# Blog Comments Sentence Level Sentiment Analysis for Estimating Filipino ISP Customer Satisfaction

Frederick F, Patacsil, and Proceso L. Fernandez

***Abstract*—**Blog comments have become one of the most common means for people to express and exchange their personal opinions.In this study, we develop a system for automated opinion retrieval from blog comments to estimate the customer satisfaction for three main Filipino Internet Service Providers (ISPs). Data was first gathered from comments located in some of the most popular blog sites discussing the Filipino ISPs. A portion of the data collected was manually labeled in order to establish ground truth. Automatic word seeding, n-gram tokenization, stemming and other Sentiment Analysis (SA) techniques were appliedto extract useful informationfrom the textual data.We experimentedwith Naïve Bayes (NB) and Support Vector Machine (SVM) language classifiers to determine the sentiment polarity of the blog comments. The results of the experimentationshowedthattheSVM generally outperformed the NB.Furthermore, the model configurations involving SVM + tri-gram+ Porter stemmer + stop word and the SVM + tri-gram+ Porter stemmereach obtaineda sufficiently high classification accuracy of 87%.These good results indicate that it can be possible for some interested parties to have a sense of the sentiments of their customers by applying some automated sentiment analysis on blog comments

***Keywords*—**Sentiment Analysis, Blogs, Translator Machine, Naïve Bayes, Support Vector Machines.

## I. Introduction

SENTIMENT Analysis (SA) refers to the application of Natural Language Processing (NLP), computational linguistics, and text analytics to identify and extract subjective information in source materials such as those discussions about certain products and services [11]. In this study, we used SA to the data gathered from the comments section of selected blogs in order to estimate the customer satisfaction for the three main Filipino Internet Service Providers (ISPs).

Blogs are one of the platforms that allow internet users to express personal opinions about a specific topic. The comments section of blogs typically contains the reaction and opinions of the readers of a blog. Naturally, such comment section contains various kinds of expressions that are either positive or negative in nature, and these can be indicative of the customers' satisfaction.

In this study, we investigate major blog sites related to internet service providers. In the Philippines, there are three major ISPs, namely Smart Communication, Globe Telecom and Sun Cellular. As of 2012, Smart has 1.73 million total subscribers followed by Globe with 1.7 million subscribers and Sun with 650,000 total subscribers [2].

With the continuous growth of internet access, online user reviews are increasingly becoming the de-facto standard for measuring the quality of products and services. Many Filipino internet customers have expressed their sentiments about the quality of services and the speed of the internet access provided by an ISP through social networks such as Facebook, Twitter and blog sites. The Internet customer comments on online reviewsare becoming an influencing factor that may affect the decision of other customers about the products and the services of certain providers. Therefore, these sentiments are very important sources of information that IPS should seriously take into account in improving their services and development of their products. However, the sheer volume of online comments makes it difficult for any human to process and extract all meaningful information. As a result, there has been a trend towardsdevelopment of systems that can automatically summarize opinions from a set of reviews and display them in an easy-to-process manner. The process of analyzing and summarizing opinions is known as Sentiment Analysis (SA), a type of natural language processing for tracking the moods and sentiments of the public about a particular service, product or topic. Furthermore, SA may involve building a system or automated method of collecting and examining opinions about the products made through blog posts, comments, reviews or tweets [17].

In this study, we trained and comparedthe performance of the NB and SVM using automatic polarity classified dataset andtested the proposed modelusing manually labeled dataset.In addition, we experimentedon different model configurations such as elimination of stop word, stemming and the use of n-grams).

## II. Related Literature

A lot of researchers of sentiment analysis have applied various approaches to predict the sentiments of words, expressions or documents. These are Natural Language Processing (NLP) and pattern-based techniques, machine learning algorithms such as Naïve Bayes (NB), and Support Vector Machines (SVM) and even unsupervised and semi-supervised learning techniques. In this section we discuss some of these approaches.

Please submit your manuscript electronically for review as e-mail attachments. When you submit your initial full paper version, prepare it in two-column format, including figures and

Frederick F, Patacsil is Assistant Professor, Pangasinan State University, Phillipines. (Email ID: frederick_patacsil@yahoo.co.uk)

http://dx.doi.org/10.15242/IIE.E0215004

tables.

### *A. Unsupervised Sentiment Analysis*

There have been several previous researches on sentiment analysis utilizing unsupervised automatic classification using word seeds to classify the polarity of adocumentwhich can be an entire article, a paragraph or even a single sentence. Some researches that were conducted on sentiment classification using an unsupervised approach and seed of words are described below.

Zagibalov and Carroll [17] described and evaluated a method of automatic seed word selection for unsupervised sentiment classification of product reviews in Chinese. Their aim was to investigate means to improve the classifier by automatically finding a better seed word. They based the selection on: their initial seed using the following observations: (1) The initial seed should always be more often used without negation in positive texts, while in negative texts it is more often used with negation and (2) the seed occurs more often in positive texts than negative, and more frequently without negation than with it. They decided “Good” as their initial word seed. The results obtained are close to those of supervised classifiers and sometimes better, up to an F1 score of 92%.

Turney [15] presented in his paper a simple unsupervised learning algorithm that utilizes two arbitrary seed words (“Poor” and ”Excellent” ) to calculate the semantic orientation of phrases. The algorithm has three steps: (1) extract phrases containing adjectives or adverbs, (2) estimate the semantic orientation of each phrase, and (3) classify the review based on the average semantic orientation of the phrases. The core of the algorithm is the second step, which uses Pointwise Mutual Information and Information Retrieval PMI-IR to calculate the semantic orientation The sentiment of a document is calculated as the average semantic orientation of all such phrases. This approach was able to achieve 66% accuracy for the movie review domain at the document level. He found movie reviews to be the most difficult because of this argument “the whole review is not necessarily the sum of the parts”[16].

