides with the following features:

a) *Eliminating Punctuations*

Elimination of Punctuation evacuates all the punctuations that show up into the RawString. However, eliminatingpunctuations is optional, it is suitable to use it, as it makes the text cleaner and consistent.

b) *Conversion into Lower and Upper Case*

Case conversion is the process to convert the raw string to an upper case or lower case. Changing over the case delivers a steady content yield.

c) *Eliminating Stop Words*

Every dialect contains a set of words that are used occasionally, these are by and large the words which essentially don't add any intent to the sentence. The dedication of such words is henceforth not required in NLP assignments. Hence, StopWords must be discarded.Most web search engines are designed to more often take out such words from the query to spare space and time.

TABLE I. SAMPLE LIST OF STOP WORDS

A	About	Be
Been	Below	Between
Cannot	Does	During
Each	Few	Further
Here	Is	It
Myself	No	Off
Ought	Over	Own
Same	So	Suh
There	Through	Under
Until	What	Yourselves

d) Substituting and Correcting Tokens

It is at times essential to change over various entities or to supplant them in the given Raw String. For example, it gets less demanding if we have an immaculate orderly string, free from punctuations, repeating characters, symbols, emoticons, grammatical tenses present in it. Substitution and Correction strategies are utilized to disentangle and make a standard authoritative shape.

The substitution technique utilized here is supporting the use of Regular Expressions and the adjustment strategy usedis for repeating characters (known as the word lengthening) in Raw String.

i. Substituting Contracted Text to Non-Contracted Text

These are the short words which are produced using more than two, one of a kind words or syllable by averting letters and sounds. Contracted words, generally, incorporate elimination a vowel and replacing it by a punctuation in forming. Such words may be either positive or negative depending upon the combination of words picked. The following table enlists some examples of Contracted Words in the English Language.

TABLE II. SAMPLE LIST OF CONTRACTED WORDS

Ain’t	Aren’t	Can’t
Could’ve	Didn’t	Don’t
Haven’t	Hadn’t	He’ll
How’ll	How’s	I’d
I’ll	Isn’t	It’ll
It’s	Ma’am	Mightn’t
Mustn’t	Must’ve	O’clock
Shan’t	Shouldn't	That’d
They’ll	We’d	Where’s
Won’t	Would’ve	Why’ll

ii. Deleting Repeating Characters

As the internet advanced so did the internet communication language, and especially the internet slang language. Word Lengthening is a technique for extending a word, particularly attaching the last letter of the word numerous circumstances. For example, heyyyy, ewww, ohh

Prevailing patterns demonstrate that especially teens on the internet use this technique for communication. This disrupts the normal language.

2. Tokenization

The first and the most important step in NLP tasks is Tokenization. The aim here is to, say, taking a String from the input and breaking it down into smaller distinct components. These components are called as tokens. Simply, this breaking is to break the string to get all the words which have been used to form that string.

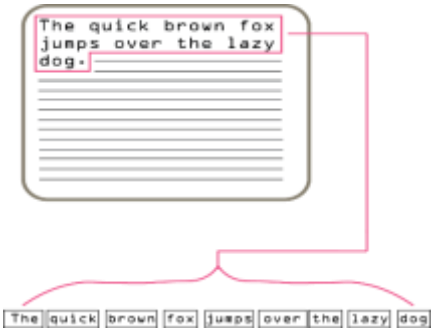

FIG 1. LEVELS OF SENTIMENT ANALYSIS.

3. Tagging-Identifying Words

POS-Tagging is the process of allotting different categories of Grammatical Parts of Speech(POS) to each generated token. It is the most fundamental step in recognizing words as nouns, adjectives, adverbs, verbs et cetera. The POS Tag can help in recognizing the words that are related to sentiments.

TABLE III. SAMPLE LIST OF POS-TAGGERS

Tag	Meaning
CC	Coordinating Conjunction
DT	Determiner
JJ	Adjective
MD	Model
NN	Noun(Singular)
RB	Adverb
SYM	Symbol
UH	Interjection
VBP	Verb

Each tag resembles a kind of part of speech. There are about 36 available tags provided by Penn Treebank for POS-Tagging.

4. Lexicon Application

Lexicons are preordained dictionaries which contain a well-organized set of words or expressions which can be positive, neutral or negative in nature. A wide variety of lexicons is available. Lexicons can either take a concept-based or rule-based approach to determine the polarity of a given token.

Different Lexicons available are:

a) Bing-Liu Opinion Lexicon contains roughly 2006 positive words and 4783 negative words.
b) General Inquirer Lexicon, build up with around has 1,915 positive and 2,291 negative words.
c) Dictionary of Affect in Language is made out of the emotional words that can be utilized for Sentiment Analysis.
d) VADER is a champion among the most understood Lexicon as it can understand slang language.
e) SentiSense has a tremendous rundown of words with more than 5,496 words and 2,190 synsets which are named with feelings from more than 14 distinctive emotional classes.
f) Sentiment Lexicon given by the University of Pittsburgh is a vocabulary of around 8000 words with positive, negative and neutral feelings.

