An Enhanced Lexical Resource for Text Mining and Sentiment Investigation

Aftab Alam*, Mohammed Irfan
Department of Computer Science
King Khalid University
Abha, KSA
*Email: aftabjh [AT] gmail.com

Abdul Mateen Ansari
Department of Information Systems
Bisha University
KSA

Abstract— **In modern era Text Mining and Sentiment investigation is a growing research area, covering over several disciplines such as data mining, text mining, etc. Text mining is a vital art of extracting the texts from the huge set of text set or reviews. Sentiment investigation is a category of natural language processing for tracking the mood of the public about a specific product or theme. The present works of text mining used Sentiwordnet as a lexical resource. The major drawback of this present Sentiwordnet is non-determination of total count, such as it doesn't offer the specific number of +1, -1 and 0 total words. These data and facts are necessary because without these details of total count if further data mining techniques are applied, it may give erroneous results. To facilitate the text mining task, this work focus on design of Developed Sentiwordnet so that it can produce the count of total words by distinguishing them into +1, -1 and 0 words. Experiments are conducted on standard movie review and product review datasets. These works also make use of Stanford POS tagger for labeling the dataset. The calculated words can be used to enhance the results comparatively better.**

Keywords-component; Text mining; Sentiment Investigation; POS tagging; Sentiwordnet;

I. INTRODUCTION

Texts are part of daily life. Billions of users share Text on different aspects of life every day. Text Mining and Sentiment Investigation go hand in hand. Sentiment investigation is a type of natural language processing for tracking the mood of the public about a particular product or topic. Sentiment investigation, involves in building a system to collect and examine Text about the product made in blog posts, comments, reviews or tweets. Poongodi and Radha define the text mining as identification of objectivity or subjectivity in statements [1]. Malik Muhammad Saad MISSEN, in his thesis work describes text mining as the task of differentiating between truthful and textured information. Bing Liu, in his work related sentiment investigation with text mining by stating "Sentiment investigation is the field of study that analyses people's Text, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes [2, 3].

Dissimilar other domains, the data mining also use datasets for the experiments. A dataset is defined as a collection of related data. Datasets consist of all of the information gathered during a survey which needs to be analyzed. The data in a dataset can be anything, like movie, fruit, flower, product, image to name a few. And even a dataset can contain numerical, binary, nominal data etc., it depends on the choice of the researcher to select the right dataset for his/her research work. In this educative era, people are very much interested in learning new, develop something new. Datasets play an important role in conducting experiments. There usage is not restricted to specific field. They are used in many fields like data mining, image processing, natural language processing techniques etc., to name a few. Comparison of results can be done by conducting experiments with different datasets. The quality of the dataset, size of the dataset and domain of the dataset are some of the factors which affect the sentiment analysis process. This work uses the standard movie review and product review datasets for the experiments of Developed SentiWordNet.

WordNet is a openly available database of words containing a semantic lexicon for the English language that organizes words into groups called synsets, (i.e., synonym sets). A synset is a collection of synonym words linked to other synsets according to a number of different possible relationships between the synsets, (e.g., are-a, have-a, are-part-of, and others). SentiWordNet is a publicly available lexical resource for research purposes providing a semi-supervised form of sentiment classification [4] based on the annotation of all the synsets of WordNet according to the notions of "positivity", "negativity", and "neutrality". Each synset s is associated to three numerical scores Pos(s), Neg(s), and Obj(s) which indicates the degree to which the terms in the synset are positive, negative, objective, (i.e., neutral), respectively. The following Figure 1 shows the result of existing SentiWordNet for a particular set of words.

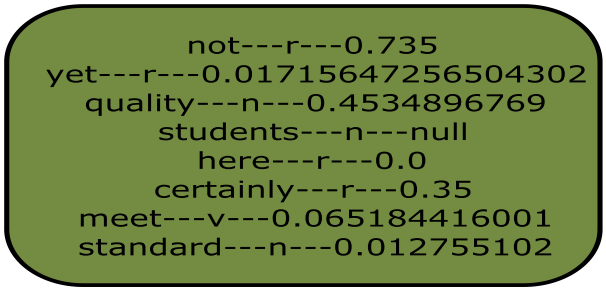

Figure 1. Sample Result of Present SentiWordNet.