Another study that made use of unsupervised system and word of seed was conducted by Rothfels and Tibshirani[12]. They examined an unsupervised system of iteratively extracting positive and negative sentiment items which can be used to classify documents. They adopted the idea of semantic orientation to choose the initial set of seeds and they hand picked two sets of reference seeds, one positive and one negative. They used for positive set, somewhat arbitrarily, the words “Good", “Excellent", “Amazing", “Incredible", and “Great" and for our negative set, “Bad", “Poor", and “Tterrible". The results of their study achieved an accuracy of 5.5%, a modest but significantly better than a near-baseline accuracy of 50.3% with the original approach.

Zhang et al.[14], pursued the analysis of product reviews using a bootstrapping method to find the product features and opinion words in iterative steps. Furthermore, a method was presented to get the initial seeds of product features and opinion words automatically.

The task of selecting seed words includes the following steps:

(i) In the product reviews, choose a small set of features and opinions as “seed words”

(ii) Count the co-occurrence of candidates and seed words in the product reviews

(iii) Use a figure of merit based on these counts to select new seed words

(iv) Return to steJp (ii) and iterate n times

The experimental results of that study are encouraging. They indicate that the proposed method and techniques are effective in performing this task of feature-level opinion mining.

### *B. Supervised Sentiment Analysis*

Sentiment detection is a task under sentiment analysis which aims to automatically label a text, sentence and document as positive and negative. Machine learning methods have been used to classify sentiments of different textual data. Following are some studies that used machine translator and machine learning classifier in the conduct of sentiment analysis.

Pang, Lee and Vaithyanathan [9], considered the problem of classifying documents not by topic, but by overall sentiment, e.g., determining whether a review is positive or negative. They employed three machine learning methods (Naive Bayes, maximum entropy classification, and support vector machines) to classify sentiment polarity of a movie review. Interestingly, the result shows that standard machine learning techniques outperformed human-produced baselines. However, they stated that this learning machine does not perform on sentiment classification on traditional topic-based categorization.

Another study that explored different pre-processing techniques and employed various features and classifiers to a newly created Czech social media dataset was conducted by Habernal, Ptacek and Steinberger[8]. They undertook to establish a large human annotated Czech social media corpus and evaluated by a state-of-the-art supervised machine learning methods for sentiment analysis. In their in-depth research on machine learning methods for sentiment analysis of Czech social media,the result of their study using a combination of features (unigrams, bigrams, POS features, emoticons, character n-grams) achieved F-measure 0.69 which significantly outperformed the baseline (unigram feature without preprocessing) in three-class classification.

A research that examined how classifiers work over Spanish Twitter data was conducted by Sidorov, Jiménez and Jiménez. They explored different settings (n-gram size, corpus size, number of sentiment classes, balanced vs. unbalanced corpus, various domains) that may affect precision of the machine learning algorithms. They made use of machine learning likeNaïve Bayes, Decision Tree, and Support Vector Machines as classifier experimental tools and came up with the best settings of parameters for opinion mining in Spanish Twitter. The results revealed that the best configuration of parameter when classifying opinion in Spanish are the following:(1) using unigrams as features, (2) using less possible number of classes: positive and negative, (3) using

http://dx.doi.org/10.15242/IIE.E0215004

at least 3,000 tweets as training set (incrementing this value does not improve precision significantly), (4) balancing the corpus with regard to the proportional representation of all classes obtained slightly worse results, and (6) Supported Vector Machine was the classifier with the best precision[15].

A similar study conducted by Shoukry[13]demonstrates an application of Arabic sentiment analysis by implementing a sentiment classification for Arabic tweets. The retrieved tweets are analyzed to provide their sentiment polarity (positive, or negative). The process starts with the translation of Arabic sentences into English using one of the standard translation softwares. Then, translated sentences were classified according to its sentiment classes "positive" and "negative" using machine learning systems.In addition, the researcher explored several settings that may affect the classification performance of the classifiers. The study proposed hybrid system that used all the identified features from the ML approach, and the sentiment lexicon from the SO approach, resulting in an accuracy and recall of 80.9%, while its precision and F-measure is 80.6%.

### C. Our Approach

In this paper, we presentedand comparedthe classification performance of NB and SVM under various configurations of stemming, eliminating stop words and n-gram tokenization. The end goal is to automatically determinethe sentiment of blog comments based on the use of English translator tools to translate online comments in Filipino into English form. We experimented on an automated polarity classifier using a bag of words and a machine learning classifier that utilized automated polarity sentences and test the classification performance usingmanually labeled sentences (ground truth).

## III. General Methodology

### A. Information Extraction

In this study, we first searched for blog articles from Google that discuss and compare the services of the three main ISP. Blog articles that contain many comments from their customers were highly considered. The following are examples of blog articles that were included in the data: "Comparing Globe Tattoo, SmartBro and Sun Broadband Wireless » tonyocruz.com", "Globe Tattoo vs. Smart Bro vs. Sun Broadband Wireless Which is the best Reader Comments TechPinas Philippines' Technology News, Tips and Reviews Blog" and "The Best ISP (Internet Service Provider) in the Philippines Jehzlau Concepts".

Blog comments from the selected blog sites that feature the services of the major internet providers were extracted using a customized PHP web scraping application which retrieves the customer comments and other important information, and then stores these in a database automatically. The unnecessary parts such as the title, articles, and other contents were not included in the mining process. Furthermore, since the comment area is composed of several HTML tags, unnecessary characters and non-textual contents were stripped out using a modified PHP application.

