Term frequency Inverse Document for Advanced and efficient "Opinion Data Mining"

Yathiraj GR[*1], T D Roopamala[*2]

[*1]Department of Computer Science and Engineering, Sri Jaya Chamarajendra college of Engineering, Mysuru, Karnataka, India-570006

[*2]Department of Computer Science and Engineering, Sri Jaya Chamarajendra college of Engineering, Mysuru, Karnataka, India-570006

Abstract

Opinion data mining is an evolving field in the current generation. With the digital age, there's a vast majority of the organizations interested in getting insights into the collected data. The Opinion Data Mining allows one to easily categorize and predict things with ease. Companies like Facebook, Amazon, Flipkart gathering data perform a heavy analysis and apply machine learning and artificial intelligence to study the behaviors of the customers why they like or dislike certain products and push related products into the market and target reaching the know customers who can buy or purchase. You will be seeing next to the preferred or efficient ways of opinion data mining that will help the organizations in getting the insights from the massive data gathered over the period.

Keywords: Data Mining, Sentiment Analysis, Opinion Mining, Machine Learning

DOI Number: 10.14704/nq.2022.20.11.NQ66187 **NeuroQuantology 2022; 20(11): 1928-1932**

I. What is Opinion Data Mining?

Opinion data mining is the science or art of analyzing the data and then come up with the mechanism to understand what drives or motivates the consumers in doing something. What are their sentiments, it could be positive, negative, or neutral? The opinion mining allows the organizations in gathering all sorts of valuable insights so one can predict things with ease. With the technology advancing and the data that's being gathered over the period, it's natural for organizations to tap into the consumer data and get a detailed insight so they can be prepared or make some corrections or improvements in their process or product.

II. Challenges with the data

One of the most common challenges of the data is due to its highly unstructured format. There are multiple ways in which the organizations are gathering the information say with the help of reviews, feedbacks, etc. However, as you can see none of these are in a structured format where one can apply some rules and get some insights.

The unstructured data comes in various formats like Text, Pictures, Audio, Video, etc. Data collection is one thing and the future performing the detailed analysis and classifying the same and predicting involves a series of steps to derive or conclude at certain aspects.

eISSN 1303-5150

III. Opinion Data Mining at High-Level

Here are the high-level steps or processes of deriving at the opinion of data mining. The first and foremost thing that's required is the "data" without the raw data, we won't be able to do anything.

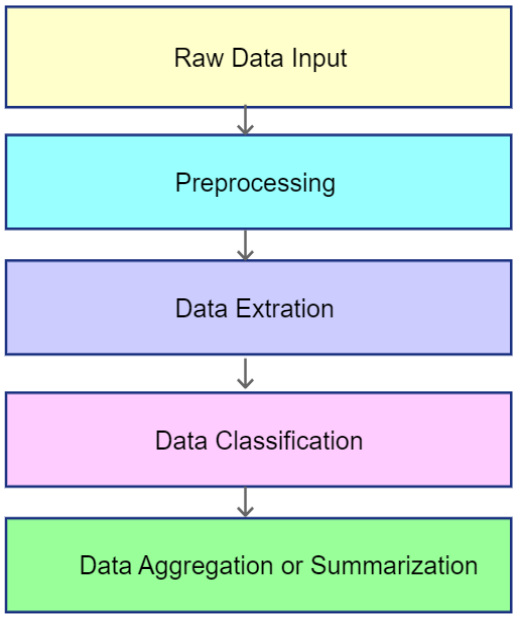


Figure 1: High-Level Diagram

1) Gather all the raw data. It could be in any format. But we are going to focus much on the textual input.
2) The preprocessing is where the raw data is tokenized and then a stop word processing is done to remove all the stop words. Further, perform stemming, etc.
3) In the Data Extraction phase, for each of the tokenized data, Shallow chunking is done to identify whether or not a token is a noun group, verbs, verb groups, etc. The parts of speech tagging are done.
4) In the Data Classification phase is where the supervised machine learning is applied to understand the consumer sentiments and identify various patterns to get more insights into the data.
5) The last phase is the Data Summarization where the summary of the sentences is done. This will in turn help one to get more insights with the metadata surrounding it. A more detailed Opinion summarization happens here.

Sentence Extractor

The sentence extractor is the process of extracting a valid sentence from a given chunk of raw data. We call it a "Sentenizer" due to the nature of sentence extraction by making use of the advanced machine learning model. The sentence extractor removes all the unwanted sentences that need to be furtheranalyzed.

Sentence Tokenizer

Tokenization is the process of breaking down the sentences into small chunks or segments. The following aspects will be performed as part of the tokenization process.

1) Convert all text to lowercase.
2) The trailing and leading spaces will be removed.
3) Clean up all the punctuations if any.
4) Strip all the special characters.
5) Filter from the ignored list of words.

