Issues and Challenges in Analyzing Opinions in Marathi Text

Neelima Mhaske[1] and Ajay Patil[2]

[1]School of Computer Sciences, North Maharashtra University, Jalgaon, Maharashtra, India

[2]School of Computer Sciences, North Maharashtra University, Jalgaon, Maharashtra, India

Abstract

Opinion mining is the process of identifying and extracting subjective information in texts and further analyzing the subjective text for deciding whether it is positive, negative or neutral. It is very popular and challenging area of research. The need for automatic opinion analysis arises from the human nature of believing others. While making decisions, it is very common for people to consult other people for their opinions and experiences. The Internet and social media have greatly influenced the way people communicate with each other. The availability of large amount of information on Internet and social media sites makes it difficult for users to read whole text and analyze opinions. Automatic opinion mining systems make the task of reading and analyzing opinions easy and time efficient. Opinion mining systems are language and domain dependent. To build opinion mining resources and systems it is necessary to understand how opinions are expressed in target language. Most of the research in the field of opinion mining has been carried out for English language. The growing applications of this field have attracted the researchers to develop resources and tools for non-English languages. In case of Indian languages resources have been developed for Bengali and Hindi languages. Work is also in progress for other languages. In this paper we discuss various issues and challenges in the opinion mining of Marathi Text.

Keywords: *Opinion Mining, Sentiment Analysis, Opinion Extraction, Marathi Language*

1. Introduction

Opinion mining or sentiment analysis is concerned with extracting subjective information from natural language texts. The subjective information may include attitudes, beliefs, emotions, opinions, evaluations etc. The sentiment analysis has many application areas like recommender systems, review analysis, blog analysis, advertising systems, etc. With the growing use of Internet, people express their opinions on various issues over the Internet using blogs, portals, comments, reviews, etc. Analysis of this information can produce interesting results for various domains like marketing, government, politics, advertising, etc. Opinion mining is gaining interest in academics as well as industry due to wide range of applications. Along with the social networking sites, almost every e-commerce and newspaper site provides feedback or review service to the users where the users can express their feelings regarding the products/news. This information provides interesting information for the industry to gain insight into customer needs and for individuals to get opinions of other people so that they can make their decisions. The opinion mining could be performed on various types of documents including reviews, newspaper articles, user comments on social networking sites, etc. Most of the research in this field focuses on English language. Many lexical resources like sentiment lexicons, opinion corpus, etc. are available for subjectivity and sentiment analysis for English language.

Since the last few of years, India is rapidly becoming a digital nation. It is one of the countries having large number of Internet users. Internet supports many languages other than English that gives people freedom to use their native languages. People generally prefer information in their mother tongue. India has many scheduled and non-scheduled languages out of which Marathi is one of the official languages ranking fourth in number of speakers in India. Many Marathi newspapers are available on Internet in the e-paper format. Leading social media sites also support Marathi language for readers to express their views. Availability of large information gives rise to the need of automated systems for natural language processing tasks. This need has motivated researchers to develop various tools for Marathi language. In their work on part of speech tagging, H.B. Patil et.al. (2014) presented a rule based system using limited training corpora [5]. They discussed the various steps involved in part of speech tagging like tokenization, stemming, morphological analysis and disambiguation. They also presented a comprehensive analysis of stemmers that are available for Indic languages [6]. V.B Patil and B.V. Pawar (2015), modelled the complex structure of Marathi language in Link Grammar framework [18]. Nita Patil et.al. (2016) presented a survey on named entity recognition techniques for Indian and non-Indian languages [12]. They also presented issues and challenges concerned with the NER system development for Indian languages [13].

In the present work, we discuss various issues and challenges that needs to be addressed while building

resources and tools for automatic opinion mining systems. Opinion mining is the language and domain dependent task. Resources for one language do not give same results for another language. In order to analyze opinions, it is necessary to study how opinions are expressed in target language. Every person has his own way of representing his views. The two major tasks in opinion mining are subjectivity analysis and sentiment analysis. The subjectivity classification task separates subjective (opinions) text from objective (facts) text. The sentiment analysis deals with deciding the overall polarity of subjective text i.e. whether the text is positive, negative or neutral.

