Sentiment Analysis using Natural Language Processing (NLP)

Miss. Neha Hasanmiya Surve

Asst Professor, Department of Information Technology, DBJ College, Chiplun-415605, Maharashtra, India.

---***---

Abstract: Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral. It played an important role in identifying what other people think and what their behavior is. Applying the Sentiment analysis on the product review on e-market helps not only the customer but also the industry people to take the decision. This paper represents the use of Opinion Mining, Natural language Processing and SentiWordNet in this Application in JAVA. This paper includes the Sentiment Analyzer which classifies the sentiment into positive, negative or neutral depending on the polarity. Based on the Scope of text there are three levels of Sentiment polarity categorization namely the document level, the sentence level, and the entity (word) level. Such type of classification can help the customers to get the reviews about the product as the reviews are classified into positive, negative and neutral using sentimental Analysis. Here, ex- mobile phones can be used as the product with features as screen, processors etc. This give a business solution for users and industries for effective product decisions.

Keywords: Sentiment Analysis, Opinion Mining, Natural Language Processing, SentiWordNet

1. INTRODUCTION

Sentiment is an attitude, thought or judgment promoted by feeling. It is also known as Opinion Mining [2], studies people's sentiments towards certain entities. Internet is a resourceful place with respect to sentiment Information. From user's perspective, people are able to post their own content through various social media, such as forums, micro blogs, or online social networking sites. Nowadays e-market have a growing business and have become revolutionary in terms of purchasing goods online. Thus peoples show their reaction and attitude towards the product in their product review. Now, other customers on the basis of previous customers Experience and reviews can think of buying the product on e-market or not.so every customer have a different attitude towards the product.

Sentiment Analysis is a process of describing whether the piece of text is positive, negative or neutral. Sentiment analysis [1] on a product helps the customer to check how many positive and negative reaction have been done on the product. It not only help the consumers to get the opinions about the product but it also help the Company to do the product analytics for the further improvement of the product. Sentiment Analysis uses a part of Natural Language Processing which thus helps in preprocessing of text.

This paper provides a way for Sentiment Analysis using JAVA. Java being an Object Oriented Language provides a better and efficient platform for Sentiment Analysis because of tools and dictionary present. The tools include NLP (Natural Language Processing), SentiWordNet dictionary which contain max number of words. SentiWordNet provides a score for every word present in that dictionary. It also involves extraction of words, text processing, text Categorization, part of Speech tagging, text Classification is required for preprocessing.

2. RESEARCH DESIGN AND METHODOLOGY

Natural Language Processing (NLP)

A formal definition of NLP [3] frequently includes wording to the effect that it is a field of study using computer science, artificial intelligence, and formal linguistics concepts to analyze natural language. A less formal definition suggests that it is a set of tools used to derive meaningful and useful information from natural language sources such as web pages and text documents.

A user query is processed using NLP techniques in order to generate a result page that a user can use. When we work with a language, the terms, syntax, and semantics, are frequently encountered. The syntax of a language refers to the rules that control a valid sentence structure. For example, a common sentence structure in English starts with a subject followed by a verb and then an object such as "Tim hit the ball". We are not used to unusual sentence order such as "Hit ball Tim". Although the rule of syntax for English is not as rigorous as that for computer languages, we still expect a sentence to follow basic syntax rules.

The semantics of a sentence is its meaning. As English speakers, we understand the meaning of the sentence "Tim hit the ball". However, English and other natural languages can be ambiguous at times and a sentence's meaning may only be determined from its context. As various machine learning techniques can be used to attempt to derive the meaning of text. Here we in our Application we use Apache OpenNLP library. The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text. It supports the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing. These tasks are usually required to build more advanced text processing services. The Apache OpenNLP library contains several components, enabling one to build a full natural language processing pipeline.

These components include: sentence detector, tokenizer, name finder, document categorizer, part-of-speech tagger, and parser. Components contain parts which enable one to execute the respective natural language processing task, to train a model and often also to evaluate a model. Each of these facilities is accessible via its application program interface (API).

Use of natural language processing

NLP is ideal for analyzing this type of information. Machine learning and text analysis are used frequently to enhance an application's utility. A brief list of application areas follows:

Searching: This identifies specific elements of text. It can be as simple as finding the occurrence of a name in a document or might involve the use of synonyms and alternate spelling/misspelling to find entries that are close to the original search string.

Machine translation: This typically involves the translation of one natural language into another.

Summation: Paragraphs, articles, documents, or collections of documents may need to be summarized.NLP has been used successfully for this purpose.

Named Entity Recognition (**NER**): This involves extracting names of locations, people, and things from text. Typically, this is used in conjunction with other NLP tasks such as processing queries.

