A Method to Overcome Misspelled Words in Reviews using Pattren Matching Technique

Ms. Manushree A. M [1]
B.E.,(M. Tech)
Department of Computer Science and Engineering
Adichunchanagiri Institute of Technology
Chikkamagaluru, Karnataka, India

Mr. Adarsh M. J [2]
B.E., M. Tech
Assistant professor,
Department of Computer Science and Engineering
Adichunchanagiri Institute of Technology
Chikkamagaluru, Karnataka, India

***Abstract*— The current era has a trend to share their opinion about any software launched, product or anything will be posted on web sites in the form of reviews. These reviews contain important data about the particular product. The reviews considered useful only when they are well processed and made fit to the analyzing system. This paper deals with the misspelled aspect problem using Pattern Matching Technique (PMT) in opinion mining.**

Keywords—Opinion mining, pattren matching, sentiment analysis, Aspect word, Training set.

I. INTRODUCTION

Data mining is a vast field that has a main concept of processing, mining of data using mathematical techniques and models to achieve and to extract desired data from a big set of information. Data extraction can achieved using different methods and pattern matching can consider as one of the best method to obtain solution to any problem. The review posted on the websites, blogs, and social media such as twitter, face book contain most use full information. This information in the form of reviews said to be useful only when they processed in well-defined procedure to make it suitable for the machine to analyze the concept. Reviews are a set of sentences contain some set of words to form an opinion about any product. Practically we cannot expect the user to write reviews, which are grammatically correct or always spelled in a right way.

Words are spelled wrongly due to many reasons like keyboard interruption or intentional. Intentional misspelled words considered as a trend in this current world. Some instances of spell mistakes committed due to keaboard interruption are “computer” has misspelled as “computre”, “good” misspelled as “bood”. These mistakes are considered to be more frequently committing spell mistakes because these are the mistakes that are not intentional but are often due while typing. Mistakes due to lack of knowledge about dictionary example the word “parallel” may become “paralell”, the word “beautiful” misspelled as “beautifull”.

Mistakes due to the influence of spoken language or wrong spelling interpretation example “mica” becomes “mika”, “rabbit” becomes “rabit” even though this mistakes are not intentional, still it affect the automatic aspect process. Mistakes that are intentional example the word nice has modified as “nce”, “tomorrow” will be modified as “2mro”, “good morning” will be changed to “gd mg”, “phone” becomes “fone”, the word “coming” becomes “cmg” are intentional and are probably as a trend in today’s world. These are popularly called as a text - messaging language that is the short form that are used while chatting.

This trend of text language concept leaves its influence both on messaging and in real world communication including reviews. The main challenge is to deal with these trendy spelling mistakes. The proposed PMT proposes a method to deal with these trendy misspelled words in sentences.

II. LITERATURE SURVEY

In [1] adopted opinion -mining concept to predict product use ability in the market. It uses a concept of conversion of unstructured data format to a structured data format using an idea of feature-opinion pair extraction and semantic orientation analysis. Here word frequency statistics considered to eliminate the words that have low frequency. This aimed to find customer view to find product evaluation.

In [2] designed for lingual separation and sentiment classification. In the first processing step is to classifying reviews into classes of languages and the next step is to classify text data into positive sentence and negative sentence. Here the analysis of online product dataset for two languages has considered.

In [3] adopted aspect level opinion mining and syntactic dependency based approach to obtain aggregate score of opinion word. Here aspect table has maintained for most frequent words, the aspect score has assigned based on the on the most frequently appearing aspect content in the set of statement where as the frequent data set.

A hybrid lexicon generator for a process of enhancement of domain knowledge a lexicon of sentence statics for word classification has demonstrated in [4] this method fits well for sentiment of machine learning than state of art for different data set supervised data for social media demine.

There exists an issue of compatibility between investigation and domains. A domain focused automated approach is generated for lexicon in sentiment of social

media data also a strategy of weighting for sentiment score for a lexicon of static version for the domain area focus has been discussed. An investment value for estate investment for the reason of mining for variety of data that are generated through user in a large scale reviews through online is in [5]. Multiple fine-grained features for different perspectives intentionally designed for estate values in the form of redundant and inter correlation has described in this paper. A ranker of pair wise has simulated in this work. Here a competitive of efficiency features for learning model.

III. PROBLEM STATEMENT

The purpose of this method is to develop a technique that has a capable of extracting aspect words in a data set irrespective of spell mistakes using a combined method of PMT and Naive Bayes training method. It aims to compare the accuracy between PMT and non-PMT and to obtain the increased accuracy a combined method of PMT and Naive Bayes has obtained.

IV. METHODOLOGY

The architecture of the proposed system has shown in the Fig 1. The proposed system has the following steps.