Many approaches in the opinion mining systems rely on the existence of sentiment lexicon. Opinions are normally expressed using subjective/sentiment words at word, sentence, paragraph or document level. Sentiment lexicon consists of words that represent subjective information. Sentiment lexicon enables the construction of efficient rule-based subjectivity and sentiment classifiers that rely on the presence of lexicon entries in the text [3]. The sentiment lexicon may also contain polarity information in the form of sentimental category of individual term e.g. highly positive, positive, etc. or polarity (positive/negative) and numeric values indicating intensity of the polarity.

In this paper we explore how opinions are expressed in Marathi language. We also discuss various issues that need to be addressed for efficiently analyzing opinions in Marathi sentences. The examples used in the paper are taken from the movie reviews domain. The reviews are collected from online archives of various Marathi newspapers.

2. Related Works

Although English is the dominant language in the field of opinion mining, non – English languages are also being studied at great extent. Many researchers opt to translate English resources into another language. Translation approach is simple and easy to implement. But translation may lead to ambiguity and loss of accuracy. So it becomes necessary to study the target language itself for opinion expression.

The lexicon based approach is the most widely used approach in the field of opinion mining. Sentiment lexicons could be constructed for any language. It has been observed that adjectives and adverbs contribute most to the subjectivity [4], [17]. The sentiment lexicons have been developed for many languages like Roman [15], Korean [8], Arabic [11], etc.

For the Japanese language Kanamaru, Murata, & Isahara (2007) implemented opinion extraction, opinion holder identification, topic relevance detection and polarity classification components using machine learning technique with support vector machine as the basis [16]. They used 1-gram to 10-gram strings, words, part of speech information as features for the classification task. In their work for language independent opinion sentence detection Zubaryeva & Savoy (2009) used statistical approach to opinion detection and its' evaluation on the English, Chinese and Japanese corpora compared with three baselines, namely Naïve Bayes classifier, a language model and an approach based on significant collocations [14].

The Indian languages are less explored in the field of opinion mining. A study in sentiment analysis of Bengali language was initiated by Das & Bandyopadhyay (2009) using the news and blog corpus of Bengali language [2]. For constructing the sentiment lexicon, they used word level translation process followed by error reduction technique. Bakliwal, Arora, & Varma (2012) developed a sentiment lexicon for Hindi language using the Wordnet based approach [1].

Although the sentiment lexicon provides an efficient way to identify opinions in text, other than the sentiment words, various linguistic features also contribute to subjectivity and affect the polarity of text. Polanyi & Zaenen (2006) discuss various valence shifters at sentence level and discourse based contextual valence shifters that affect polarity orientation of the text [10]. Valence shifters are the constructs that alter the text polarity in various ways. For opinion mining in any language, study of such valence shifters greatly affect the classification results.

3. Opinion Indicators

To analyze opinions, first we need to understand how opinions are expressed in target language. Opinions are mostly expressed using specific opinion words that explicitly convey sentiment information. The collection of such words i.e. sentiment lexicon can be used to identify and extract subjective text. The sentiments can be expressed at various levels like word, phrase, sentence, paragraph, etc.

Like in other languages, sentiment bearing words exist in Marathi language. Most of the research in opinion mining has focused on using adjectives and adverbs as major indicators of subjectivity [4],[17]. However it is observed that along with adjectives and adverbs, nouns and verbs are also used to express opinions. Table 1 list some of the subjective words that occur in Marathi.

Table 1. List of subjective words in Marathi

POS Category	***Positive***	***Negative***
Nouns	आनंद, उत्साह, कौतुक,चांगुलपणा	अतिशयोक्ती, अपयश,दुर्लक्ष

Adjectives	अप्रतिम, छान, अनुपम, उत्तम	कमकुवत, तद्दन, निकृष्ट
Adverbs	आनंदाने, चांगलं	जेमतेम, विनाकारण
Verbs	भावणे, आवडणे	खटकणे

These words are used in sentences to convey sentiments about various aspects. Although sentiment lexicon is a good resource for identifying sentiments, the richness of languages allows a single word to be used in different context with different senses. So the presence of sentiment words does not always indicate opinions. e.g.