Information grouping: This is an important activity that takes textual data and creates a set of categories that reflect the content of the document. You have probably encountered numerous websites that organize data based on your needs and have categories listed on the left-hand side of the website.

Parts of Speech Tagging (**POS**): In this task,

Text is split up into different grammatical

Elements such as nouns and verbs.

This is useful in analyzing the text further.

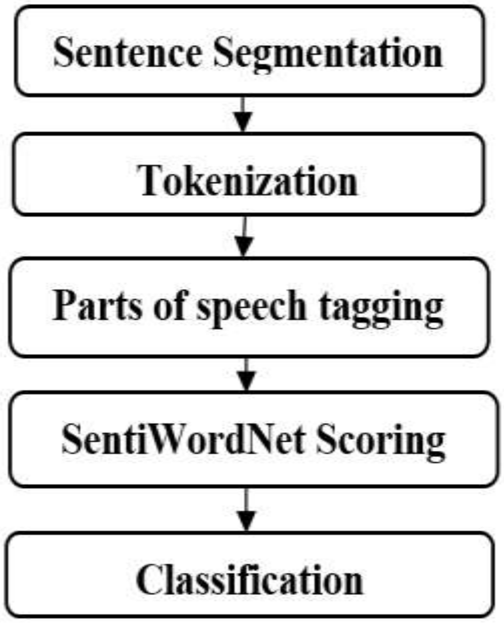


Fig 1: Basic step of Sentence Analysis.

3. COMPONENTS USED

3.1 Components for Data Processing

1. **Sentence Detection:** The OpenNLP Sentence Detector can detect that a punctuation character marks the end of a sentence or not. In this sense a sentence is defined as the longest white space trimmed character sequence between two punctuation marks. The first and last sentence make an exception to this rule. The first non-whitespace character is assumed to be the begin of a sentence, and the last non whitespace character is assumed to be a sentence end.

i. Sentence Detection Tool :

For Sentence Detector the English sentence detector model **en-sent.bin** is used and start the Sentence Detector Tool .The Sentence Detector can be easily integrated into an application via its API. To instantiate the Sentence Detector the sentence model must be loaded first

ii. Sentence Detection API :

The Sentence Detector can be easily integrated into an application via its API. To instantiate the Sentence Detector the sentence model must be loaded first. After the model is loaded the SentenceDetectorME can be instantiated. The Sentence Detector can output an array of Strings, where each String is one sentence.

2. Tokenization :

Tokenization is the process of breaking text down into simpler units. For most text, we are concerned with isolating words. Tokens are split based on a set of delimiters. These delimiters are frequently whitespace characters. The OpenNLP Tokenizer segment an input character sequence into tokens. Tokens are usually words, punctuation, numbers, etc.

EX: Sentence:

Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.

Tokens Output the individual tokens in a whitespace separated representation:

Pierre Vinken , 61 years old , will join the board as a nonexecutive director Nov. 29.

i. Tokenizer Tools: OpenNLP possesses a Tokenizer interface that is implemented by three classes:SimpleTokenizer, TokenizerME, and WhitespaceTokenizer. This interface supports two methods:

Tokenize: This is passed a string to tokenize and returns an array of tokens as strings.

TokenizePos: This is passed a string and returns an array of Span objects. The Span class is used to specify the beginning and ending offsets of the tokens. For Tokenizing the English token model **en-token.bin** is used and start using the Tokenizer Tool.

Tokenizer API:
The Tokenizer can be integrated into an application by the defined API. The shared instance of the WhitespaceTokenizer can be retrieved from a static field WhitespaceTokenizer.INSTANCE. The shared instance of the SimpleTokenizer can be retrieved in the same way from SimpleTokenizer.INSTANCE. To instantiate the TokenizerME (the learnable tokenizer) a Token Model must be created first. The tokenizer offers two tokenize methods, both expect an input String object which contains the untokenized text. If possible it should be a sentence, but depending on the training of the learnable tokenizer this is not required. The first returns an array of Strings, where each String is one token.

3. Tagging : The Part of Speech Tagger marks tokens with their corresponding word type based on the token itself and the context of the token. A token might have multiple pos tags depending on the token and the context. The OpenNLP POS Tagger uses a probability model to predict the correct pos tag out of the tag set. To limit the possible tags for a token a tag dictionary can be used which increases the tagging and runtime performance of the tagger.

POS Tagger Tool: For Tagging the English maxent pos model en-pos-maxent.bin is used and start the POS Tagger Tool.

