*J. Basic. Appl. Sci. Res.*, 4(3)181-186, 2014
© 2014, TextRoad Publication
ISSN 2090-4304
**Journal of Basic and Applied Scientific Research**
*www.textroad.com*

# A Review of Feature Extraction in Sentiment Analysis

**Muhammad Zubair Asghar[1], Aurangzeb Khan[2], Shakeel Ahmad[1], Fazal Masud Kundi[1]**

[1]Institute of Computing and Information Technology, Gomal University, D.I.Khan, Pakistan,
[3]Institute of Engineering and Computer Sciences, University of Science and Technology Bannu, Pakistan,

*Received: January 18 2014*
*Accepted: February 18 2014*

## ABSTRACT

**Abstract:** Rapid increase in internet users along with growing power of online review sites and social media has given birth to Sentiment analysis or Opinion mining, which aims at determining what other people think and comment. Sentiments or Opinions contain public generated content about products, services, policies and politics. People are usually interested to seek positive and negative opinions containing likes and dislikes, shared by users for features of particular product or service. Therefore product features or aspects have got significant role in sentiment analysis. In addition to sufficient work being performed in text analytics, feature extraction in sentiment analysis is now becoming an active area of research. This review paper discusses existing techniques and approaches for feature extraction in sentiment analysis and opinion mining. In this review we have adopted a systematic literature review process to identify areas well focused by researchers, least addressed areas are also highlighted giving an opportunity to researchers for further work. We have also tried to identify most and least commonly used feature selection techniques to find research gaps for future work.
**KEYWORDS:** Sentiment Analysis, Feature Extraction, Opinion Mining, Feature Selection, Text Mining.

## INTRODUCTION

Sentiment is a view, feeling, opinion or assessment of a person for some product, event or service [1, 2, 3].Sentiment Analysis or Opinion Mining is a challenging Text Mining and Natural Language Processing problemfor automatic extraction, classification and summarization of sentiments and emotions expressed in online text [1,2].Sentiment analysis is replacing traditional and web based surveys conducted by companies for finding public opinion about entities like products and services [1]. Sentiment Analysis also assists individuals and organizations interested in knowing what other people comment about a particular product, service topic, issue and event to find an optimal choice for which they are looking for.

By the end of 2013,over 181 million blogs were tracked with 6.5 million personal blogs and 12 million blogs written on social networks with majority of users seeking opinions on products and services[3, 4].Sentiment analysis is of great value for business intelligence applications, where business analysts can analyze public sentiments about products, services, and policies[5].Sentiment Analysis in the context of Government Intelligence aims at extracting public views on government policies and decisions to infer possible public reaction on implementation of certain policies [6].

Feature based sentiment analysis include feature extraction, sentiment prediction, sentiment classification and optional summarization modules [9]. Feature extraction identifies those product aspects which are being commented by customers, sentiment prediction identifies the text containing sentiment or opinion by deciding sentiment polarity as positive, negative or neutral and finally summarization module aggregates the results obtained from previous two steps. Feature extraction process takes text as input and generates the extracted features in any of the forms like Lexico-Syntactic or Stylistic, Syntactic and Discourse based [7, 8].

In this review, we focus on state-of-art paradigms used for feature extraction in sentiment analysis. We will discuss and evaluate existing techniques in intuitive way, which was not found in previous reviews and finally issues and challenges to be solved in this area are highlighted. The rest of the paper is divided into different sections. Section II discusses methodology adopted for the review process; Section III gives framework for feature extraction in sentiment analysis, presenting the strengths and weaknesses of the existing methods. Section IV evaluates and discusses issues and challenges faced in feature extraction. Section V concludes and reports opportunities for further research.

## METHODOLOGY

In this section we present methodology adopted for conducting literature review, steps taken for this purpose include searching, criteria for inclusion & exclusion and presentation.

### *Searching*

We have searched the required papers having frequent citations from different online repositories including IEEE-explore, Google Scholar, ACM, Springer and Science Direct. Key words like opinion mining, feature extraction in opinion mining, feature extraction in sentiment analysis, metrics for feature extraction in opinion mining etc. were used as search terms, which resulted in access to all of our potential required papers.