- Sentence Extraction.
- Pos –tagging.
- Processing.
- Training set interaction.
- PMT technique.
- Combined model analysis.

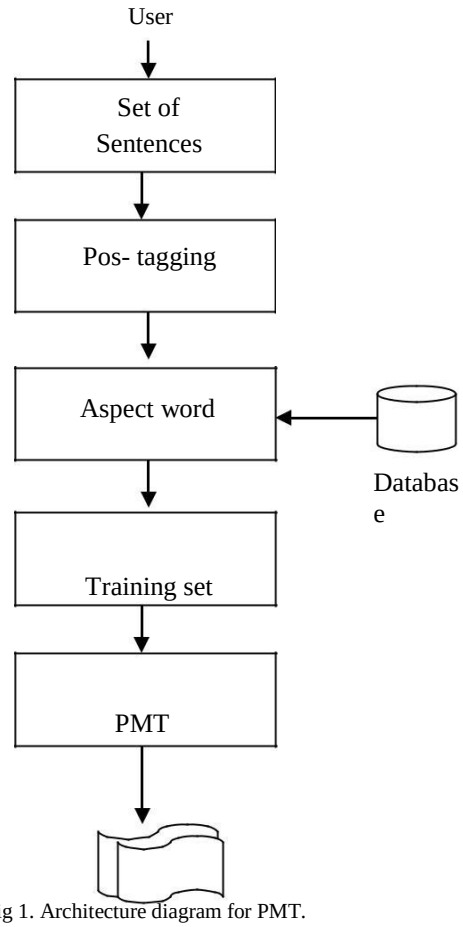

Fig 1. Architecture diagram for PMT.

A. User interaction

The system has designed to allow user to load a set of sentences for the system for analysis. In the present work, a set of 500 different sentences considered for analysis the user is responsible for loading this set of sentences to the system.

B. Processing

The set of sentences are processed using pos -tagging process. The pos- tagging refers to parts of speech tagging which implies identification of parts of speech in the sentence so as to identify all available parts of speech to get pos interrelation between different words in a sentence.

Aspect word allowed extracting from a set of database containing the words that spelled in a right form also which not spelled in a right form for the analysis. These words obtained from a set of sentences or from predefined database.

C. PMT technique

Pattern Matching Technique deals with the identification of all possible words from a set of misspelled words from a database containing training sentences. These training set has been processed using Naive Bayes text training method where it accepts all possible occurrence of words from the database. Combined model gives the increased in accuracy due to the increase in extraction of aspect words from a group

D. Methods

A. Natural Language Tool Kit (NLTK)

NLTK is a tool kit mainly designed for processing text in sentences. Which is best suited for Python platform. The paragraph has segmented in to sentences, up next these sentences can segmented in to words. The segmented words used for pos- tagging, where pos tagging refers to parts of speech tagging which involves identification of different parts of speech in a sentence and to tag its type. The pos-tag mainly used to identify relationship within a sentence to identify sentiment of any particular sentence.

The sentiment score has been assigned to each sentence based on Naïve Bayes algorithm where as these algorithm are all available as a built in libraries in NLTK and the accuracy with respect to aspect has also been increased when compared to SentiWordNet and it is very convenient to deal with twitter data due to available libraries.

B. TextBlob

TextBlob is a super set NLTK. This indicates that TextBlob has all feature of NLTK in addition it has some more additional features to make text processing more efficient. The sentiment score has assigned based on polarity subjectivity relation. The polarity subjectivity relations in TextBlob increase the accuracy of the model in sentiment analysis to provide score for the sentences for the analysis

purpose. The result obtained is much more nearer to the real world. It has designed to compute with even multiple sentences and sentences of moderate complexity.

Language identification facility has provided in TextBlob through this facility different languages can identified and languages can translated to any other languages because of a new facility called language translation. The translation method employees the platform to handle with the sentences, which are in other languages and this, helps to make a single computational environment for processing the data written in different languages.

Filtering of input data in the form of sentences involves removal of unwanted data from the set of input data such as smiles, new line characters, null characters white spaces. Normalization of text data involves identification of words which are used in its short form by the user and converting it to a formal computational acceptable way example "u're" will be normalized in to "you are" the normalization process 1.

E. Combined model analysis

A combined method of PMT and Naive Bayes gives a result, which is more accurate than the result of individual methods this for the reason that Naive Bayes [12] training method is used for training a machine to make it capable to extract any dynamic update in a system. The Bayesian rule given in Eq 1 P(c|x) represents the discriminative of posterior probability, P(x|c)P(c) represents the product of likelihood and prior. Finally, P(x) represents evidence for the system obtained through experience.

$$P(c|x) = \frac{P(x|c)\,P(c)}{P(x)} \qquad \text{……(Eq 1)}$$

The main of combining both PMT and Naïve Bayes training set is to make a machine more robust. The machine has a capability to act smartly for some situations, which are not predefined. The increase is robustness also increases the hit rate of a machine in a most considerable way by increasing the chances of obtaining the sentence, which contains aspect.