VI. RESULTS

1. Considerations and Statistics

To demonstrate the working, the environment setup was:
Programming Language: Python
Toolkit: NLTK
Crawler/Scraper Framework: Selenium

a) The informational collection of hotel reviews was extracted with the assistance of a Crawler and a Scraper designed in Python with the assistance of Selenium Framework. The information was recorded in excel sheets. The information recorded comprised of User ID, Review Date, Title of the Review, Review. Of this information, the Review segment was additionally separated, which for the most part was immediately utilized for analysis.
b) The data was normalized and then tokenized. The text tokenization technique used was the RegexpToknizer, which is used to tokenize the given set of strings into words. The Regular Expression created is **r'\w+'** which will tokenize the String into tokens at each whitespace occurrence. The use of RegexpTokenizer ensures that the words which might contain punctuation marks, must not get separated. RegexpTokenizer uses the re.split() function to split the words with matching whitespaces and gaps. Later, different functions were developed to perform text normalization steps.
c) The NLTK suite provides a stopwords corpus. This corpus is an instance of NLTK.corpus.reader. It consists of stopwords of upto 14 different languages including English. Expulsion and Substitution of Contracted Words rely on upon which kind of Tokenization technique is used. [The Tokenizer function utilized as a part of this project depends on Regular Expressions(RegexpTokenizer) working on whitespaces. If a Tokenizer of another kind like PunktwordTokenizer or TreebankwordTokenizer is utilized, then the substitution of Contracted Words is very prescribed. For this a function can be made which will hold the contracted words and their substitutions; from that point, the string can be passed through the function and the contracted words can thus be supplanted. Eradication of Repeating Characters can be refined by methods for the help of wordnet package in NLTK suite
d) Penn Treebank has been used for POS tagging, as already discussed it contains 36 different types of POS tags.
e) VADER lexicon has been used for analysis as it provides a lot of words and its ability to support the slang language.
f) To lead this analysis, lexicon approach was utilized to decide the polarity score. The data set used in this project was the reviews of the hotel scraped from different websites across the internet. The dataset contained about 500 reviews. The opinions in the data set were a combination of all reviews and opinions and not favoring a single polarity view.
g) Finally, the output was plotted in the form of word clouds, generated for both positive and negative sentiments.

XYZ Hotel Review
The accompanying table shows the insights of the extracted and analyzed data:

TABLE IV. DETAILED LIST OF STATISTICS FOR XYZ HOTEL, INDIA

Element	Quantity
Total Number of Reviews	500
Total Number of Words (Unabridged)	75356
a. Total Number of Words (Negative)	3278
b. Total Number of Words (Positive)	6028

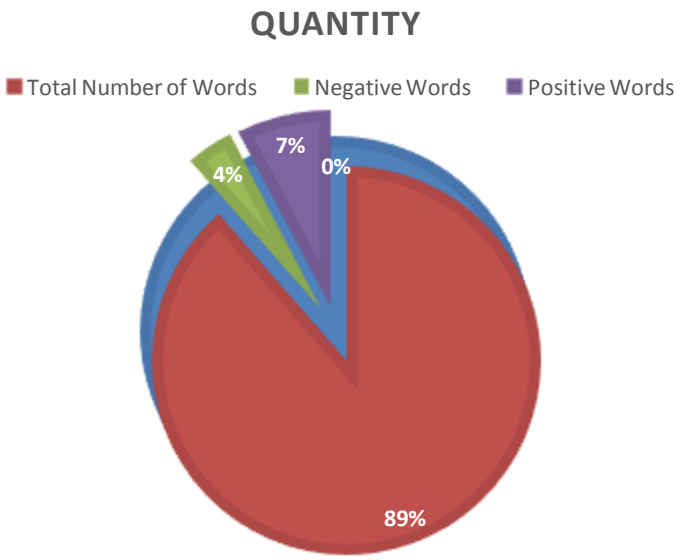

FIG 2. PIE CHART FOR DETAILED STATISTICS REPRESENTING PERCENTAGES OF NEGATIVE AND POSITIVE WORDS

2. Results

Of the 75356 words in 500 reviews, the negative and positive words that were extracted numbered to be 135 and 201 respectively. A portion of the extracted words appears in the accompanying figure. Be that as it may, this figure does not look like to be the most utilized words. They are recorded in the order of extraction.
The two pictures delineate initial 10 words extracted in both polar classifications.

