A Sentiment Analysis and Classification Algorithm Utilizing an Independent Term Matching Scheme Sensitive to Negations and Word Count Patterns

Dr. Asoka Korale, C.Eng. MIET, Chanuka Perera, Dip. ,ABE(UK) Eranda Adikari, BSc., C.Eng. MIESL, Nadeesha Ekanayake, B.Sc.

ABSTRACT

The "Sentiment" or "Emotion" contained in a short comment particularly when expressed by a customer in relation to a product or level of service provides valuable feedback to an organization on customer satisfaction enabling timely actions to resolve current and emerging issues of varying degrees of importance to the business. The ability to measure the "Sentiment" and particularly its "degree" allows the severity of the feedback to be quantified. Independent Term Matching Algorithms (ITMA) while allowing for the calculation of an overall sentiment score for a comment made up of several emotion bearing words or phrases typically do not take into account negations in the comment and therefore miss important meaning contained therein [3]. Typically these algorithms cannot also be trained to recognize patterns within the comments and so are unable to classify comments as to their inherent content.

In light of this, we propose a novel algorithm that allows rating multiple emotion is bearing words in a comment using the ITMA while also accounting for negations. This is effected by modifying the sentiment score of the emotion bearing word occurring subsequent to the negation so that its value is adjusted in a direction opposite in polarity to the affected word. The proposed model is also able to classify comments according to their content so they can be directed to the relevant authority for necessary action. This classification is vital as it allows an assessment of the severity of the issue via its sentiment score while also allowing steps for its resolution in a timely and effective manner by individuals with the relevant domain expertise.

We use a process of Association Rule Discovery to find groups of words that are likely to occur together in a comment to facilitate the classification. This allows the generation of word combinations or word sets that are used in training a Naive Bayes Classifier. The rules are selected based on support, confidence and coverage criteria so that only those that are meaningful to the classification are chosen. This process of keyword generation and association rule prioritization allows us to classify a diverse set of comments relating to a number of sources and a range of issues emanating from a varied group of customers concerning sentiment, its degree, and its content.

1. INTRODUCTION

Any Organization that has a client interface or touch point will invariably have to deal with customer requests and feedback on the offered products and services. These touch points may be in the form of Corporate Websites, FaceBook, and Social Media pages, Call Centers and regular Sales and Support outlets where customers will interface and interact with the organization either through its representatives or via electronic means. In every such instance, it is important for an organization interested in meeting its customer expectations to assess the degree of customer satisfaction by soliciting feedback. In fact, customer satisfaction surveys often form the cornerstone of organizational strategy providing indications as to where targets are being met and where improvements are needed in a wide variety of operational areas of the business usually focusing on its Product Offering and Service Quality.

This customer feedback is usually in the form of short comments and answers to specific questions put to the customer that is usually recorded in electronic form. The feedback that is received via

structured questionnaires having a choice of predetermined responses is rather straightforward to analyze and arrives at its conclusions using statistical methodology. The comments are however unstructured and may relate to any topic and any aspect of the organization the customer has interacted with and is thus more challenging to analyze and quantify.

The comments, however, are a richer source of information on what customers wish to convey to the organization. A means to successfully analyze this aspect of the feedback and distil the essence of what is conveyed provides an organization with valuable insights into its day to day operations and a way to map out its future strategies.

The sentiment or emotional content of the comments and the specific topic or areas to which it relates are the broad aspects that usually concern a business. Evaluating customer sentiment then allows an organization to assess its standing in the eyes of its customers and the public and when observed in the temporal dimension, enables valuable feedback as to how the company and its operations are viewed on a day to day basis. The time dimension is a key attribute as certain types of feedback and customer perceptions vary with time of day and day of the week as some issues are inherently time-bound and occur intermittently or with some regularity.

"Sentiment" can be defined as a view, an opinion held or expressed [4]. It is a representation of the attitudes, emotions, and opinions that is held by an individual and is concerned with impressions and not fact. Thus words bearing emotion convey the sentiment. The analysis of such words in isolation and groups enables us to quantify the emotional content inherent in a particular corpus. Detailed dictionaries have been compiled through large surveys conducted across different demographics, and the emotional content of words have been measured on numerical scales and tabulated. These dictionaries enable algorithms to analyze and rate the sentiment content and its degree of a sequence of words.