### *Inclusion Criteria*

Papers and doctoral thesis published from 2011 to 2012 are included only, as publications prior to these dates are discussed in earlier surveys. Most cited publications prior to 2001 are also considered and included in this survey, as they provide basic and clear understanding to issues under consideration.

### *Exclusion Criteria*

Abstracts, editorials and unpublished material like reports and thesis are excluded. Non-English or publications having no English translation are also not considered.

### *Presentation*

In first phase, titles and abstracts are taken from above listed electronic repositories on the basis of key word searching and topic relevance. Selected abstracts are discussed in group meeting and a list of eligible publications is

***Corresponding Author:** Muhammad Zubair Asghar, Institute of Computing and Information Technology, Gomal University, D.I.Khan, Pakistan.

prepared. Each publication included in the eligibility list is downloaded and reviewed by conducting detailed study of methodology, results and conclusion sections are also evaluated. Findings extracted from all sections of each publication are summarized in excel work sheets, supported by graphs for visual analysis. Thorough critics made on each cited paper helped in identifying further research directions for future researchers.

## REVIEW FRAMEWORK OF FEATURE EXTRACTION IN SENTIMENT ANALYSIS

This section presents survey of the related work performed on feature extraction in Sentiment Analysis. We have reviewed more than sixty publications and categorically summarized their main techniques and contributions in different sections. Major feature extraction and manipulation steps and techniques, identified from cited publications are summarized in below sections.

### *Pre Processing*

Pre-processing analyzes the opinions from syntactical point of view and original syntax of sentence is not disturbed[22].In this phase, the several techniques like POS tagging, Stemming and Stop word removal are applied to data set for noise reduction and facilitating feature extraction.

*Part of speech (POS) tagging:* Parts of speech or POS tagging is a linguistic technique used since 1960 and has recently got particular attention of NLP researchers [9, 12, 23, 24, 25, 26, 27,28, 29] for product feature extraction as product aspects are generally nouns or noun phrases. POS tagging [22] assigns a tag to each word in a text and classifies a word to a specific morphological category such as norm, verb, adjective, etc. POS taggers are efficient for explicit feature extraction in terms of accuracy they achieved [9, 24, 27, 30,], however problem arises when review contains implicit features [9].

Hidden Markova Models are widely used for developing POS taggers due to accuracy as compared to other techniques like rule based, statistical and machine learning[1].Different English language POS taggers like NL Processor linguistic parser, Stanford POS tagger, Gate ANNIE POS Tagger and Claws POS tagger are used for this purpose. Python based NLTK toolkit [65] has a rich collection of all modules including POS, needed by NLP researchers and text miners. ICTCLAS is a Chinese lexical analyzer for performing POS tagging and many other functions. GENIA tagger [66] is specially designed for biomedical text tagging such as MEDLINE abstracts.

*Stemming and Lemmatization:* Stemming and Lemmatization are two essential morphological processes of pre-processing module during feature extraction [9, 23, 24, 25,26, 28,31,32].The stemming process converts all the inflected words present in the text into a root form called a stem[22]. For example, ‘automatic,’ ‘automate,’ and ‘automation’ are each converted into the stem ‘automat.’ Stemming gives faster performance in applications where accuracy is not major issue [33].

The first stemmer was published by Julie Beth Lovins in 1968.Martin Porter designed and published his stemmer in the July 1980. Porter and Lancaster are the stemming algorithms, supported by python NLTK.RSLP Stemmer[1], ISRI Stemmer[2] and SnowballStemmer[3]are non-English plugins.

The lemma of a word includes its base form plus inflected forms [34]. For example the words “plays”, “played and “playing” have “play” as their lemma. Lemmatization groups together various inflected forms of word into a single one[35].Stemming removes word inflections only whereas; Lemmatization replaces words with their base form. For example, the words “caring” and “cars” are reduced to “car” in a stemming process whereas lemmatization reduces it to “care” and “car” respectively, hence lemmatization is considered to be more accurate.

Unlike stemming, lemmatization needs additional dictionary support for searching and indexing, which enhances its accuracy in feature extraction applications, but degrades speed of Lemmatizer [36]. Word Net Lemmatizer with Word Net Database is used to lookup for lemmas; while Bio Lemmatizer is used for lemmatization of biomedical text [33, 37]. Balahur and Montoya [38] in their work extracted product features using Word Net; whereas Concept Net [39] was used for extracting detail on new technological products.

