SENTIMENT ANALYSIS USING HARN ALGORITHM

D V Nagarjana Devi
Assistant Professor,
IIIT, RGUKT,Nuzvid
devi.duvvuri@rgukt.in

Dr.T.V.Rajanikanth,
Professor,SNIST,
Hyderabad.TG
rajinitv@gmail.com

Pantangi Rajashekar,
UG Student,
IIIT, RGUKT,Nuzvid
pantangi.shekar95@gmail.com

Gangavarapu Akhil,
UG Student,
IIIT,RGUKT,Nuzvid
akhil.rgukt976@gmail.com

***Abstract*—Sentiment Analysis, is the field which has been looked into a great depth recently. There has been a lot of work in identification of polarities. There are many techniques developed by researchers to classify the opinionated data. During our research we have found out that there is no technique which could give 100% accuracy. There are many challenges to Sentiment Analysis like semantic ambiguity statements, comparison sentences, domain specific adaptation, sarcastic statements etc. We have found a solution to one of the limitations, thereby contributing to the elevation of accuracy level. The basic problem that we have identified is polarity switching of a word based on the domain in which it is being used (Domain Specific Adaptation). To solve this problem we have developed an algorithm which could correctly classify the statement based on the domain in which the word is being used. The proposed algorithm is HARN algorithm. It is an unsupervised learning method which uses basic structure of the sentences, domain dictionaries and pre-defined polarities to classify the given sentence. The below report discusses the present existing approaches to sentiment classification, HARN algorithm and its implementation details.**

***Keywords*: sentiment analysis; Stanford parser, NV dictionaries, data mining.**

I. INTRODUCTION

Natural language processing (NLP) and computational linguistics have a rich historical background but still there was minimum research in the field of opinion mining before the year 2000 except for some earlier work on subjectivity, sentiment adjectives, viewpoints and interpretation of metaphors. Apart from its applications in the field of data mining, web mining, and information retrieval, the sentiment analysis task has spread to the management sciences. Insights and applications from sentiment analysis have been useful in other areas including politics, law making, sociology and psychology. The terms opinion mining and sentiment analysis were first introduced in 2003 by Nasukawa et al and Dave et al respectively. However the researches in the field of sentiments and opinions started earlier. Based on the regularities in text, J.M. Wiebe presents an algorithm to identify subjective characters in narratives. To identify the point of view of the authors Wiebe suggests an algorithm, he has extensively studied naturally occurring narratives to propose this algorithm.To refine the information access task, M.A.Hearst has developed an approach for intelligent systems which uses a direction based text interpretation. Semantic orientation of adjectives and polarity similarity of two conjoined adjectives are achieved with 82% accuracy through a method proposed by Hatzivassiloglou and McKeown.

The research in the field of opinion mining has grown rapidly after the year 2000. The major reason behind this explosive growth is the World Wide Web. Earlier there was little opinionated text for study but today there is large amount of user generated content in the form of comments, reviews and debates.

In our work we have observed that there are many limitations to sentiment analysis. We tried to solve one of the limitations. We have developed a novel algorithm namely HARN, using which we can overcome the limitation of sentiment analysis.

In this paper, we would like to discuss about the methods used so far, their implementations and our new proposed algorithm, it's working. Finally we are going to conclude with the results of our algorithm and future implementations.

II RELATED WORK

There are many works done so far on sentiment analysis. There are many novel approaches which are being developed.

From [1], multimodal sentiment analysis is the analysis which uses audiovisual input to analyze emotions, attitude and opinion of the individual which is a novel method of sentiment analysis.

There are many limitations to sentiment analysis which are thoroughly discussed in [2] and also they tried to solve the limitation using Fuzzy Inference Method (FIM).

Siti Rohaidah, in [3] discussed the importance of feature selection in sentiment analysis. They proposed different ways of selecting features. After analyzing all the methods, he concluded that using metaheuristic algorithms, which is a combination

978-1-5090-3765-0/16/$31.00 ©2016 IEEE

of ACO & GA and so on would help in selection of the best features.

In [4], there is a brief overview of all the methods used so far to implement sentiment analysis. Different approaches like machine learning and lexicon based and its sub methods.

