www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 4 April 2017, Page No. 20895-20902
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i4.16

Sentiment Analysis in Today's Trend: A Review

Surendra Kumar[1], Dr.Harsh Kumar[2], Dhruba Jyoti Kalita[3]
[1]Assistant Prof Dr.K.N.Modi University Newai Rajasthan
surendra.ftp@gmail.com
[2]Associate Prof Dr.K.N.Modi University Newai Rajasthan
drharshkumar@hotmail.com
[3]Assistant Prof Dr.K.N.Modi University Newai Rajasthan
mestop12@gmail.com

Abstract: Sentiment analysis a branch of text analysis that is widely used to increase the reliability of the product and used to analyze the people's emotions, opinion towards the products, hot topic, any event and any organization etc. There are various social media as well as e-commerce site which are generating a large amount of data in the form of tweets, blogs, status, reviews etc. In this survey paper we have presented an elaborative hierarchy of sentiment analysis techniques and sub-techniques or mechanisms and algorithms along with their pros and cons. We have also presented here the applications of sentiment analysis in different fields along with various challenges in it to implement. The main target of this survey paper is to give full description of sentiment analysis.

Keywords: Sentiment Analysis, Hot Topics, Text Analysis, opinion, Tweets, Blogs

I. Introduction

We are in an age of internet where we can express our views and opinion for a product or a hot topic with some interfaces like blogs, social networking sites, review sites, discussion forum etc. We can say that Sentiment analysis is mechanism through which it is used to identify and classify the opinion for a particular topic or product. It is used for findings the opinion of a group about particular entity. Entity can be anything like any topic, any event or individual. Sometimes sentiment analysis is also called as opinion mining. In opinion mining we extract and analyze the opinion about any entity given by people online and sentiment analysis identify the emotion given in the text then analyze it. The aim of SA is to find out the opinion and identify the sentiment given by the people and find the polarity as described in fig 1.

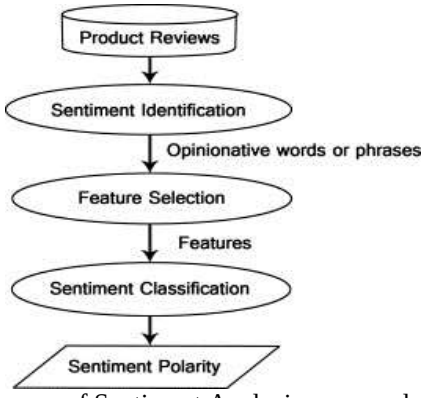


Fig: 1 Process of Sentiment Analysis on a product review

This decision making process is formed using opinion of ordinary people, celebrities and leaders opinion. When a person wants to buy a product online, he/she search some review regarding a product and then he/she takes some decision based on review posted by the users but some time it is hard to take decision on the huge number of reviews so sentiment analysis can be used [1]. It is the one of the interesting research area in which lots of work is going on the problem of NLP i.e Natural language Processing that is the findings of opinion in different languages [2].Natural Language Processing is a domain of computer science and computational linguistics concerned with the linking between computers and human

(natural) languages for communication" [3]. In Sentiment analysis "The match was interesting" is different from "the match was bogus". People write writes contradictory. Review may be positive or negative. It is easy to understand by human but some time it is not easy to analyze these reviews by the machines.

II. Application of sentiment analysis

- The one of the most important application in the field of business that a company collects the reviews of a product given by the customers and they decide efficiency and effectiveness of the product [4] and improve the aspect of product from which customers are not satisfy.
- When someone wants to buy a product online then we have to take to decision on that product on basis of the reviews given by the customers and then we decide that product should buy or not buy.
- In matrimonial sites to match groom and bride according to their interest one can use sentiment analysis. Here we can collect their accessed data on the web and classify them according to those data.
- In news channels are carry out exit or opinion poll based on the voter's opinion in particular region. In this process it is very much needed to analyses the sentiment of the voters to decide their interesting parties.
- From raw information which is large in quantity by using mining techniques one can extract important information regarding a topic in case of sentiment analysis the raw information lies in the form of tweets and specifications of an object and product. This is to summarize that information [5].
- Another application of sentiment analysis is to give recommendation and choice of product on the basis of our previous history.
- For rating movies so that people can choose their kind of movies sentiment analysis can be used. Here users can be clustered in to multiple clusters according to their interest.