*Stop Word Removal:* Stop word concept was first introduced by Hans Luhn, H.P[40]. Stop words are common and high frequency words like “a”, “the”, “of”, “and”, “an”. Different methods available for stop-word elimination [22]; ultimately enhance performance of feature extraction algorithm[9,10, 22,24,35].

The stop words removal reduces dimensionality of the data sets and thus key words left in the review corpus can be identified more easily by the automatic feature extraction techniques. Words to be removed are taken from a commonly available list of stop words. Savoy [41] had given huge collection stop word list. At simplest level stop words are iterated in chosen word list and removed from text. This technique can be implemented by using languages like Java, python, Perl, supported by machine learning toolkits like NLTK, WEKA and GATE.

### *Feature Categorization*

Different types of features, identified from literature review on sentiment analysis are categorized as under.

*Morphological Types:* There are three types of morphological features i.e. semantic, syntactic and lexico structural [2, 7, 8]. Semantic type of features works on contextual information and semantic orientation (SO). Contextual information method is used to add text at sentence level [44]. On the other hand, semantic orientation (SO) technique make use of latent semantic analysis (LSA) and point wise mutual information (PMI), which assigns polarity score to every word or phrase [42, 43]. Syntactic class of feature use POS tagging, chunk labels, dependency depth feature and Ngram word. Lexico structural feature consist of special symbol frequencies, word distributions and word level lexical features, rarely used in opinion mining [8].

*Frequent Features:* Frequent product features, also called hot features, are the features in which people have more interest [9]. Apriori Association rule mining also called frequent pattern mining [45] is widely used in text mining

[1]*Stemmer for Portuguese,* [2]*Arabic, and* [3]*Non English*

literature [9, 10,24, 46] for frequent feature extraction in transactional data. The Apriori algorithm operates in twophases. In the first phase, it finds all frequent item sets in the transaction database that fulfill a user-defined

threshold. In phase two rules are generated from identified frequent item sets. Hu and Liu [9] run the associate rule miner CBA [47], which just uses first step of Apriori Algorithm. Bei F et al [48] designed clustering technique on the basis of frequent pattern mining and the work of [49] focused on phrase extraction for frequent feature selection.

Major problem associated with clustering based frequent feature selection approaches is their domain dependency in terms of heuristics and threshold setting [50].

*Implicit Features:* Implicit features are the features which are not apparent in review [33]. For example, in review "The hotel is expensive", user is referring to the feature "price" although word "price" is not clearly mentioned. Adjectives and adverbs are the most common implicit feature indicators [22, 51]. For example, the adjective "heavy" shows the weight feature, but it needs certain level of domain knowledge for reviews like "the traffic is heavy", where the adjective "heavy" doesn't represent weight.

Zhang W et al [27] extended association rule mining adopted by [9] by introducing collocation selection method for implicit feature identification in Chinese corpus. They found that implicit features occur very close to the explicit features**.** Ghani R et al [52] used classifier to determine whether certain feature is discussed implicitly in review or not. In addition to explicit feature extraction, Wang and H. F [15] identified implicit features by using a mapping function from opinion words to product features.

***Feature Selection***

Proper feature selection techniques in sentiment analysis have got significant role for identifying relevant attributes and increasing classification accuracy [53].Feature selection methods are grouped into four main categories NLP or heuristic based, Statistical, Clustering based and Hybrid.

Natural language processing based techniques mainly operate on three basic principles:(a) Noun, noun-phrases, adjectives, adverbs usually express product features[9,12,23,24].(b)Terms occurring near subjective expressions can act as features[57]. (c) P is product and F is feature in phrases like 'F of P' or 'P has F' [58]. They have got high accuracy, but low recall with dependency on accuracy of part of speech of tagging. Clustering or Machine Learning based feature extraction techniques are implemented by [18, 21, 27,59, 62,63, 64], requiring few parameters to tune[50]. Key weakness of clustering is that only major features can be extracted and it is difficult to extract minor features [13].

