classified and the hyper plane is as far as possible from the points closest to it. It has a drawback of dimensionality [6].

ii. Naive Bayes Method

Naive Bayes is suitable when the size of training data is less. It is a probabilistic classifier based on applying Bayes theorem with strong independence assumptions. Its running time is linear in the size of input. Maximum entropy classifiers do not assume statistical independence of the independent variables (as in Bayes classifier) that serve as predictors. Learning in this model is slower than for a Naive Bayes classifier, and thus may not be suitable if the number of classes to learn is very large [6].

iii. Maximum Entropy Classifiers

Learning in a Naive Bayes classifier involves counting the number of co-occurrences of features and classes. In a maximum entropy classifier the weights, which are typically maximized using Maximum-a-Posteriori (MAP) estimation, must be learned. Unsupervised methods make use of PMI (Point wise mutual information) for co-occurrence of a word with positive or negative word. Dictionary based methods can be used for expanding the set of positive and negative sentiment words and classifying the words. But the drawback here is that a word may be used positively in some domain and negatively in some other domain. Further, in case of sentence level classification, knowing the main aspect in the sentence is important otherwise sentence level classification becomes futile. Another issue is that, it is not always a case that a sentence will talk only about one topic or have only one opinion. In such cases, feature level classification is better than sentence level classification [6].

3. Lexicon based Techniques

In unsupervised technique, classification is done by comparing the features of a given text against sentiment lexicons whose sentiment values are determined prior to their use. Sentiment lexicon contains lists of words and expressions used to express people's subjective feelings and opinions. For example, start with positive and negative word lexicons, analyse the document for which sentiment need to find. Then if the document has more positive word lexicons, it is positive, otherwise it is negative. The lexicon based techniques to Sentiment analysis is unsupervised learning because it does not require prior training in order to classify the data [6].

The elementary steps of the lexicon based techniques are outlined below

1. Preprocess each text (i.e. remove HTML tags, noisy characters).
2. Initialize the total text sentiment score: s _ 0.
3. Tokenize text. For each token, check if it is present in a sentiment dictionary.

(a) If token is present in dictionary,
   i. If token is positive, then s _ s + w.
   ii. If token is negative, then s _ s − w.

4. Look at aggregate text sentiment score s,
   (a) If s > threshold, then classify the text as positive.
   (b) If s < threshold, then classify the text as negative.

There are three methods to construct a sentiment lexicon: manual construction, dictionary-based methods and corpus-based methods. The manual construction of sentiment lexicon is a difficult and time-consuming task.

In dictionary based techniques the idea is to first Collect a small set of opinion words manually with known attitudes, and then to grow this set by searching in the WorldNet dictionary for their synonyms and antonyms. The newly founded words are added to the seed list. The next iteration starts. The iterative process stops when no further new words are found. Opinion words share the same orientation as their synonyms and opposite orientations as their antonyms. Hu and Liu use this technique to find semantic orientation for adjectives. The dictionary based approach have a limitation is that it can't find opinion words with domain specific orientations.

Corpus based techniques rely on syntactic patterns in large corpora. Corpus-based methods can produce opinion words with relatively high accuracy. Most of these corpus based methods need very large labeled training data. This approach has a major advantage that the dictionary-based approach does not have. It can help find domain specific opinion words and their orientations. Ting-Chun Peng and Chia-Chun Shih uses part-of-speech (POS) patterns for extracting the sentiment phrases of each review, they used unknown sentiment phrase as a query term and get top-N relevant phrases from a search engine. Next, sentiments of unknown sentiment phrases are computed based on the sentiments of nearby known relevant phrase using lexicons.

The idea behind the centroid classification algorithm is extremely simple and straightforward (Songho tan, 2008). Initially the prototype vector or centroid vector for each training class is calculated, then the similarity between a testing document to all centroid is calculated, finally based on these similarities, document is assigned to the class appropriate to the most similar centroid. The K-nearest neighbor (KNN) is a typical example based classifier that does not build an explicit, declarative representation of the category, but depends on the category labels attached to the training documents similar to the test document. Given a test document d, the system finds the k nearest neighbors among training documents. The similarity score of each nearest neighbor document to the test document is used as the weight of the classes of the neighbor document (Songho tan, 2008).