III. Problem and Challenges in Sentiment Analysis

- The very first important challenge is the problem in Natural Language Processing. It is because of the variety of languages present in the complete domain of Natural Language Processing. It is not possible to focus on all the languages that are present in a multilingual world [6]. So it should be our aim to develop such kind of a model which will be able to deal with all the languages.
- The second challenge is that there is no standard for expressing our views and people express their view in their own style, by creating a little difference in text and it creates huge difference of the sentences. For example the story was unpredictable; steering of the car is unpredictable, go read the book. Above given three sentences has different sentiment, the first sentence conveyed a positive sentiment and second sentence conveyed a negative sentiment and third sentence also conveyed positive sentiment [7].
- Dynamism in data is another challenge in sentiment analysis. Even though on static data we can build sentiment analysis model in an efficient way it is hard to build the same model on dynamic data. Here dynamic data means the data which are streaming kind of data where data flows continuously. For example twitter provides a platform to expose continuous data. Here data grows at each and every minutes and no limit is there in the growth of data.
- Unstructuredness in certain text makes it impossible for the sentiment analysis model to cope up with upcoming text data. For this we need to build up models which uses modern approaches for data cleaning and collecting where classical model of data cleaning such as part of speech tagging, named entity recognition etc.
- Subjectiveness of a sentence or opinion can be destroyed by the use of Utterances, Clauses, and Fragments. These normally brings to much abstractions to the subject matter for which the

information that emphasis the subject gets hidden .So it is a challenge to bring the subject matter through sentiment analysis model that we design.

- In sentiment analysis one opinion can be positive in some situation and can be negative in some situation. In traditional sentiment analysis technique the problem of classification was consider as two class classification problem. Here positive class was used to represent the positive emotions of the user on some objects in web or some other plate form like Facebook, twitter and amazon, IMDB and Netflix etc. and the same way the negative class for the negative emotions of users. But in today's trends classical classification technique like above in not adequate. We also need to classify users whose answers of interaction lies in between the positive and negative class. So it is a challenging task to design one sentiment analysis technique that considers both positive and negative class as well as the intermediate class between these two.

IV. Classification of Sentiment analysis:

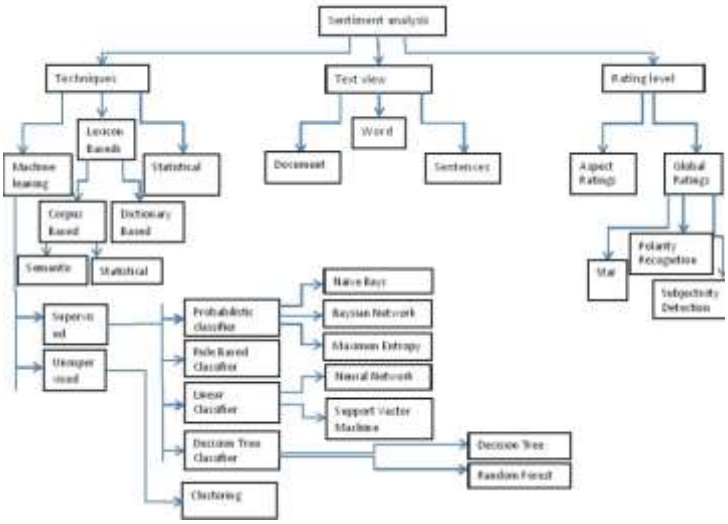


Fig 1: Hierarchy of sentiment analysis techniques

The categorization of sentiment analysis can be done based on three factors and they are techniques used, level of detail and level of rating. By considering the first factor that is techniques used sentiment analysis can be classified in to following three basic categories:

- Sentiment Analysis through Machine Learning
- Sentiment Analysis through Lexicon Based
- Sentiment Analysis through Statistical Based