Text 1. सिनेमातले संवाद खूपच साधारण आहेत. (The dialogs in the movie are too ordinary)

Text 2. गोष्टीची पार्श्वभूमी साधारण २५ वर्षांपूर्वीची आहे. (The background of the story is approximately 25 years ago)

In Text 1 the word 'साधारण' is used with meaning ordinary which is negative term, while in Text 2 same word is used with meaning approximately which is objective.

There could also be situations where a single word or phrase may have different polarity in different domains or contexts. As shown in Text 3 and 4, the term 'वेड' (madness) can be used as positive opinion indicator in movie reviews domain and it has negative valence in health domain.

Text 3. जादूई संगीतानं तरुणाईला वेड लावणारा संगीतकार ए. आर. रेहमान बॉलिवूड बरोबरच परदेशी सिनेसृष्टीतही ठसा उमटवतोय. – A statement in movie reviews admiring the music of a musician to be maddening

Text 4. सर्वसाधारण भाषेत ज्याला वेड लागणे म्हणतात त्या विकृतीला 'चित्तभ्रम' किंवा 'बुद्धिभ्रष्टता' म्हणतात. – A statement in health domain describing a mental disorder.

Use of multiword expressions and sayings is also very common in natural language text. In case of such expressions group of words needs to be analyzed as a single unit. Multiword expressions can be added to the sentiment lexicon as separate entries.

Indirect Opinions

Identification of indirect opinions is one of the most challenging issues in the opinion mining systems. Sentiment words are not the only option to express opinions. Sentiments can also be expressed without using any sentiment word. Such opinions are indirect opinions. When writing short comments, the writers use specific words. But when writing descriptive reviews, writers have freedom to extend their views in multiple sentences. In such descriptive texts sentiments could span among multiple sentences, where each sentence does not necessarily contain sentiment words, but the sentences are related to each other as a single unit. e.g.

Text 5. संगीत, छायांकन, संकलन या सर्वच बाजुंनी चित्रपट परिपूर्ण बनला आहे. एकच अपवाद तो मध्यंतराचा. (The movie is perfect in all the aspects like music, photography, editing, etc. Interval is the only exception.)

In this example, the first sentence is a direct positive opinion about the music, photography, and editing aspects of movie. While the second sentence provides an information about another aspect, i.e. interval, stating that it is an exception to the first sentence. The second sentence is indirect negative opinion about the interval of the movie. The second sentence itself does not convey any type of sentiments explicitly. It will be meaningful only when used with first sentence and the polarity of it will be dependent on the polarity of the first sentence.

The indirect opinions are also encountered with common beliefs and indicators in the target domain. e.g.

Text 6. या चित्रपटाने पहिल्याच दिवशी १०० कोटी कमावले. (The movie earned 100 crore on first day)

This is a factual sentence giving information about the collection of a movie. However earning large amount is a positive indicator for movies. So this statement conveys a positive sentiment for the particular movie.

Many such situations can be found in natural language texts where the authors are expressing implicit opinions. While analyzing indirect opinions, it becomes necessary to study the domain for different contexts and situations and to increase the scope of the sentiment unit to be larger than a single sentence.

4. Sentence Structure

Every natural language processing system relies on the basic grammar of the target language. Grammar gives us insight into how the text is constructed using words. Words are the basic units in opinion mining systems. Although single sentiment words are capable of expressing opinions, words are normally used in bigger unit i.e. a sentence.

In case of direct opinions, sentiment words are a part of the sentence. The position of the sentiment word depends on the sentence structure. The principal word order in Marathi is SOV (subject–object–verb). The Marathi grammar has three types of sentences:

a. केवलवाक्य (Simple Sentence)
   Single principle clause.
   सिनेमा तांत्रिकदृष्ट्या उत्कृष्ट झाला आहे.

b. मिश्रवाक्य (Complex Sentence)

Single principal clause and one or more subordinate clauses.
उत्तम तंत्रकौशल्य, चांगले कलावंत असूनही हा चित्रपट फारसे समाधान करत नाही.

c. संयुक्तवाक्य (Compound Sentence)
Two or more sentences of above types connected by various connectors.
सिनेमाची कथा रोचक आहे आणि काही ठिकाणी ती मजेशीरसुद्धा वाटते.

