MINI PROJECT
(2020-21)

SENTIMENTAL ANALYSIS USING SCIKIT LEARN

MID-TERM REPORT

Institute of Engineering & Technology

Submitted by
Akanksha Mishra
(181500051)
Aman Kumar
(181500073)

Supervised By: -
Mr. Piyush Vasistha
Asst. Professor
Department of Computer Engineering & Applications

Contents

Abstract

Sentiment analysis or opinion mining is the computational study of people’s opinions,sentiments,attitudes, and emotions expressed in written language. It is one of the most active research areas in natural language processing and text mining in recent years. Its popularity is mainly due to two reasons. First ,it has a wide range of applications because opinions are central to almost all human activities and are key influences of our behaviors. Whenever we need to make a decision , we want to hear others' opinions. Second ,it presents many challenging research problems, which had never been attempted before the year 2000.Part of the reason for the lack of study before was that there was little opinionated text in digital forms. It is thus no surprise that the inception and rapid growth of the field coincide with those of the social media on the Web.In fact,the research has also spread outside of computer science to management sciences and social science due to its importance to business and society as a whole. In this talk,I will start with the discussion of the mainstream sentiment analysis research and then move on to describe some recent work on modeling comments, discussions, and debates, which represents another kind of analysis of sentiments and opinions.

Sentiment classification is a way to analyze the subjective information in the text and then mine opinion . Sentiment analysis is the procedure by which information is extracted from the opinions, appraisals and emotions of people in regards to entities,events and their attributes. In decision making,the opinions of others have a significant effect on customer ease, making choices with regards to online shopping ,choosing events, products,entities. The approaches of text sentiment analysis typically work at a particular level like phrase,sentence or document level. This paper aims at analyzing a solution for the sentiment classification at a fine-grained level, namely the sentence level in which polarity of the sentence can be given by three categories as positive , negative and neutral.

Introduction

Sentiment Analysis refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials.Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document.The attitude may his or her judgements or evaluation affective state, or the intended emotional communication. Sentiment analysis is the process of detecting a piece of writing for positive,negative or neutral feelings bound to it.Humans have the innate ability to determine sentiment; however,this process is time consuming,inconsistent, and costly in a business context. It’s just not realistic to have people individually read tens of thousands of user customer reviews and score them for sentiment .

For example if we consider Semantria’s cloud based sentiment analysis software . Semantrias’s cloud based sentiment analysis software extracts the sentiment analysis software extracts the sentiment of a document and its component through the following steps:

- A document is broken in its basic parts of speech, called POS tags, which identify the structural elements of a document, paragraph, or sentence (ie Nouns, adjectives, verbs, and adverbs) .
- Sentiment-bearing phrases, such as “terrible service”, are identified through the use of specifically designed algorithms .
- Each sentiment-bearing phrase in a document is given a score based on a logarithmic scale that ranges between -10 and 10 .
- Finally, the scores are combined to determine the overall sentiment of the document or sentence Document scores range between -2 and 2 .

Semantria’s cloud-based sentiment analysis software is based on Natural Language Processing and delivers you more consistent results than two humans. Using automated sentiment analysis, Semantria analyzes each document and its components based on sophisticated algorithms developed to extract sentiment from your content in a similar manner as a human – only 60,000 times faster.

Existing approaches to sentiment analysis can be grouped into three main categories:

- Keyword spotting
- Lexical affinity
- Statistical methods

Keyword spotting is the most naive approach and probably also the most popular because of its accessibility and economy .Text is classified into effect categories based on the presence of fairly unambiguous affect words like ‘happy’, ‘sad’, ‘afraid’, and ‘bored’ .The weaknesses of this approach lie in two areas: poor recognition of affect when negation is involved and reliance on surface features .About its first weakness, while the approach can correctly classify the sentence “today was a happy day” as being happy, it is likely to fail on a sentence like “today wasn’t a happy day at all” About its second weakness, the approach relies on the presence of obvious affect words that are only surface features of the prose .

In practice, a lot of sentences convey affect through underlying meaning rather than affect adjectives For example, the text “My husband just filed for divorce and he wants to take custody of my children away from me” certainly evokes strong emotions, but uses no effect keywords, and therefore, cannot be classified using a keyword spotting approach .

Lexical affinity is slightly more sophisticated than keyword spotting as, rather than simply detecting obvious affect words, it assigns arbitrary words a probabilistic ‘affinity’ for a particular emotion For example, ‘accident’ might be assigned a 75% probability of being indicating a negative affect, as in ‘car accident’ or ‘hurt by accident’ These probabilities are usually trained from linguistic corpora .Though often outperforming pure keyword spotting, there are two main problems with the approach First, lexical affinity, operating solely on the word-level, can easily be tricked by sentences like “I avoided an accident” (negation) and “I met my girlfriend by accident” (other word senses) Second, lexical affinity probabilities are often biased toward text of a particular genre, dictated by the source of the linguistic corpora This makes it difficult to develop a reusable, domain-independent model .