Sentiment Analysis through Machine Learning: This kind of sentiment analysis uses traditional learning based model to do sentiment analysis which is in corporate in two ways of learning either supervised or unsupervised. In supervised learning technique the raw data in hand is divided in to two basic parts i.e. training set and test set. The training set is to make the machine trained enough to classify new sample coming from the test set. Most powerful supervised learning models that can be used with sentiment analysis are support vector Machine, Neural Network and LDA (Linear Discernment Analysis) based model. On the other hand the unsupervised learning is completely based on pattern recognition model. Most of the clustering techniques come under unsupervised learning and few of them are K-Mean algorithm, Naïve-Bays, Fuzzy C-means algorithm and many other probabilistic algorithms

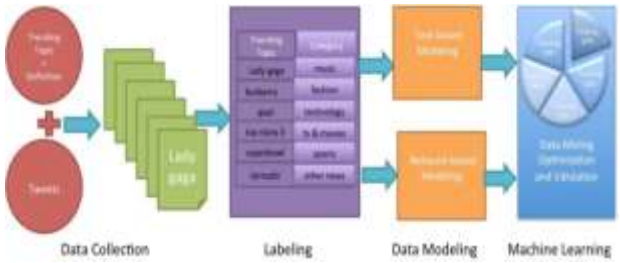


Fig 2: Sentiment Analysis through Machine Learning

Support Vector Machine:

"Support Vector Machine [8] is most powerful a supervised machine learning a model which can be used for classification. In this algorithm, we plot each sample in a data set as a point consisting of dimensions as there coordinates. Support Vector Machine can also be considered as an optimization problem where we try to find an optimal hyper plane in between two classes, positive class and negative class respectively. SVM can deal with both linear and nonlinear data in linear data we normally don't map the input

space in to feature space where dimensions increases. In case of nonlinear data we use kernal function to map input space to a higher dimensional feature space. The building of linear SVM in the feature space is equivalent to building nonlinear SVM in the input space.

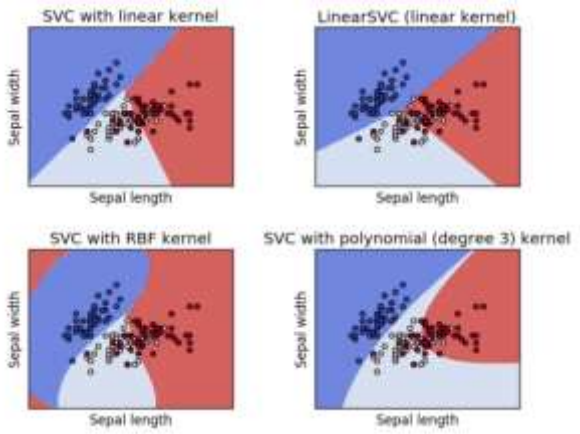


Fig 3: Support Vector Machine

Neural Network Model:

Neural networks are another branch supervised learning which mimics the human brain. It is modeled and based on the human brain. Here something called Neurons is used as nodes for signal accumulation. Neural networks can use a layered approach. Three basic layers that can be incorporate with NN are input layer, hidden layer and output layer. Deep Neural network is the most recent advancement done in the neural networks which consist of multiple hidden layers. Due to the complexity of identifying the processing in the hidden layers neural networks fails in many cases.

LDA model:

Linear discriminant analysis[9] (LDA) also known as Fisher's linear discriminant analysis or as Canonical variety analysis is a which is used to find linear combinations of observed features which distinguish the boundary of two or more classes precisely. The resulting combinations can be used for dimensionality reduction, before classification. LDA is based on principal component analysis (PCA) in the way of finding linear combinations of variables. LDA is different from PCA in the fact that it explicitly attempts to model the difference between the classes

Unsupervised:

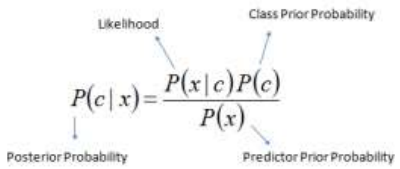
K-Mean algorithm:

K-means is one of the popular unsupervised learning algorithms to solve clusteri ng problem through a certain number of clusters (assume k clusters) fixed apriori the algorithm tries to classify the data. The main aim is to define k centers, for each cluster. By calculating the distance of the sample from those centers the algorithm decides its cluster. Each cluster should be distant enough so that we can avoid overlapping of clusters [10]. This algorithm aims in minimizing the objective function knows as squared error function and given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\|x_i - v_j\| \right)^2$$

