Survey on Sentiment Analysis and Sentiment Classification

[1] S.J.Veeraselvi, [2] M.Deepa
[1]*PG Scholar, Kalaignar Karunanidhi Institute of Technology, Coimbatore, Tamil Nadu, India*
[2]*Assistant Professor, Kalaignar Karunanidhi Institute of Technology, Coimbatore, Tamil Nadu, India*

Abstract

Opinions are the fundamental aspect to almost all decision making activities. The increased usage of internet and the exchange of user opinions through social media and public forums on the web has become the motivation for sentiment analysis. Due to the infinite amount of user opinions available throughout the web it is necessary to automatically analyze and classify sentiment expressed in opinions to make the decision making process an easy task. Opinion Mining or Sentiment Analysis is a Natural Language Processing technique that attempts the system to automatically identify and extract sentiments expressed in user reviews. The basic task of sentiment analysis is sentiment classification which classifies a user review as positive, negative, neutral. This survey gives an overview of sentiment analysis, sentiment classification, methods used for sentiment classification.

Key words: Sentiment Analysis, Sentiment Classification, Cross Domain Sentiment Classification.

1. Introduction

In traditional time, whenever we need to make decision, we would like to hear others suggestion where an individual get suggestion from family or friends. To improve a business customer opinion is needed so it collects feedback, conduct survey and so on. With the explosive growth of internet, people express their opinion or reviews in forums, blogs, etc. Each site contains huge volume of opinioned text and it is difficult for users to examine all reviews and summarize the orientation. Hence the idea behind sentiment analysis lies in providing summarization of opinions to develop a system where by opinions can be classified into positive, negative, neutral reviews based on the sentiment expressed in the documents [3], [19]. Automated opinion mining uses machine learning approach, a component of artificial intelligence. Thus sentiment analysis also known as opinion mining, is a technique that try to find and understand the opinion and sentiment by analyzing the opinion data and it also support in human decision making. By definition an opinion is a quintuple, $O = (e_i, f_{ij}, so_{ijkl}, h_k, t_i)$ where, e_i = target entity , $f_{i\ j}$ =feature of the target entity, so_{ijk} =sentiment value of the opinion(positive, negative, neutral), h_i =opinion holder, t_i =the time when opinion expressed [19]. The five components are very essential. The object on which the opinion is given is the target entity(e_i) which is the first component. We need a target entity because without knowing the target entity the piece of opinion has little or no value. The target entity may be a product, service, person or organization. The second component is feature (f_{ij}) of the target entity on which the sentiment is expressed. The third component in the opinion is the sentiment (so_{ijk}) which helps to classify the opinion as positive, negative, neutral. The fourth component is opinion holder (h_k) the one who express the opinion. The fifth component is time(t_i) which helps to analyze the opinion based on the time it is expressed. Thus the value of the opinion articulated before a year ago and the opinion articulated an hour ago has different magnitude. *For example "The **games** (f_{ij}) in the **iphone** (e_i) are **pretty funny** (so_{ijk}) ".* The time and opinion holder is available in the web [3], [19]. Fig 1. Depicts the sentiment analysis model. The web users post their views, comments and feedbacks about a

product or thing through blogs, forums, social networking sites etc. The data preparation step performs necessary data pre-processing and cleaning on the dataset for the successive analysis. Basic preprocessing steps tokenizing, stop words filtering and stemming. Then the review analysis step analyzes the linguistic features of reviews so that interesting information, including opinions and/or product features, can be identified. Two commonly adopted tasks for review analysis are POS tagging and negation tagging. Then various machine learning techniques can be applied in order to classify the polarity (positive and negative opinions) using the obtained reviews and finally the result will summarize the opinion impact based on the sentiment of the web users expressed in the reviews