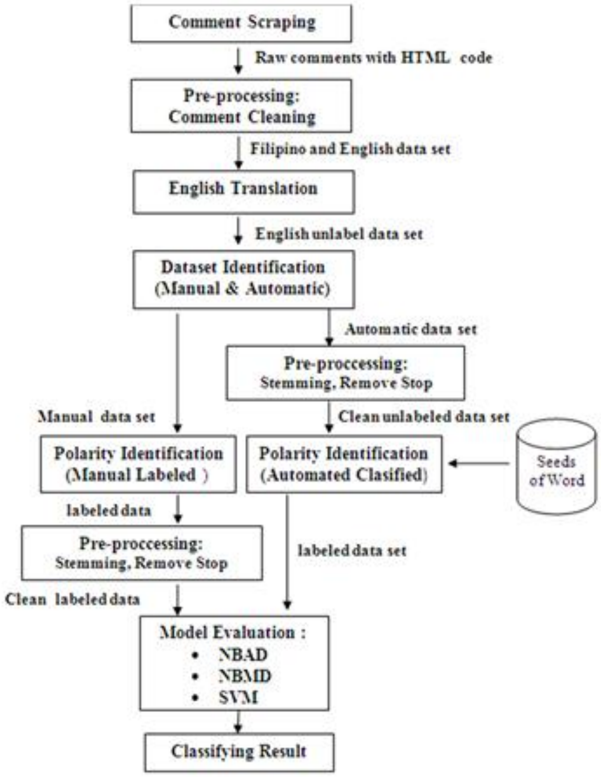

Comment Scraping
Raw comments with HTML code
Pre-processing: Comment Cleaning
Filipino and English data set
English Translation
English unlabel data set
Dataset Identification (Manual & Automatic)
Automatic data set
Pre-proccessing: Stemming, Remove Stop
Manual data set
Clean unlabeled data set
Polarity Identification (Manual Labeled )
Polarity Identification (Automated Clasified)
Seeds of Word
labeled data
Pre-proccessing: Stemming, Remove Stop
labeled data set
Clean labeled data
Model Evaluation :
• NBAD
• NBMD
• SVM
Classifying Result

Fig.1 General Methodology of the Study

### B. Machine Translation

The comments were not all written in English. There were some Filipino words (in some instances even sentences) that were observed from the collected data. Thus, after data cleansing, a machine translation using Google Translate was employed. Specifically, a modified application in PHP was used to automatically convert Filipino sentences into their English equivalent using the Google Translate API. Filipino words that were not recognized by the machine were manually corrected and converted using the same tool.

### C. Building datasets

We utilized 14,000 sentences derived from 5280 blog comments. Such data were used to identify the sentiments of the customers, whether they are satisfied with the internet services provided by Globe, Smart, and Sun.

### 1. Building dataset for NBAD

#### a) Data preprocessing

Before the 14,000 sentences were fed into the automated polarity classifier system for polarity identification and labeling, preprocessing of the sentences is required. Removing stop words (common words that have a little value in the process of identifying sentiment analysis e.g. "a", "the", etc.) and stemming (words that carry similar meanings, but in different grammatical forms such as "connect","connects" and "connected" was combined into one word "connect") was also applied in the sentence preprocessing. In this way, the sentences can show a better representation (with stronger correlations) of these terms, and even the dataset can be reduced for achieving smallerer processing time.

*b) Polarity Identification*

The identification of sentiment polarity was done using a modified PHP application specifically for sentiment analysis. The application counted the number of positive and negative wordsand then computed the total score. If the total score (positive score-negative score) was greater than 0, it would be considered as a positive sentiment and if it was smaller than 0, it would be considered as a negative sentiment. The PHP code snippet for this is shown in Fig. 2.

```
for($x = 0; $x < count($comment); $x++){
  $seeddummy = "";
echo $comment[$x];
   $sdummy = explode(" ", $comment[$x]);
    $ssdummy = array();
for ($x1 = 0; $x1 < count($sdummy); $x1++){
       $ssdummy[$x1]  Stemmer::Stem(strtolower($sdummy[$x1]));
   }
foreach ($seeds as $key => $seed) {
    if(in_array(trim($seed),$ssdummy,true)){
            $seeddummy = $seeddummy." ".$seed;
    $count = $count + $seedid[$key];}  // count sentence polarity
```

Fig 2. Sample PHP Code That Identify Sentiment Polarity of the Sentences

The result of polarity identification by the PHP polarity application program is shown in table 1. Out of 14,000 sentences, only 3956 were found to be positive and negative sentences. Furthermore, there are 2585 negative sentences compare to 1371 positive sentences, indicating that there are more negative sentences than positive sentences based on automatic polarity identification.

TABLE I
DISTRIBUTION OF AUTOMATICALLY LABELED POSITIVE AND NEGATIVE SENTENCES

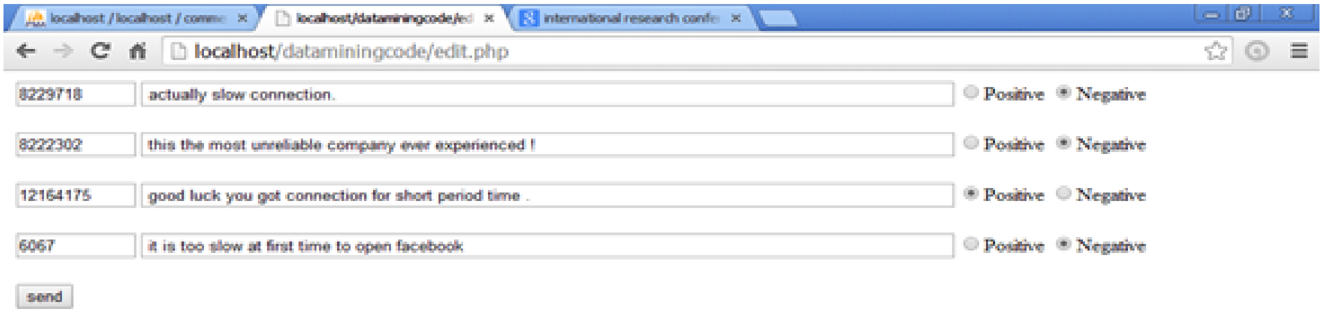
| **Sentences** | **No. of Sentences** |
|---|---|
| Negative | 2585 |
| Positive | 1371 |
| **Total** | **3956** |

*2. Building dataset for NBMD*

*a) Polarity Identification*

The same 14,000 sentences used in the automated sentiment classifier were also used as experimental data set for our machine learning classifiers.Four groups with three members each were tasked to manually label the polarity of the sentences. The formulation of the groups was based on the research conducted by Bogart. The results of his study suggest that one rater can rate pretty well, but three can rate better and there is not much gain after increasing to three[4].Group of raters were oriented and trained to apply the criteria foran event that has phenomenon (e.g. "slow internet, bad services" – Negative and "fast internet, good services"– Positive) should be applied. Then, members of the groups were selectedfrom the group of raters that were oriented and trained to identify the polarity of the sentences using systematic random sampling.