Naïve-Bays:

It is a classification technique based on Bayes' Theorem. Normally, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naïve Bays method concentrates on finding related features for some object to cluster them For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'. Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below [11]:



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Fuzzy C-means:

Fuzzy c-means (FCM) [12] is a clustering algorithm which allows one sample of data to be clustered in to two or more clusters. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m}\left\|x_i - c_j\right\|^2 \quad , \quad 1 \le m < \infty$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the i_{th} of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\|x_i - c_j\right\|}{\left\|x_i - c_k\right\|}\right)^{\frac{2}{m-1}}} \quad , \quad c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

This iteration will stop when $\max_{ij}\left\{\left|u_{ij}^{(k+1)} - u_{ij}^{(k)}\right|\right\} < \varepsilon$, where ε is a termination criterion between 0 and 1, whereas k is the iteration steps. This procedure converges to a local minimum or a saddle point of J_m.

Sentiment Analysis through Lexicon Based

Lexicon based sentiment analysis consider the semantic orientation of words or phrases that occur in a text [13] to judge the sentiment. In this approach a dictionary of positive and negative words needs to be maintained, where each word is associated with either positive or negative value. To create dictionaries there are many proposed methods available which include both manual [14] and automatic [15] approaches. Lexicon-based approach normally uses traditional bag of words to represent a piece of text message. To predict the overall sentiment for the text message a function is required that combines the sentiment of the words, such as sum or average. The predicted sentiment of the message also combines negation or intensification of the local context of a word.

Sentiment Analysis through Statistical Based

Statistical methods include classical methods such as Bayesian inference and support vector machines which are popular for effective and efficient classification of texts. By using a machine learning algorithm a large training set of corpus of annotated texts the system not only learn the affective valence of keywords but also consider the valence of other arbitrary keywords, punctuation, and frequency of word co-occurrence. It is possible to decide the polarity of a word by studying the occurrence frequency of the word in that corpus of texts [16]. If a word appears in the positive text frequently, then polarity of the word is decided as positive polarity and in the same way we can decide the negative polarity of a word. Words that occur both in the positive and negative text with the same frequency have neutral polarity. The state of the art methods for sentiment analysis are based on the above principle. Words appearing with in the same context with same frequency are going to have same polarity. So the basic way of deciding the polarity of an unknown word is to calculate the frequency which is relative to the co-occurrence of a word with another word. Traditional statistical methods are weak semantically in the sense that obvious affect keywords, other lexical or co-occurrence elements in a statistical model bear little predictive value individually. Due to the above limitations statistical classification method works well on user's text on page or paragraph level but not on small units like sentences or clauses [17].

Sentiment Analysis is a subdomain of text analysis which can be carried out in three basic levels namely document level, sentence level and word level. The expressive power of the sentiments that we draw from documents maximizes the judgment about the opinion that someone gives for some subject.

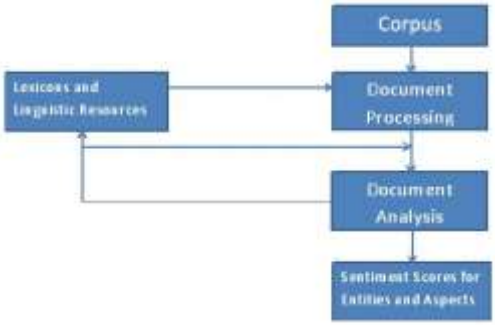


Fig 4: Statistical Sentiment Analysis

- **Document level sentiment analysis**

The document-level Sentiment Analysis model analyzes of text message in whole document and marks the positive or negative sentiment label for the text. It can do sentiment analysis of any sentiment oriented text which is not too short, this model can do sentiment analysis for the input text which has more than 40 characters in length and ha support for following languages:

- English (US and UK)
- French
- German
- Italian
- Spanish

To run The document-level Sentiment Analysis model one Has to wrap it with a transform API which contains a function getSentiment() that produces output as positive or negative sentiment label for the text. For any null or empty input it returns null as output [18]. Note that NULL is returned for any input which is either null or empty.