Statistical techniques are further divided into three sub types, univariate, multivariate and hybrid[8,53,54]. Univariate methods, also called feature filtering methods, take attributes separately, examples of this type include information gain (IG), chi–square, occurrence frequency, log likely-hood and minimum frequency thresholds. Univariate techniques have computational efficiency, but they ignore attribute interactions [53].Decision tree models, recursive feature elimination and genetic algorithms are the examples of multivariate methods, which consider group of attributes and use wrapper model for attribute selection [54]. As compared to univariate, multivariate methods are expensive in terms of computational efficiency, as they evaluate attribute interactions. Hybrid techniques combine univariate multivariate and other methods for achieving accuracy and efficiency [8, 55,56].

Hu et al [9] applied hybrid techniques such as POS Tagging with WordNet dictionary. Frequent aspect set identification was performed using Association Miner CBA. Compactness and redundancy pruning methods were used for eliminating irrelevant features. In contrast to [9], Ly Dk et al [10] incorporated Sentence level syntactic information for isolating real product features from a bulk of unnecessary features using Stanford Dependency Parser [11]. Somprasertsri G et al [12] combined lexical and syntactic features with a maximum entropy model for product feature extraction. Zhang H et al [13] combined association rules and point-wise mutual information for identifying product features with added advantage of utilizing HowNet [14] sentiment dictionary. Wang and H. F [15] identified product features by using bootstrapping iterative learning strategy with additional linguistic rules for extracting low occurring features and opinion words. Zhang et al. [16] proposed two steps for product feature mining. In first step features were extracted by using part-whole relation patterns and a "no" pattern for performance enhancement. In second step, they ranked feature candidates by feature importance.

Features were extracted by Conditional Random Fields [17, 18, 19] and Maximum Entropy Model [12, 20]. Hadano M et al [21] used Bayon[1] clustering algorithm for feature identification, high accuracy was achieved when they compared results with baseline method.

***Feature Cleansing***

Large numbers of unnecessary features are produced during frequent feature set generation phase, which need to be removed. Feature cleansing process removes such surplus features by applying feature pruning algorithms. Irrelevant features are eliminated by [9, 10] using compactness pruning method. Jeong et al [25] combined features with identical characteristics and then most representative features are highlighted out of candidate feature set, at last redundant features are cleansed. Wie et al [30] proposed different cleansing rules for noise reduction from huge feature set. During first step candidate features are removed from given feature set by using part of speech(POS) mapping. In next step noun and noun phrases are cleansed. In final step, sentences with nearest adjective acting as feature terms are identified for redundancy elimination.

We have reviewed more than hundred papers and categorized the papers according to the use of feature selection algorithms for sentiment analysis, graph in figure 1 shows that NLP based feature selection techniques are used more frequently than other techniques, hybrid techniques have not become much popular, still more work is required to be performed to check their efficiency.

[1]http://code.google.com/p/bayon

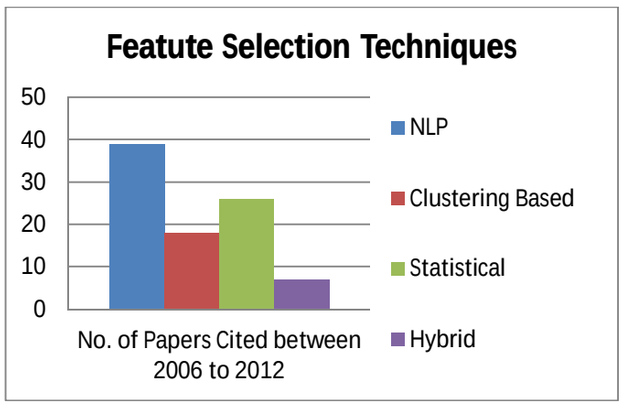


Figure 1 Usage of Feature Selection Techniques

## ISSUES AND CHALLENGES

Feature extraction in sentiment analysis is facing different issues like large feature space problems, redundancy, domain dependency, difficulty in implicit feature identification, and limited work on Lexico-structural features. Following are the general challenges in feature extraction, identified by different researchers [8,13, 54].