The raters used a customized PHP application to label the sentiment polarity of the sentences. A screenshot of the application is shown in Fig. 3


localhost / localhost / comm
localhost/dataminingcode/ed
international research confe

localhost/dataminingcode/edit.php

| | | | |
|---|---|---|---|
| 8229718 | actually slow connection. | ○ Positive | ● Negative |
| 8222302 | this the most unreliable company ever experienced ! | ○ Positive | ● Negative |
| 12164175 | good luck you got connection for short period time . | ● Positive | ○ Negative |
| 6067 | it is too slow at first time to open facebook | ○ Positive | ● Negative |

send

Fig. 3 Sample Screenshot Of The Application Used For Manually Labeling The Polarity Of The Sentences

The data were divided equally into four dataset and each group were given an equal number dataset to label manually. Three members of each group independently labeled same set of dataset. In order to check the raters consistency test, the results of the labeling process were tested using Kappa inter-rater reliability. Reliability of dataset is very important component of overall confidence in the accuracy results of this research study.

The interrater reliability of each group is shownintable2. Group 1 has an "almost perfect agreement" with a kappa value of $\kappa = 0.857$. Group 2 and 3 have "substantial agreement" with kappa values of $\kappa=0.76$ and $\kappa = 0.73$ respectively. $\kappa = 0.586$ is the lowest kappa value and was obtained by group 4. The average interrater reliability for the sentence polarity raters was found to be $\kappa = 0.74$. According to the definition of the Fleiss' Kappa statistic, the accuracy of the inter-rater reliability is considered to be "Substantial agreement". By convention, a $\kappa>0.60$ to $<=0.75$ is considered acceptable inter-rater reliability for applied test[3][7][10].

http://dx.doi.org/10.15242/IIE.E0215004

TABLE II
FLEISS' KAPPA AGREEMENT RESULTS OF THREE GROUP SENTIMENT POLARITY RATER

| Raters Group | Expected Agreement | Observation Agreement | Fleiss' Kappa |
|---|---|---|---|
| Group 1 | 0.965 | 0.754 | **0.857** |
| Group 2 | 0.892 | 0.547 | 0.76 |
| Group 3 | 0.914 | 0.681 | 0.73 |
| Group 4 | 0.815 | 0.552 | 0.586 |
| **Average** | | | **0.74** |

Legend: Landis and Koch [8] suggest the following interpretations:

| Kappa | Agreement |
|---|---|
| < 0 | Less than chance agreement |
| 0.01–0.20 | Slight agreement |
| 0.21– 0.40 | Fair agreement |
| 0.41–0.60 | Moderate agreement |
| 0.61–0.80 | Substantial agreement |
| 0.81–0.99 | Almost perfect agreement |

After the Fleiss' Kappa inter-rater reliability was determined, the three raters on the group agreed on the sentiment polarity of other sentences that were labeled differently to come up with a common polarity for each sentence. Furthermore, only sentences with positive and negative polarity were included as part of the experimental dataset.

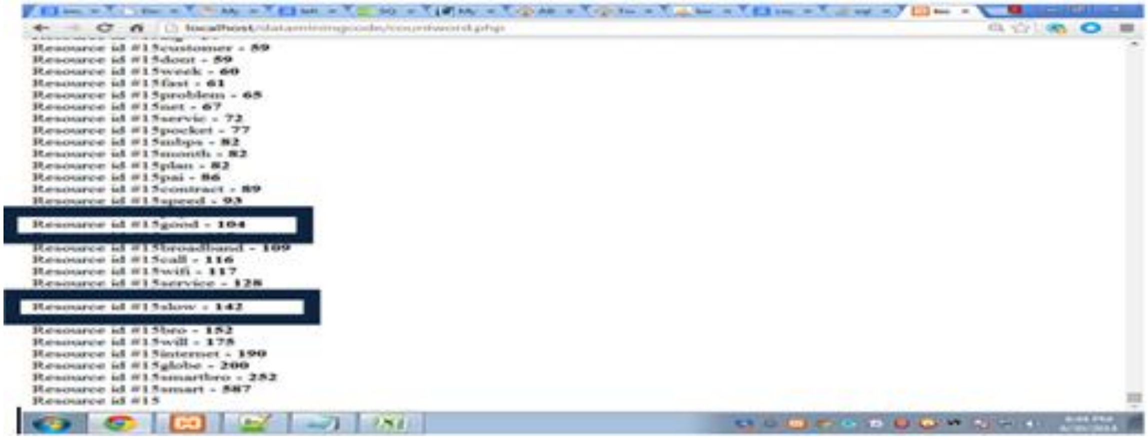
TABLE III
DISTRIBUTION OF MANUALLY LABELED POSITIVE AND NEGATIVE SENTENCES

| Sentences | No. of Sentences |
|---|---|
| Negative | 2793 |
| Positive | 539 |
| **Total** | **3332** |

### *D. Determination of Word Seeds*

Before carrying out a classification algorithm on the dataset, there was a need to create a word dictionary to be used as the base for sentiment analysis. Several research works utilized automatic seed selection using a set of words as initial seed for the dictionary. The pioneer work was initiated by Turney's [5] which classifies a document using two human-selected seed words (the word "Poor" as negative and "Excellent" as positive). Zagibalov and Carroll also utilized this approach and they described it as 'almost-unsupervised' system that starts with only a single, human-selected seed ("Good"). They also claimed that the 'almost-unsupervised system produces a better result [17].

