BILINGUAL SENTIMENT ANALYSIS USING NATURAL LANGUAGE PROCESSING

Indrajeet Patel, Research Scholar, Rewa Institute of technology, India
Parikshit Tiwari, Head of Department, Computer Science & Engineering, Rewa Institute of technology, India

Abstract:

The world is moving from manually operated system towards automation. Sentimental Analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral. Sentiment analysis (sometimes known as opinion mining or emotion AI) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.

Generally speaking, sentiment analysis aims to determine the attitude of a speaker, writer, or other subject with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event. The attitude may be a judgment or evaluation (see appraisal theory), affective state (that is to say, the emotional state of the author or speaker), or the intended emotional communication (that is to say, the emotional effect intended by the author or interlocutor).

I. INTRODUCTION

Natural language processing (NLP) is a field of computer science, artificial intelligence and computational linguistics concerned with the interactions between computers and human (natural) languages, and, in particular, concerned with programming computers to fruitfully process large natural language corpora.

The ultimate growth of any business requires the satisfaction level of customers. Any business entity whether it's based on product based or service requires the interaction of human at some level. It may include the collection customer of customer feedback or their experiences. This whole process is actually adds the major cost of the business operations moreover this is something which cannot be easily replaced with the computer system.

Sentiment analysis, also known as opinion mining, is the analysis of the feelings (i.e. attitudes, emotions and opinions) behind the words using natural language processing tools. Sentiment Analysis Algorithms are used in these tools to find out the relevant information related to sentiments. Sentiment Analysis is the use of natural language processing, statistics, and text analysis to extract, and identify the sentiment of text into positive, negative, or neutral categories.

The NLP Algorithm actually Identify and extract sentiment in given string. Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. This algorithm takes an input string and assigns a sentiment rating in the range [-1 t0 1] (very negative to very positive).

II. COMPONENTS OF NATURAL LANGUAGE PROCESSING

There are two components of NLP as given –

1. **Natural Language Understanding (NLU):**

Understanding NLU involves the following tasks

a. Mapping the given input in natural language into useful representations.
b. Analyzing different aspects of the language.

II. **Natural Language Generation (NLG):**

It is the process of producing meaningful phrases and sentences in the form of natural language from some internal representation.

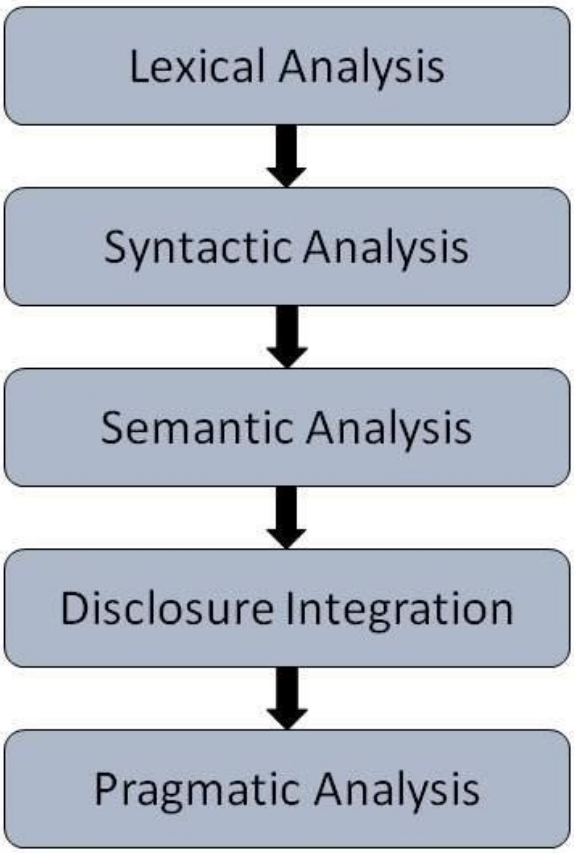
It involves −

a. Text planning − It includes retrieving the relevant content from knowledge base.
b. Sentence planning − It includes choosing required words, forming meaningful phrases, setting tone of the sentence.
c. Text Realization − It is mapping sentence plan into sentence structure.

The NLU is harder than NLG.