A. *Dictionary Structure*

SentiWordNet dictionary can be downloaded online as a text file. In this file, the term scores are grouped by the synset and the relevant part of speech. Each entry in the dictionary is made of seven fields: the part of speech, the offset, the positive total, the negative total, the objective total, the associated synset terms, and an example of the context in which the term may be used. The parts of speech used in this dictionary are limited to adjective, noun, verb, and adverb. The offset is a numerical value which uniquely identifies a synset in the database. Associated synset terms are list of terms included in that particular synset. The Developed SentiWordNet overcomes the drawback of existing SentiWordNet by providing better accuracy in data mining techniques. As it produces the count of the total words along with the scores, this can be used as an efficient lexical resource in Text mining and sentiment analysis.

II. LITERATURE SURVEY

There is an increasing number and variety of research papers in the area of sentiment analysis and classification. We chose to consider only related work which makes use of the following: document-level classification, n-gram features (such as unigrams), Part-of-speech tagging, lexical resources (especially SentiWordNet), and the review domain, such as movie reviews and product reviews. Pang and Lee applied machine learning techniques to classify movie reviews according to sentiment [5]. They employed POS Tagger for tagging of reviews and Sentiwordnet for scoring and analysis of movie reviews. Their work doesn't consider the total count, i.e., number of positive total words, number of negative total words, and number of neutrally total words and directly the result of SentiWordNet is used for classification.

Ohana and Tierney studied sentiment classification using features built from the SentiWordNet database of term polarity scores [6]. Their approach started by building a data set of relevant features using SentiWordNet and a machine learning classifier. They have used a three-fold classification approach and results obtained were similar to those obtained using manual lexicons seen in the literature. Their work demonstrated that the results obtained from SentiWordNet could be used as an important resource for sentiment classification tasks. Here also, they didn't consider the count of positive, negative and neutrally total words in detail.

A simple unsupervised learning algorithm for classifying a review as recommended or not recommended was presented by Turney [7]. The algorithm takes a review as input and produced classification. They followed a three step approach: using a part-of-speech tagger to identify phrases in a review that contain adjectives or adverbs, estimating the semantic orientation of each phrase extracted, and assigning the review to a class, either recommended or not recommended, whose decision is based on the average semantic orientation of the extracted phrases. If the average is positive, the review is assumed to recommend the item, otherwise, the item is not recommended. The point wise mutual information and information retrieval algorithm is used to measure the similarity of pairs of words or phrases to estimate the semantic orientation of a phrase.

Dave et al. developed ReviewSeer, a document level Text classifier that uses statistical techniques and POS tagging information for sifting through and synthesizing product reviews, essentially automating the sort of work done by aggregation sites or clipping services [8]. They first used structured reviews for testing and training, identifying appropriate features and scoring methods from information retrieval for determining whether reviews are positive or negative. These results performed as well as traditional machine learning methods. They then used the classifier to identify and classify review sentences from the web, where classification is more difficult. They were able to obtain fairly good results for the review classification task through the choice of appropriate features and metrics, but they identified a number of issues that make this problem difficult like rating inconsistency, sparse data, skewed distribution, and ambivalence comparison.

III. DESIGN AND IMPLEMENTATION

Even though existing SentiWordNet gives the scores of the words, the major drawback of it is non-determination of total count, i.e., it doesn't provide the details of number of positively total, number of negatively total and number of neutrally total words. To facilitate the Text mining task, the Sentiwordnet is slightly developed so that it can produce the count of total words by distinguishing them into positive, negative and neutral words. The count of the total words is important because it helps the researcher to recognize how many words are positively total, how many are negatively total and how many are neutrally total words.

Knowledge of total word count is important. Because, without this knowledge, if the work is continued by applying other data mining techniques, there is a chance of obtaining wrong results. Hence, Developed SentiWordNet is developed to get more accurate results. For example, the sample result generated for one document by Developed SentiWordNet is shown in Figure 2. Each line represents a term, its part-of-speech and total, where, "a" represents adjective, "v" represents verb, "r" represents adverb, and, "n" represents a noun.