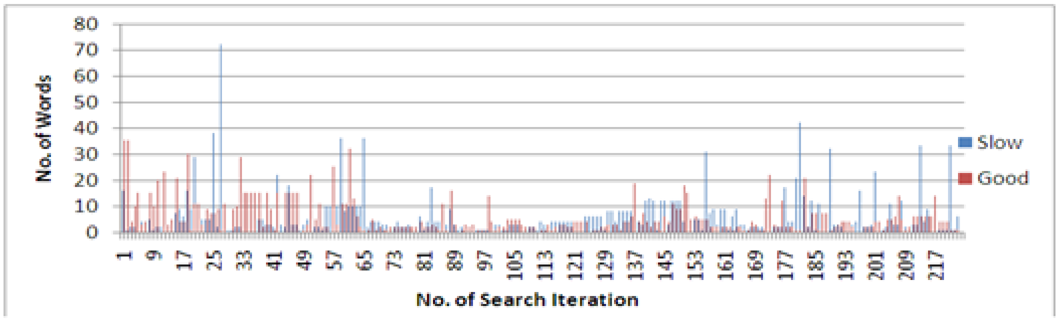
This research also utilized automatic identification of seed of words for the dictionary based on the data set. The adjective that had the most number of occurrences as highlighted in Fig. 5were used as the initial seeds.



```
Resource id #15customer - 59
Resource id #15dont - 59
Resource id #15week - 60
Resource id #15fast - 61
Resource id #15problem - 65
Resource id #15net - 67
Resource id #15servic - 72
Resource id #15pocket - 77
Resource id #15mbps - 82
Resource id #15month - 82
Resource id #15plan - 82
Resource id #15pai - 86
Resource id #15contract - 89
Resource id #15speed - 93
Resource id #15good - 104
Resource id #15broadband - 109
Resource id #15call - 116
Resource id #15wifi - 117
Resource id #15service - 128
Resource id #15slow - 142
Resource id #15bro - 152
Resource id #15will - 175
Resource id #15internet - 190
Resource id #15globe - 200
Resource id #15smartbro - 252
Resource id #15smart - 587
Resource id #15
```

Fig. 5 Word Count Based On the Training Data Set

The initial seed set consisted of the keyword {"Good" and "Slow"}. An application program automatically searches for and retrieves the synonyms and antonyms of the initial seeds from an online thesaurus dictionary using http://thesaurus.altervista.org/thesaurus/v1 as part of the word seeds of this research. After the first process, the application retrieved the first synonym word found and repeats the process of searching and retrieving of synonyms and antonyms from the online thesaurus dictionary. The process is repeated until no more new words have been added to the word collection.



*** No. of words searched (repeated word was automatically disregarded by the application system)
Positive – 930
Negative – 1351

Fig. 6 Result of Word Seeds Searching Using an Online Thesaurus Dictionary

## IV. Experimentation And Result

The proposed automated polarity classifier was trained to identify the polarity of unlabeled sentences and it uses its own predictions to teach itself to classify unlabeled sentences using a positive and negative bank word seeds. We conducted experiments to evaluate and compare the performance of NBAD and NBMD. Furthermore, we also tried SVM n-grams features in the classification of manually labeled sentences. The classification experimentations were carried out using 10-fold cross validation using Rapid Miner 5.3.In terms of comparing their performance, wemade use of the confusion matrix that contains the precision, recall, accuracy and F-measure measurement[1].

TABLE IV
CONFUSION MATRIX FOR TWO-CLASS CLASSIFIER

| Predicted Class | True Negative | True Positive | Class Precision | F-Measure |
|---|---|---|---|---|
| Negative | tn | fp | Negative | F-Measure Negative |
| Positive | fn | tp | Positive | F-MeasurePositive |
| Class Recall | Negative | Positive | | |

Accuracy:

The accuracy (AC) is the proportion of the total number of predictions that were correct. It is determined using the (1) :

$$AC = \frac{tn + tp}{tn + fp + fn + tp}$$

The recall (in the case of positive cases) is the proportion of positive cases that were correctly identified, as calculated using the (2):

$$\text{Recall} = \frac{tp}{tp + fp}$$

The precision n (in the case of positive cases) is defined as the proportion of negative cases that were classified correctly, as calculated using the (3):

$$\text{Precision} = \frac{tp}{tp + fn}$$

The F-Measure considers both precision and recall providing a single measurement for a system avoiding having two independent measures. It is computed using the (4):

$$FM = 2\frac{(\text{precision} * \text{recall})}{(\text{precision} + \text{recall})}$$

### *A .Experimenting Dataset*

a) Experimenting automated labeled dataset using Naïve Bayes and SVM.

We utilized the NB and SVM machine learning system to test the classification performance of the proposed automated polarity classifier. Table 5 reports the performance in terms of precision, recall, F-measure and accuracy.

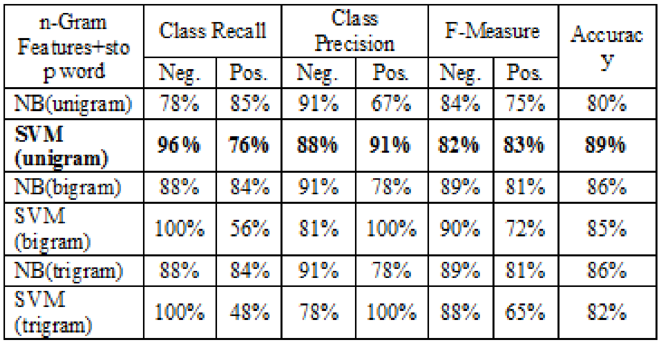
TABLE V-A
10-FOLD CROSS VALIDATION OF THE PROPOSED SENTIMENT CLASSIFIER USING AUTOMATED LABELED CLASSIFIED DATASET USING G-GRAM FEATURES

| | Class Recall | | Class Precision | | F-Measure | | Accur acy |
|---|---|---|---|---|---|---|---|
| am Features | Neg. | Pos. | Neg. | Pos. | Neg. | Pos. | |
| NB(unigram) | 79% | 83% | 90% | 68% | 84% | 75% | 81% |
| **SVM (unigram)** | **94%** | **83%** | **91%** | **89%** | **93%** | **86%** | **91%** |
| NB(Bi-gram) | 90% | 84% | 92% | 85% | 92% | 84% | 89% |
| SVM (bi-gram) | 100% | 62% | 83% | 100% | 91% | 76% | 87% |
| NB(tri-gram) | 88% | 86% | 92% | 80% | 90% | 83% | 87% |
| SVM (trim-gram) | 100% | 51% | 78% | 100% | 89% | 68% | 83% |