Sentiment Analysis is thus critical to organizational success in that it even influences how speculators trade in the stock market taking advantage of the emotional content of news items appearing in print and electronic media concerning different aspects of an organization and its products. It is not unusual for a company to monitor social media and other electronic channels to assess customer satisfaction levels on a new product or the public response to a particular course of action. Governments too are now in the habit of mining media channels for feedback on popular sentiment to its policies and leadership.

Mobile Telecommunications Service providers receive customer feedback from many of the physical and electronic touch points in the form of customer service centres, call centres, and social media pages. Service quality is particularly a key differentiator in this space making customer feedback a critical component of devising an organizational and operational strategy. In fact, Service Quality Issues raised from customer feedback is sometimes the only way in which Network performance issues are initially identified and remedial actions were taken. Feedback on service quality is also key business driver helping to determine where future investments in infrastructure are to be directed. The day to day performance of the network can also be monitored in this context. In light of this, the feedback comments received via the different channels and touch points must be directed to the appropriate authority for action soon as possible via some automated means.

These factors are then the principal drivers for an organization to employ Sentiment Analysis and Classification of customer feedback comments. In light of this, we propose a novel technique that analyzes short comments, rates their emotional content and classifies them into a set of predetermined categories based on a set of pre-classified comments used as a training sample.

2. GENERAL APPROACH AND FEATURES OF THE ALGORITHM

By this modelling, an organization is able to classify and direct comments to the relevant authority while assigning a sentiment score to each comment. The sentiment score of the comment is an indication of the severity of the customer feedback which also allows a certain priority to be assigned to each class of comment. The classifier is trained using word associations (rules) derived from Association Rule Mining of the comments prioritized by confidence level.

2.1 The proposed Sentiment Analysis Algorithm

Independent Term Matching [3] is a technique that is amenable to the analysis of short comments. The comments analyzed and modelled in this particular use case are typical of the length of one or two dozen words and so do not present a large corpus of text for analysis via methods that require training.

We enhance the Independent Term Matching algorithm by incorporating the facility to handle negations in the comment that change the emotional content of the word following the negation.

2.2 Approaches to the Classification Scheme

A classification scheme modeled on the Naïve Bayes technique is employed where a set of pre-classified comments are analyzed for word frequencies. In this approach, it is the occurrence of words and not the order of their occurrence in a particular comment that is important. In other words, the set of words in a comment are considered to be independent of each other given the particular class to which that comment belongs. The algorithm proceeds by determining the conditional probability of a word given a particular class and obtains the predicted class via Bayes theorem and the assumption of the conditional independence of each word given its class.

2.3 Association Rule Mining in the Classification

Association Rules provide an insight into the occurrence of words and word combinations in a comment. It acts as a guide in picking certain keywords that are useful in classifying the initial training comments.

In this modeling, we consider each comment as a transaction and the particular words in the comment as the items in that transaction. The Association Rules are then mined for the aggregation of all comments (transactions) which then is cast in the light of a traditional market basket analysis giving association rules (relationships) between words in a comment.

The rules are quantified via performance measures derived from Confidence and Lift to determine which words are most likely to occur together and provide a guide in determining the class of the comment based on the occurrence of certain keywords in the initial classification-training phase.

3. CURRENT STATE OF THE ART IN SENTIMENT ANALYSIS AND CLASSIFICATION SCHEMES

Existing Sentiment Analysis techniques usually classify the sentiment of a text into three subjective categories “positive”, “negative” and “neutral”. These techniques rely on a supervised learning algorithm that uses pre-classified texts to learn the patterns of words in comments that bear a particular category of sentiment. The classification result is only as good as the training samples offered to the learning algorithm to discover the emotional content of words and their grouping given a particular class of comment.

Given the wide variety of comments and words therein, the number of pre-classified samples would also have to be large in order to use this technique successfully in the context of our particular use case.