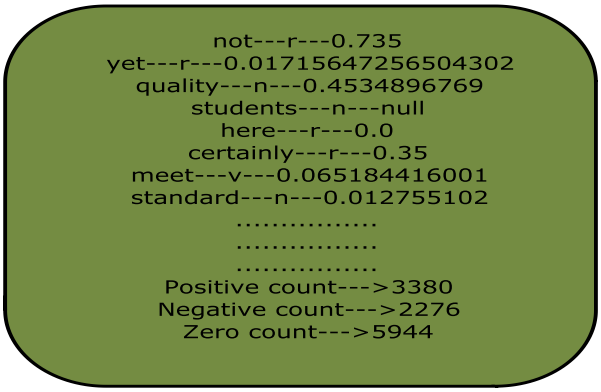

Figure 2. Sample Result of Upgraded SentiWordNet

The Developed Sentiwordnet is implemented on standard movie and product review datasets. After the preparation of dataset it needs to be pre-processed before experimenting it for the task of Text mining. Over all Implementation Procedure of Proposed Work in Text mining and Sentiment Analysis is explained here. The Figure 3 shows various steps of processing the dataset.

A. Data Preparation

This is the first step of pre-processing. It can be regarded as the most important step because the quality of the dataset also depends on Text mining and sentiment analysis process. Following Figure 3 shows the pre-processing steps.

B. Stop Words Removal

Stop words are words which are filtered out prior to, or after, processing of natural language data (text). There are some of the most common, stop words such as the, is, at, which, and on. These stop words can cause problems when searching for phrases that include them, particularly in names such as 'The Who', 'The The', or 'Take That'. So these are removed by simple coding.

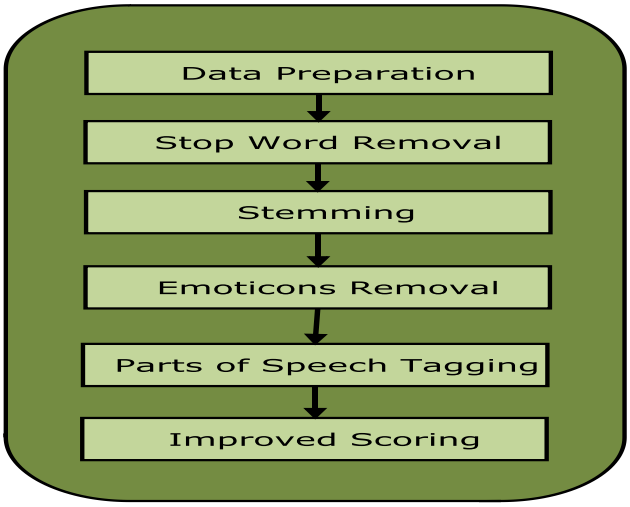

Figure 3. Phases of Pre-processing of Datasets

C. Stemming

Online reviews are generally used with informal language and they include internet jargons, slang and contemporary spellings say, e.g., use of apostrophes, ing form of words to name a few. So such words must be re-visited and stemmed for correct data retrieval. Basic stemming is employed.

D. Emoticons Removal

There is a multitude of emoticons that are used frequently in online reviews. Since this work does not focused on emoticons they are ignored and discarded.

E. Parts of Speech Tagging

The reviews are tagged by their respective parts of speech. For this the POS Tagger is used. A POS tagger parses a string of words, (e.g., a sentence) and tags each term with its part of speech. For example, parsing the following text which is taken from our dataset: Every term has been associated with a relevant tag indicating its role in the sentence, such as VBZ (verb), NN (noun), JJ (adjective), etc; The entire list of tags and their meaning is based on the Penn Treebank Tagset, an annotated corpus which seems to be the most popular standard used in most POS tagging.

F. Developed Scoring

To facilitate the Text mining task, the Sentiwordnet is slightly developed so that it can produce the count of total words by distinguishing them into positive, negative and neutral words. The count of the total words is important because it helps the researcher to recognize how many words are positively total, how many are negatively total and how many are neutrally total words. Without this knowledge, if the work is continued by applying other data mining techniques, there is a chance of obtaining wrong results. Hence, Developed Sentiwordnet is developed to get more accurate results.

G. Implementation with Movie Review Dataset

The pre-processing steps are applied for movie review datasets. From the dataset, stop words are removed. Tokenisation is performed. POS tagger is used to tag the dataset with their parts of speech. For example of parsing the following text which is taken from Movie Review dataset: "This movie is very interesting. The unique point about this movie is the Villon has played superb when compared to hero" produces the result as shown in Figure 4.

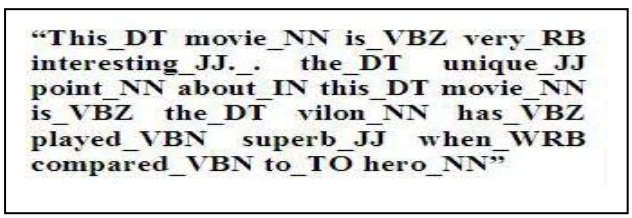

Figure 4. Sample POS Tagger result for Film Review Dataset.