- High Dimensionality: High dimensionality or large feature sets cause performance degradation due to computational problems and thus proper feature selection methods are essentially required [54].
- Redundancy: N-grams are highly redundant, causing redundancy problems in both univariate and multivariate methods [54]
- Domain dependency: Performance of clustering based feature extraction techniques is domain dependent, creating cross domain and generalization problems[59].
- Hybrid method's performance: Ability of hybrid methods to overcome problems arising from redundancy, is still not confirmed, needing further experiments [54].
- Lexico-structural features: Unlike syntactic and semantic features, limited work is carried out on Lexico-structural features in feature extraction algorithms [8].
- POS tagging problem: Accuracy of heuristic based feature selection techniques depend on the accuracy of POS tagging [56], so designing an efficient POS staggers is still an issue to be resolved, especially for non-English languages.

## V. EVALUATION AND DISCUSSION

In this section we focus on the research gaps found during review of existing studies on feature extraction methods for sentiment analysis. Figure 2 shows that most of the research in feature extraction for sentiment analysis is focused on explicit feature extraction, syntactic and semantic features. Less work is performed on implicit feature extraction, Lexico-structural features. Graph in Figure 2 shows the following research gaps for future researchers.

1)Hidden Markov Model, due to its accuracy, is the most widely used technique for POS tagging as compared to rule based, statistical and machine learning algorithms. There is potential for research to be performed on developing non rule based POS taggers for non-English languages like Urdu and other regional languages of Pakistan, as work in these languages is still not mature enough.

2)In contrast to stemming, researchers prefer to use lemmatization for Inflection removal as it gives more accurate results [34]. Development and evaluation of lemmatizes for un-segmented languages like Urdu and Thai is still in its infancy, needing further work, as these languages have no clear boundaries [60,61].

3) Less work is performed on implicit features identification as they are comparatively hard to identify than explicit features [9, 15, 27], creating an opportunity for further work. Lexico-structuralor stylistic features are explored little for sentiment analysis, further research is needed to verify their effectiveness in product and movie reviews domain[8,54].

4)Most of existing features selection methods are unable to address relevant features in redundant feature space, so dimensionality reduction of huge feature spaces is another target area for further research[54].

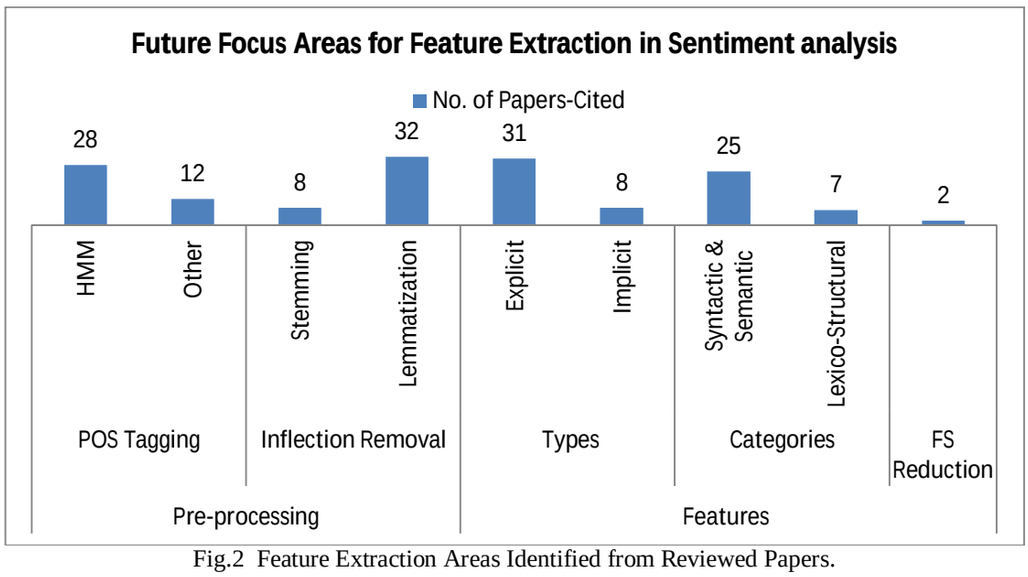


Fig.2 Feature Extraction Areas Identified from Reviewed Papers.

