AUTOMATIC EMPLOYEE RATING SYSTEM USING SENTIMENTAL ANALYSIS

Rauoof Rajeev Hameed*1 , Reshma Elza Ulahannan*2 , Renitta Mariya Solomon*3 , Treesa Joy*4

Jerin Thomas'5

*UG Students, 'Assistant Professor'Department of Computer Science and

Engineering,Amal Jyothi College Of Engineering, Kottayam, India.

ABSTRACT : Many sentimental analysis methods for the classification of reviews use training and test data based on star rating provided by the reviewers. However when reading reviews it appears that the reviewers rating do not always give an accurate measure of the sentiment of the review. We performed an annotation study which showed that reader perceptions can also be expressed in ratings in reliable way and that they are closer to the text than the reviewer ratings.

Keywords-

1.INTRODUCTION

Sentiment-analysis tools attempt to discover user opinions in these reviews by changing the text to numerical ratings. Building these tools requires a large set of annotated data to train the classifiers. Most developers Building these tools requires a large set of annotated data to train the classifiers. Most developers compile a training and test corpus by collecting reviews from web sites on which customers post their reviews and give a star rating. They check Associate in training their tools against these reviewer ratings assumptive that they're correct live of the sentiment of the review.

2. RELATED WORK

They include reviewer's ratings ranging from strong negative to strong positive. Each review contains the following information:

Reviewer rating: a user rating given by there viewer translated to a scale ranging from 0 to 10 (very negative to very positive) describing the overall opinion of the user.

Review text : a brief text describing there viewer's opinion of the employee.

Reader ratings : ratings of two users on a scale ranging from 0 to 10. These ratings are described in more detail in the next section.

Reader ratings and agreement scores

Two annotators (R1 and R2), both native speakers of Dutch and with no linguistic background added a reader rating to each review. They were asked to read the review and rate the text on a scale from 1-10 (very negative to very positive), answering the question whether the reviewer would advise them to choose the hotel, or not. As correlation may be high while not essentially high agreement on absolute values, we tend to perform evaluations on categorical values. A 2-class evaluation was performed by translating 1 to 5 ratings to 'negative' and 6 to 10 ratings to 'positive'; a 4-class evaluation is performed by translating 1-3 ratings to 'strong negative', 4 to 5 ratings to 'weak negative', 6 to 7 ratings to 'weak positive' and 8 to 10 to 'strong positive'. Agreement was measured between each annotator pair in terms of percentage of agreement and kappa agreement (κ).

raters	1/10	2-class		4-class	
REV-R1	0.82 *r*	0.81 κ	0.90%	0.51 κ	0.63%
REV-R2	0.83 *r*	0.82 κ	0.91%	0.53 κ	0.65%
R1-R2	0.92 *r*	0.92 κ	0.96%	0.71 κ	0.78%

Table 1. Inter-annotator agreement.

Table (1) shows that inter-annotator agreement is quite high between all raters both when correlation is measured on the 10-point-scale(r>=0.82) and when agreement is measured with the 2-class annotation sets (κ >= 0.81). Agreement on the 4 class annotations is

much lower($\kappa >= 0.51$) showing that polarity strength is difficult to annotate. However, given the purpose of this study, we are not interested in agreement as such. Our focus is on the differences in agreement between readers and reviewers. From that perspective it is interesting to note that, according to all measures, the re-viewer is an outlier. Agreement between each individual reader and the reviewer (REV-R1and REV- R2, respectively) is consistently lower than agreement between both readers (R1-R2). The differences already become important when measuring agreement on 2-class annotations, but even more prominent when measuring agreement on 4-class annotations. All observed differences ranging from 5 up to15%, are statistically significant ($p < 0.01$).