III. STEPS IN NATURAL LANGUAGE PROCESSING

There are general five steps –



Lexical Analysis − It involves identifying and analyzing the structure of words. Lexicon of a language means the collection of words and phrases in a language. Lexical analysis is dividing the whole chunk of txt into paragraphs, sentences, and words.

Syntactic Analysis (Parsing) − It involves analysis of words in the sentence for grammar and arranging words in a manner that shows the relationship among the words. The sentence such as "The school goes to boy" is rejected by English syntactic analyzer.

Semantic Analysis − It draws the exact meaning or the dictionary meaning from the text. The text is checked for meaningfulness. It is done by mapping syntactic structures and objects in the task domain. The semantic analyzer disregards sentence such as "hot ice-cream".

Discourse Integration − The meaning of any sentence depends upon the meaning of the sentence just before it. In addition, it also brings about the meaning of immediately succeeding sentence.

Pragmatic Analysis − During this, what was said is re-interpreted on what it actually meant. It involves deriving those aspects of language which require real world knowledge.

IV. SENTIMENTAL ANALYSIS PROCESS:

Sentiment Analysis identifies the sentiment expressed in a text then analyzes it. Therefore, the target of SA is to find opinions, identify the sentiments they express, and then classify their polarity as shown in figure 1.

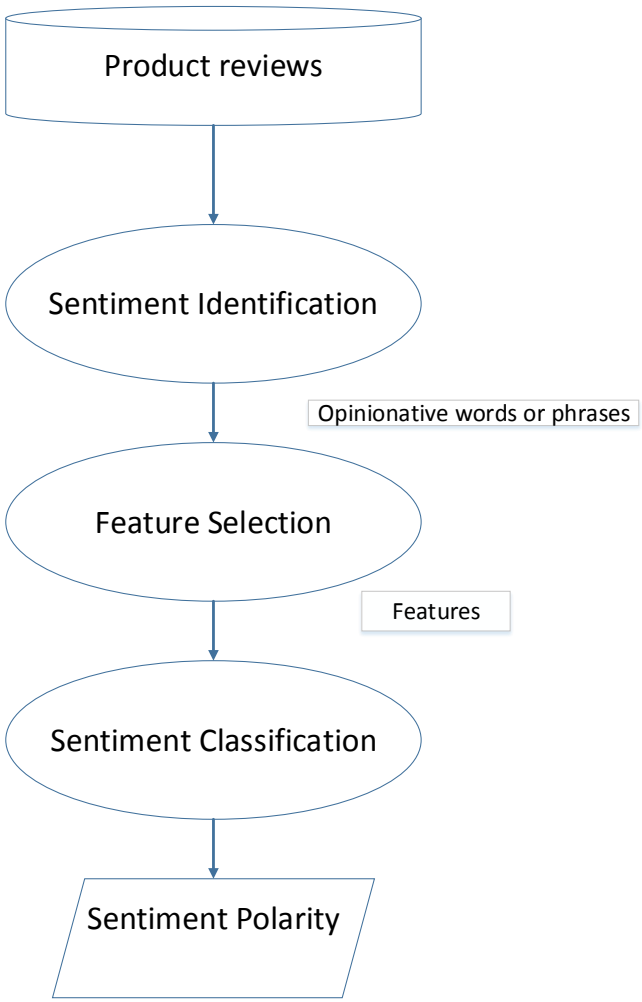


Figure 1

The Sentimental Analysis (SA) can be broadly classified into 3 main categories.

1. Lexical Based Approach
2. Machine Based Approach
3. Hybrid Approach

There are many applications and enhancements on SA algorithms that were proposed in the last few years. This survey aims to give a closer look on these enhancements and to summarize and categorize some articles presented in this field according to the various SA techniques.

We have discussed the Feature Selection (FS) techniques

and their classification which includes all types of approaches including Artificial Intelligence Algorithm, Statically Modeling and Rule Based Approaches. The Sentiment Classification (SC) Algorithms are shown in Figure 2.