Stop word Removal

The stop words are the most common words such as "a", "an", "in", "the" etc. that occur in a given text or a sentence(s). The stop words are removed or cleaned up by looking up against the bag of stop words. Once we do the word tokenization, the stops words removal process is then applied to the tokenized sentences to get the clean filtered list of words for data mining. [1]

Getting the word frequency

The word frequency is evaluated using the input text, tokenizer, and with the help of the stop provider which is responsible for removing all the stop words in a given text.

eISSN 1303-5150

1. Tokenize the given text. The Tokenization process will return to return the collection of tokens.
2. Get unique words.
3. For each of the unique words, if it's a word and not a stop word.
4. Get the word count.

Associate the word count with the unique text.

Parts of Speech Tagging

The part of speech is the process of labeling each word in a given sentence with the relevant part of speech with say – Noun, Pronoun, Verb, Adverb, Adjective, etc. Penn Treebank Project is the best reference for evaluating the parts of speech.[3]

Tag	Description	Example	Tag	Description	Example	Tag	Description	Example
CC	coordinating conjunction	*and, but, or*	PDT	predeterminer	*all, both*	VBP	verb non-3sg present	*eat*
CD	cardinal number	*one, two*	POS	possessive ending	*'s*	VBZ	verb 3sg pres	*eats*
DT	determiner	*a, the*	PRP	personal pronoun	*I, you, he*	WDT	wh-determ.	*which, that*
EX	existential 'there'	*there*	PRP$	possess. pronoun	*your, one's*	WP	wh-pronoun	*what, who*
FW	foreign word	*mea culpa*	RB	adverb	*quickly*	WP$	wh-possess.	*whose*
IN	preposition/ subordin-conj	*of, in, by*	RBR	comparative adverb	*faster*	WRB	wh-adverb	*how, where*
JJ	adjective	*yellow*	RBS	superlatv. adverb	*fastest*	$	dollar sign	*$*
JJR	comparative adj	*bigger*	RP	particle	*up, off*	#	pound sign	*#*
JJS	superlative adj	*wildest*	SYM	symbol	*+,%, &*	"	left quote	' or "
LS	list item marker	*1, 2, One*	TO	"to"	*to*	"	right quote	' or "
MD	modal	*can, should*	UH	interjection	*ah, oops*	(	left paren	[, (, {, <
NN	sing or mass noun	*llama*	VB	verb base form	*eat*	)	right paren	],), }, >
NNS	noun, plural	*llamas*	VBD	verb past tense	*ate*	,	comma	,
NNP	proper noun, sing.	*IBM*	VBG	verb gerund	*eating*	.	sent-end punc	. ! ?
NNPS	proper noun, plu.	*Carolinas*	VBN	verb past part.	*eaten*	:	sent-mid punc	: ; ... – -

Penn Treebank POS Tag Set [2]

A target dataset is taken into consideration for POS tagging. There are two main approaches to building POS tagged sentences. Rule-based and Probabilistic Taggers.

Data Classification Phase

The data classification phase is where the Supervised Machine Learning algorithms are utilized for classifying the given sentences and get insights on the sentiments. The sentiment analysis works best when we have a large corpus of trained data. An Evidence-based mechanism is used for classifying the dataset and predict the percentage of positive and negative scores.

The supervised machine learning for sentiment analysis requires a dedicated set of Positive and Negative keys words with a specific weight assigned to it.

Here's the high-level algorithm of the "Opinion Mining" Classifier.

1. Build the Evidence for the Positive Dataset.
2. Build the Evidence for the Negative Dataset.
3. Read the test dataset for evaluating the scores for the Positive and Negative.

The classifier takes Positive and Negative Evidence for building positive and negative scores.

IV. Opinion Scoring and feature detection using HAC

The opinion scoring is evaluated using the **High Adjective Count Algorithm** (HAC) Algorithm [4].

One may gain a thorough grasp of user opinions by using opinion mining with scoring. provide additional information about the facts, user purpose, and user feelings regarding many areas. The automated feature collection process made possible by opinion scoring aids in the development of machine learning models.

1. Set a threshold of opinion score for building the features.
2. Get the raw dataset for building the opinion score.
3. Perform the data cleaning and the necessary preprocessing to remove all stop words.
4. Apply the stemming process.
5. Perform the tokenization on the cleaned data set.
6. Perform the Parts of Speech (POS) on the tokenized data.
7. Get all the adjectives from the POS.

eISSN 1303-5150

8. Get all the closest noun for each of the adjectives in Step 6.
9. Get the unique nouns
10. Create a hash map of all of the nouns and initialize the value to 0.
11. Loop through the adjectives and for each of the nearest noun, increment the hash mapped noun count by 1.
12. Filter all the hash mapped nouns with a value greater than or equal to the threshold value.
13. The threshold filtered nouns with their adjectives describe the potential features.

V Term Frequency Inverse Document

paper with a frequency-inverse Frequency is typically used to assess or comprehend a word's importance relative to other words in a corpus or collection of words. Occasionally, a term will appear more than once in a document. Therefore, such words may have more significance than others. Even now, the tf-idf is one of the most well-known word weighting algorithms.