- **Sentence level sentiment analysis**

Instead of document, sentence level sentiment analysis does the sentiment analysis at the level of sentence and decides positive, negative and neutral sentiment for the sentence. Subjectivity classification that is regarding the expression of some facts attached with sentence can be achieved by sentence level sentiment analysis which in turn gives subjective views and opinion. Subjectivity is not same as sentiment as many objective sentences can have opinions. [19].

- **Word level sentiment analysis**

Word-level sentiment classification considers a lower granularity of the input text and expresses sentiment at that level of granularity. Sentiment analysis of the entities at this level flows from a lower level to higher level. Let

X: Random variable over data sequences to be labeled

Y: Random variable over corresponding label of sequences

y_i, is a component of Y representing a range over a finite label alphabet (positive, negative, neutral), The notion of sentiment flow can be represented as $Y = [y_1, y_2 \dots \dots y_k]$ for words $x = [x_1, x_2, \dots \dots x_k]$. x_i corresponds to natural language words and other tokens found in text, while y_i to labels from the restricted set.

Let us take the following example

"The rose is beautiful only in a clean garden not in jungle".

In the above sentence we have sentiments associate with multiple entities like "rose is beautiful", "clean garden" and "jungle". The subjectivity or the sentiment orientation can be analyzed for the above piece of text by considering all the entities.it is better not to consider only one entity which hides the subjectivity of the piece of text.

Finally sentiment analysis can also be categorized into few basic categories based on rating level. They are

- **Aspect Rating based sentiment analysis**

Aspect level analyses the entity over which opinion is buildup. Sentiment words are basic units in identifying the sentiments associated with a sentence or a document. This sentiment words are combined together to form something called as sentiment lexicon [20]. Sentiment lexicon is further used by the algorithm for data analysis. It is difficult to identify the opinions which are objective and subjective text in nature. Subjective sentences express opinion about a subject or they represent a person's perspective. Objectivity and the irony in sentences is another issue which decides the complexity of the algorithm used in data analysis. For the Basic Understanding of the concept there should be a

proper vocabulary to understand terms and definitions that is used sentiment analysis and opinion mining. According to [20] opinion is a tuple with two components target g and sentiment s. Here target g is an entity for which opinion is given or it is an aspect of the entity. The following example illustrates the concept of entity and aspect.

"The picture quality of Lenovo k6 power is excellent".

In the above piece of text Lenovo k6 power is an entity and picture quality is the aspect. If we want to write this in (g, s, h, t) format then,

g: Picture quality of Lenovo k6 power (aspect)

s: Excellent (it can be positive, negative or neutral)

h: A person or machine who gave the review

t: Time stamp at which the opinion is given

Using't' we can analyses the time constraint of an opinion whether the opinion is older or new.

Multiple conflicting sentiments make it difficult to predict the output. To resolve the problem it is necessary to connect sentiments with the aspects.

- **Global rating based sentiment analysis**

In global rating based sentiment analysis rates reviews on global level. Classification which is based on global rating considers two polarities, positive and negative. Most of the machine learning techniques can be deployed for global rating based sentiment analysis. It can be further divided into three categories

- **Stars**
- **Subjectivity detection**
- **Polarity Recognition**

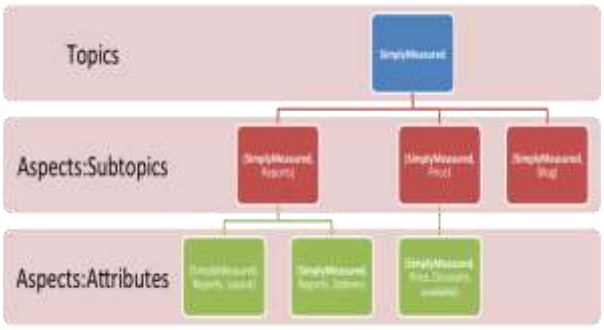


Fig 5: Global rating based Sentiment Analysis

Situation where the polarity of a classification problem not clearly defined we use star based rating for reviews of product. We can identify neutrality of a review by 3 stars, positivity by more than 3 stars and negativity by less than 3 stars.

Star Level	Meaning
☆	I hate it.
☆☆	I don't like it.
☆☆☆	It's ok.
☆☆☆☆	I like it.
☆☆☆☆	I love it.