Combining rule-based classification, machine learning and supervised learning into a new combined method. For each sample set, they carried out 10-fold cross validation. For each fold, the related samples were divided into training and a test dataset. For each test sample, a hybrid classification is carried out, i.e., if one classifier fails to classify a document, the classifier passes the document onto the next classifier, till the document is classified or no other classifier exists. Given a training dataset, the Rule Based Classifier (RBC) used a Rule Generator to generate a set of rules and a set of antecedents to represent the test sample and used the rule set derived from the training set to classify the test sample. If the

test sample was unclassified, the RBC forewords the associated antecedents onto the Statistic Based Classifier (SBC), if the SBC could not classify the test sample; the SBC passed the associated antecedents onto the General Inquirer Based Classifier (GIBC), which used the 3672 simple rules to determine the consequents of the antecedents. The SVM was given a training dataset to classify the test sample if the three classifiers failed to classify the same. An ensemble technique is one which combines the outputs of several base classification models to form an integrated output. Rui Xia (2011) used this approach and made a comparative study of the effectiveness of ensemble technique for sentiment classification by efficiently integrating different feature sets and classification algorithms to synthesize a more accurate classification procedure. In his work, two types of feature sets are designed for sentiment classification, namely the part-of-speech(POS) based feature sets and the word-relation based feature sets. Then, three text classification algorithms namely support vector machines ,naive Bayes and maximum entropy are employed as base-classifiers for each of the feature sets to predict classification scores. Three types of ensemble methods, namely the weighted combination, fixed combination and meta-classifier combination, are evaluated for three ensemble strategies namely ensemble of classification algorithms, ensemble of feature sets and ensemble.

4. Hybrid Techniques

Few research techniques have indicated that the combination of both the machine learning and the lexicon based approaches improve sentiment classification performance. Mudinas et al. presents concept-level sentiment analysis system, pSenti, which is developed by combining lexicon based and learning-based approaches. The main advantage of their hybrid approach using a lexicon/learning symbiosis is to attain the best of both worlds-stability as well as readability from a carefully designed lexicon, and the high accuracy from a powerful supervised learning algorithm. Their system uses a sentiment lexicon constructed using public resources for initial sentiment detection. Currently the sentiment lexicon consists of 7048 sentiment words including words with wildcards and sentiment values are marked in the range from − 3 to +3. They used sentiment words as features in machine learning method. The weight of such a feature is the sum of the sentiment value in the given review. For those adjectives which are not in sentiment lexicon, their occurring frequencies are used as their initial values.. Experiment results show that while a general purpose sentiment lexicon provides only minor accuracy improvement, incorporating Domain specific dictionaries leads to more significant improvement. With the help of the new opinionated indicators, additional Opinionated tweets can be identified. Afterwards, a sentiment classifier is trained to assign sentiment polarities for entities in the newly identified tweets[10].

5. Semantic Orientation based Techniques-

The Semantic orientation approach to Sentiment analysis is “unsupervised learning” because it does not require prior training in order to mine the data. Instead, it measures how far a word is inclined towards positive and negative. Much of the research in unsupervised sentiment classification makes use of lexical resources available. Kamps et al (2004) focused on the use of lexical relations in sentiment classification. Andrea Esuli and Fabrizio Sebastiani (2005) proposed semi-supervised learning method started from expanding an initial seed set using WordNet. Their basic presumption is terms with similar orientation tend to have similar glosses. They find out the expanded seed term’s semantic orientation through gloss classification by statistical technique. When the review where an opinion lies in, cannot provide sufficient contextual information to determine the orientation of opinion, Chunxu Wu(2009) proposed an approach which resort to other reviews discussing the same topic to mine useful related information, then use semantic similarity measures to judge the orientation of opinion. They attempted to tackle this problem by getting the orientation of context independent opinions , then consider the context dependent opinions using linguistic rules to infer orientation of context distinct-dependent opinion ,then extract contextual information from other reviews that comment on the same product feature to An unsupervised learning algorithm by extracting the sentiment phrases of each review by rules of part-of-speech (POS) patterns was investigated by Ting-Chun Peng and Chia-Chun Shih (2010). For each unknown sentiment phrase, they used it as a query term to get top-N relevant snippets from a search engine respectively. Next, by using a gathered sentiment lexicon, predictive sentiments of unknown sentiment phrases are computed based on the sentiments of nearby known sentiment words inside the snippets. They consider only opinionated sentences containing at least one detected sentiment phrase for opinion extraction. Using the POS pattern opinion extraction is done. Gang Li & Fei Liu (2010) developed an approach based on the k-means clustering algorithm. The technique of TF-IDF (term frequency – inverse document frequency) weighting is applied on the raw data. Then, a voting mechanism is used to extract a more stable clustering result. The result is obtained based on multiple implementations of the clustering process. Finally, the term score is used to further enhance the clustering result. Documents are clustered into positive group and negative group.

