Domain Driven Word Sense Disambiguation

Deeksha S[1] Niranjan S[2] Nithin S[3] Bhoomika P[4] Dr. Paramesha. K[5]
[1,2,3,4,5]Department of Computer Science and Engineering
[1,2,3,4,5]Vidyavardhaka College of Engineering Mysore, India

Abstract— The aim of our project is to resolve the ambiguation in sentences based on its domain. Most of the times sentences used by us will have different meanings. The difference in the meaning of such sentences depends on the prefix, suffix as well as the subject in the sentence. Understanding the meaning of the sentence based on the subject or situation is easy for humans. But it is not same with respect to a machine. We will work to provide solution to this problem.

Keywords: Natural Language Processing, Data Mining, Text Mining, Sentimental Analysis, Opinion Mining

I. INTRODUCTION

Ambiguity is a state in which there are multiple possible directions for an element. And in most of the cases currently existing systems fail to decide the current direction. In our case the ambiguity exists in adjective which defines the subject. The entire meaning of the sentence depends on the prefix and suffix of the ambiguous word, domain and the subject of the sentence. But the currently existing systems are failing to analyze them. Our aim is to resolve this problem.

Natural language processing is how computer programs are able to make sense of words in the surrounding context. For example, you could write a computer program to pick up on sarcasm such as "That's funny... not." Or to understand "The world will end!" as an exclamation verses "The world will end?" as a question. But machines can't simply read and interpret language innately like humans can. They do this through the calculations. And so calculations on words and textual features is what allows machines to determine if a piece of text contains sarcasm, or if it's more negative than positive in a sentiment, or contains more rhetoric rather than factual statements.

Counting the frequency of words and taking into account the surrounding context and then doing calculations is the basis of how machines make sense of natural language. So then in order to count or calculate words and textual features the raw text, or natural language itself, first needs to be processed in a way that allows machines to work with the more structured data format. This basically means cleaning up the text and then organizing it into tables of word counts across documents. It could also mean tabling pairs of words that occur together taking into account surrounding context of the words. Cleaning up raw text and organizing it into a table is absolutely an essential step in natural language processing. The word processing should be emphasized in natural language processing. Without processing you're just left with natural language which is mentioned machines cannot easily interpret like you and I. They needed to be processed first, and then do calculations on. Some key applications of natural language processing are categorizing texts into negative or positive sentiment to automatically identify unsatisfied customers from satisfied customers.

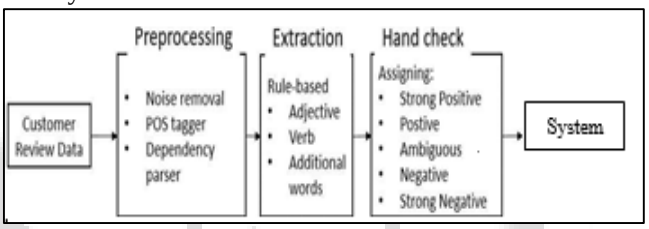

Fig. 1: Process Diagram

II. METHODOLOGY

In fitting algorithm to train the set, it is difficult to find a good or even a well-performing machine learning algorithm for a particular dataset. We went through a process of trial and error to settle on a short list of algorithms that provides better result. We studied a couple of algorithms. In our work we are going to show and discuss the performance of Gaussian Naive Bayes and Decision Tree algorithms.