Most of the sentences follow this order and belong to any one of the above mentioned types making it possible to design extraction patterns for opinions. However all sentences in a document might not follow the correct grammatical order. Every writer has his own writing style that makes similar opinions to be expressed in different manner. e.g.

Text 7. संवाद प्रभावी आहेत. (The dialogs are effective)
Text 8. संवादांमधून मनोरंजनही होते.
(The dialogs entertain us too)
Text 9. चित्रपटातील प्रसंगांवर संवादांचे प्रभुत्व दिसते.
(Dominance of the dialogs could be seen on movie events)

The incorrect grammar issue causes more problems in social media and blog contents. On social media the users mostly use short comments instead of full sentences. The statements need not be grammatically correct. Handling such data is a difficult task.
The simple sentences are comparatively easy to analyze. But the complex and compound sentences are difficult especially if contradictory opinions are expressed in same sentence. The complex and compound sentences make use of different types of connectors to connect various clauses in a sentence.

Connectors
The study of the connectors is necessary while deciding the polarity of the sentences. The connectors are used to connect sentence clauses in compound and complex sentences. Connectors can combine multiple sentiment clauses of same polarity or of different polarity. When used with contradictory opinions, the polarity of the whole sentence will be dependent on the type of connector used. Following are the types of connectors:

a. Coordinating (प्रधानत्वसूचक): links two words, phrases, clauses or sentences that are grammatically equivalent. In this case the clauses can be separated and polarity of each part can be calculated separately.
सिनेमाची कथा रोचक आहे आणि काही ठिकाणी ती मजेशीरसुद्धा वाटते. (The story of the movie is interesting and at some places it is funny too.)

b. Subordinating (गौणत्वसूचक): are the conjunctions that link dependent clause to an independent clause. The polarity of the sentence can be set to that of independent clause.
सिनेमाचं कथानकच एवढ जबरदस्त आहे की, सिनेमाला संगीताची फारशी गरज जाणवली नाही. (The story of the movie is so strong that it does not need music)

Marathi is a free order language. So the principal word order is not always followed in texts. We often encounter cases where compound and complex sentences are constructed without using explicit connectors. Consider the example in Text 10,

Text 10. उत्तम तंत्रकौशल्य, चांगले कलावंत असूनही हा चित्रपट फारसे समाधान करत नाही. (In spite of good technical skills and actors, the movie doesn't satisfy)

In this sentence opinion about three features तंत्रकौशल्य (technical aspects), कलावंत (actors) and चित्रपट (movie) are expressed. First two features have positive polarity. The scope of negation word नाही is limited to the last feature. Presence of the word असूनही splits the sentence into two clauses where one clause is of positive polarity while the other is negative. In this sentence the word असूनही makes the second clause of the sentence dominant over the first clause. So the overall polarity of the sentence could be set as negative. While processing complex and compound sentences, the position and type of connectors play an important role in deciding the overall polarity of the sentences.

5. Negation

Negation is another challenging aspect in opinion mining. The problem of negation needs to be addressed for opinion mining system of any language. Negation is a grammatical construction that contradicts or negates part or all of a sentence's meaning. In case of English negative clauses and sentences commonly include the negative particle "not" or the contracted negative "n't". Similarly in Marathi, words can be negated using prefixes like "अ", "गैर", etc.
अ : अप्रतिष्ठा, अप्रासंगिक
गैर : गैरवर्तन, गैरसमज
Negating terms like न, नसून, नसल्याने, etc. are also used to negate word polarities. Negation can be applied at word or sentence level. e.g.

Text 11. हा संपूर्ण चित्रपट मनोरंजनात्मक आहे. (The movie is recreational)

Text 12. हा चित्रपट मनोरंजनात्मक नसून माहितीपर आहे. (The movie is not recreational but informative)
Negation 'नसून' is applied on subjective word 'मनोरंजनात्मक' (recreational)

When applied at sentence level, the negating term reverses the polarity of the whole sentence e.g.