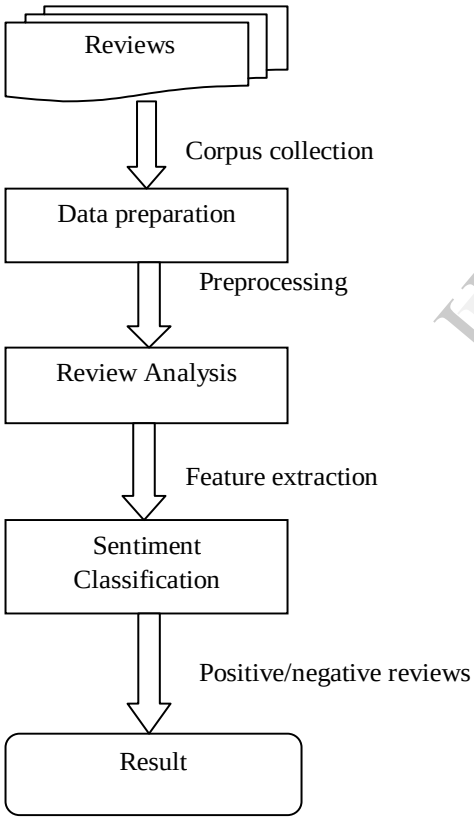

Figure 1 Sentiment analysis model

2. Sentiment Analysis

Textual information is any information presented using words and characters. Two main types of textual information are Facts (using keywords) and Opinions (emotions). Ordinary keyword search will not be suitable for mining all kinds of opinions. Hence it becomes necessary that the sophisticated opinion extraction methods are used. Automated extraction of subjective content from digital text and predicting the subjectivity such as positive or negative. Opinion extraction is the process of extracting inputs for opinion mining. Sentiment classification and sentiment clustering are the two sub tasks of opinion extraction. Sentiment analysis is one of the sub tasks in text mining. It aims to determine the thoughts of the writer with respect to some topic or object or an article. Sentiment analysis is a NLP problem because it touches every aspect of NLP like negation handling, word sense disambiguation which are the problems that are not yet solved. Thus sentiment analysis provides a great platform for NLP researchers. One of the challenges in sentiment analysis is to obtain the subjectivity content. Subjectivity analysis classifies content is objective or subjective. Thus the subjective content describe a person opinion (eg: ram is a good boy) while an objective sentence contain many point of views (eg: I bought a iphone a few days ago). Thus emotions are peculiar to each person. Sentiment analysis can be achieved by including several sub tasks. The subtasks consist of subjectivity detection, polarity classification. They are presented in the following sections.

2.1. Subjectivity Detection

Subjectivity analysis classifies content is objective or subjective. Thus the subjective sentence express some personal feelings (eg: Ram is a good boy) while an objective sentence present some factual information about the world (eg: I like iphone). Thus emotions are peculiar to each person. Hence the objective sentence can't play a role in sentiment analysis. Subjectivity content can be obtained from two main types of opinions namely regular and comparative. Regular opinion has two subtypes as direct and indirect opinion. The direct opinion refers to an opinion expressed directly on an object (eg: the picture quality is high). The indirect opinion is expressed indirectly on an object (eg: though he takes the medicine, he has not recovered from typhoid).A comparative opinion expresses the similarity or differences between two or more objects.(eg: Slice

tastes better then coke). Sentiment is expressed differently in different domain so the subjective detection is a domain and context dependent problem [6], [21], [22].

2.2. Polarity Classification

Another important task in sentiment analysis is polarity classification. The goal is to classify the subjective sentences as positive or negative. A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level — whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. The polarity classification is a binary classification task where an opinionated document is labeled with an overall positive or negative sentiment. Sentiment Polarity Classification can also be termed as a binary decision task. "*Not only do I not approve Supernova 7200, but also hesitate to call it a phone"* has a positive polarity word *approve*; but its effect is negated by many negations. Polarity classification considered as binary classification task where a opinion is labeled with positive, negative tags [3], [11].

2.3. NEGATION

Negations which tend to be disregarded in text analysis , play an important role in sentiment analysis by flipping a positive term into negative and vice versa. Negation can change the text polarity. For example I like this film and I don't like this film are considered similar when using similarity measures but the negation term *not* classifies the opinion as a negative opinion. The scope of the negation expression determines which sequence of words in the opinions is affected by negation words such as no, not, never, etc. The presence of negation words in a opinion does not mean that the opinion specify a negative sentiment (eg: the film is not boring). In linguistics a morpheme is the smallest grammatical unit in a language. The study of morphemes is known as morphology (eg; unbreakable comprises of three morphemes un, break, able). Negation can be morphological where it is either denoted by a prefix ("dis-", "non-") or a suffix ("-less").the use of specific part-of-speech tags pattern to identify the negations makes the sentiment analysis task more efficient and accurate [3], [5], [7], [9].