Independent Term Matching Algorithms like the one used in this paper has no recourse to a sequence of training comments but utilize ratings for emotion bearing words computed from surveys. In this regard, this technique is more amenable to handle short comments or when the amount of text is too small to train a model. A shortcoming of this technique is its inability to handle the effect of negations. This class of algorithm, however, can produce a continuum for the sentiment rating depending on the precision of which the words in the Dictionary are rated by combining the effect

While the Naïve Bayes “bag of words” technique is widespread, this classification typically does not use keywords generated via Association rule mining of the comments. In fact, the two techniques of Naïve Bayes and Association Rules can be considered complementary in that both look for patterns of words in a comment which are then aggregated across all comments to derive probabilities and rules respectively.

4. MAIN CONTRIBUTIONS OF THE PAPER

1. *A novel Association Rule Mining Algorithm*

A flexible technique designed to generate only the required set of rules at a particular level or depth with rules sorted and prioritized by confidence

level. The important rules provide insight into keywords that should be used in the classification of comments used in model training.

2. *Sentiment Scoring incorporating the effect of Negations*
The Independent Term Matching Algorithm is enhanced with the ability to rate the sentiment of words modified by preceding Negations giving a more complete, holistic and accurate rating of the emotional content of a comment. The sentiment of such affected words is adjusted in the direction opposite to its current polarity by an amount that reflects the uncertainty in its rated sentiment.

3. *Classification via Naïve Bayes technique utilizing Key Words derived from Association Rules*
Classification of comments via Naïve Bayes while an established method benefits from the added insights gained from the keywords used in the classification of the training data comment set. The use of word associations derived from the rules displaying high confidence allows a higher accuracy and greater consistency in the classification of the training data resulting in better overall performance in the prediction.

5. COMPONENTS OF THE PROPOSED ALGORITHM

Association Rule Mining, Comment Classification, and Sentiment Rating are the three principal aspects of the proposed model. Although Sentiment Analysis and Classification may be considered in isolation, our use case demands that we rate the sentiment of the comments in a particular predicted class for prioritization of handling.

5.1 Association Rule Mining

The Association Rules are mined from an Incidence Matrix (IM) by placing apriori limits on the support for each Item which is the column sum of IM. IM is formed as a comment vs. word (Item) matrix, by allocating unity for every word that is found in each comment. If a word appears in a comment more than once, the entry in the corresponding word column in the row allocated for that comment is still unity.

Once the column sum (IMsum_min) is formed the particular columns that do not meet the minimum support criteria (min_supp) are not considered and eliminated from the analysis as a rule that incorporates that item cannot, in any case, meet the minimum support requirement.

The rule discovery then proceeds to the first stage by determining all pairs of items that meet the minimum support criteria. The support for a pair of items is then simply the dot product of the corresponding columns of IM. In the next step, the pairs of items that meet the minimum support criteria are selected, and the support and confidence of the rule are also determined.

In the second stage, only the rules that met the minimum support criteria from the first stage are selected to form the rules. The antecedent in the second stage is then formed by merging antecedent and consequent terms from the rules (that met the minimum support criteria) of the first stage. The three-item rules (2 in items in the antecedent derived from stage 1, and one new item in the consequent) that can be formed that meet the minimum support criteria can only come from the set of terms selected from stage 1 and the set IMsum_min.

For example, let A, B, C, D be items that meet the minim support criteria in IM.
In the first stage all pairs {A,B}, {A,C}, {A,D}, {B,C}, {B,D}, {C,D} are tested for minimum support. (There will be 4C2 = 6 such terms to be tested.)

A->B is a valid rule, only if the support for AUB meets min_supp criteria

In the second stage A and B (and all other pairs that met min_supp) (are merged to form the antecedent of the rule at stage 2 if they meet the min_supp criteria. (let us assume that {A,B}, {A,C} and {C,D} meet the min_supp criteria)

Thus in stage 2
A&B->C is a valid rule only if the support for AUBUC meets min_supp criteria (and similarly we test support for rules A&B->D, A&C->B, A&C->D and C&D->A and C&D->B)

The rule discovery proceeds in this manner until all unique combinations of items have been created in the rules.