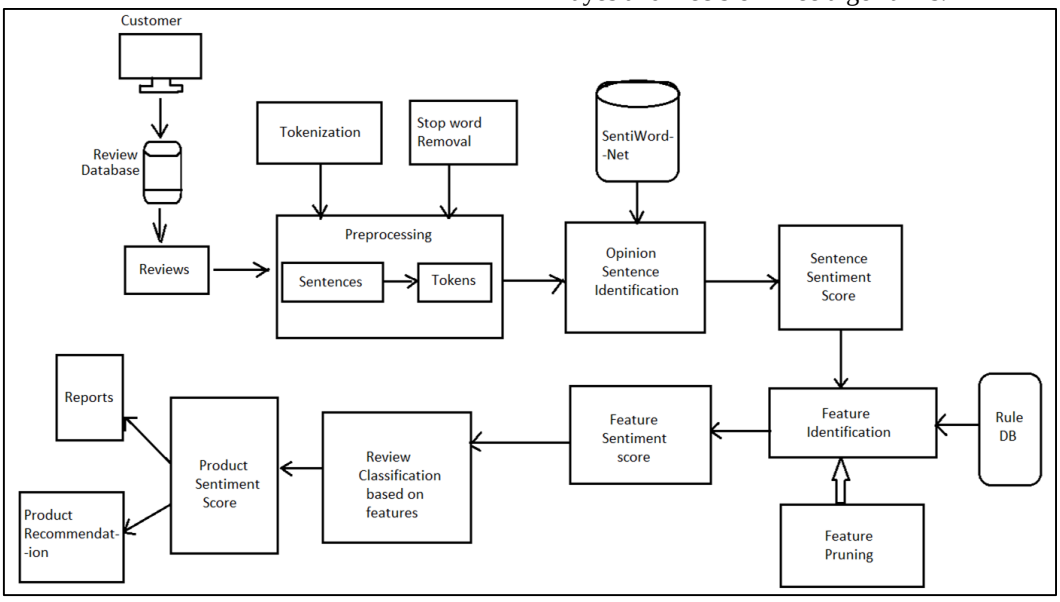

Fig. 2: System Architecture

The Machine Learning system uses the training data to train models to see patterns, and uses the test data to evaluate the predictive quality of the trained model. In the terminology of machine learning, classification is considered an instance of supervised learning, learning where a training set of correctly identified observations is available. We use Gaussian Naive Bayes classifier and Decision Tree classifier to predict the restaurant reviews whether it is good or bad. Naive Bayes classifier is widely used in Spam filtering (identify spam e-mail) and Sentiment Analysis (in social media analysis, to identify positive and negative customer sentiments). The process that our system undergoes is shown in the figure. First stage involves data collection which is followed by data-preprocessing. We are suing Naïve Bayes as well as Decision tree algorithms to calculate the test result.

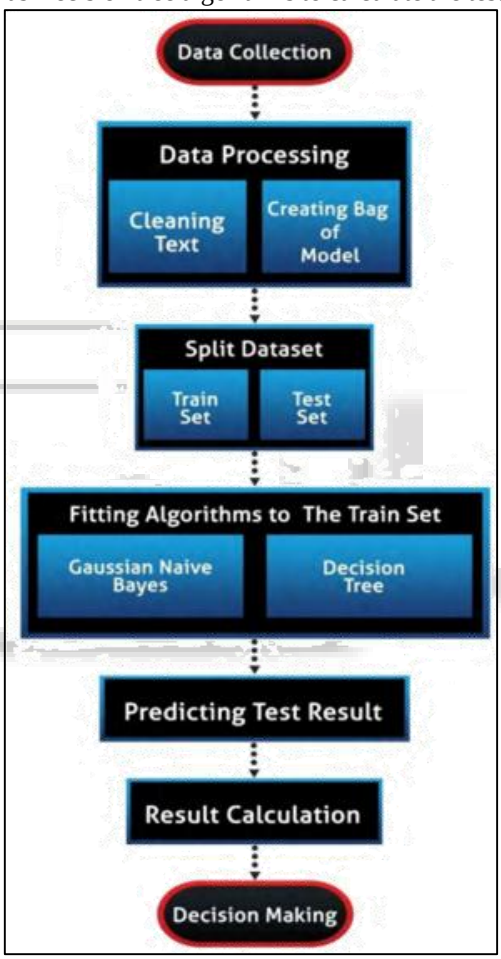

Fig. 3: FlowChart

III. RELATED WORK

[1] In this paper author proposed a genetic word sense disambiguation algorithm, they use wordnet to determine the possibility senses for certain in words or sentences, here a novel conceptual word is used for that domain information also given for identifying disambiguation. proposed algorithm performs well for set of domains and based on weighted words wgwsd algorithm is used which performs well in general corpus, but this system proposed works for only nouns.

[2] In this paper author proposed an enhanced knowledge-based word sense disambiguation (WSD) algorithm which works finding similarities between the actual text and target meaning, here they proposed both semantic and syntactic between the sentences. this algorithm performs well compare to previous proposed systems this approach takes advantage of syntactic and semantic information to solve natural language ambiguity.

[3] In this paper they present a methodology for Word Sense Disambiguation based on domain information whereas domain means a set of words in which all those words are related to each other. Here they used unsupervised approaches and they used wordnet domains as lexical database. here they classify similar words to different sense of domains based on that they give result.