3. Sentiment Classification

Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects. Sentiment classification is the task of classifying a given review with respect to the sentiment expressed in the review. Sentiment classification has been applied to numerous tasks such as opinion summarization, opinion mining, market analysis, etc. In opinion summarization, first classify the opinion as positive or negative and then create a summary for each sentiment type (positive, negative). In [19] sentiment classification is mainly studied at three levels namely document level, sentence level, aspect level.

3.1. Document Level Sentiment Classification

Sentiment classification is essentially a text classification problem. Tradition text classification mainly classifies document of different topics (eg: politics, science, sports, economics, etc) based on topic related words (key words). A document typically contains numerous opinions. The task at this level is to examine and classify whether the whole opinion document express a positive or negative sentiment.

3.2. Sentence Level Sentiment Classification

This level classifies the sentiment expressed in each sentence. Sentence level sentiment classification can be decipher as a three class or two class classification problem. In three class classification the sentence is classified as positive, negative or neutral and in two class classification a sentence is classified as positive or negative. The basic step in sentence level sentiment classification is subjective sentence extraction then the next step is to extract features and then classify the sentence as positive or negative or neutral. Sentence level sentiment classification also deals with comparative and sarcastic sentences.

3.3. Aspect Level Sentiment Classification

Aspect level sentiment classification is also known as feature based sentiment classification. For example "the picture quality of NOKIA phone is great", here the aspect is "picture quality". This type of sentiment classification formulates the sentiment

classification more efficient. The objective of aspect level sentiment classification is to discover the quintuple (e_i,f_{ij},so_{ijkl},h_k,$t_{i)}$ of the opinion. The quintuple is explained in section I. the two main task in aspect level sentiment classification is aspect extraction and aspect sentiment classification. In aspect extraction the features are extracted (picture quality). Aspect sentiment classification determines whether the opinion on different aspects is positive, negative or neutral. In most cases sentiment classification is termed as a binary classification where a sentiment of a opinion is classified as two classes namely positive or negative. The two main tasks of sentiment classification are polarity assignment and intensity assignment. Polarity is binary values that signify either positive or negative. Sentiment polarity assignment deals with analyzing, whether a text has a positive, negative polarity. Sentiment intensity assignment deals with analyzing the strength of sentiments. For example consider two sentences "I don't read history" and "I hate reading history", where both specify a negative polarity but the succeeding opinion has extra negative semantic orientation than the first opinion

.

4. SUPERVISED LEARNING METHOD

Classification is a form of data analysis that can be used to extract models describing important data classes. Many classification methods are based on machine learning techniques. In classification learning a classifier is called supervised learning method. Learning a classifier is called supervised learning method. Data classification is a two step process: Training phase-learning the model from a corpus of training data. Classification phase-classifying the unseen data based on the trained model. As in [3], [18] the following are the some of the machine learning techniques which is used to determine whether the review is positive or negative. It uses movie reviews as input for sentiment classification.

4.1. Simple Bayesian Classification

Bayesian classifier known as navie Bayesian classifier. Bayesian classifier is applied for large databases. in Bayesian classification a tuple X belongs to a class c based on the probability of the tuple X. for text classification the document d belongs to a class C, c^* = arg max$c$ $P(c \mid d)$. The *Naive Bayes* (NB) classifier uses the Bayes' rule

$$P(c/d) = \frac{P(c)P(d/c)}{P(d)} \qquad (1)$$

where $P(d)$ plays no role in selecting c^*. To estimate the term $P(d \mid c)$, Naive Bayes decomposes it by assuming the fi's are conditionally independent given d's class

$$P_{NB}\langle c|d\rangle := \frac{P(c)\prod_{i=1}^{m} P(f_i/c)^{n_i(d)}}{P(d)} \qquad (2)$$

Where m is the no of features and fi is the feature vector. Consider a training method consisting of a relative-frequency estimation $P(c)$ and $P(fi \mid c)$. Despite its simplicity and the fact that its conditional independence assumption clearly does not hold in real-world situations, Naive Bayes -based text categorization still tends to perform surprisingly swell indeed.