## VI.CONCLUSION

Feature extraction in sentiment analysis is an emergent research field so in this paper we have concentrated on related work performed in this area to identify directions for future work. As described in section III, many feature selection techniques, NLP based, Machine learning or clustering based, Statistical, Hybrid, are discussed. Features are categorized as syntactic, semantic, lexico-structural, implicit, explicit and frequent, making it easy for future researchers to work on. Different pre-processing modules like POS tagging, stop word removal, stemming and lemmatization are discussed with potential research areas focused on. Finally we conclude that feature space reduction, redundancy removal and evaluating performance of hybrid methods of feature selection can be the future direction of research work for all researchers in the field of feature extraction in sentiment analysis.

**Acknowledgment**:

The authors declare that they have no conflicts of interest in this research.

## VI. REFERENCES

1. Indurkhya N., Damereau F.J., (Eds). 2010. Handbook of Natural Language Processing. 2nd Ed., Chapman & Hall/CRC, Boca Raton.
2. K. Khan, B.B. Baharudin, A. Khan, F. e-Malik, Mining opinion from text documents: a survey, Proc. of the 3rd IEEE International Conference on Digital Ecosystems and Technologies, pp.217-222, 2009.
3. Bo Pang and Lillian Lee, Opinion mining and sentiment analysis, , Foundations and Trends in Information Retrieval, Vol. 2, No 1-2 (2008) 1–135, c, 2008.
4. NM Incite, Social Media Intelligence Company, http://www.nmincite.com/?page_id=210, last accessed 18, April, 2013.
5. A. Funk, Y. Li, H. Saggion, K. Bontcheva, and C. Leibold. Opinion analysis for business intelligence applications. In First international workshop on Ontology-Supported Business Intelligence (at ISWC), Karlsruhe, October 2008. ACM.
6. Stylios George et al, *"Public Opinion Mining for Governmental Decisions" Electronic* Journal of e-Government Volume. 8 Issue 2 2010, (pp203-214)
7. Amitava Das et al , 2008, Topic-Based Bengali Opinion Summarization, Coling 2008: Poster Volume, pages 232–240, Beijing, August 2010.
8. Ahmed Abbasi, etl. "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums" ACM Transactions on Information Systems, Vol. 26, No. 3, Article 12, 2008.
9. Hu, M., and Liu, B. 2004. Mining Opinion Features in Customer Reviews. *AAAI'04*, 2004.
10. D. K. Ly, K. Sugiyama, L. Ziheng, and M.-Y. Kan. Product Review Summarization from a Deeper Perspective. In Proc. of the 11th CM/IEEE Joint Conference on Digital Libraries(JCDL 2011), pages 311–314, 2011.
11. Stanford Dependency Parser http://nlp.stanford.edu/software/lex-parser.shtml
12. Somprasertsri, G., &Lalitrojwong, P. (2008). Automatic Product Feature Extraction from Online Product Reviews using Maximum Entropy with Lexical and Syntactic Features. In *IEEE International Conference on Information Reuse and Integration*, Las Vegas, 250-255
13. H. Zhang, Z. Yu, M. Xu, & Y. Shi. Feature-level sentiment analysis for Chinese product reviews. In *3rd International Conference on Computer Research and Development*, pages 135-140, Shanghai, 2011.
14. Z. Dong and Q. Dong. HowNet And The Computation Of Meaning. World Scientific, 2006.
15. Wang, Bo., Houfeng Wang. Bootstrapping both Product Features and Opinion Words from Chinese Customer Reviews with Cross-Inducing In *Proceedings of IJCNLP 2008.*
16. L. Zhang, S.H. Lim, B. Liu, and E. O'Brien-Strain. Extracting and Ranking Product Features in Opinion Documents. Proceedings of COLING. 2010.
17. Xu B et al, Product Features Mining Based On Conditional Random Fields Model, Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, Qingdao, 11-14 July 2010.
18. Zhang S et al, Product Features Extraction and Categorization in Chinese Reviews, ICCGI 2011 : The Sixth International Multi-Conference on Computing in the Global Information Technology.
19. J. Lafferty, A. McCallum and F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of ICML. 2001.
20. H. L. Chieu and H. T. Ng. 2002. Named entity recognition: a maximum entropy approach using global information. In Proceedings of the 19th Coling, pages 190–196.
21. Hadano Masashi, Kazutaka Shimada, and Tsutomu Endo. Aspect identification of sentiment sentences using a clustering algorithm (in japanese). In Proceedings of FIT 2010.
22. C. C. Aggarwal and C.-X. Zhai, *Mining Text Data*, Springer, 2012.
23. Kobayashi N et al, Extracting Aspect-Evaluation and Aspect-of Relations in Opinion Mining, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1065–1074, Prague, June 2007.
24. Archak, N., Ghose, A., AndIpeirotis, P. G. 2007. Show me the money!: deriving the pricing power of product features by mining consumer reviews. In *KDD '07:* Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, New York, NY, USA, 56–65.
25. Jeong H et al, FEROM: Feature Extraction and Refinement for Opinion Mining, ETRI Journal, Volume 33, Number 5, October 2011.
26. Mishne, G. Experiments with mood classification in blog posts. In Workshop on Stylistic Analysis of Text for Information Access, August, 2005
27. Zhang, W., et al. Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis. Expert Systems with Applications (2012), http://dx.doi.org/10.1016/j.eswa.2012.02.166
28. Sara Stymne et al, Pre- and Post-processing for Statistical Machine Translation into Germanic Languages, Proceedings of the ACL-HLT 2011 Student Session, pages 12–17,Portland, OR, USA 19-24 June 2011.
29. Elena Lloret et al, Experiments on Summary-based Opinion Classification, Proceedings of the NAACL HLT 2010.
30. Wei, W., Liu, H., He, J., Yang, H., Du, X.: Extracting Feature and Opinion Words Effectively from Chinese Product Reviews. In: Proceedings of the 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery, vol. 4,pp. 170–174 (2008).