TABLE V-B
10-FOLD CROSS VALIDATIONOF THE PROPOSED SENTIMENT CLASSIFIER USING AUTOMATED LABELED CLASSIFIED DATASET USING N-GRAM AND PORTER STEMMER FEATURES

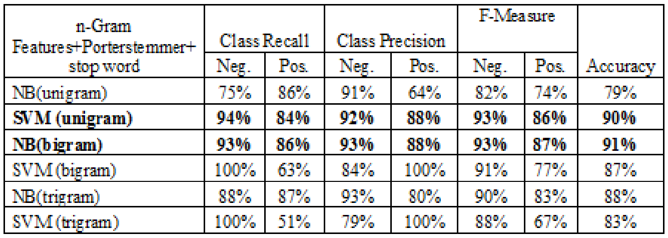
| n-Gram Features+Porter stemmer | Class Recall | | Class Precision | | F-Measure | | Accuracy |
|---|---|---|---|---|---|---|---|
| | Neg. | Pos. | Neg. | Pos. | Neg. | Pos. | |
| NB(unigram) | 78% | 86% | 91% | 67% | 84% | 75% | 81% |
| **SVM (unigram)** | **97%** | **89%** | **94%** | **93%** | **95%** | **91%** | **94%** |
| NB(bigram) | 93% | 87% | 93% | 88% | 93% | 87% | 91% |
| SVM (bigram) | 100% | 62% | 83% | 100% | 91% | 77% | 87% |
| NB(trigram) | 89% | 87% | 93% | 81% | 91% | 84% | 88% |
| SVM (trigram) | 100% | 51% | 79% | 100% | 89% | 68% | 83% |

TABLE V-C
10-FOLD CROSS VALIDATION OF THE PROPOSED SENTIMENT CLASSIFIER USING AUTOMATED LABELED CLASSIFIED DATASET USING N-GRAM AND STOP WORDS FEATURES



| n-Gram Features+stop word | Class Recall | | Class Precision | | F-Measure | | Accuracy |
|---|---|---|---|---|---|---|---|
| | Neg. | Pos. | Neg. | Pos. | Neg. | Pos. | |
| NB(unigram) | 78% | 85% | 91% | 67% | 84% | 75% | 80% |
| **SVM (unigram)** | **96%** | **76%** | **88%** | **91%** | **82%** | **83%** | **89%** |
| NB(bigram) | 88% | 84% | 91% | 78% | 89% | 81% | 86% |
| SVM (bigram) | 100% | 56% | 81% | 100% | 90% | 72% | 85% |
| NB(trigram) | 88% | 84% | 91% | 78% | 89% | 81% | 86% |
| SVM (trigram) | 100% | 48% | 78% | 100% | 88% | 65% | 82% |

TABLE V-D
10-FOLD CROSS VALIDATION OF THE PROPOSED SENTIMENT CLASSIFIER USING AUTOMATED LABELED CLASSIFIED DATASET USING N-GRAM, PORTER STEMMER AND ELIMINATION OF STOP WORDS FEATURES



| n-Gram Features+Porterstemmer+stop word | Class Recall | | Class Precision | | F-Measure | | Accuracy |
|---|---|---|---|---|---|---|---|
| | Neg. | Pos. | Neg. | Pos. | Neg. | Pos. | |
| NB(unigram) | 75% | 86% | 91% | 64% | 82% | 74% | 79% |
| **SVM (unigram)** | **94%** | **84%** | **92%** | **88%** | **93%** | **86%** | **90%** |
| **NB(bigram)** | **93%** | **86%** | **93%** | **88%** | **93%** | **87%** | **91%** |
| SVM (bigram) | 100% | 63% | 84% | 100% | 91% | 77% | 87% |
| NB(trigram) | 88% | 87% | 93% | 80% | 90% | 83% | 88% |
| SVM (trigram) | 100% | 51% | 79% | 100% | 88% | 67% | 83% |

Table 5a reveals the classifying accuracy results of n-gram feature using NB and SVM. SVM unigram achieved the highest classification accuracy of 91%. Furthermore, it also obtained the highest f-measure with 93% for negative classes and 86 percent for positive classes.

Table 5bshows that SVM obtained the highest classification accuracy of 94% under unigram with the application of Porter stemmer features. Moreover, the substantial increase of the accuracy has been seen only on SVM + unigram, but the remaining features have a minimal increase in their accuracy performance.

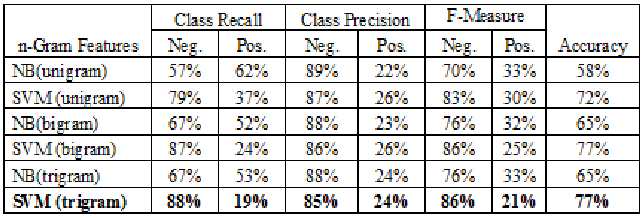
The results of n-gram and stop word is shown in table 5c, in which SVM unigram obtained the highest classification accuracy of 89%. Comparing the accuracy performance to other features, this combination obtained the lowest in all measure area. Moreover, table 5d reveals that NB bigram achieved 91%, which is the highest classification accuracy result when we implemented Porter stemmer and stop word elimination.