Table 2. Negation applied at sentence level

Sentence	संपादन चांगले नाही. (Editing is not good)
Subjective Word	चांगले– Positive
Sentence polarity	Negative

After identifying the negation indicators, it is necessary to determine the scope of the negation. Scope of the negation word defines the part of the subjective text which is negated with that word. A negation word may contradict part or the entire sentence. Jia et.al. (2009) show that the identification of the scope of negation improves both the accuracy of sentiment analysis and the retrieval effectiveness of opinion retrieval [9]. They have used the parse tree and typed dependencies along with heuristic rules. Councill, McDonald, & Velikovich (2010) presented a negation detection system based on a conditional random field modeled using features from an English dependency parser including the lowercased token string, token POS, token-wise distance from explicit negation cues, POS information from dependency heads, and dependency distance to explicit negation cues [7].

6. Grammatical Moods

The grammatical mood conveys speaker's attitude about the state of being of what the sentence describes. Moods play an important role in opinion mining systems. In Marathi there are four types of moods: स्वार्थ,आज्ञार्थ, विध्यर्थ, संकेतार्थ.Here we discuss only the संकेतार्थ(subjunctive)mood which is related to opinion mining. With this mood, author can set up a context of possibility or necessity in texts. It is used to explore conditions that are contrary to facts. e.g.

Text 13. जर कथा आणि पटकथा चांगली असती तर, चित्रपट आणखी चांगला बनु शकला असता. (If the story and screenplay was good, the movie would have been better)

The above sentence contains two clauses:

- First clause "कथा आणि पटकथा चांगली असती" indirectly expresses negative sentiment about the story and screen play suggesting that the story and screenplay are not good
- Second clause "चित्रपट आणखी चांगला बनु शकला असता" suggests that the movie would have been better if the first clause was true.

It is observed that the second clause is dependent on the first clause, both the clauses are contradictory to the real situation. Thus the polarity of the whole sentence could be set to negative. The subjunctive mood can be identified by use of specific constructs like जर - तर, तथापि, or verb phrases like हवेहोते, झालेअसते,etc.

7. Irony

People do not always use formal or plain words in expressing their opinions, especially when writing informal text. In case of newspaper articles the language used is formal and grammatically correct. But when writing reviews or comments, the language of the author is not always formal and grammatically correct. While working on informal text we often encounter ironic statements that are difficult to analyze due to the use of words to express the opposite of their literal meaning. e.g.

Text 14. कथाच नसल्याने सारा आनंदी आनंद आहे. (Since there is no story, everything is out of order.)

Here the word आनंदीआनंद(all is well) basically has positive polarity. But the way it is used in this sentence, gives it negative sense.
Situational irony statements also express contradictory sentiments in a single sentence. Such types of statements are used when the situation is contrary to the author's expectations. e.g.

Text 15. '…'सारखा चांगला चित्रपट बनवणारा दिग्दर्शक या चित्रपटात मात्र घोर निराशा करतो. (The director who has made good movies like '…' disappoints in this movie.)

In this sentence the first clause '…' सारखा चांगला चित्रपटबनवणारा दिग्दर्शक suggests a positive opinion about a particular director who has directed a good movie. The author is expecting the other movies directed by the same director to be good. The second clause, however expresses negative polarity about the director as he has failed to fulfill the expectations. In other words this sentence expresses a negative opinion about a director who is otherwise considered to be very good.

8. Possibilities

Different people may have different opinions on same topic. The author cannot always be sure that the reader will agree with his assessments. Sometimes the author

may not be sure about making a particular opinion. In such cases the author may express possibility of opinions. e.g.

Text 16. कदाचित हा सिनेमा तुम्हाला निराश करु शकतो. (Perhaps this movie will disappoint you.)

Here the author indicates that the movie might disappoint you. It is not sure whether the author himself is disappointed or not, so the decision depends on the reader's experience. Such types of sentences could be classified as neutral sentences.

9. Conclusion

Opinion mining is a language and domain dependent task. When constructing opinion mining resources, it becomes necessary to study the target language. Here in this paper we discuss various issues and challenges that need to be addressed while building automatic opinion mining system for Marathi language. This study helps to efficiently identify and extract subjective text that later can be analyzed for polarity orientation. We observe that mere presence of sentiment words does not always indicate presence of opinions but other linguistic features also contribute to subjectivity and affect the polarity of the text. We intend to develop automated rule based opinion detection system for Marathi language. The study of linguistic features will help to develop rules and extraction patterns for extracting subjective information from text and analyzing this subjective text for polarity classification.