V. EXPERIMENTAL RESULTS

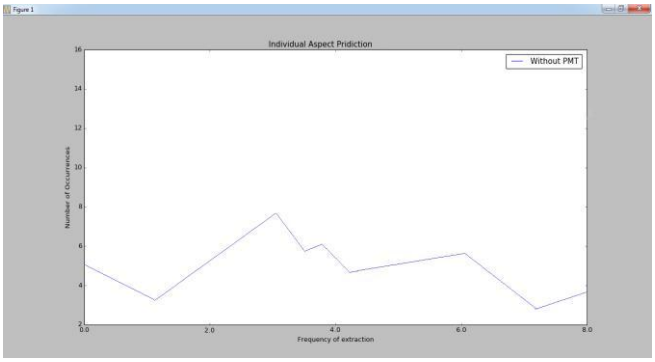

Fig. 2 Result for non-PMT method.

Fig 2. Shows the result obtained for non–PMT method. A data set of 500 sentences has considered for the testing processing. These sentences are stored in a database for further process. The result obtained through non-PMT method gives a less frequency of extraction of words from the data set. The main reason for the less frequency is non-PMT is capable for extracting the word that spelled correctly and it fails in identifying the words that are misspelled, even though that word has the same meaning.

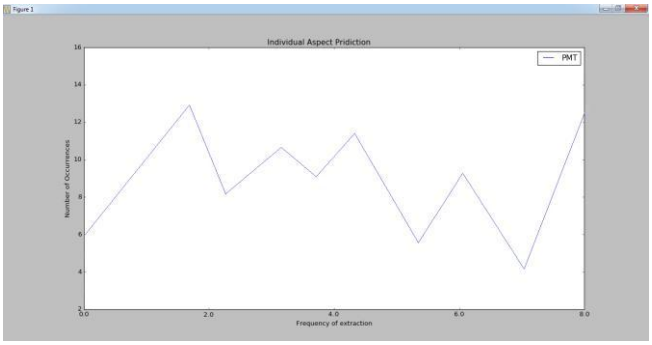

Fig. 3 Result for PMT technique.

Fig 3. Shows the result obtained after applying PMT technique. The graph shows a high value compared to that of non-PMT method this is for the reason that PMT is capable of identifying the words even though they are misspelled this increases the accuracy rate. A comparative graph shown in Fig 4 to prove PMT method can extract more aspect than non-PMT technique. The comparative graph gives increased and modified

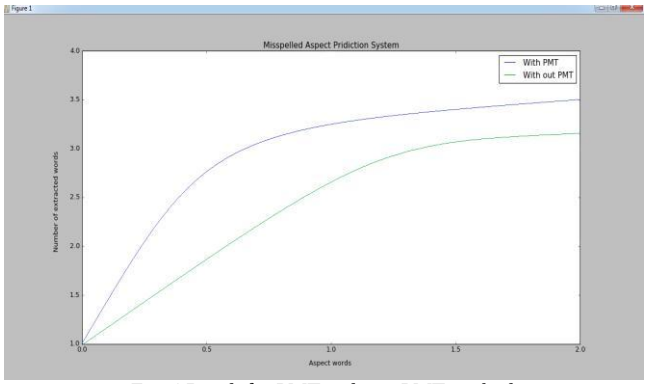

Fig. 4 Result for PMT and non-PMT method.

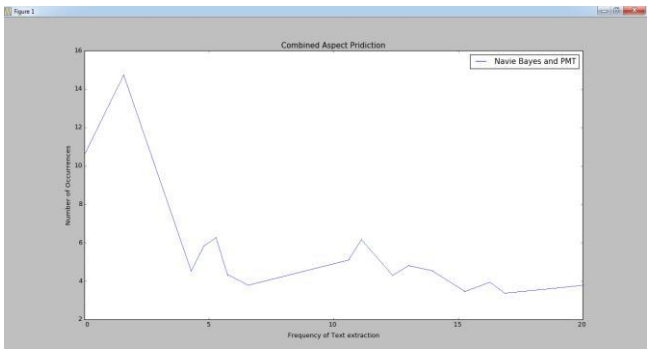

Fig. 5 Result for combined approach.

Fig 5 shows the result of a combined approach obtained through an aggregate method obtained using PMT and Naive Bayes text training method. The accuracy has increased with respect to number of words occurrence. However, the increase in word extraction is due to PMT the PMT increases the chance of hit in a set of huge sentences. The basic idea used is to use sub string matching in a comparatively large document in the search of aspect or required word in a document to increase number of sentences for analysis. The increase in accuracy in frequency is due to Naive Bayes training set while evaluation of the data set. The accuracy is directly dependent on the training set considered for training the machine.