Statistical methods, such as Bayesian inference and support vector machines, have been popular for affect classification of texts .By feeding a machine learning algorithm a large training corpus of affectively annotated texts, it is possible for the system to not only learn the affective valence of affect keywords (as in the keyword spotting approach), but also to take into account the valence of other arbitrary keywords (like lexical affinity), punctuation, and word co-occurrence frequencies. However, traditional statistical methods are generally semantically weak, meaning that, with the exception of obvious effect keywords, other lexical or co-occurrence elements in a statistical model have little predictive value individually .As a result, statistical text classifiers only work with acceptable accuracy when given a sufficiently large text input .So, while these methods may be able to effectively classify user's text on the page- or paragraph- level, they do not work well on smaller text units such as sentences or clauses .

- **Application of Sentimental Analysis**

1.) Online Services
2.) Fashion
3.) Health & Food
4.) Electronics
5.) Travel
6.) Entertainments

1.3 Hardware Requirements

- Memory [4GB RAM (or higher)]
- Intel core i3 64-bit Processor (or higher)

1.3 Software requirements

- Any OS
- Python and its libraries

Objective

Sentiment Classification is a way to analyze the subjective information in the text and then mine the opinion. Sentiment analysis is the procedure by which information is extracted from the opinions , appraisals and emotions of people in regards to entities, events and their attributes. In decision making, the opinions of others have a significant effect on customer ease, making choices with regards to online shopping , choosing events,products ,entities.

Implementation Details

Naive Bayes:

This method mainly concentrates on the attributes Training Phase involves the Elimination of Special Characters and Conversion to Lower case and Word Count stages are used in this method .The DataSet obtained from the WordCount is passed to another stage where all the neutral words are eliminated by using Positive and Negative words .Finally the obtained DataSet is given as a input to the Naïve Bayes method and along with that Sentiment polarities are calculated carefully and given as input.

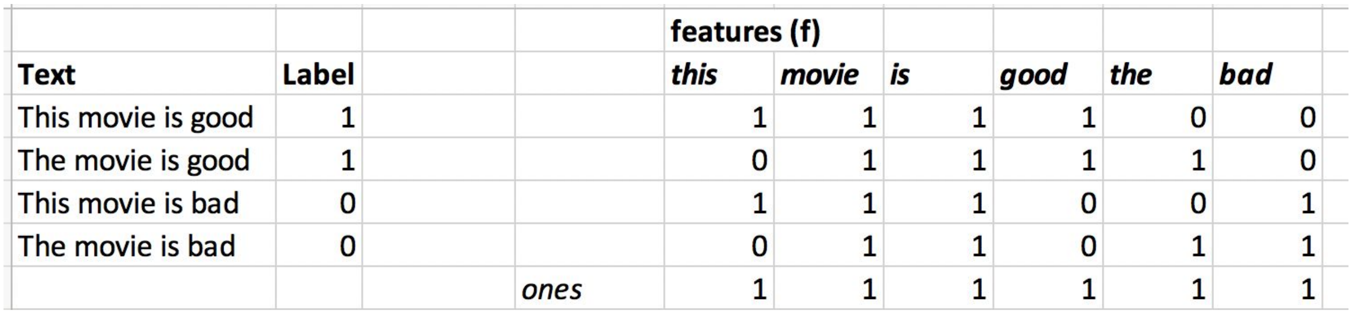
In the testing phase the test data is passed through the Elimination of Special Characters and Conversion to Lowercase stage .The prior,conditional and posterior probabilities are calculated using the input data .

Bag of Words:

The Elimination of Special Characters and Conversion to Lowercase and Word Count stages are used in this method .The DataSet obtained from the WordCount is passed another stage where all the neutral words are eliminated by using Positive and Negative words.Finally the obtained DataSet is given as a input to the Bag of Words method .

In the Testing Phase, the Special characters and upper case letters are eliminated from the test data . In this method to eliminate the Zero probability problem each word is considered as repeated once .The prior,conditional and posterior probabilities are calculated using the input data .

SCREENSHOTS



				features (f)					
Text	**Label**			***this***	***movie***	***is***	***good***	***the***	***bad***
This movie is good	1			1	1	1	1	0	0
The movie is good	1			0	1	1	1	1	0
This movie is bad	0			1	1	1	0	0	1
The movie is bad	0			0	1	1	0	1	1
			ones	1	1	1	1	1	1

References

- https://www.courseera.com

- https://www.google.com

- https://www.kaggle.com