Fig 6: General Rating System

Subjectivity detection means distinguishing opinionativity from non-opinionativity (for the phrases). The subjectivity detection process contain three basic steps identify, classify and aggregate. A topic mentioned with the input data has to be identified first, to associate opinionative sentences with these them. In the classification phase we classify the documents or sentences into topics identified in the identification phase. After doing identification and classification aggregation is needed to decide the polarity.

Polarity can be defined as two extremes that can be assigned to an opinion in sentiment analysis. In above polarity recognition is the process of identifying positiveness/negativeness of a particular opinion. Neutrality lies in between positiveness and negativeness of an opinion.

Conclusion: Sentiment analysis is a sub branch of text analysis which has applications in many of the domains of technology. It is an analysis tool to dig the human opinion on data. It has applications in reviews classification, recommendation, predictive models and business statistics evolution. Many platforms in the web like websites, blogging, and social networking site play a vital role in sentiment analysis as a source of data. There are many techniques available that can be used as sentiment analysis tools ranging from machine learning algorithms to mining algorithms. In this paper we have presented a detail survey on these techniques and applications of sentiment analysis. In future sentiment analysis has a scope in NLP where we will be able to do mining of sentiments not only for a particular language but for many different languages.

References:

1. Neethu, M. S., and R. Rajasree. "Sentiment analysis in twitter using machine learning techniques." In Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on, pp. 1-5. IEEE, 2013.
2. Vinodhini, G., and R. M. Chandrasekaran. "Sentiment analysis and opinion mining: a survey." International Journal 2, no. 6 (2012): 282-292.
3. Kumar, Ela. Natural language processing. IK International Pvt Ltd, 2011.
4. Das, Sanjiv, and Mike Chen. "Yahoo! for Amazon: Extracting market sentiment from stock message boards." In Proceedings of the Asia Pacific finance association annual conference (APFA), vol. 35, p. 43. 2001.
5. Liu, Bing, Minqing Hu, and Junsheng Cheng. "Opinion observer: analyzing and comparing opinions on the web." In Proceedings of the 14th international conference on World Wide Web, pp. 342-351. ACM, 2005.
6. Seki, Yohei, David Kirk Evans, Lun-Wei Ku, Le Sun, Hsin-Hsi Chen, Noriko Kando, and Chin-Yew Lin. "Overview of Multilingual Opinion Analysis Task at NTCIR-7." In NTCIR. 2008.
7. Mukherjee, Subhabrata, and Pushpak Bhattacharyya. "Sentiment analysis: A literature survey." arXiv preprint arXiv:1304.4520 (2013).
8. https://www.analyticsvidhya.com/blog/2015/10/understaing-support-vector-machine-example-code/
9. Montanari, Angela, Daniela G. Calò, and CinziaViroli. "Independent factor discriminant analysis." Computational Statistics & Data Analysis 52, no. 6 (2008): 3246-3254.
10. https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm
11. https://www.analyticsvidhya.com/blog/2015/09/naive-bayes-explained/
12. https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html.
13. Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. "Lexicon-based methods for sentiment analysis." Computational linguistics 37, no. 2 (2011): 267-307.
14. Tong, Richard M. "An operational system for detecting and tracking opinions in on-line discussion." In Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification, vol. 1, p. 6. 2001.
15. Turney, Peter D., and Michael L. Littman. "Measuring praise and criticism: Inference of semantic orientation from association." ACM Transactions on Information Systems (TOIS) 21, no. 4 (2003): 315-346.
16. Read, Jonathon, and John Carroll. "Weakly supervised techniques for domain-independent sentiment classification." In Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion, pp. 45-52. ACM, 2009.
17. Banker, Shreya, and Rupal Patel. "ABrief REVIEW OF SENTIMENT ANALYSIS METHODS." International Journal of Information 6, no. 1/2 (2016).
18. http://docs.oracle.com/cd/E64107_01/bigData.Doc/data_processing_bdd/src/rdp_de_sentiment_svm.html
19. Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. "Recognizing contextual polarity in phrase-level sentiment analysis." In Proceedings of the conference on human language technology and empirical methods in natural language processing, pp. 347-354. Association for Computational Linguistics, 2005.
20. Liu, Bing. "Sentiment analysis and opinion mining." Synthesis lectures on human language technologies 5, no. 1 (2012): 1-167.