4.2 Maximum Entropy

Another classification technique, which has proven effective in a number of natural language processing applications, is maximum entropy. Sometimes, it outperforms Naive Bayes at standard text classification. Its estimate of $P(c \mid d)$ takes the exponential form

$$P_{ME}(c/d) := \frac{1}{Z(d)} \exp\left(\sum_i \lambda_{i,c} F_{(i,c)}\ (d,c)\right) \qquad (3)$$

where Z(d) is a normalization function. $Fi_{,c}$ is a feature/class function for feature f_i and class c, defined as follows,

$$F_{i,cs}(d,c') := \begin{cases} 1, & n_i(d) > 0 \text{ and } c' = c \\ 0, & otherwise \end{cases} \qquad (4)$$

For instance, a particular feature/class function might fire if and only if the bigram "still hate" appears and the document's sentiment is hypothesized to be negative. Maximum Entropy makes no assumptions about the relationships between features and so might potentially perform better when conditional independence assumptions are not met.

4.3. Support Vector Machines

Support vector machine is a supervised machine learning classification technique for oth linear and non linear data. Support vector machines (SVMs) have been shown to be highly effective at traditional text categorization. In SVM the data is mapped to high dimension. SVM searches for hyperlane with the largest margin, that is, the maximum marginal hyperlane. The associated margin gives the largest separation between classes. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output. In the two-category case, the basic idea behind the training procedure is to find a maximum margin hyper plane, represented by vector $\vec{w}$, that not only separates the document vectors in one class from those in the other, but for which the separation, or *margin*, is as large as possible.

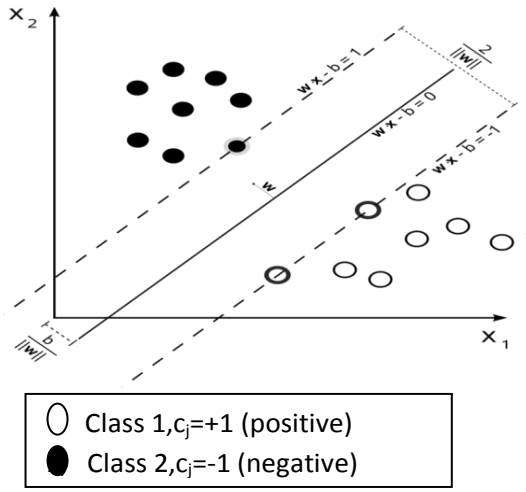

Figure 2 SVM trained with samples from two classes. Samples on the margin are called the support vectors.

This corresponds to a constrained optimization problem; letting *cj* {1, −1} (corresponding to positive and negative) be the correct class of document *dj*, the solution can be written as,

$$\vec{w} = \sum_j \alpha_j c_j \vec{d_j}, \alpha_j \geq 0 \qquad (5)$$

where the *j* α's (Lagrange multipliers) are obtained by solving a dual optimization problem. Those for which dj are greater than zero are called *support vectors*, since they are the only document vectors contributing to. Classification of test instances consists simply of determining which side off s hyper plane they fall on.

5. UNSUPERVISED SENTIMENT CLASSIFICATION

[20] proposed a simple unsupervised learning algorithm for classifying reviews as positive or negative. The review is classified based on the semantic orientation of the phrases. The steps are as follows

Step1: Extract phrases containing adjectives or adverbs. Adjectives are the good indicators of subjective sentences. The part of speech tagger is used to extract two consecutive phrases from the review

Step2: The PMI-IR(Point Mutual Information-Information Retrieval) is used to obtain the semantic orientation of the extracted phrases. The PMI-IR algorithm uses mutual information as measure of the strength of semantic association between two words. The PMI-IR between two words $word_1$, $word_2$ is,

$$\text{PMI(word1,word2)} = \log_2 \left[\frac{p(word_1 \,\&\, word_2)}{p[(word]_1)\, p\,[(word]_2)} \right] \qquad (6)$$

$P(word_1 \,\&\, word_2)$ is the probability that $word_1$, $word_2$ co-occur. If the words are statistically independent, then the probability that they co-occur is given by the product $p(word_1)p(word_2)$. The semantic orientation of a phrase is calculated as,

SO (phrase) = PMI (phrase, "excellent")
- PMI (phrase, "poor")

SO is positive when phrase has s a good association, SO is negative when phrase has a bad association.