The comparisons of different features in different classifiers are shown in table 5a,table 5b, table 5cand table 5d. The best results were obtained using SVM both unigram + Porter stemmer achieving a classification accuracy of 94% and the lowest results were obtained by n-gram and stop word. It is also notable that the classifying accuracy of SVM features is almost the same both in bigram and trigram in the three experiments namely: n-gram, n-gram+stemmer, and n-gram+stemmer+stop word. Finally, combining features increase the accuracy performance of SVM except in the case of n-gram and stop word.

b) Testing the proposed method using the manually labelled sentences as test dataset.

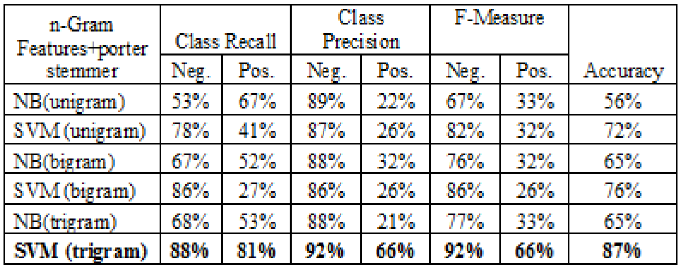
To evaluate the performance of the proposed method on sentiment classification, we use the manually labeled dataset with 2793 positive and 539 negative sentences.

TABLE VI-A
TESTING RESULTS OF THE PROPOSED SENTIMENT CLASSIFIER USING MANUALLY LABELED DATASET IMPLEMENTING N-GRAM FEATURES



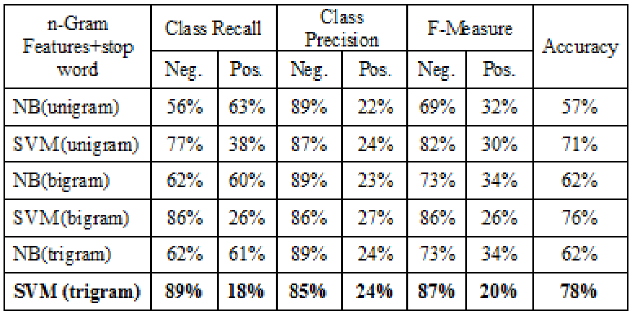
| n-Gram Features | Class Recall | | Class Precision | | F-Measure | | Accuracy |
|---|---|---|---|---|---|---|---|
| | Neg. | Pos. | Neg. | Pos. | Neg. | Pos. | |
| NB(unigram) | 57% | 62% | 89% | 22% | 70% | 33% | 58% |
| SVM (unigram) | 79% | 37% | 87% | 26% | 83% | 30% | 72% |
| NB(bigram) | 67% | 52% | 88% | 23% | 76% | 32% | 65% |
| SVM (bigram) | 87% | 24% | 86% | 26% | 86% | 25% | 77% |
| NB(trigram) | 67% | 53% | 88% | 24% | 76% | 33% | 65% |
| **SVM (trigram)** | **88%** | **19%** | **85%** | **24%** | **86%** | **21%** | **77%** |

TABLE VI-B
TESTING RESULTS OF THE PROPOSED SENTIMENT CLASSIFIER USING MANUALLY LABELED DATASET IMPLEMENTING N-GRAM AND PORTER STEMMER FEATURES



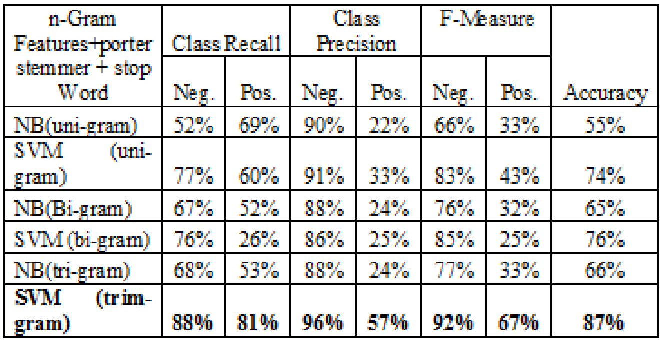
| n-Gram Features+porter stemmer | Class Recall | | Class Precision | | F-Measure | | Accuracy |
|---|---|---|---|---|---|---|---|
| | Neg. | Pos. | Neg. | Pos. | Neg. | Pos. | |
| NB(unigram) | 53% | 67% | 89% | 22% | 67% | 33% | 56% |
| SVM (unigram) | 78% | 41% | 87% | 26% | 82% | 32% | 72% |
| NB(bigram) | 67% | 52% | 88% | 32% | 76% | 32% | 65% |
| SVM (bigram) | 86% | 27% | 86% | 26% | 86% | 26% | 76% |
| NB(trigram) | 68% | 53% | 88% | 21% | 77% | 33% | 65% |
| **SVM (trigram)** | **88%** | **81%** | **92%** | **66%** | **92%** | **66%** | **87%** |

TABLE VI-C
TESTING RESULTS OF THE PROPOSED SENTIMENT CLASSIFIER USING MANUALLY LABELED DATASET IMPLEMENTING N-GRAM AND REMOVING STOP WORDS FEATURES



| n-Gram Features+stop word | Class Recall | | Class Precision | | F-Measure | | Accuracy |
|---|---|---|---|---|---|---|---|
| | Neg. | Pos. | Neg. | Pos. | Neg. | Pos. | |
| NB(unigram) | 56% | 63% | 89% | 22% | 69% | 32% | 57% |
| SVM(unigram) | 77% | 38% | 87% | 24% | 82% | 30% | 71% |
| NB(bigram) | 62% | 60% | 89% | 23% | 73% | 34% | 62% |
| SVM(bigram) | 86% | 26% | 86% | 27% | 86% | 26% | 76% |
| NB(trigram) | 62% | 61% | 89% | 24% | 73% | 34% | 62% |
| **SVM (trigram)** | **89%** | **18%** | **85%** | **24%** | **87%** | **20%** | **78%** |