31. Danuta Ploch, Exploring entity relations for named entity disambiguation. In Proc. of ACL 2011Student Session, pages 18–23.
32. Lloret, E., Saggion, H. & Palomar, M. Experiments on summary-based opinion classification, *in* 'Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysisand Generation of Emotion in Text', p. 107115.
33. Mining Product Opinions and Reviews on the Web, Jordão F and Brazil C, 2010, Master Thesis.
34. D. M. Christopher, R. Prabhakar, and S. Hinrich.*Introduction to Information Retrieval*. Cambridge University Press, 2008.
35. IBM Cognos Content Analytics Information Center, available at http://publib.boulder.ibm.com/infocenter/analytic/v2r1m0/index.jsp?topic=%2Fcom.ibm.discovery.es.ta.doc%2Fiiysalgstopwd.htm
36. Opinion Mining Of Political Views, By: Sheenam Bajaj, Department of Computer Science The University of Sheffield, Master thesis, 2011.
37. BioLemmatizer: a lemmatization tool for morphological processing of biomedical text. Liu H, Christiansen T, Baumgartner WA, Verspoor K. Journal of Biomedical Semantics. 2012 Apr 01;3:3./doi:10.1186/2041-1480-3-3.
38. Balahur, A. and Montoyo, A, A Feature Dependent Method for Opinion Mining and Classification. International Conference on Natural Language Processing and Knowledge Engineering, Oct 2008, China.
39. ConceptNet: A Practical Commonsense Reasoning Toolkit. BT Technology Journal 22 (2004).
40. Luhn, H.P, The automatic creation of literature abstracts. IBM Journal of Research and Development, 1958, pp 159-165.
41. Savoy, J. (2005). IR Multilingual Resources at UniNE. Available at http://members.unine.ch/jacques.savoy/clef/index.html. Last access - 21/01/2011.
42. Turney, P. D. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, In Proceedings of the 40th Annual Meetings of the Association for Computational Linguistics,Philadelphia, Pennsylvania, 417-424.
43. P. D. Turney and M. L. Littman, "Measuring praise and criticism: Inferenceof semantic orientation from association," *ACM Transactions on Information Systems (TOIS)*, vol. 21, pp. 315–346, 2003.
44. Riloff, E., Wiebe, J., &Wilson, T. (2003). Learning subjective nouns using extraction pattern bootstrapping. *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003)*.
45. R. Agrawal and R. Srikant. Fast algorithms for mining association rules. Proc. 20th Int. Conf. Very Large Data Bases, VLDB, 1215:487–499, 1994.
46. Zhang H, Yu Z, XU M, Shi Y, Feature-level Sentiment Analysis for Chinese Product Reviews, IEEE, 2011.
47. Liu, B., Hsu, W., Ma, Y. 1998. Integrating Classification and Association Rule Mining. KDD-98, 1998.
48. F. Beil, M. Ester, X. Xu. Frequent term-based text clustering, ACM KDD Conference, 2002.
49. Y.-B. Liu, J.-R. Cai, J. Yin, A. W.-C. Fu. Clustering Text Data Streams, Journal of Computer Science and Technology, Vol. 23(1), pp. 112–128, 2008.
50. Opinion-Based Entity Ranking, Ganesan, Kavita A., and ZhaiChengXiang , Information Retrieval, Volume 15, Issue 2, (2012)