4. Implications for sentiment analysis

We investigated how automated sentiment analysis methods perform with the different sets of annotations by applying two widely used approaches to document- level sentiment classification. Classifier accuracy is measured against the three sets of ratings (R1, R2 and REV) we described in the previous section

4.1 The lexicon-based approach

The first method is a lexicon-based approach which starts from a text which is lemmatized with the Dutch Alpino-parser1.The approach is similar to the "vote-flip-algorithm" proposed by Choi and Cardie (2008). The intuition about this algorithm is simple: for each review the number of matched positive and negative words from the sentiment lexicon are counted. If polar words square measure preceded by anegator their polarity is flipped, if polar words square measure preceded by Associate in training modifier, their polarity is doubled.

We then assign the bulk polarity to the review. In the case of a tie (being zero or above zero), we have a tendency to assign neutral polarity . The sentiment lexicon used in this approach is an automatically derived general language sentiment lexicon obtained by Word Net propagation.

4.2 The machine-learning approach

The second method is a machine learning approach that also starts from a text that is lemmatized by the Dutch Alpino parser. After lemmatization the text is transformed to a word-vector representation by applying Weka's StringToWord Vector with frequency representation (instead of binary). We used Weka's Naïve Bayes Multinominal (NBM) classifier to classify the reviews. The NBM was chosen because our review texts are rather short (with an average of 68 words) and, according to Wang and Manning (2012), NBM classifiers perform well on short snippets of text. Results reported are average of ten-fold cross- validation-accuracies using R1, R2 and REV ratings as training and test data.

4.3 Results on different types of ratings

Results are evaluated against the whole set of 1,172 reviews (cf. table 2 'all'). As many approaches to sentiment analysis do not use the class of weak sentiment (Liu, 2012), we also evaluated against a subset of strong negative (ratings 1 to 3) and strong positive (ratings 8 to10) reviews (cf. table 2, 'strong'). Table (2) shows the classification results in terms of accuracy, obtained by the lexicon-based approach (LBA, row 1, 2, 3) and the machine-learning approach (NBM, row 4, 5, 6).

	name	ratings	all	strong
1	LBA	REV	78.3	85.0
2	LBA	R1	80.5	88.1
3	LBA	R2	80.7	88.1
4	NBM	REV	83.6	86.4
5	NBM	R1	86.9	92.2
6	NBM	R2	86.7	92.2

Table 2. Results of sentiment analysis.

Star Level	General Meaning
★	I hate it.
★★	I don't like it.
★★★	It's okay.
★★★★	I like it.
★★★★★	I love it.

Figure 1
Rating System for Amazon.com.

The rating is based on a star-scaled system, where the highest rating has 5 stars and the lowest rating has only 1 star.

RESEARCH DESIGN AND METHDOLOGY

Negation phrases identification

Algorithm 1 Negation phrases identification
Input: Tagged Sentences, Negative Prefixes
Output: NOA Phrases, NOV Phrases

```
for every Tagged Sentences do
  for i/i + 1 as every word/tag pair do
    if i + 1 is a Negative Prefix then
      if there is an adjective tag or a verb tag in next pair then
        NOA Phrases ← (i, i + 2)
        NOV Phrases ← (i, i + 2)
      else
        if there is an adjective tag or a verb tag in the pair after next then
          NOA Phrases ← (i, i + 2, i + 4)
          NOV Phrases ← (i, i + 2, i + 4)
        end if
      end if
    end if
  end for
end for
return NOA Phrases, NOV Phrases
```

Words like adjectives and verbs square measure able to convey opposite sentiment with the assistance of negative prefixes. For instance, take

into account the subsequent sentence that was found in associate in training electronic devicers review:"The in-built speaker conjointly has its uses however the phase "nothing revolutionary" provides a lot of or less negative feelings. Therefore, it is crucial to identify such phrases. In this work there are work, there square measure 2 styles of nothing of phrases are known, particularly negation of adjective (NOA) and negation of verb (NOV). Most common negative prefixes like not, no or nothing square measure treated as adverbs by POS tagger. Hence, we have a tendancy to propose formula one for phrases identification. The formula was able to determine twenty one ,586 totally different phrases with total prevelance of over .68 million.