5.1.1 Definitions and Performance Measures

Support (A->B) = N(AUB)/N (1)

where *N*(AUB) represents the number of times items A and B are brought together in the whole transaction set and N is the number of rows in IM (corresponds to the number of unique subscribers or transactions)

Confidence(A->B)

$$= \text{Support}(AUB)/\text{Support}(A) \quad (2)$$

$$= P(E_A \& E_B)/P(E_A) \quad (3)$$

$$= P(E_B/E_A) \quad (4)$$

Where $P(E_X)$ is the probability of word (item) X being present in a particular comment.

The confidence of a rule (A->B) is a measure of the validity of the rule; it has a conditional probability interpretation in that, it is the probability that B occurs when A has already occurred. In market basket parlance it is the probability of word (item) B being in a comment given that A is in the same comment.

Loosely speaking, we use the confidence measure in our reasoning as an indication of the degree of correlation between words when we have rules of the form word1 -> word2. It is also used to rank the association rules particularly when one common antecedent gives rise to many consequent items.

The Lift is a popular measure used to measure how likely a rule is if the items are independent.

Lift(A->B)

$$= \text{Confidence}(A->B)/\text{Support}(B) \quad (5)$$

$$\text{Lift}(A->B) = P(E_A \& E_B)/P(E_A)P(E_B) \quad (6)$$

5.2 Sentiment Rating via Independent Term Matching while accounting for Negations

The overall sentiment of a comment is determined by aggregating the effect of the individual words that bear emotional content or sentiment. The words that bear sentiment are identified by the existence of a match in a dictionary of 14,000 words compiled by Warriner et al. This dictionary contains a mean and standard deviation value for each emotion bearing word in the dimensions of Valence (Happiness), Activation (Arousal) and Dominance. These values are the result of a survey of raters where it is assumed that the distribution of the ratings for each word follows a Normal Distribution. The means and standard deviations for each emotional dimension of each word provided in the Dictionary are then the means and standard deviations of the ratings given by the population of raters who rated each word.

Thus by aggregating the effect of the all such matched words in a comment by summing the corresponding Normal Probability Density functions corresponding to each matched word we are in effect treating the overall density of a particular sentiment dimension of a comment as a sum of Gaussians each with a particular mean and standard deviation. In effect, this is a Gaussian Mixture Model (GMM) where the overall sentiment rating of a particular dimension is then determined as the value that gives the highest probability of the cumulative sum of the Normal individual densities.

In this modeling it is assumed that each matched word is equally likely to occur in particular comment and that a particular emotional dimension of each word is independent of its other emotional dimensions. Thus we can consider the cumulative effect of all matched sentiment-bearing words via the sum of the individual probability densities.

$$f(x;\theta) = \sum_{k=1}^{N} p_k g(x; m_k \sigma_k) \ , \quad (7)$$

where x represents the sentiment score, N the number of matched words in a comment m_k, σ_k the mean and standard deviation of the Normal Distribution of the ratings of each matched word,

$$p_k = \frac{1}{N} \quad (8)$$

as the words are of equal probability of occurrence and are also independent of each other, and

$$g(x; m_k, \sigma_k) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{\|x-m_k\|}{\sigma_k}\right)^2} \quad (9)$$

The overall sentiment of a comment then is determined as

$$x_{comment} = \max_x f(x;\theta) \quad (10)$$

Which is the point at which the probability of the mixture of distribution is a maximum, and so is the most likely value for the overall sentiment of a comment composed of several words.

Thus a comment of the form "the service was good but was late" finds a match in the Dictionary for words "service", "good" and "late". The cumulative effect of the three words in the comment and therefore the overall sentiment of the comment in the Valence and Activation dimensions are displayed in Figure 1 and Figure 2. The Valence and Activation ratings for each word matched in the Dictionary and their cumulative effect via GMM is found in Table 1.

Table 1: Dictionary Sentiment Ratings for each matched word in the comment.