51. Liu, B., Hu, M., Cheng, J.: Opinion observer: analyzing and comparing opinions on the web. In: WWW '05: Proceedings of the 14th international conference on World Wide Web, New York, NY, USA, ACM (2005) 342-351.
52. Asghar, M. Zubair, et al. "Systemized Approach for Software Corrective Maintenance Effort Reduction." (2011).
53. Peter Koncz and Jan Paralic, An approach to feature selection for sentiment analysis,, INES 2011 • 15th International Conference on Intelligent Engineering Systems • June 23–25, 2011, Poprad, Slovakia.
54. A. Abbasi, et al., "Selecting Attributes for Sentiment Classification Using Feature Relation Networks," IEEE Transactions on Knowledge and Data Engineering, vol. 23, pp. 447-462, 2011.
55. P.V. Balakrishnan, R. Gupta, and V.S. Jacobs, "Development of Hybrid Genetic Algorithms for Product Line Designs," IEEE Trans. Systems, Man, and Cybernetics, vol. 34, no. 1, pp. 468-483, Feb. 2004.
56. H. Liu and L. Yu, "Toward Integrating Feature Selection Algorithms for Classification and Clustering," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 4, pp. 491-502, Apr. 2005.
57. S.-M. Kim and E. Hovy, "Automatic identification of pro and con reasons in online reviews," in Proceedings of the COLING/ACL Main Conference Poster Sessions, pp. 483–490, 2006.
58. Popescu, A.-M. And Etzioni, O. 2005. Extracting product features and opinions from reviews. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Morristown, NJ, USA, 339–346.
59. Y. Lu, C. Zhai, and N. Sundaresan. Rated aspect summarization of short comments. In Proceedings of the18th international conference on World wide web, pages 131–140, Madrid, Spain, 2009, ACM.
60. Badam-Osor Khaltar Atsushi Fujii, A Lemmatization Method for Mongolian and its Application to indexing for Information Retrieval, Information Processing & Management, 2009 – Elsevier.
61. DURRANI, N. AND HUSSAIN, S. 2010. Urdu Word Segmentation. In Proceedings of the 11th Annual Conference of Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics (HLT-NAACL'10).
62. Zhai, Z., Xu, H. & Kang, B. Exploiting Effective Features for Chinese Sentiment Classification. *Expert Systems with Applications*, 38(8), pp. 9139-9146, 2011.
63. Q. Mei, X. Ling, M.Wondra, H. Su, and C. X. Zhai, "Topic sentiment mixture: Modeling facets and opinions in weblogs," in *Proceedings of WWW*, pp. 171–180, New York, NY, USA: ACM Press, 2007. (ISBN 978-1-59593-654-7).
64. Q. Su, X. Xu, H. Guo, X. Wu, X. Zhang, B. Swen, and Z. Su. Hidden Sentiment Association in Chinese Web Opinion Mining. *Proceedings of WWW'08*, pp. 959-968, 2008.
65. Natural Language Toolkit, http://www.nltk.org/Home.
66. Biomedical Text POs Tagger, GENIA Pos tagger ,http://www.nactem.ac.uk/tsujii/GENIA/tagger/.
67. Mohammed Albared, Nazlia Omar, Mohd. Juzaiddin Ab Aziz and Mohd Zakree Ahmad Nazri, Automatic Part of Speech Tagging for Arabic: An Experiment Using Bigram Hidden Markov Mode, Lecture Notes in Computer Science, Publisher: Springer Berlin / Heidelberg, Isbn: 978-3-642-16247-3, Subject: Computer Science.