TABLE V-D
TESTING RESULTS OF THE PROPOSED SENTIMENT CLASSIFIER USING MANUALLY LABELED DATASET IMPLEMENTING N-GRAM, PORTERSTEMMER AND REMOVING STOP WORDS FEATURES



| n-Gram Features+porter stemmer + stop Word | Class Recall | | Class Precision | | F-Measure | | Accuracy |
|---|---|---|---|---|---|---|---|
| | Neg. | Pos. | Neg. | Pos. | Neg. | Pos. | |
| NB(uni-gram) | 52% | 69% | 90% | 22% | 66% | 33% | 55% |
| SVM (uni-gram) | 77% | 60% | 91% | 33% | 83% | 43% | 74% |
| NB(Bi-gram) | 67% | 52% | 88% | 24% | 76% | 32% | 65% |
| SVM(bi-gram) | 76% | 26% | 86% | 25% | 85% | 25% | 76% |
| NB(tri-gram) | 68% | 53% | 88% | 24% | 77% | 33% | 66% |
| **SVM (trim-gram)** | **88%** | **81%** | **96%** | **57%** | **92%** | **67%** | **87%** |

http://dx.doi.org/10.15242/IIE.E0215004

The best results were obtained using the SVM + tri-gram+ Porter stemmer + stop word and SVM + tri-gram+ Porter stemmer achieving a classification accuracy of 87%. In fact the use of Porter stemmer improves the classification results in SVM from unigram to bigram experiment, but eliminating stop words contribute nothing or no significant contribution to improve the classification performance in any experiments as shown in the above tables. The SVM classifier also got a significant increase in the accuracy in the implementation of n-gram+stemmer features. Pruning non-relevent prefix and postfix in the words contributed in the increase of performance specially in the case of SVM trigram.

Majority of values for accuracy obtained during the testing period were above 50%, which represent a quite accurate classification, even though the values of some precision and recall tend to be lower. Furthermore, the accuracy classification performance of 87% during experimentation is quite promising because this is very close to those obtained by opinion mining researches conducted in recent years. We would like to conclude that this result obtained is inspiring and they encourage us to continue improving the model.

## V. Conclusions

This research paper presents a method of evaluating Filipino sentiments using an automatic polarity classifier, language translation machine and machine learning systems to classify the sentiment polarity of the sentences. Furthermore, automated labeled dataset were experimented using NB and SVM machine learning system incorporating the features such as unigram, bigram and trigram + Porter stemmer and elimination of stop word. The proposed method was tested using manually labeled dataset utilizing the same tools and features used during training period.

The SVM machine learning system outperformed the NB machine learning classifier during training and testing periods. In addition, best results were obtained using the SVM + tri-gram+ Porter stemmer + stop word and SVM + tri-gram+ Porter stemmer achieving a classification accuracy of 87 %.

In future work, we plan to further study other methods to improve the classification performance of the proposed classifier by experimenting and employing other polarity estimation methods. Furthermore, we will explore the possibility of utilizing other machine translators and to find ways to reduce the impact of the translation errors.

## References

[1] http://www2.cs.uregina.ca/~dbd/cs831/notes/ confusion_matrix /confusion_matrix.html

[2] http://www.internetworldstats.com/asia/ph.htm, Philippines Internet Usage Stats and Marketing Report.

[3] Barrett, P. Assessing the Reliability of Rating Data. http://www.pbarrett.net/ presentations/ rater.pdf

[4] Bogartz, R. S., Interrater Agreement and Combining RatingsUniversity of Massachusetts, Amherst. http://people.umass.edu /~bogartz /Interrater%20Agreement.pdf

[5] Chesley, R. K. Srihari, B. Vincent and Li Xu. Using verbs and adjectives automatically classify blog sentiment, American Association for Artificial Intelligence (www.aaai.org), 2005

[6] Landis, J.R., Koch, G.G. (1977). The measurement of observer agreement for categorical data. Biometrics. http://dx.doi.org/10.2307/2529310

[7] Ling, Y. How to Conduct Inter-rater Reliability Tests?. Overseas Chinese Association for Institutional Research An AIR Affiliate That Supports IR Professionals Since 1996.

[8] Habernal, I., Ptacek, T. and Steinberger, J., Sentiment Analysis in Czech Social Media Using Supervised Machine. Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 65–74,Atlanta, Georgia, 14 June 2013. c2013 Association for Computational Linguistic.

[9] Pang, B., Lee, l. and Vaithyanathan, S. Thumbs up? Sentiment Classification using Machine Learning Techniques, Proceeding, EMNLP '02 Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10, Pages 79-86.

[10] Polit, D. F. and Beck, C. T., Nursing Research: Generating and Assessing Evidence for Nursing Practice, 8th edition, Lippincott Williams & Wilkins. 2008

[11] Roebuck, Kevin. Sentiment Analysis: high-impact strategies - what you need to now: definitions, adoptions, impact, benefits, maturity, vendors, Emereo Publishing, Nov 05, 2012

[12] Rothfels, John and Tibshirani, Julie. Unsupervised sentiment classification of English movie reviews using automatic selection of positive and negative sentiment items, http://nlp.stanford.edu/courses/cs224n/2010

[13] Shoukry, Amira, Collaboration Technologies and Systems (CTS), 2012 International Conference technologies and systems ,21-25 May 2012, Page(s):546 - 550

[14] Shu Zhang, Yingju Xia, Yao Meng, and Hao Yu, A Bootstrapping method for finer-grained opinion mining using Graph model, 23rd Pacific Asia Conference on Language, Information and Computation, pages 589–595, 2009

[15] ]Sidorov, G., Miranda-Jiménez, S. and Viveros-Jiménez, F., Empirical Study of Machine Learning Based Approach for Opinion Mining in Tweets. Proceeding MICAI'12 Proceedings of the 11th Mexican international conference on Advances in Artificial Intelligence - Volume Part I, Pages 1-14

[16] Turney, Turney D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 417-424

[17] Zagibalov, Taras and Carroll, John. Automatic seed word selection for unsupervised sentiment classification of Chinese text, Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 1073–1080, Manchester, August 2008.