Comment Words	Valence Rating		Activation Rating	
Dictionary Value	Mean	Std Dev	Mean	Std Dev
'service'	6.83	1.54	2.95	2.09
'good'	7.89	1.24	3.66	2.72
'late'	3.32	1.17	5.57	2.56
Simple Average	**6.01**	**1.32**	**4.06**	**2.46**

Table 2: Cumulative Rating for entire comment via Gaussian Mixtures.

Word	Valence Rating	Activation Rating
max- GMM	**7.5**	**3.7**

Thus the overall sentiment ratings found via a maximum probability criteria is somewhat different to a simple average of the values. But in most instances the simple average and the max-GMM results are very close as short comments usually only contain words of a single polarity and do not in general feature words bearing emotions contrary to one another (like "late" vs. "good" and "service"). One can readily observe that the distributions for "service" and "good" are close to one another as they in general express similar sentiment. The presence of "late" in the comment alters the Gaussian mixture such that the Simple Average and the max-GMM value are at variance.

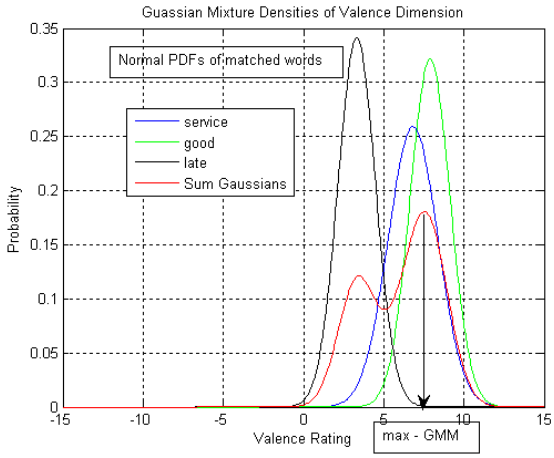

Figure 1: Gaussian Mixtures of matched words in the Valence Dimension

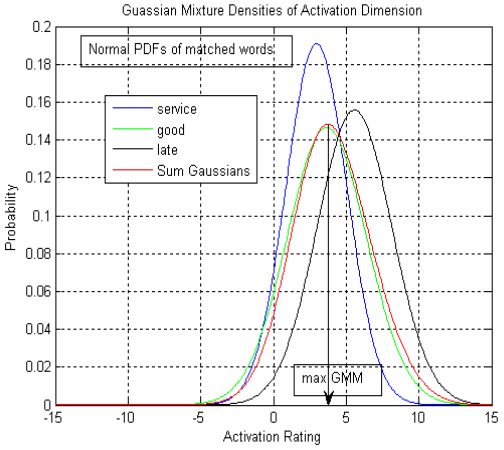

Figure 2: Gaussian Mixtures of matched words in the Activation Dimension

5.2.1 Accounting for Negations

In this paper, we also account for Negations that modify the meaning of word immediately following the negation by adjusting the sentiment score of the word immediately following the negation in a direction opposite in polarity to its matched directory sentiment value. The magnitude of the adjustment made corresponds to the standard deviation of the particular rating value being adjusted. Thus analysis of the comment "the service was not good and late" results in the following.

Table 3: Dictionary Sentiment Ratings for each matched word in the comment.

Comment Words	Valence Rating		Activation Rating	
Dictionary Value	Mean	Std Dev	Mean	Std Dev
'service'	6.83	1.54	2.95	2.09

not 'good'	6.65	1.24	6.38	2.72
'late'	3.32	1.17	5.57	2.56
Simple Average	**5.6**	**1.32**	**4.97**	**2.46**

Table 4: Cumulative Rating for entire comment via Gaussian Mixtures.

Word	Valence Rating	Activation Rating
max- GMM	**6.7**	**4.5**

Our reasoning here is that while "not good" is not the same as saying that it is "bad," it, however, has the effect of diminishing the positive sentiment expressed in "good." The same thinking applies to the activation dimension. This adjustment is user definable and can be set in consideration with the particular types of comment and analysis task at hand.