6. Role of Negation

Negation is a very common linguistic construction that affects polarity and therefore necessity to be taken into consideration in sentiment analysis. Negation is not only conveyed by common negation words (not, neither, nor) but also by other lexical units. Research in the field has shown that there are many other words that invert the polarity of an opinion expressed, such as valence shifters, connectives or modals. “I find the functionality of the new mobile less practical”, is an illustration for valence shifter, “Perhaps it’s a good phone, but I fail to see why”, shows the effect of connectives. An illustration sentence using modal is, “In theory, the phone should have worked even under water”. As can be noticed from these illustrations, negation is a difficult yet important aspect of sentiment analysis. Kennedy and Inkpen (2005) evaluate a negation model which is fairly

identical to the one proposed by Polanyi and Zaenen (2004) in document-level polarity classification. A simple scope for negation is chosen. A polar expression is thought to be negated if the negation word immediately precedes it. Wilson et al. (2005) carry out more advanced negation modeling on expression-level polarity classification. The work uses supervised machine learning where negation modeling is mostly encoded as features using polar expressions. Jin-Cheon Na (2005), reported a study in automatically classifying documents as expressing positive or negative. He investigated the use of simple linguistic processing to address the problems of negation phrase. In sentiment analysis, the most prominent work examining the impact of different scope models for negation is Jia et al. (2009). They proposed a scope detection method to handle negation using static delimiters, dynamic delimiters, and heuristic rules focused on polar expressions Static delimiters are unambiguous words, such as because or unless marking the beginning of another clause. Dynamic delimiters are, however, rules, using contextual information such as their pertaining part-of-speech tag. These delimiters suitably account for various complex sentence types so that only the clause containing the negation is considered. The heuristic rules focus on cases in which polar expressions in specific syntactic configurations are directly preceded by negation words which results in the polar expression becoming a delimiter itself.

Feature based Sentiment Classification Technique–

Due to the increasing amount of opinions and reviews on the internet, Sentiment analysis has become a hot topic in data mining, in which mining opinion features is a key step. Sentiment analysis at both the document level and sentence level has been too coarse to determine precisely what users like or dislike. In order to address this problem, sentiment analysis at the attribute level is aimed at mining opinions on products' specific attributes from reviews. Hu's work in (Hu, 2005) can be considered as the pioneer work on feature-based opinion summarization. Their feature mining algorithm is based on heuristics that depend on feature terms respective occurrence counts. They use association rule mining based on the Apriori algorithm to extract frequent item sets as explicit product features. Popescu et al (2005) developed an unsupervised information extraction system called OPINE, which mined product features and opinions from reviews. OPINE first mines noun phrases from reviews and retains those with frequency greater than an experimentally set threshold and then assesses those by OPINE‟s feature assessor for extracting explicit features. The assessor evaluates a noun phrase by calculating a Point-wise Mutual Information score between the phrase and meronymy discriminators associated with the product class. Popescu et al apply manual mining rules in order to find the opinion words. Khairullah Khan et al (2010) developed a method to find features of product from user review in an efficient way from text through auxiliary verbs (AV) {is, was, are, were, has, have, had}. From the results of the experiments, they found that 82% of features and 85% of opinion-oriented sentences include AVs. Most of existing methods utilize a rule-based mechanism or statistics to extract opinion features, but they failed to notice the structure characteristics of reviews. The performance is not good . Yongyong Zhail (2010) proposed a approach of Opinion Feature Extraction based on Sentiment Patterns, which takes into account the structure characteristics of reviews for higher values of precision and recall. With a self-constructed database of sentiment patterns, sentiment pattern matches each review sentence to obtain its features, and then filters unnecessary features regarding relevance of the domain, statistics and semantic similarity. Gamgarn Somprasertsri (2010) dedicated their work to properly identify the semantic relationships between product features and opinions. His approach is to mine product feature and opinion based on the consideration of syntactic information and semantic information by applying dependency relations and ontological knowledge with probabilistic based model.

Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, Jiajun Bu,Chun Chen, Xiaofei and Xifeng Yan proposed a techniques to to analyse and interpret the public sentiment variations in microblogging services. The Proposed system is based on two Latent Dirichlet Allocation (LDA) models to analyse tweets in significant variation periods, and conclude possible reasons for the variations.