[4] In this paper their approach is to WSD using wsd specific wordnet of polysemy words, here they developed a new model of wordnet and differentiate polysemy words based on some clue words these are nouns, verb, adjective, adverbs called synonym sets. this results in increasing in accuracy compare to previous systems.

[5] In this paper they proposed a Max-Probability Density based Clustering (MPDC) algorithm which helps to resolve the problem of Word Sense Disambiguation in semantic document. here they take input from wordnet and based on density of concept they sense the max probabilities algorithm results in good efficiency

Many of the sentiment analysis systems only determines the polarity of the sentences without any sense disambiguation. Although it is considered as one of the most important tasks in sentiment analysis, we can find only few methods proposed in literature for this purpose.

Due to improper sense disambiguation there are plenty of errors in sentiment analysis [6]. In some research works [7][8] the researchers had started their work by first creating the lexicon dictionaries where the words are already associated with the prior polarity. However, this contextual polarity of words that has been appeared in the phrase may be different from that word's prior polarity, this is because the words may appear in different senses [6].

In [6] to determine the contextual polarity of the given phrase a method has been proposed. This is based on a fact that polar clauses must identified first since the neural clauses are frequent. To determine the polarity of the polar clauses, context of word must be used with linguistic features such as sentence feature, modification feature and structure feature [9]. However, in this approach the word context is basically used to determine the effect of enhancers, negation and modifiers instead of disambiguation the word sense [9]. A word sense disambiguation method at sentence level is considered in [10], this is based on classifying the parts of speech presented in the sentences where it will be matched.

While determining the polarity in this approach, the parts of speech pattern of the sentences is extracted and tried to match with the dictionaries framed, to identify the appropriate sense associated with the word. However, this may fail to produce satisfactory results, since the same parts of speech pattern may not have the same sense.

For example, let us consider two sentences, "the mobile is small", "the hotel is small" [9], here both the sentences have included the word "small". However, in the first sentence it is used to indicate positive sense so that the mobile may fit in to the pockets, but in second sentence it is

used to indicate negative sense, that the hotel is small to stay. And many of the sentiment analysis systems are based on the pre-built lexicon dictionaries, where this includes the opinionated words. The issue in these lexicon dictionaries is they only lists the common sense words. Hence, it is difficult them to transform in to sentiment orientation sense and also, they do not support the matching mechanism in order to disambiguate the word sense.

IV. CONCLUSION

Main objective of our project is to resolve the ambiguation in analysis of users' review. By studying various journals and views of the authors we got to know how we should approach the problem. The ambiguation depends on the prefix and suffix of the ambiguous word as well as the subject of the sentence. We will classify the reviews based on the positive and negative datasets using sentimental analysis. But with respect to ambiguous words we need to consider the prefix and suffix. We will create a separate dataset for the ambiguous words, which will be referred only if the prefixes and suffixes direct the sentence towards ambiguity.

ACKNOWLEDGMENT

The authors express gratitude towards the assistance provided by our mentors and faculty members who monitored us throughout the research and helped us in achieving desired results in given time

REFERENCES

[1] ChunHui Zhang,Yiming Zhou,Trevor Martin, "Genetic Word Sense Disambiguation Algorithm".
[2] Lin Han,Xinjie Deng,Guohua Wu, "A Knowledge-Based Word Sense Disambiguation Algorithm Utilizing Syntactic Dependency Relation".
[3] S.G Kolte,S.G Bhirud "Word Sense Disambiguation using WordNet Domains".
[4] Udaya Raj Dhungana,Subarna Shakya,Kabita Baral and Bharat Sharma, "Word Sense Disambiguation using WSD Specific WordNet of Polysemy Words".
[5] Bin Shi, Liying Fang, Jianzhuo Yan, Pu Wang,"Word Sense Disambiguation of Semantic Document".
[6] Theresa Wilson, Janyce Wiebe, Paul Hoffmann, "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis".
[7] Soo-Min Kim, Eduard Hovy, "Determining the Sentiment of Opinions".
[8] Hong Yu, Vasileios Hatzivassiloglou, "Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences".
[9] Umar Farooq, Tej Prasad Dhamala, Antoine Nongaillard, Yacine Ouzrout and Muhammad Abdul Qadir, "A Word Sense Disambiguation Method for Feature Level Sentiment Analysis".
[10] A.Khan, B.Baharudin, K.Khan, "Sentiment Classification using Sentence-Level Lexical Based Semantic Orientation of Online Reviews".