5.3 Simplifying Assumption of Naive Bayes Technique

A Naïve Bayes technique forms the basis for the classification scheme. The technique assumes the conditional independence of the words $\{X_i\}$ given the class $\{C_j\}$. Thus the conditional probability of a particular set of words $\{X_i\}$ occurring in a particular comment class C_j has a simpler form amenable to fast implementation via a computer algorithm. The Naïve Bayes relationship can be derived as follows

$$P(X_1, X_2, ...X_N / C_j) = P(X_1, X_2, ..., X_N, C_j) / P(C_j) \quad (11)$$

$$= P(X_1 / X_2, ..., X_N, C_j) P(X_2, X_3, .., X_N, C_J) / P(C_j) \quad (12)$$

$$= P(X_1 / X_2, .., X_N, C_j) .....P(X_n / C_j) P(C_j) / P(C_j) \quad (13)$$

Under the assumption of conditional independence of X_i given C_j

$P(X_i / X_2, ...X_N, C_j) = P(X_i / C_j)$ for all,

Moreover, the relationship simplifies to

$$P(X_1, X_2, ..., X_N / C_j) = P(X_1 / C_j) P(X_2 / C_j) ...(X_N / C_j)$$

5.3.1 Classification via Naive Bayes

By this technique, we assume that the order of words in a comment is independent of each other given the class. Thus a class is determined sorely on the specific words in a comment and their frequency of occurrence in that comment. Thus the algorithm treats the words in a comment as an unordered set or a "bag of words".

In the training phase, we classify a set of comments based on the aggregate effect of the words in that particular remark. From the training data, we estimate the vocabulary $V = \{X_1, X_2, \ldots, X_T\}$ the set of all unique words across all of the comments.
From this classification, we are able to determine the conditional probabilities of a particular word given a particular class $P(X_i / C_j)$ as

$P(X_i / C_J) = N(X_i \cap C_j) / N(C_j)$ the proportion of the number of occurrences of word X_i in class C_j across the number of words in class C_j

$P(C_j) = N(C_j) / N(C)$ the proportion of the specific class Cj across all comment classes C that occurs in the overall data set

The posterior probability $P(C_j / X_i, X_2, ..., X_N)$ which is the probability of a class given a particular set of words Xi, i = 1…N is obtained via the traditional Bayes identity

$$P(C_j / X) = P(C_j, X) / P(X) \quad (15)$$

$$= P(X / C_j) P(C_j) / P(X) \quad (16)$$

ignoring P(X) which is a common denominator, results in a measure of the maximum likelihood of a particular word set $X = \{X_1, X_2, \ldots, X_N\}$ belonging to a particular class C_j

$$P(C / X) = \max_{C_j} \{P(X / C_j) P(C_j)\} \quad (17)$$

$$= \max_{C_j} \{P(X_1 / C_j) P(X_2 / C_j) ...P(X_N / C_j) .P(C_j)\} \quad (18)$$

6. RESULTS AND PERFORMANCE

A sample of the comments analyzed to demonstrate the algorithm in this paper is found in Table 5.

Table5. Sample Comments

1.HOTLINE ISSUES - DELAY IN ANSWERING - CX SERVICE ASSISTANCE Today morning CX has called to the 444 H\L for Movie Ticket & he has waited for more than 10 mins in the line, regarding this now CX was very disappointed on our service. So pls be kind enough to chk on ths & give the call back to the CX ASAP. * Note: - Regarding this issue CX need the call back from one of our manager & CX has requested not to charge a single rupee from his no for this issue.
2.Yes,man magea prshnaya kiyapu gaman eyaa magea prshnea

wisaduwaa he's a good

3.Yes kad pin nambar signal

4.Wenath ayathana wala mema pahasukam nomati nisa

5.very good service

6.uparimaya

7.Uparima

8.think so

9.thanks

10.Super

11.Solved

12.She resolved my problem.

13.Service nallam

14.Sambanda weemata boho welawak giya nisa

15.recharge

16.Prashnayata pilithura hodin pahadili kara dima

17. Payak athulatha gataluwa nirakaranaya karanwa kiuwa. Thawamath gataluwa nirakaranaya kara natha.

18.oba ayathanaya sewawan sadaha ihala mudalak ayakarana nisa

19.no mms setting laba dunnada save kala nohaka

20.nam apahu e tika ewanna

21.Mata awashshaya u pilithurau pahadili lesa laba ganemata hakiuna.