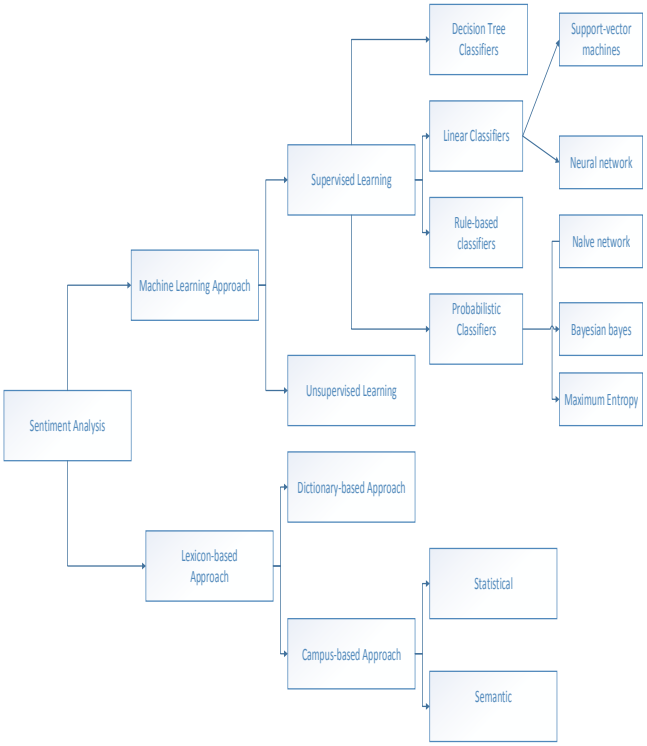


Figure 2

This survey uniquely gives a refined categorization to the various SA techniques

It discusses also new related fields in SA which have attracted the researchers lately and their corresponding articles. These fields include Emotion Detection (ED), Building Resources (BR) and Transfer Learning (TL). Emotion detection aims to extract and analyze emotions, while the emotions could be explicit or implicit in the sentences. Transfer learning or Cross-Domain classification is concerned with analyzing data from one domain and then using the results in a target domain. Building Resources aims at creating lexica, corpora in which opinion expressions are annotated according to their polarity, and sometimes dictionaries. In this paper, we have summarized authors a closer look on these fields.

V. Feature Selection in Sentiment Classification:

Sentiment Analysis task is considered a sentiment classification problem. The first step in the SC problem is to extract and select text features. Some of the current features are

***Terms presence and frequency*:** These features are individual words or word n-grams and their frequency counts. It either gives the words binary weighting (zero if the word appears, or one if otherwise) or uses term frequency weights to indicate the relative importance of features.

***Parts of speech* (*POS*):** finding adjectives, as they are important indicators of opinions.

***Opinion words and phrases*:** these are words commonly used to express opinions including *good or bad*, *like or hate*. On the other hand, some phrases express opinions without using opinion words. For example: *cost me an arm and a leg*.

***Negations*:** the appearance of negative words may change the opinion orientation like *not good* is equivalent to *bad*.

VI. Feature Selection Methods:

Feature Selection methods can be divided into lexicon-based methods that need human annotation, and statistical methods which are automatic methods that are more frequently used. Lexicon-based approaches usually begin with a small set of 'seed' words. Then they bootstrap this set through synonym detection or on-line resources to obtain a larger lexicon. Statistical approaches, on the other hand, are fully automatic.

The feature selection techniques treat the documents either as group of words (Bag of Words (BOWs)), or as a string which retains the sequence of words in the document. BOW is used more often because of its simplicity for the classification process. The most common feature selection step is the removal of stop-words and stemming (returning the word to its stem or root i.e. flies → fly).

In the next subsections, we present three of the most frequently used statistical methods in FS and their related articles. There are other methods used in FS like information gain and Gini index.

1. Point-wise Mutual Information (PMI)

The mutual information measure provides a formal way to model the mutual information between the features and the classes. This measure was derived from the information theory. The point-wise mutual information (PMI) $M_i(w)$ between the word w and the class i is defined on the basis of the level of co-occurrence between the class i and word w. The expected co-occurrence of class i and word w, on the basis of mutual independence, is given by $\boldsymbol{P_i \,.\, F(w)}$, and the true co-occurrence is given by $\boldsymbol{F(w) \,.\, P_i(w)}$.