7. Applications and Tools

Some of the applications of sentiment analysis includes hotspot detection in forums, online advertising etc. Online advertising has become one of the major revenue sources of today's Internet ecosystem. Sentiment analysis find its latest application in Dissatisfaction oriented online advertising Guang Qiu(2010) and Blogger-Centric Contextual Advertising (Teng-Kai Fan, Chia-Hui Chang ,2011), which refers to the assignment of personal ads to any blog page, chosen in according to bloggers interests.

In order to identify potential risks, it is important for companies to gather and analyse information about their competitors' products and plans. Sentiment analysis find a major role in competitive intelligence (Kaiquan Xu , 2011) to extract and visualize comparative relations between products from customer reviews, with the interdependencies between relations taken into consideration, to help enterprises determine potential risks and further design new products and marketing strategies.
Opinion summarization summarizes opinions of articles by telling sentiment polarities, degree and the correlated events. With opinion summarization, a customer can easily see how the existing customers feel about a product, and the product maker can get the reason why different stands people like it or what they complain about.. Other applications includes online message sentiment filtering-mail sentiment classification, web blog author's attitude analysis etc.

Red Opal is a tool that enables users to find products based on features. It scores each product based on features from the customer reviews (Christopher Scaffidi, 2007). Opinion observer is a sentiment analysis system for analysing and comparing opinions on the web. The system shows the results in a graph format showing opinion of the product feature by feature (Bing Liu, 2005).

Besides these automated tools, various online tools like SentiStrength, TwitterSentiment, Twitrratr, Twendz ,Social mention, and Sentimetrics are available to track the sentiment in social networks.

8. Challenges

Sentiment Analysis is a problem of text based analysis, but there are some challenges that make it difficult as compared to traditional text based analysis. Some of these are handling Negations, polysemy, slangs and domain generalization.

1. Mapping Slangs
Slangs are short forms of original words often used in online texting. Ex gr8 is slang for opinionated word great and 5n for fine. These not defined in traditional dictionaries but are often used in online texting. Grammatically incorrect words are generally filtered out as they are not found in the language lexicons. The results of sentiment analysis can be improved if slangs can be mapped to original words or opinionated slangs can be added to lexicons.

2. Handling Negation
If a negation occurs near an adjective the polarity is estimated to the opposite of the polarity of adjective. E.g. the resolution is good should be classified positive. The resolution is not good should be classified negative. To classify this, the polarity is set to opposite of polarity of the occurring adjective when accompanied by 'not'. But review, I don't say that the resolution is not good will be classified as negative when it should be classified as positive in a problem with two class labels and should be classified as neutral in a problem with three class labels. Mathematical model can address this problem. Polarity = (-1) n X Polarity of adjective. But again this model has limitations also for sentences like. No wonder the resolution is good.
Pure language processing techniques or purely mathematical models fail to handle negation.

3. Handling Polysemy
Polysemy is a word with multiple meanings. For example the adjective 'idle' means unemployed or unoccupied, not operating or not being used, not wanting to work or lazy and ineffective or powerless. Finding out the correct context sensitive meaning of polysemy is a challenging task.

4. Domain Generalization
Sentiment analysis is generally carried out targeting a particular domain for good results. But a generalized sentiment analyser remains a challenge because word / sentence that means positive in a domain may mean negative in other.

5. Language Generalization
Although a few sentiment analysers do classify reviews for Chinese and Vietnamese languages, most of them are implemented targeting English language. A broad view of opinion towards a product could be made available if Multilanguage opinion miners are developed.

6. Maintaining Opinion Time
Some sentiment analysers calculate and store opinion towards a product and anytime queried this result is returned. Over a period of time opinion towards a product may not remain the same.

7. Opinion Object Identification
The object towards which the opinion is expressed is mentioned in the first sentence or the early portions of the documents. The later parts may also have opinionated information about the object but the objected may referenced by pronoun like it, this and that.

8. Feature Matrix Construction
Opinion towards a product may be directly specified or a review may include opinion towards various features of product. In such cases opinion towards a product is actually an aggregation of opinion towards the features. Generally features for which the opinions should be calculated are predefined but the reviewers may also comment on other features. This opinion is not counted but it should be. This can be achieved by construction of dynamic feature matrix.

9. Hidden Sentiments Identification
Identifying opinion towards an object is classification problem but identifying hidden sentiments like anger, disgust and joy is a challenging task.