The tagged documents are fed as input to Sentiwordnet for scoring and count of total words is determined. Figure 5 shows the total words with count.

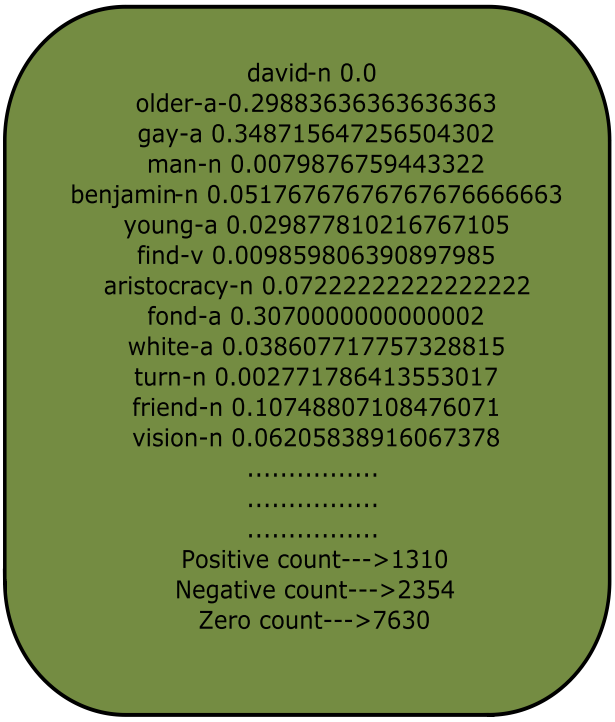

Figure 5. Sample SentiWordNet Result for Movie Review Dataset

H. Implementation with Product Review Dataset

The pre-processing steps are applied to product review dataset. From the dataset, stop words are removed. Tokenization is performed.

POS tagger is used to tag the dataset with their parts of speech. For example, parsing the following text which is taken from Product Review dataset: "Quality of this product is very bad, to look it is good but features are very worst" produces the result as shown in Figure 6

Quality_NNP of_IN this_DT product_NN is_VBZvery_RB bad_JJ ._. to_TO look_VB it_PRP is_VBZ good_JJ but_CC features_NNS are_VBP very_RB worst_JJS

Figure 6. Sample POS Tagger Result forProduct Review Dataset

Figure 7 shows the total words with count. The tagged documents are fed as input to Developed Sentiwordnet. The total and count of total words is determined.

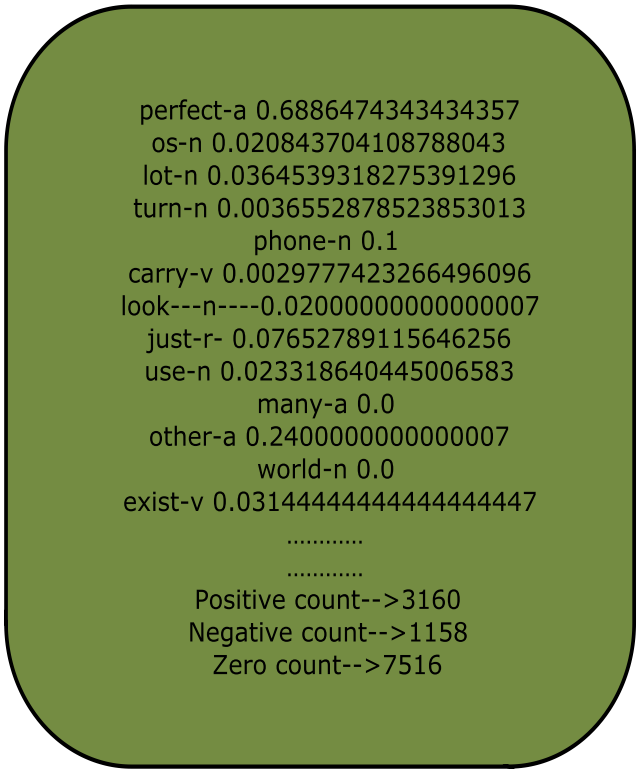

Figure 7.Sample SentiWordNet Result for Product Review Dataset

IV. RESULTS AND DISCUSSIONS

This section gives the results of the experiments. The final result of the preprocessing step is the word and its respective total (positive, negative, neutral) and count of the total words. Experiments are conducted on different datasets namely movie review and product review to get a better understanding scoring method applied for Text mining.