The mutual information is defined in terms of the ratio between these two values and is given by the following equation:

$M_i(w) = \log(F(w) \cdot p_i(w)F(w)\cdot P_i) = \log(p_i(w)P_i)$

The word *w* is positively correlated to the class *i*, when $M_i(w)$ is greater than 0. The word *w* is negatively correlated to the class *i* when $M_i(w)$ is less than 0.

PMI is used in many applications, and there are some enhancements applied to it. PMI considers only the co-occurrence strength. Y_u and W_u have extended the basic PMI by developing a contextual entropy model to expand a set of seed words generated from a small corpus of stock market news articles. Their contextual entropy model measures the similarity between two words by comparing their contextual distributions using an entropy measure, allowing for the discovery of words similar to the seed words. Once the seed words have been expanded, both the seed words and expanded words are used to classify the sentiment of the news articles. Their results showed that their method can discover more useful emotion words, and their corresponding intensity improves their classification performance. Their method outperformed the (PMI)-based expansion methods as they consider both co-occurrence strength and contextual distribution, thus acquiring more useful emotion words and fewer noisy words.

2. Chi-square ($\chi 2$)

Let *n* be the total number of documents in the collection, $p_i(w)$ be the conditional probability of class *i* for documents which contain *w*, P_i be the global fraction of documents containing the class *i*, and $F(w)$ be the global fraction of documents which contain the word *w*. Therefore, the $\chi 2$-statistic of the word between word *w* and class *i* is defined as

$$\chi^{i2}=n\cdot F(w)2 \cdot (p_i(w)-P_i)2F(w)\cdot(1-F(w))\cdot P_i\cdot(1-P_i)$$

χ^2 and PMI are two different ways of measuring the correlation between terms and categories. χ^2 is better than PMI as it is a normalized value; therefore, these values are more comparable across terms in the same category.

χ^2 is used in many applications; one of them is the contextual advertising as presented by Fan and Chang . They discovered bloggers' immediate personal interests in order to improve online contextual advertising. They worked on real ads and actual blog pages from ebay.com, wikipedia.com and epinions.com. They used SVM (illustrated with details in the next section) for classification and χ^2 for FS. Their results showed that their method could effectively identify those ads that are positively-correlated with a blogger's personal interests.

3. *Latent Semantic Indexing (LSI)*

Feature selection methods attempt to reduce the dimensionality of the data by picking from the original set of attributes. Feature transformation methods create a smaller set of features as a function of the original set of features. LSI is one of the famous feature transformation methods. LSI method transforms the text space to a new axis system which is a linear combination of the original word features. Principal Component Analysis techniques (PCA) are used to achieve this goal. It determines the axis-system which retains the greatest level of information about the variations in the underlying attribute values. The main disadvantage of LSI is that it is an unsupervised technique which is blind to the underlying class-distribution. Therefore, the features found by LSI are not necessarily the directions along which the class-distribution of the underlying documents can be best separated.

VII. SENTIMENT CLASSIFICATION TECHNIQUES:

Sentiment Classification techniques can be roughly divided into machine learning approach, lexicon based approach and hybrid approach. The *Machine Learning Approach* (*ML*) applies the famous ML algorithms and uses linguistic features. The *Lexicon-based Approach* relies on a sentiment lexicon, a collection of known and precompiled sentiment terms. It is divided into dictionary-based approach and corpus-based approach which use statistical or semantic methods to find sentiment polarity. The *hybrid Approach* combines both approaches and is very common with sentiment lexicons playing a key role in the majority of methods. The various approaches and the most popular algorithms of SC are illustrated in Figure 2 as mentioned before.

Machine Learning Approach: Machine learning approach relies on the famous ML algorithms to solve the SA as a regular text classification problem that makes use of syntactic and/or linguistic features.

Text Classification Problem Definition: We have a set of training records D = {X1, X2, …, Xn} where each record is labeled to a class. The classification model is related to the features in the underlying record to one of the class labels. Then for a given instance of unknown class, the model is used to predict a class label for it. The hard classification problem is when only one label is assigned to an instance. The soft classification problem is when a probabilistic value of labels is assigned to an instance.