10. Updating / Down-dating Lexicons
Lexicons should be updated to accommodate new words and suppress the words which are no more used. Performance of sentiment analyzer depends majorly on the accuracy of the lexicon. Larger the lexicon, the results are accurate. Smaller the lexicon, the performance is faster. So lexicon should be fitted properly as per requirement of the hour

9. Analysis and Comparison

Supervised machine learning techniques have shown relatively better performance than the unsupervised lexicon based methods. However, the unsupervised methods is important too because supervised methods demand large amounts of labeled training data that are very expensive whereas acquisition of unlabeled data is easy. Most domains except movie reviews lack labeled training data in this case unsupervised methods are very useful for developing applications. Most of the researchers reported that Support Vector Machines (SVM) has high accuracy than other algorithms. The main limitation of supervised learning is that it generally requires large expert annotated training corpora to be created from scratch, specifically for the application at hand, and may fail when training data are insufficient. The opinion words that are included in the dictionary are very important for the lexicon based approach. If the dictionary contains less words or thorough, one risks the chance of over or under analysing the results, leading to a decrease in performance. Another significant challenge to this approach is that the polarity of many words is domain and context dependent. For example, 'funny movie' is positive in movie domain and 'funny taste' is negative in food domain. Such words are associated with sentiment in a particular domain.

Current sentiment lexicons do not capture such domain and context sensitivities of sentiment expressions. Without a comprehensive lexicon, the sentiment analysis results will suffer. The lexicon-based approach can result in low recall for sentiment analysis. The main advantage of hybrid approach using a lexicon/learning combination is to attain the best of both worlds, high accuracy from a powerful supervised learning algorithm and stability from lexicon based approach. There is need develop techniques to analyse and interpret the public sentiment

10. Conclusion

Among the surveyed techniques many techniques use a lexicon for opinion aggregation and some use training set and some use both. Different types of techniques should be combined in order to overcome their individual drawbacks and benefit from each other's merits, and enhance the sentiment classification performance. Further no work is done on the sentiment variations and the reasons behind the sentiment variations .So, there is a need to develop efficient techniques to analyse and interpret the public sentiment variations and reasons behind the variations in Microblogging services.

References

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inform. Retrieval*, vol. 2, no. (1–2), pp. 1–135, 2008.
[2] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment in twitter events," *J. Amer. Soc. Inform. Sci. Technol.*, vol. 62, no. 2, pp. 406–418, 2011.
[3] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," Computational Linguistics, vol. 37, 2011, pp. 267-307
[4] M. Hu and B. Liu, "Mining and summarizing customer reviews," Proceedings of the tenth ACM international conference on Knowledge discovery and data mining, Seattle, 2004, pp. 168-177.
[5] Chaovalit,Lina Zhou, Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches, Proceedings of the 38th Hawaii International Conference on System Sciences – 2005.
[6] Hajmohammadi, M., Ibrahim, R., Ali Othman, Z.. Opinion Mining and Sentiment Analysis: . International Journal of Computers & Technology, North America, 2, jun. 2012
[7] Keke Cai; Spangler, S.; Ying Chen; Li Zhang; , "Leveraging Sentiment Analysis for Topic Detection," Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08. IEEE/WIC/ACM International Conference on , vol.1, no., pp.265-271, 9-12 Dec. 2008.
[8] Sowmya Kamath S, Anusha Bagalkotkar, Ashesh Khandelwal, Shivam Pandey, Kumari Poornima"Sentiment Analysis Based Approaches for Understanding User context in web content". International conference on Communication systems and network technologies.
[9] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R., Sentiment analysis of Twitter data. In Proceedings of the Workshop on Languages in Social Media (LSM '11). Association for Computational Linguistics, Stroudsburg, PA, USA, 30-38, 2011.
[10] Bahrainian, S.A., Dengel, A., Sentiment Analysis using Sentiment Features, In the proceedings of WPRSM Workshop and the Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence, Atlanta, USA, 2013.
[11] A. Mudinas, D. Zhang, M. Levene, "Combining lexicon and learning based approaches for conceptlevel sentiment analysis", Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, ACM, New York, NY, USA, Article 5, pp. 1-8, 2012

Author Profile

Pankaj Bhalerao Research Scholar RMD Sinhgad of Engineering, Savitribai Phule Pune University. He Received B.E. in Computer Engineering from University of Pune. Currently he is pursuing M.E. in Computer Engineering from RMD Sinhgad School of Engineering, Warje, Savitribai Phule Pune University, Pune.

Prof. Trupti Dange received the B.E. and M.Tech in Computer Engineering from University of Mumbai. She is working as Assistant Professor in Department of Computer Engineering, RMD Sinhgad School of Engineering Pune, India. She is having more than four year experience.