TABLE 1. PERFORMANCE MEASURE OF MISSPELLED ASPECT PREDICTION SYSTEM.

Misspelled Aspect Prediction System			
Method		Frequency	Accuracy
Aspect word	Without PMT	18.6732	56%
	PMT	25.9964	67%
	Modified	36.7836	78%

Table 1 depicts the variation in the frequency and accuracy for the results obtained for the comparative method of without using PMT, with using PMT and a combined method obtained using PMT and Naive Bayes training center. The result obtained without using PMT has a low accuracy rate due to less hit rate frequency for the words. PMT method is a bit more advanced method than previous method PMT has a increased accuracy to 56% to 67% as it has a capability to overcome and extract misspelled words from a set of data. A new modified approach has obtained a still more increased accuracy from 67% to 78% as is uses the combined method of PMT and Naive Bayes training set.

VI. CONCLUSION AND FUTURE WORK

Misspelled words in reviews is a most common problem which often affect the accuracy of aspect words form a huge set of data set. The purpose of this paper is to obtain a comparative result of using the system without PMT technique and using PMT technique to show the increase in accuracy from 56% to 63%. A hybrid technique has demonstrated to show a more increased accuracy from 63% to 74% using a combined approach of PMT and Naive Bayes text classification training set.

Future scope

This technique can used for extraction of aspect form an immense set of data.

A large set of training data can used to obtain increased accuracy.

REFERENCES

[1] Mingxing Wu and Liya Wangand Li Yi, "A Novel Approach Based on Review Mining for Product Usability Analysis," National Natural Science Foundation of China IEEE 2013.
[2] Jantima Polpinij, "Multilingual Sentiment Classification on Large Textual Data," Fourth International Conference on Big Data and Cloud Computing IEEE 2014.
[3] Chinsha T C and Shibily Joseph, "A Syntactic Approach for Aspect Based Opinion Mining," 9th International Conference on Semantic Computing, IEEE, 2015.
[4] Hsiang Hui Lek and Danny C.C. Poo, "Aspect-based Twitter Sentiment Classification," Computational Linguistics, International Conference on Tools with Artificial Intelligence, 2013.
[5] S.Nirmala Devi, Dr.S.P Rajagopalan and Dr.V.Anuradha, "Index Based Multiple Pattern Matching Algorithm Using Frequent Character Count in Patterns," International Journal of Advanced Research in Computer Science and Software Engineering,2013.
[6] Dr. Muhammad Shahbaz, Dr. Aziz Guergachi and Rana Tanzeel ur Rehman, "Sentiment Miner: A Prototype for SentimentAnalysis of Unstructured Data and Text," IEEE 2014.
[7] Hsiang Hui Lek and Danny C.C. Poo, "Aspect-based Twitter Sentiment Classification", International Conference on Tools with Artificial Intelligence, IEEE 2013.
[8] Hongyu Mao, Keqin Wang, Rui Ma, Yifan Gao, Yuanzhi Li Kun Chen, Dejun Xie, Wei Zhu, Ting Wang and Huaiqing Wang, "An Automatic News Analysis and Opinion Sharing System for Exchange Rate Analysis", 11th International Conference on e-Business Engineering, IEEE 2014.
[9] Y. Jo and A. H. Oh, "Aspect and sentiment unification model for online review analysis," in *WSDM'11*. Hong Kong, China: ACM, 2011, pp. 815–824.
[10] C. Li, J. Zhang, J.-t. Sun, and Z. Chen, "Sentiment Topic Model with Decomposed Prior," in *SDM'13*. Austin, TX, USA: SIAM, 2013, pp. 767–776.D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *ICML'06*. Pittsburgh, PA, USA: ACM, 2006, pp. 113–120. T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, pp. 5228–5235, Apr. 2004.
[12] X. Wang and A. McCallum, "Topics over time: a non-Markov continuous-time model of topical trends," in *KDD'06*. Philadelphia, PA, USA: ACM, 2006, pp. 424–433.
[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *JMLR*, vol. 3, pp. 993–1022, 2003.

BIOGRAPHIES

Ms. Manushree A. M is a student of Computer Science in, Adichunchanagiri Institute of Technology, Chikkamagaluru, Presently pursuing M. Tech (CS & E).

She has received B.E from Adichunchanagiri Institute of Technology, affiliated to Visvesvaraya Technological University, Belagavi in the year 2015.

Mr. Adarsh .M J is working as an Assistant professor in the Department of computer science & Engineering, AIT college, Chikkamagaluru. He has 11 years of teaching experience and he has obtained his M. Tech from N. I. E, Mysore in the year 2009. His research fields are Data Science and Data Analytics in Social Networks.