Lexicon Based Approach: Opinion words are employed in many sentiment classification tasks. Positive opinion words are used to express some desired states, while negative opinion words are used to express some undesired states. There are also opinion phrases and idioms which together are called *opinion lexicon*. There are three main approaches in order to compile or collect the opinion word list. *Manual approach* is very time consuming and it is not used alone. It is usually combined with the other two automated approaches as a final check to avoid the mistakes that resulted from automated methods. The two automated approaches are presented in the following subsections.

Hybrid Approach: There are techniques that cannot be roughly categorized as ML approach or lexicon-based Approach. Formal Concept Analysis (FCA) is one of those techniques. FCA was proposed by Wille as a mathematical approach used for structuring, analyzing and visualizing data, based on a notion of duality called Galois connection. The data consists of a set of entities and its features are structured into formal abstractions called *formal concepts*. Together they form a concept lattice ordered by a partial order relation. The concept lattices are constructed by identifying the objects and their corresponding attributes for a specific domain, called *conceptual structures*, and then the relationships among them are displayed. Fuzzy Formal Concept Analysis (FFCA) was developed in order to deal with uncertainty and unclear information

VIII. Conclusion and Future Work:

This survey paper presented brief overview on the recent updates in SA algorithms and applications. These techniques contributions to many SA related fields that use SA techniques for various real-world applications. After analyzing these articles, it is clear that the enhancements of SC and FS algorithms are still an open field for research. Naïve Bayes and Support Vector Machines are the most frequently used ML algorithms for solving SC problem. They are considered a reference model where many proposed algorithms are compared to.

The interest in languages other than English in this field is growing as there is still a lack of resources and researches concerning these languages. The most common lexicon source used is WordNet which exists in languages other than English. Building resources, used in SA tasks, is still needed for many natural languages.

Information from micro-blogs, blogs and forums as well as news source, is widely used in SA recently. This media information plays a great role in expressing people's feelings, or opinions about a certain topic or product. Using social network sites and micro-blogging sites as a source of data still needs deeper analysis. There are some benchmark data sets especially in reviews like IMDB which are used for algorithms evaluation.

In many applications, it is important to consider the context of the text and the user preferences. That is why we need to make more research on context-based SA. Using TL techniques, we can use related data to the domain in question as a training data. Using NLP tools to reinforce the SA process has attracted researchers recently and still needs some enhancements.

References:

[1] Wilson T, Wiebe J, Hoffman P. Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of HLT/EMNLP; 2005.

[2] Liang-Chih Yu, Jheng-Long Wu, Pei-Chann Chang, Hsuan-Shou Chu
Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news

[3] Michael Hagenau, Michael Liebmann, Dirk Neumann. Automated news reading: stock price prediction based on financial news using context-capturing features. Decis Supp Syst; 2013.

[4] Xu Tao, Qinke Peng, Yinzhao ChengIdentifying the semantic orientation of terms using S-HAL for sentiment analysisKnowl-Based Syst, 35 (2012), pp. 279-289

[5] Zhou L, Li B, Gao W, Wei Z, Wong K. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In: Presented at the 2001 conference on Empirical Methods in Natural Language Processing (EMNLP'11); 2011.

[6] Heerschop B, Goossen F, Hogenboom A, Frasincar F, Kaymak U, de Jong F. Polarity Analysis of Texts using Discourse Structure. In: Presented at the 20th ACM Conference on Information and Knowledge Management (CIKM'11); 2011.

[7] Sunil Kumar Gupta, Dinh Phung, Brett Adams, Svetha VenkateshRegularized nonnegative shared subspace learningData Min Knowl Discov, 26 (2012), pp. 57-97

[8] Hanhoon Kang, Seong Joon Yoo, Dongil Han**Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews**Expert Syst Appl, 39 (2012), pp. 6000-6010

[9] Ester Boldrini, Alexandra Balahur, Patricio Martínez-Barco, Andrés Montoyo**Using EmotiBlog to annotate and analyse subjectivity in the new textual genres**Data Min Knowl Discov, 25 (2012), pp. 603-634

[10] Josef Steinberger, Mohamed Ebrahim, Maud Ehrmann, Ali Hurriyetoglu, Mijail Kabadjov, Polina Lenkova, Ralf Steinberger, Hristo Tanev, Silvia Vázquez, Vanni Zavarella**Creating sentiment dictionaries via triangulation**Decis Support Syst, 53 (2012), pp. 689-694