The POS Tagger now reads a tokenized sentence per line:

Pierre Vinken , 61 years old , will join the board as a nonexecutive director Nov. 29. The POS Tagger will now echo the sentences with pos tags:

JJ	Adjective
NN	Noun, singular or mass
RB	Adverb
VB	Verb, base form
VBD	Verb, past tense

Pierre_NNP Vinken_NNP ,_, 61_CD years_NNS old_JJ ,_, will_MD join_VB the_DT board_NN as_IN a_DT nonexecutive_JJ director_NN Nov._NNP 29_CD ._.The tag set used by the english pos model is the Penn Treebank tag set.

i. **POS Tagger API:** The POS Tagger can be embedded into an application via its API. First the pos model must be loaded into memory from disk or another source. In the sample below it's loaded from disk.

After the model is loaded the POSTaggerME can be instantiated.The POS Tagger instance is now ready to tag data. It expects a tokenized sentence as input, which is represented as a String array, each String object in the array is one token.

The tags array contains one part-of-speech tag for each token in the input array. The corresponding tag can be found at the same index as the token has in the input array.

4. SentiWordNet

SentiWordNet is a lexical resource for opinion mining. It is based on word net synsets. Each synset is assigned with three scores: positivity, negativity, and objectivity. SentiWordNet Contains English nouns, verbs, adverbs, adjectives which are also called as Synsets. Sentiment Analysis of a phrase, text may require detecting and processing sentiments. If sense the word measure the relatedness to others, define & describe their meaning.

It consist of synset can be associated with a value concerning the negative or positive or neutral. SentiWordNet consist of synset i.e positive or negative synset consist of positive or negative score. the process is running iteratively on the wordnet. SentiWordNet is a lexical resource in which each entry is a set of lemma-PoS pairs sharing the same meaning, called "synset". Each synset associated with the numerical scores Pos(s) and Neg(s), which range from 0 to 1.

5. To Calculate the polarity

Obtaining the prior-polarity sentiment scores of adjectives, adverbs and verbs: To determine the sentiment scores of the extracted adjectives, adverbs and nouns, it uses a sentiment-based lexicon use WordNet as lexicon. WordNet is a lexical resource for sentiment analysis. It assigns to each synset of WordNet three sentiment scores: positivity, negativity, and objectivity.

It has been used as the lexicon in sentiment classification studies. WordNet is determining the polarities of adjectives which then lead to the document polarities for multilingual sentiment analysis. Since each word in WordNet has multiple senses, then calculate the average polarity scores (i.e., positive, negative, and objective scores) for its adjective, adverb, and verb senses separately.

4. Formulae's Used to calculate

Total positive= totalPOS*2;

Total negative= totalNeg*2;

Positive Accuracy= totalPOS/(totalPOS+ totalNeg);

Negative Accuracy= totalNeg/(totalPOS+ totalNeg);

Documents Processed=(totalPOS+totalNeg)*2;

For positive->

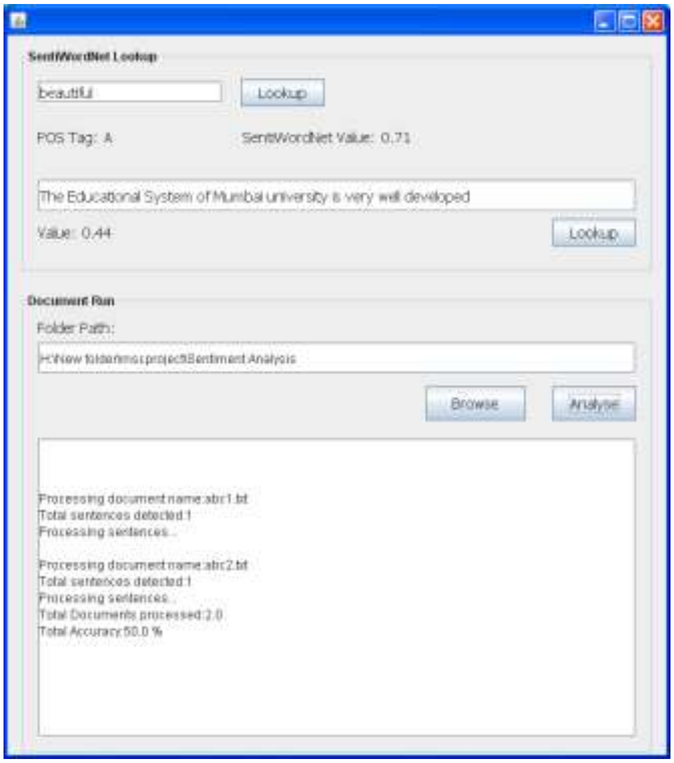
Accuracy= Positive Accuracy/2;

For Negative->

Accuracy= Negative Accuracy/2;

Total Accuracy=Accuracy*100;