The limitations of sentiment analysis are a major drawback and due to these limitations the accuracy of the developed methods are also declining. So, we have decided to remove one of these limitations by proposing an algorithm, which uses predefined structure definitions

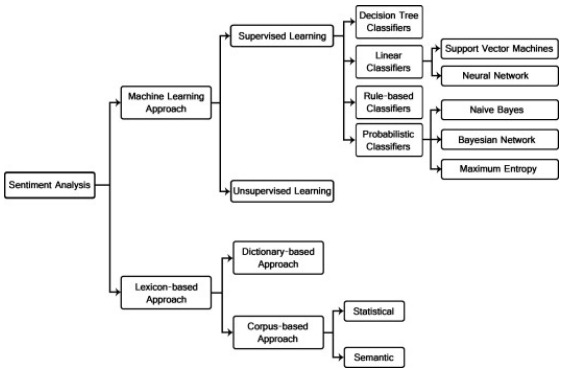

Figure 1: Techniques of sentimental analysis

III. LIMITATIONS OF SENTIMENT ANALYSIS.

There are five main factors that give trouble in finding their polarity to all the tools of sentiment analysis:

1.*Context:* A word can act either positively or negatively based on the context. Positivity or negativity of the word depends upon the context in which the word is being used. We cannot decide the polarity of some words. They are decided based on context."*Train is going fast*" here in this context *fast* is a positive word whereas in "*My battery is draining fast*" it has a negative sense.

2. *Semantic Ambiguity:*Presence of positive word or negative word cannot decide the polarity of the sentence. The usage of these words can be different. For example, "*Why don't you suggest me a good book to read?*". This sentence is not a positive sentence even if it uses the word *good*. Sometimes, there might be sentences which do not contain any sentiment words yet give us a sentiment. In, "*this game in mobile phones costs us a lot of battery, so it can be avoided*", there are no negative words, yet the sentence is negative. In [8], a novel approach has been discussed to solve this problem.

3. *Sarcasm:*Sarcastic sentences are very hard to catch and this is a huge limitation to Sentiment Analysis. For example, "*Sure, I'm very **happy** for the power loss while I was watching my favorite basketball match*." is a negative sarcastic sentence even though it has a positive word (very happy). Even positive or negative words switch their polarities in sarcastic sentences. In [9][10] a two novel approaches are discussed to solve sarcasm in sentiment analysis.

4. *Comparatives:*When it comes to comparative sentences like "*Samsung is much better than Blackberry*". For this sentence, present day sentiment analyzers declare it as positive sentence since it contains the word *much better*. However when you pick up sides then the scenario changes. For example, if you are reporting the customer reviews to your managers, in this case, the sentence is negative for blackberry, but positive for Samsung. The analyzers are not that intelligent to pick sides and declare the polarity of the sentence. The Solution for this problem is presented in this paper. By defining the structures of the sentences we can eliminate this limitation. In [11] an innovative method for Vietnamese language, to solve comparative sentences is presented.

5 Regional Variations: Language used is also a major concern for sentiment analysis, but why? It is because usage and meaning of words may vary with language to language. This can be observed more often in language variations, slang and dialects. For example,*sick*is the word which has different meaning in different contexts, tone and language. When we observe the examples "*That is a sick game!*" vs. "*I'm not at all feeling well, I think I am sick*". This variation is generally found between American and British English. If we consider the words like '*pretty*', '*quite'*, '*rather*',in British English those words mean "*fairly*", while in American English it means "*very*". Sometimes day-to-day conversations are also misunderstood because of these regional variation words. So, How can it not be a limitation to Sentiment analysis?

IV PROPOSED METHOD

The usage of HARN algorithm increases the probability of finding the polarity of sentences. Till now, there is no specific rule or technique to solve the problem of domain specific adaptation. But, by using this algorithm we could easily deal with domain specific sentences and comparison sentences.

The algorithm states that, we have to define basic structures of sentences. From these sentences we pick few features which will be useful for the determination of polarity of the sentence. For these features, we must build a dictionary of words for these features along with their polarity. The algorithm checks for the words in dictionary and finds their polarity and displays them.

According to algorithm if the sentence structure is not found in the defined structures. The algorithm automatically updates the structure, and prompts user to select features from the input sentence. If the polarity of the entered features are not found in the NV dictionaries, the computer will again ask user to enter the polarity values. This way the algorithm learns itself new structures and can judge them next time when they are given

HARN Algorithm:

1. Take sentence/paragraph as input
2. Do POS tagging
3. Search for given sentence structure
4. If (structure_found)
5. seletion of features based on structur
6. If (features contain noun verb)
7. Polarity check for features from NV Dictionaries
8. If found
9. Give polarity
10. Else
11. Request user to add verb polarity
12. Else
13. Collect polarity from sentiwordnet & display
14. Else
15. Adds structure
16. Prompts user for feature selection for that structure.