A. Results of Movie Review dataset

The Table 1 shows the count of positively, negatively and neutrally total words, for Movie review dataset.

TABLE I. RESULT OF MOVIE REVIEW DATASET.

WORD	COUNT SCORED
1500	Positive
905	Negative
3000	Neutral

B. Results of Product Review dataset

The Table 2 shows the count of positively, negatively and neutrally total words, for Product review dataset.

TABLE 2. RESULT OF PRODUCT REVIEW DATASET.

WORD	COUNT SCORED
2500	Positive
1500	Negative
7000	Neutral

V. COMPARISON OF RESULTS

The experiment is conducted and tested for movie reviews and product reviews dataset. Table 3 shows the analysis of the proposed method results.

TABLE 3: PROPOSED METHOD SCORE DETAILS OF DIFFERENT DATASETS.

	MOVIE	WORD
Positive	1500	2500
Negative	905	1500
Neutral	3000	7000

The comparison of proposed work (with total count) and previous works (without total count) is as shown in Table 4. In the other works the count of the scores are not considered, and directly after obtaining the Sentiwordnet results classification techniques are applied. This may result in poor accuracy and inaccurate results in classification step as the knowledge of how many words are total will not be there. To overcome this, we employed scoring along with calculation of total count which will be very helpful in further classification process and produce better results in Text mining.

TABLE 4: RELATIVE RESULT.

S. No.	METHOD	DATASET	SCORE COUNT	REMARKS
1	Poongodi S, Radha N	Tweets	Not available	Inaccurate results
2	Pang and Lee	Movie	Not available	Cant consider as accurate results
3	Malik Muhammad Saad MISSEN	Product	Not available	Inaccurate results
4	Proposed	Movie	Available	Accurate results
5	Proposed	Product	Available	Accurate results

VI. CONCLUSION

Due to the spectacular development of web environment in recent years, the expression of Texts of users in specialized sites for evaluation of a topic, and also on social networking platforms, has become one of the main ways of communication. The large amount of information on these platforms make them viable for use as data sources, in applications based on Text mining and sentiment analysis. This paper discusses an approach of designing developed Sentiwordnet and analyses the performance on the same on different datasets. The future enhancement can be done by,

- Applying different classification techniques like Support Vector Machine (SVM), Naïve Bayes (NB) classifier etc.
- Applying the various feature selection techniques for classification approach say, Information gain, Chi Square, Categorical Proportional Difference to name a few.

REFERENCES

[1] Poongodi S., Radha N., Classification of User Text from Tweets using Machine Learning Techniques, Int. J. 2013; 3(9): ISSN 2277 – 128.

[2] Missen M. M. S., Boughanem M., Cabanac G. et al., Combining Granularity-based Topic-dependent and Topic-independent Evidences for Text Detection, [Ph.D.] Dissertation, Université Paul Sabatier-Toulouse III, 2011.

[3] Liu B., Sentiment Investigation and Text Mining, Synthesis Lectures on Human Language Technologies, San Rafael, Calif.: Morgan & Claypool, 2012; 5(1): 1–167p.

[4] Baccianella S., Esuli A., Sebastiani F., Sentiwordnet 3.0: An Enhanced Lexical Resource for Sentiment Investigation and Text Mining. in LREC, 2010, 10: 2200–2204p.

[5] Pang B., Lee L., Vaithyanathan S., Thumbs up?: Sentiment Classification using Machine Learning Techniques, In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10. Association for Computational Linguistics, 2002, 79–86p.

[6] Ohana B., Tierney B., Sentiment Classification of Reviews using Sentiwordnet, in 9th. IT & T Conference, Dublin Institute of Technology, Dublin, Ireland, 2009, 13p.

[7] Turney P. D., Thumbs up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews, In Proceedings of the 40th Annual

[8] Meeting on Association for Computational Linguistics. Association for Computational Linguistics, Philadelphia, 2002, 417–424p.

[9] Dave K., Lawrence S., Pennock D. M., Mining the Peanut Gallery: Text Extraction and Semantic Classification of Product Reviews, in Proceedings of the 12th International Conference on World Wide Web. ACM, 2003, 519–528p.