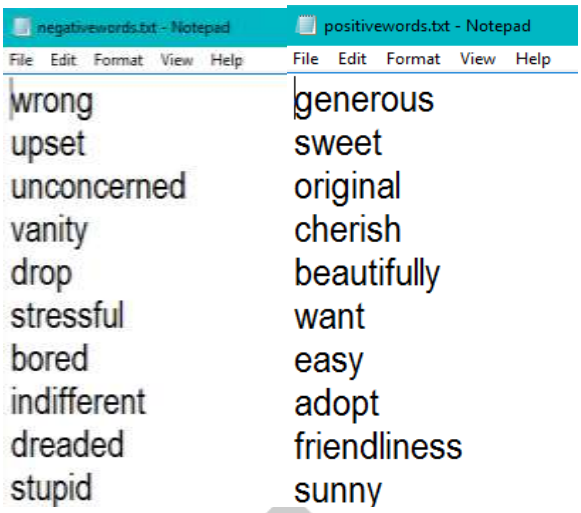

FIG 3. EXTRACTED NEGATIVE AND POSITIVE WORDS

The Word Cloud is a visual portrayal that is made out of words in a specific context or topic. The higher the occurrence frequency of a word, the greater and bolder the text style of the word is in the word cloud. The most commonly used words for XYZ Hotel review can be used to make word clouds to represent different words with their different frequency.

FIG 4.WORD CLOUD FOR NEGATIVE AND POSITIVE SET OF WORDS

As observed over, the negative word cloud conveys a few words like torture, hesitation, reluctance, struggle, stall et al. From this, we can undoubtedly make out, what are the most well-known things or words that are utilized as a part of the negative point in this specific situation. These words help to comprehend what bothered the user of the service, what disappointed them. A few words specifically depict the quality of the organization, for example, racist, criticism, messy, insult, rude and so on.The same implies with the positive word cloud, denoting the most commonly positive words used for the service was, nice, safe, thank, good, improved, friendliness et al.

VII. OPINION SPAM

Positivity and negativity are the perspectives that characterize the present and eventual fate of product or a service. At the point when an item or service gets positivity, it is certain to get distinction and gratefulness, this may come about into the solid improvement of fake surveys for brand advancement. Opinion Spamming is a method for producing fake suppositions exclusively with the end goal of advancing the brand notoriety and its services. Individuals who perform such spamming for a brand are normally identified with and work with them and are called as Opinion Spammers.

It is important to recognize and dispense with the presence of such spamming. In any case, the issue with Opinion Spam Detection is that is difficult to distinguish it. Sensibly, it is difficult to recognize a fake opinion or a review. A man can post a positive fake review for their organization and post a negative fake review for their opposition. [4]

The different types of Opinion Spams are defined as: [4]

a) Fake Reviews

These are the fake reviews generated by opinion spammers for a product, service or an organization. These reviews can be positive or negative depending upon the motive of the spammer.

b) Reviews about Brands Only

These reviews don't talk about a product, but about a certain brand.

c) Non-Reviews

These are not reviews but some random data to produce spam.

VIII. Conclusion

Sentiment Analysis is the most captivating subject in the field of NLP. Organizations are presently using it to decide their administration quality that is being given to the customer. As talked over about the example cases, not only to the organizations but rather sentiment analysis has now permitted us to make a special effort of deciding organization quality by utilizing it in the fields of education, healthcare, therapy, literature, shopping, enhancing client encounter el at. Alongside the progressing time, it's getting off the most well-known and requested on a more significant scale

Sentiments are the basis of human feelings. We cannot communicate without showing our sentiments and emotions, we are humans and will undoubtedly do that. we are equipped for understanding feelings, examining them and, taking essential steps, be that as it may, making a machine do comprehend human assessments is a significant complex assignment.

Machine learning helps in finding out about various sentiments and feelings of an individual, by understanding their natural language. Sentiment analysis ends up being a share in the progression of an organization or a product. It chooses the information nature given by the customer, which in the long run allows a relationship to make basic steps as required. Executing Sentiment Analysis has diverse approaches and techniques available. The planners have endeavored their best to surrender clients to class tools, lexicons, taggers and distinctive modules.

This paper concentrated on to show how sentiment analysis can simply be performed with the assistance of the Lexicon, aside from this approach, there are a wide number of techniques available, however, the most well-known approach is to utilize the machine learning concepts.

References

[1] Bernard Marr, Social Media and the Power of Sentiment Analysis

[2] Mohammad and Yang, Tracking Sentiment in Mail: How Gender Differ on Emotional Axes, 2011

[3] Saif Mohammad, From Once Upon a Time Happily Ever After: Tracking Emotions in Novels and Fairy Tales.

[4] Bing Liu, Sentiment Analysis and Opinion Mining,2012

[5] Farah Benerma, Carmine Cesarno, Antonio Picariello, Diego Reforgiato, VS Subrahmanian, Sentiment Analysis: Adjectives and Adverbs are better than adjectives alone.

[6] [6]Hu and Liu, Sentiment Analysis,2004 and Liu, Sentiment Analysis, 2010.