.

Top 10 sentiment phrases based on occurrence

Phrase	Type	Occurrence
not worth	NOA	26329
not go wrong	NOA	15446
not bad	NOA	15122
not be happier	NOA	14892
not good	NOA	12919
don't like	NOV	42525
didn't work	NOV	38287
didn't like	NOV	21806
don't work	NOV	10671
don't recommend	NOV	9670

A se s sentiment. Given those sentiment words projected in a word token consists of a positive (negative) word and its part-of-speech tag. In total, we tend to elect eleven, 478 word tokens with every of them that happens a minimum of thirty times throughout the dataset. For phrase tokens,3,023 phase were elect of the twenty one ,586 known sentiment phrases that every of the three phrases additionally has an incident phrase that's no but thirty. Given a token *t*, the formula for *t*'s sentiment score (SS) computation is given as:

$$SS(t) = \frac{\sum_{i=1}^{5} i \times \gamma_{5,i} \times Occurrence_i(t)}{\sum_{i=1}^{5} \gamma_{5,i} \times Occurrence_i(t)}$$

Occurence i (t) is t's number of occurrence in i- star reviews, where i=1,...,5. According to Figure 3. our dataset is not balanced indicating that different number of reviews were collected for each star level. Since 5-star reviews take a majority quantity through the whole dataset. we tend to herewith introduce a magnitude relation γ 5,i , that is outlined as:

$$\gamma_{5,i} = \frac{|5 - star|}{|i\text{-}star|}$$

In equation 3, the numerator is the number of 5-star reviews and the denominator is the number of *i*-star reviews, where *i*=1,...,5. Therefore, if the dataset were balanced, $\gamma_{5,i}$ would be set to 1 for every *i*. Consequently, every sentiment score should fall into the interval of [1,5]. For positive word tokens, we expect that the median of their sentiment scores should exceed 3, which is the point of being neutral according to Figure 1. For negative word tokens, it is to expect that the median should be less than 3.

CONCLUSION

We performed an annotation study that showed that the observed mismatch between reviewer ratings and review's sentiment is a rather frequent phenomenon. Considerable part of the reviews (ranging from 9 to 37. depending on the granularity of the classification) is classified by the reviewer in the wrong sentiment class. The annotation study also showed that reader ratings are a more accurate measure. We already expected reader ratings to be closer to the text because they are exclusively based on it. In addition, the annotation study shows that readers agree in their ratings and that the review's sentiment can be reliably annotated by readers. Our experiments in section 4 show that sentiment-analysis tools perform better with reader ratings than with reviewer ratings. This should probably not surprise us as sentiment analysis behaves like a reader whose only source of information is the review text. As such, this is a promising result. However, since reviewer ratings are widely available and come for free with the text, they will often be used to evaluate the tools. Likewise, training and fine-tuning will be done with reviewer ratings rather than with reader ratings.

REFERENCE

1.Ando, M. and S. Ishizaki (2012) Analysis of travel review data from Reader's point of View. In Proceedings oF WASSA-2012. Jeju, South Korea.

2.Carrillo de Albornoz, J., L. Plaza, P. Gervás and A. Diaz (2011). A joint model for feature mining and sentiment analysis f or product review rating. In Proceedings of ECIR-2011. Dublin, Ireland.

3.Choi, Y. and C. Cardie (2008). Learning with Compositional Semantics as Structural Inference for Subsentential Sentiment Analysis. In Proceedings of EMNLP '08. Hawaii, USA.

4.Ghose, A., G. Ipeirotis and B. Li (2012). Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content. Marketing Science, Vol. 31.

5. Liu, B. (2012). Sentiment Analysis and Opinion Mining.
Morgan & Claypool Publishers, USA. Mahony, M., P. Cunningham and B. Smyth (2010). An assessment of machine learning techniques for review recommendation.

6.https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0015-2

7.https://nevonprojects.com/sentiment-based-movie-rating-system/