22.mage parshnata pilithuru dunna.

23.lotari SMS stop

24.Its professional

25.ing tone sewawa ain kirima

26.I submitted Dtv reg form on 27th oct at yr crescat arcade. They told to call me on 28th wed to give the AC No

27.Hot line eka answer karapu girlge voice eka and care eka good

28.Hi kohomada? Mama mea dawas wala plan karagena yanawa mage next music video eka karanna. Song eka "Mata Rawana" :-)

29.harima pehediliwa mage getaluwa nirakaranaya kala thanks dailog

30.Good service but shortcomings due to some arrogant customer care officers

31.good men

32Good

33.getaluwa hadunagenimata noheki wiya..

34.First of all its great to be treated as a privilege customer. The reason is simple. I'm using dialog mobile connection and DTV, because dialog has the better

35.durakathanayata pilithuru denda epai eke hoda naraka kiyanna.

36.Cx need to add the CHU CHU TV which is a kids channel to the channel list.Since this channel is available on another TV connection.Cx need this channel to activate for DTV aswell.Please check on this and do the needfull. Thank you

37.Customer service personal have to be trained better cause they can't think out of the box.

38.bashawa wenaskaranna

6.1 Sentiment Rating via Enhanced Algorithm incorporating Independent Term Matching

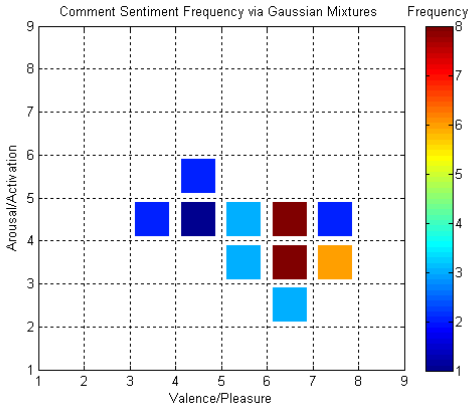

Figure 3: Overall Sentiment of the comments and their frequencies

Figure 3, depicts the distribution of the overall sentiment of each comment across all comments. It is clear the at the sentiment is positive as they occur more to the RHS of the plot while the emotion conveyed is perhaps not high in the activation dimension. A glance at the comments in Table 5, would convey such an idea when taken in a holistic sense.

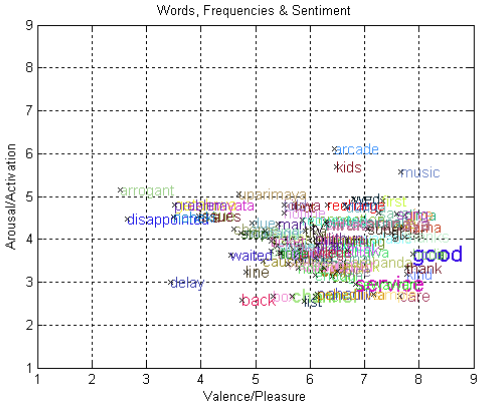

Figure 4: Sentiment of each Directory matched word and their frequency

Figure 4, Depicts the emotional content of each word appearing across the comments that have a match in the Dictionary. Words are occurring more frequently appear larger.

While in these results the sentiment was rated if there was at least one match of a comment word with the Dictionary, we should rate the sentiment if the number of matches is higher than a certain user-defined a threshold. This will further improve the accuracy of the rating process giving a better overall picture of the sentiment expressed by the customers.

6.2 Association Rule Mining words in comments

The Association Rule algorithm devised enabled rules with apriori confidence and support levels to be mined. These rules provided insight into relationships between words that are likely to occur in a comment. We cast the problem in the fashion of a market basket analysis with words as items and comments taking the place of transactions.

The rules with high support and confidence were used to obtain "keywords" that formed the antecedent and consequent terms of the selected rules. These keywords provide insight into the likelihood of a particular group of words occurring in a comment. The underlying principle of the Naïve Bayes technique is also similar where the words in a comment are considered as an unordered group, but where the word frequencies are accounted for.