There is no training set given to this algorithm. We define some basic structures, and if the given sentence matches with the structure provided, then it searches for NV dictionaries for the word, if found, gives the polarity and if it doesn't find, it prompts user to enter the word to be chosen, and its polarity in that context. Our algorithm memorizes the given input, sentence and words because, next time when it encounters a similar sentence it directly gives the output. The detailed step by step process of the HARN algorithm is given below.

1. Stanford parser: Initially, the input sentence/paragraph are sent through Stanford parser to have POS tagging [4] [5]. We have included Stanford parser to our code, the output of Stanford parser is given as input to the code, which undertakes next process. We have used online Stanford parser during development of the code [12].

For example.

Train is going fast

Input to Stanford parser, it gives output as follows

Train/NNP is/VBZ going/VBG fast/RB

Now this sentence containing both word and its POS tagging, is given as input to our code

2. Defined Structures: After taking the about sentence as input, the code initially divides the sentences into two sentences, i.e.

1. Train is moving fast

2. NNP VBZ VBG RB

According to HARN algorithm, we have to define structure of the sentences. Defining the structures is nothing but writing an if-else code with only parts of speech terms.

From the above input of sentence with tagged elements. We split the sentence basing / and divide the elements into two arrays **a** and **b**. where **b** contains the POS tagged structures and **a** contains the actual words. i.e.

From our above given example.

a[0]=Train, a[1]=is, a[2]=going, a[3]=fast.

b[0]=NNP, b[1]=VBZ, b[2]=VBG, b[3]=RB

For example: if NNP VBZ VBG RB, is the input and if the code contains the defined structure, it looks as follows.

if (b[0] = "NNP" and b[1] ="VBZ" and b[2] = "VBG" and b[3] = "RB")

If this is found, then under the if condition, we will mention what are the features to be selected for these kind of sentences, like follows,

c.append(a[0]) *(Train)*

c.append(a[3]) *(Fast)*

So by this level, the features will be selected

3. *NV dictionaries:* This is the heart of the algorithm. NV(Noun Verb) dictionaries. Here the dictionaries for nouns are maintained, which contains verb and its polarity. The above selected features if contain a noun then enters these NV dictionaries, searches for the verb in its dictionary, and if found takes the polarity associated with it. If the verb is not found associated with it, it asks the user, whether the given verb in the sentence is a

positive or negative according to the noun in the context. In this way it learns from the user. And next time, if the same kind of sentence with the verb is given, it doesn't asks for user, it directly gives output.

NV dictionaries are maintained as follows

train = {'stop':-1, 'fast':1,'delay':-1,}

battery = {'working':1, 'draining':-1}

similarly we have given many verbs along with their polarity as dictionary elements to the corresponding nouns. This is the main area where we are trying to avoid the limitation of domain specific adaptation.

4.Deciding the Polarity:

Now from our example of *Train is moving fast*The code takes features as *train* and *fast*, and checks for NV dictionaries, which contains *train* noun and also its element *fast* with its polarity of *1*.

Thus after the execution of the code, the output of the sentence, *train is moving fast* is, +1, which is positive. So the sentence is classified as positive.

Similarly if the sentence is

Battery is draining fast

As this is same as *Train is moving fast* i.e. both structures are same, i.e. both have the same POS tagging as,

Train/NNP is/VBZ going/VBG fast/RB

Battery/NNP is/VBZ draining/VBG fast/RB

As this structure is defined it takes battery and draining as features, it searches for NV dictionaries and finds draining as negative, so it displays output as -1.

5 HARN's special rule

An other rule in HARN algorithm is, *any adjective present after the verb, only enhances the polarity either positively or negatively*.

So this rule can be used to solve the problem of domain specific adaptation.

So we search for any adjectives after the verb, in our example of *battery is draining fast*, as *fast* is an adjective, which is not given a polarity, according to HARN algorithm, it just enhances the polarity of the verb either if it is positive or negative.

So, draining is -1, and fast → adjective, so it must enhance the negative polarity in this case, so the total polarity of the sentence will be -2. Therefore declaring it to be negative.

Thus we have solved the problem of domain specific adaptation. Earlier, if the sentences containing words like *fast,* they would be classified incorrectly, as *fast* is a word which would change according to the domain it is being used. But by using HARN algorithm we have solved this problem.