Step3: The third step is to calculate the average semantic orientation of the phrases. Then classify the review as recommended if the average is positive and otherwise negative

.

6. CROSS DOMAIN SENTIMENT CLASSIICATION

One of the main challenges for Sentiment Classification is the domain adaptation problem. That is, a sentiment classifier trained with the labeled data from one domain normally performs badly in another domain because it fails to learn the sentiment of unseen words [19]. The two main challenges in cross domain sentiment classification are. First we must identify which source domain features are related to target domain features. Second we require a learning framework to incorporate the information regarding the relatedness of source and target domain features. Various methods used in order to achieve sentiment classification in multiple domain. In [17], the task of sentiment classification is to learn an accurate classifier to predict the polarity of unseen sentiment data from $D_{tar.}$. It uses two subtasks (1) to identify domain independent features (2) to align domain-specific features. The mutual information criterion between features and domain as follows.

$$I(X^i;D)=\sum_{d\in D}\sum_{x\in X^i,x\neq 0} p(x,d)\log_2\left(p(x,d)/p(x)p(d)\right) \quad (7)$$

ALGORITHM

Input: labeled source domain data $D_{src}=\{(x_{src_i}, y_{src_i})\}_{i=1}^{n_{src}}$, unlabeled target domain data $D_{tar}=\{x_{tar_j}\}_{j=1}^{n_{tar}}$,the number of clusters K and the number of domain independent features m.

Output: adaptive classifier f: X→Y

Steps

1. Apply the criteria on D_{src}, D_{tar} to select l domain independent features. The remaining m-l features are treated as domain specific features.
   $\emptyset_{DI}=[\emptyset_{DI}(x_{src})]$ and $\emptyset_{DS}=[\emptyset_{DS}(x_{src})]$
   $\emptyset_{DI}(x_{tar})$ $\emptyset_{DS}(x_{tar})$
2. By using $\emptyset_{DI}$ and $\emptyset_{DS}$,calculate (DI-word)-(DS-word) co-occurrence matrix M$\in R^{(m-1)\times l}$s
3. construct matrix L=$D^{-1/2}AD^{-1/2}$, where A=$\begin{bmatrix}0 & M\\ M^T & 0\end{bmatrix}$
4. Find the K largest eigen vectors of L, $u_1,u_2,..,u_K$,and from the matrix U=$[u_1,u_2\ldots u_k]\in R^{m\times K}$Let mapping $\varphi(x_i)=x_iU_{[1:m-l,:]}$, where $x_i\in R^{m-l}$.
5. Return a classifier f, trained on $\left\{\left(\left[x_{src_i}\ \gamma\varphi\left(\varphi_{DS}(x_{src_i})\right)\right],y_{src_i}\right)\right\}_{i=1}^{n_{src}}$

7. APPLICATION

Sentiment analysis and classification is applied for many business intelligent applications. Many applications such as question answering make use of sentiment classification. Many organization use sentiment analysis as an internal work for marketing to improve their profits. Sentiment analysis can be applied for all fields directly or indirectly because opinions are the central to all human activities. It makes the decision making process easy.

8. FUTURE CHALLENGES

The challenges in sentiment classification and sentiment analysis are as follows. Sentiment classification can be categorized into single domain and cross domain. In single domain sentiment classification, a classifier is trained using labeled data of a particular domain, and the classifier is applied to the same domain for sentiment classification. Mostly sentiment classification is applied to labeled data where the labels are derived either through manual annotation effort. Thus labeled data sentiment classification is costly and requires human effort. Another challenge is entity based sentiment classification in both single and cross domain. A word has a positive sentiment in one situation and negative sentiment in another situation. Features also play a major role in single domain sentiment classification because “the laptop start up time was long” has negative context while “the laptop’s battery life is long” has positive context. In cross domain sentiment classification a classifier trained using labeled data for a particular domain is applied to classify reviews on different domain. Sentiment classification system to a new target domain in the absence of large amounts of labeled data may often results in low performance. Cross domain sentiment classification has received attention with the advancement in domain adaptation. Therefore, in the case of cross-domain classification,

it is also challenging to design a powerful classification algorithm which could fully take advantage of the unlabeled data [14].