Positivness



beautiful

POS Tag: A SentiWordNet Value: 0.71

The Educational System of Mumbai university is very well developed

Value: 0.44

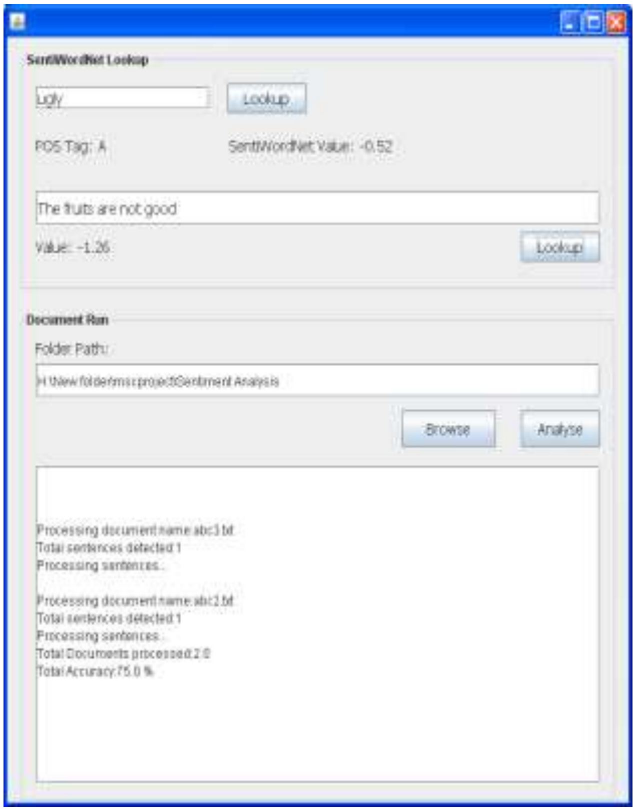
Folder Path:

H:\New folder\msc project\Sentiment Analysis

```
Processing document name:abc1.txt
Total sentences detected:1
Processing sentences...

Processing document name:abc2.txt
Total sentences detected:1
Processing sentences...
Total Documents processed:2.0
Total Accuracy:50.0 %
```

Negativness



Ugly

POS Tag: A SentiWordNet Value: -0.52

The fruits are not good

Value: -1.26

Folder Path:

H:\New folder\msc project\Sentiment Analysis

```
Processing document name:abc3.txt
Total sentences detected:1
Processing sentences...

Processing document name:abc2.txt
Total sentences detected:1
Processing sentences...
Total Documents processed:2.0
Total Accuracy:75.0 %
```

5. CONCLUSIONS

Opinion miming is an emerging field of data mining used to extract the pearl knowledge from huge volume of customer comments, feedback and reviews on any product or topic etc. A lot of work has been conducted to mine opinions in form of document, sentence and feature level sentiment analysis It is examined that now opinion mining trend is moving to the sentimental reviews of twitter data, comments used in Facebook on pictures, videos or Facebook status.

In future, Opinion Mining can be carried out on a set of reviews and set of discovered feature expressions extracted from reviews. The natural language text can be processed based on machine learning toolkit called as OpenNLP library. The advanced text processing services are built using these tasks. OpenNLP also includes perceptron and maximum entropy based machine learning. After POS tagging, opinion retrieval can be performed by extracting product candidate feature, related opinion and producing opinion feature pairs.

The keywords extracted from Opinion Retrieval. Module can be used to perform similarity check with the database dictionary. The similarity check can use semi supervised learning

6. REFERENCES

[1]https://www.academia.edu/39771248/SENTIMENT_ANALYSIS_ON_PRODUCT_FEATURES_BASED_ON_LEXICON_APPROACH_USING_NATURAL_LANGUAGE_PROCESSING

[2]https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0015-2

[3]https://books.google.co.in/books?id=q7y4BwAAQBAJ&lpg=PA2&ots=fIxSgApWhw&dq=%22The%20syntax%20of%20a%20language%20refers%20to%20the%20rules%20that%20control%20a%20valid%20sentence%20structure.For%20example%2C%20a%20common%20sentence%20structure%20in%20English%20starts%20with%20a%20subject%20followed%20by%20a%22&pg=PA2#v=onepage&q=%22The%20syntax%20of%20a%20language%20refers%20to%20the%20rules%20that%20control%20a%20valid%20sentence%20structure.For%20example,%20a%20common%20sentence%20structure%20in%20English%20starts%20with%20a%20subject%20followed%20by%20a%22&f=false

[4]https://en.wikipedia.org/wiki/Apache_OpenNLP

7. BIOGRAPHIES

Miss. Neha Hasanmiya Surve
Asst Professor, Department of Information Technology, DBJ College, Chiplun-415605, Maharashtra, India.