*6. Problem of Comparison sentences:*For comparison sentences, most of the present day sentiment analyzers just give whether the sentence is positive or negative, but they won't have clarity on which has a positive view and which has a negative view. But our algorithm clearly states from the comparison sentence to whom it is positive and to whom it is negative from the compared objects.

For example:
1.*iPhone is better that Samsung.*

If this sentence is given as input to normal sentiment analyzer, it gives output as positive, But HARN algorithm on other hand, gives clarity on to which is good and which is bad. i.e. the output of the above sentence from HARN algorithm will be

According to iPhone – Positive
According to Samsung – Negative.
2. Blackberry is not as good as Samsung

For normal analyzer the output would be only positive, but using our code the output will be

According to Blackberry – Negative

According to Samsung – Positive.

Thus, by using this HARN algorithm we can solve the problem of domain specific adaptation and comparison sentences.

7. Other that NV sentences: Sentences which do not contain NV (noun - verb), if given as input, firstly POS tagging is done by Stanford parser, then checks for NV form and also defined structures, if structure is found and no NOUN VERB combination is found, then the algorithm uses the SentiWordNet[6][7] to identify the polarities of each word and then sums up all the obtained polarities resulting in an overall polarity.

This is done only when sentence structure is found and there is no NV combination in the sentence, else if the sentence is not defined then, it updates automatically by asking the user key features in the sentence.

V. CONCLUSION AND FUTURE WORK

We have analyzed the techniques so far present for sentiment analysis. We found that there are few limitations like semantic ambiguity, comparison sentences, Domain dependency problem and

specific to English language. We have found a solution to one of the limitations, thereby contributing to the elevation of accuracy level. The basic problem that we have identified is polarity switching of a word based on the domain in which it is being used (Domain Specific Adaptation). To solve this problem we have developed an algorithm which could correctly classify the statement based on the domain in which the word is being used.The proposed algorithm is HARN algorithm. It is an unsupervised learning method which uses basic structure of the sentences, NV dictionaries and pre-defined polarities to classify the given sentence. The future work of this includes, adding more structures, NV dictionaries.

In future research we would like to mainly focus on sarcastic sentences, which are very difficult to detect. We believe that to detect a sarcastic sentence, it has to make use of the sentences given so far. Thus achieving sentiment analysis with 100% accuracy.

VI. REFERENCES

1. Sumit k yadav, Mayank Bhusan, Swathi Gupta, *Multimodal Sentiment Analysis using audiovisual format*, 2015 2nd International conference onComputing for Sustainable global development(INDIA com)

2. Zhaoxia Wang, Victor too Chaun Tong, David chan, *Issues of social data analytics with a new method for sentiment analysis of social media data*,2104 IEEE 6th International Conference on Cloud Computing Technology and Science.

3.Siti Rohaid Ahmad, Azuraliza Abu Bakar, Mohid Ridzwan Yakub, *Metaheuristic algorithms for Feature Selection in Sentiment Analysis*, Science and Information Conference 0152 July 28-30, 2015, London,UK.

4.Marie-Catheirne de Marneffe and Christopher D Manning, 2008, *Generating Typed Dependency Parses from Phase Structure Parses*, In 5th International Conference on Language Resources and Evaluation (I.REC 2006)

5.Marie-Catheirne de Marneffe and Christopher D Manning, 2008, *The Stanford typed dependencies representation,*in CLONING 2008 workshop on Cross framework and crossdomain parser evaluation.

6. Andrea Esuli and Fabrizio Sebastiani, *SentiWordnet: A publicly Available Lexical Resource for Opinion Mining*, in LREC 2016.

7. Krestin Denecke, Using SetiWordNet for multilingual sentiment analysis, Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on 7-12 April 2008.

8. Yunfang Wu, *Disambiguating sentiment ambiguous adjectives*, Natural language processing and knowledge Engineering, 2008. NLP-KE'08.

9. Santosh kumar bharti, *Parsing-based sarcasm sentiment recognition in Twitter data*, 2015 IEEE/ACM International Conference on Advances in Social Network Analysis and Mining (ASONAM).

10. Ngo Xuan Bach, Mining Vietnamese Comparative Sentences for Sentiment Analysis, Knowledge and Systems Engineering (KSE), 2015 Seventh International Conference on 8-10 Oct.2015

11. Edwin Lunando, *Indonesian social media sentiment analysis with sarcasm detection,* Advanced Computer Science and Information Systems (ICACSIS), 2013 International Conference on 28-29 Sept, 2013.

12.http://nlp.standford.edu:8080/parser/index.jsp