* tf(t)= the term frequency is the number of times the term appears in the document
* idf(d, t) = the document frequency is the number of documents 'd' that contain term 't'
* tf(t,d) = (Number of times term t appears in a document) / (Total number of terms in the document)

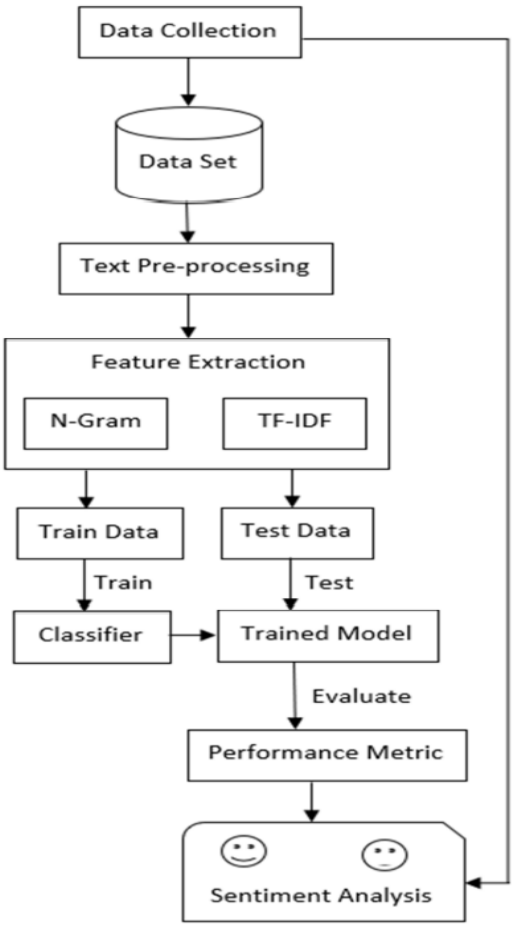
t = term
d = document
idf(t) = log [n / df(t)] + 1
idf(t) - Inverse Document Frequency
n - Total number of documents
df(t) is the document frequency of term t;
tf-idf(t, d) = tf(t, d) * idf(t)

Here is one instance.

Let's imagine there are 100 words in a paper, and five of those words are the term "dog." Therefore, (5 / 100) = 0.05 is the term frequency (i.e., tf) for "dog". Let's say we have 10 million papers, and thousands of them include the phrase "dog." Then, log(10,000,000/1,000) = 4 is used to determine the inverse document frequency (idf).The Tf-idf weight is the product of these quantities: 0.05 * 4 = 0.2.

The TFIDF algorithm has several applications, including searching, text summarization, and feature extraction.

The method of doing sentiment analysis or opinion mining using the TF-IDF based supervised learning based trained model is illustrated in the following flowchart.



eISSN 1303-5150

Below is the high-level algorithm for automatic summary generation using the TF-IDF and word frequency algorithm.

1. Tokenize the Sentences.
2. Create the frequency matrix of the words in each sentence.
3. Calculate Term Frequency and build a matrix.
4. Create a table of documents per words.
5. Perform the IDF calculation and build the matrix.
6. Perform the TF-IDF calculation and build the matrix.
7. Calculate the sentence score.
8. Find the threshold based on the average sentence score.
9. Generate or build the text summary.

References

[1] M.F. Federico, L.L. Pier, Mining interesting knowledge from weblog: a survey, J. Data Knowledge Eng. 53(2005) (2005) 225–241.

[2]S.Kaviarasan,K.Hemapriya,K.Gopinath,Semantic Web Usage Mining Techniques for Predicting Users' Navigation Requests, International Journal of Innovative Research in Computer Science and Communication Engineering,Vol. 3,Issue 5,[ISSN:2320-9801],2015.

[3]B.Lalithadevi, A. Mary Ida,W.Ancy Breen, A New Approach for Improving World Wide Web Techniques in Data Mining, International Journal of Advanced Research in Computer Science and Software Engineering,Vol. 3,Issue 1,[ISSN:2277 128X],2013

[4]Paul, N. Kenta, Better Prediction of Protein Cellular Localization Sites with the K-Nearest Neighbor Classifier, ISMB-97, Proceeding of America Association for Artificial Intelligence, USA, 1997, pp. 147–152.

[5]Qing Yang, Haining Henery Zhang, Web log mining for predictive web caching, IEEE Transactions on Knowledge and Data Engineering, Vol. 15 NO. 4. [104 1-4347/03/$17.00],2003.

[6] Stop word removal related
https://www.geeksforgeeks.org/removing-stop-words-nltk-python/
[7] Penn Treebank POS
https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
[8] Parts of Speech Tagging
https://www.nltk.org/book/ch05.html
[9] Sentiment Analysis and Feature based opinion mining
https://ijarcce.com/wp-content/uploads/2016/10/IJARCCE-68.pdf

eISSN 1303-5150