9. CONCLUSION

The survey on different applications and potential challenges of sentiment classification in single and cross domain are presented. Subjectivity analysis, negation handling and polarity classification which are the tasks to be performed before sentiment analysis is investigated. Some of the supervised and unsupervised methods are discussed. The future challenge gives the way in for the upcoming research workers.

ACKNOWLDGEMENT

Our sincere gratitude to the Almighty for the blessings and thanks to our family members for their support.

References

[1] Abbasi, Ahmed, Hsinchun Chen, and Arab Salem. "Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums."*ACM Transactions on Information Systems (TOIS)* 26.3 (2008): 12.

[2] Benamara, Farah, et al. "Sentiment Analysis: Adjectives and Adverbs are Better than Adjectives Alone." *ICWSM*. 2007.wi

[3] ChandraKala, S., and C. Sindhu. "OPINION MINING AND SENTIMENT CLASSIFICATION: A SURVEY."

[4] Chawla, Nitesh V., and Grigoris I. Karakoulas. "Learning From Labeled And Unlabeled Data: An Empirical Study Across Techniques And Domains." *J. Artif. Intell. Res.(JAIR)* 23 (2005): 331-366.

[5] Dadvar, Maral, Claudia Hauff, and F. M. G. de Jong. "Scope of negation detection in sentiment analysis." (2011).

[6] Das, Amitava, and Sivaji Bandyopadhyay. "Subjectivity detection using genetic algorithm." *the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA10), Lisbon, Portugal, August*. 2010

[7] de Albornoz, Jorge Carrillo, Irina Chugur, and Enrique Amigó. "Using an Emotion-based Model and Sentiment Analysis Techniques to Classify Polarity for Reputation." *CLEF (Online Working Notes/Labs/Workshop)*. 2012.

[8] Devitt, Ann, and Khurshid Ahmad. "Sentiment polarity identification in financial news: A cohesion-based approach." *ACL*. 2007.

[9] Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." *CS224N Project Report, Stanford* (2009): 1-12.

[10] Hassan, Ahmed, et al. "Identifying the Semantic Orientation of Foreign Words."*ACL (Short Papers)*. 2011.

[11] Hatzivassiloglou, Vasileios, and Kathleen R. McKeown. "Predicting the semantic orientation of adjectives." *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 1997.

[12] Jia, Lifeng, Clement Yu, and Weiyi Meng. "The effect of negation on sentiment analysis and retrieval effectiveness." *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 2009.

[13] Kim, Soo-Min, and Eduard Hovy. "Determining the sentiment of opinions."*Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, 2004.

[14] Li, Shoushan, et al. "Active Learning for Cross-omain Sentiment Classification."

[15] Mejova, Yelena. "Sentiment Analysis: An Overview." *Comprehensive exam paper, available http://www.cs.uiowa.edu/~ymejova/publications/CompsYelenaMejova. pdf [2010-02-03]* (2009).

[16] Ormándi, Róbert, István Hegedűs, and Richárd Farkas. "Opinion mining by transformation-based domain adaptation." *Text, Speech and Dialogue*. Springer Berlin Heidelberg, 2010.

[17] Pan, Sinno Jialin, et al. "Cross-domain sentiment classification via spectral feature

alignment." *Proceedings of the 19th international conference on World wide web*. ACM, 2010.

[18] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 2002.

[19] Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis."*Foundations and trends in information retrieval* 2.1-2 (2008): 1-135.

[20] Turney, Peter D. "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews." *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002.

[21] Wiebe, Janyce M., Rebecca F. Bruce, and Thomas P. O'Hara. "Development and use of a gold-standard data set for subjectivity classifications."*Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Association for Computational Linguistics, 1999.

[22] Wiebe, Janyce, and Ellen Riloff. "Creating subjective and objective sentence classifiers from unannotated texts." *Computational Linguistics and Intelligent Text Processing*. Springer Berlin Heidelberg, 2005. 486-497.