Thus the rules can be used as a means to train the Naïve Bayes model with classification classes indicated by the occurrence of certain groups of keywords

6.3 Classification via Naive Bayes

Cross-validation across the training sample generated a match more than 90%.

We believe that certain comments should be excluded from the classification as they do not convey any meaningful category information that can be used to classify them. The performance will further increase with this initial preprocessing and filtering.

On predicted data, the performance was more than 75%.

6.4 Adding local language support to the Dictionary

As this use case necessitated the analysis of Sinhala and Tamil language words in addition to the English, the Dictionary was updated with such words by replicating the sentiment rating values of English words that were similar in meaning / emotional content to the non-English words and entering them against the non-English words.

6.5 Variance in max-GMM and Simple Average

Figure 5, depicts the variance between max-GMM and Simple Average methods for estimating the overall sentiment of a comment. It is seen that 90% of the time the measures are within +/- 0.5 in the case of the Valence Attribute. The CDF of the difference in the Activation attribute is tightly centred on the origin indicating hardly any variance. As discussed earlier this is also an indication that most comments convey sentiments of a single polarity and only a few comments (less than 10%) have words with conflicting emotional content.

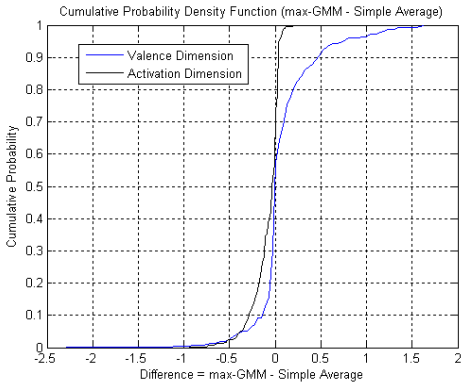

Figure 5: Variance between GMM and Simple Average measures for estimating overall comment sentiment

7. CONCLUSION AND FUTURE WORK

In this work, we demonstrated that short comments could be analyzed for their sentiment via an Independent Term Matching Scheme [3]. This technique together with its enhancements provides a way for an organization to classify the comments and also rank them by the degree of severity which other techniques using machine learning cannot achieve due to the need for a large corpus of text to train such algorithms.

Through our enhancements to the Independent Term Matching Algorithm [3], we were successfully able to rate the sentiment of comments containing negations by appropriately translating the sentiment ratings of words affected by such negations. This led to further improving the measurement accuracy of the overall emotional content of a comment.

It was also found that the sentiment ratings calculated for a typical comment in our particular application were similar whether a simple average of the sentiment ratings in each dimension was used or whether the maximum probability point of the Mixture of Gaussians was used. This is because these short comments typically have only words of a particular polarity and do not in general express more than one sentiment that's contrary to one another.

The use of Association Rule generated keywords enabled more accurate, and consistent classification of comments used to train the Naïve Bayes model resulting in improved classification / prediction performance.

Local language support in the form of an updated Dictionary enabled the algorithm to be more versatile concerning analyzing the sentiment and classifying comments containing non-English words or a mix or English and other non-English words.

We plan to improve classification performance by iteratively classifying the miss-classified comments in the prediction and adding them to the training samples. In this way we endeavour to assist the algorithm to learn from its mistakes and improve its performance with each iterative step.

8. REFERENCES

[1]. Turney, Peter (2002). "Thumbs Up or Thumbs

[2]. Down? Semantic Orientation Applied to Unsupervised Classification of Reviews".Proceedings of the Association for Computational Linguistics. pp. 417–424. arXiv:cs.LG/0212032.

[3]. Pang, Bo; Lee, Lillian; Vaithyanathan, Shivakumar (2002)."Thumbs up? Sentiment Classification using Machine Learning Techniques". Proceedings of the Conference on Empirical Methods

[4]. Ramasamy, Siddarth; "Visualization of the Sentiment of Tweets", M.Sc. Thesis, Department of Computer Science, North Carolina University, 2011

[5] web access: Sentiment_analysis https://en.wikipedia.org/wiki/Sentiment_analysis, extracted on 29 May 2016.