Acknowledgments

The authors are thankful to the University Grants Commission, New Delhi, India for supporting this research under the Special Assistance Program (SAP) at the level of DRS-I.

References

[1] Bakliwal, P. Arora and V. Varma, "Hindi Subjective Lexicon: A Lexical Resource for Hindi Polarity Classification", The eighth international conference on Language Resources and Evaluation, 2012, pp. 1189-1196

[2] Das and S. Bandyopadhyay, "Subjectivity Detection in English and Bengali: A CRF-based Approach", 7th International Conference on Natural Language Processing, Macmillan Publishers, 2009

[3] Banea, R. Mihalcea and J. Wiebe, "A Bootstrapping Method for Building Subjectivity Lexicons for Languages with Scarce Resources", Proceedings of the Sixth conference on International Language Resources and Evaluation, 2008, pp. 2764-2767

[4] F. Benamara, C. Cesarano, A. Picariello, D. Reforgiato, and V. Subrahmanian, "Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone", Proceedings of the International Conference on Weblogs and Social Media, 2007

[5] H.B. Patil, A.S. Patil, B.V. Pawar, "Part-of-Speech Tagger for Marathi Language using Limited Training Corpora", IJCA Proceedings on National Conference on Recent Advances in Information Technology NCRAIT(4), 2014, pp. 33-37

[6] H.B. Patil, A.S. Patil, B.V. Pawar, "A Comprehensive Analysis of Stemmers Available for Indic Languages", International Journal on Natural Language Computing (IJNLC) Vol. 5, No.1, 2016, pp. 45-55

[7] I. G. Councill, R. McDonald and L. Velikovich, "What's Great and What's Not: Learning to Classify the Scope of Negation for Improved Sentiment Analysis", Proceedings of the Workshop on Negation and Speculation in Natural Language, 2010, pp. 51–59

[8] J. Kim, H.Y. Jung, S. Nam, Y. Lee and J. Lee, "Found in Translation: Conveying Subjectivity of a Lexicon of One Language into Another Using a Bilingual Dictionary and a Link Analysis Algorithm", ICCPOL 2009, LNAI 5459, 2009, pp. 112–121

[9] L. Jia, C. Yu and W. Meng, "The Effect of Negation on Sentiment Analysis and Retrieval Effectiveness", CIKM'09, ACM, 2009, pp. 1827-1830

[10] L. Polanyi and A. Zaenen, "Contextual Valence Shifters", Computing attitude and affect in text: Theory and applications, 2006, pp. 1-10, Springer.

[11] M. Abdul-Mageed and M. T. Diab, "Subjectivity and Sentiment Annotation of Modern Standard Arabic Newswire", Proceedings of the Fifth Law Workshop, 2011, pp. 110–118

[12] N.V. Patil, A.S. Patil, and B.V. Pawar, "Survey of Named Entity Recognition Systems with respect to Indian and Foreign Languages", International Journal of Computer Applications 134(16), 2016, pp.21-26

[13] N. V. Patil, A. S. Patil and B. V. Pawar, "Issues and Challenges in Marathi Named Entity Recognition", International Journal on Natural Language Computing (IJNLC) Vol. 5, No.1, 2016, pp. 15-30

[14] O. Zubaryeva and J. Savoy, "Investigation in Statistical Language-Independent Approaches for Opinion Detection in English, Chinese and Japanese", Third International Cross Lingual Information Access Workshop, 2009, pp. 38–45

[15] R. Mihalcea, C. Banea and J. Wiebe, "Learning Multilingual Subjective Language via Cross-Lingual Projections", Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007, pp. 976–983

[16] T. Kanamaru, M. Murata and H. Isahara, "Japanese Opinion Extraction System for Japanese Newspapers Using Machine-Learning Method", Proceedings of NTCIR-6 Workshop Meeting, 2007, pp. 301-307

[17] V. Hatzivassiloglou and K. R. McKeown, “Predicting the Semantic Orientation of Adjectives”, Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL, 1997, pp. 174-181

[18] V.B. Patil, B.V. Pawar, “Modeling Complex Sentences for parsing through Marathi Link Grammar Parser”, International Journal of Computer Science Issues, Vol.12, Issue 1, No. 2, 2015, pp. 108-112