Classification of Short Text Comments by Sentiment and Actionability for VoiceYourView

William Simm*, Maria-Angela Ferrario, Scott Piao, Jon Whittle and Paul Rayson
Computing Department, Lancaster University, Lancaster, UK
*w.w.simm@lancaster.ac.uk

I. INTRODUCTION

In this paper, we report on the application of four methods of sentiment analysis and apply two of those to actionability analysis on short sentence based data. The aim is to compare the results of the analysis to manually rated scores to determine the degree of success of each method at estimating the sentiment of the dataset. This paper contributes an independent comparative analysis of sentiment and actionability estimation accuracy, applying the chosen methods without additional data pre-processing. Although figures exist in the literature for individual techniques, these figures are typically produced by the inventors of the techniques and have not been independently verified.

The concept of ‘actionability’ is a rather novel one in the context of automated text analysis and opinion mining. In our case we consider as ‘actionable’ statements that provide a clear suggestion on how a product or a service (or a feature of both) could be improved.

The test data is taken from the first trial of the VoiceYourView project, a project which aims to develop a kind of intelligent kiosk which uses speech recognition and natural language processing to gather the public’s creative input on the public space designs [15]. A public library was the first trial of the VoiceYourView system, and the data comprises over 2000 individual short single-sentence comments from library users on the refurbishment of the library. Both written and spoken comments were captured form the public as part of the trial. Written comments were captured on postcards and spoken comments were captured using the handset system pictured in Figure 1.

VoiceYourView is novel in that it allows unstructured voice and text input so that users can comment on anything. VoiceYourView shares some similarities with social media commenting systems—such as Twitter and Facebook, where users are free to comment on anything and everything. However in the case of the first VoiceYourView trial, we were interested in comments specifically about the library in which it was installed. The VoiceYourView comments are processed automatically in real-time using speech recognition and natural language processing (NLP) and summaries are automatically presented on public display screens to set up a conversation between users of the space. These summaries contain automatically extracted theme and sentiment summaries. The system used in the Library trial for determining sentiment, was a tagger-based approach described here. The theme of comments was determined and categorized by extracting the semantic category of the first noun or adjective in the comment [15]. Reports produced from the data intended for stakeholder feedback contain manual measures of sentiment, theme and actionability of the comments, although it is the intention to perform this analysis automatically.

Here data from VoiceYourView is used as it consists of nominally well structured sentences since it originates from spoken and hand written sources. There is scope for

Abstract— Much has been documented in the literature on sentiment analysis and document summarisation. Much of this applies to long structured text in the form of documents and blog posts. With a shift in social media towards short commentary (see Facebook status updates and twitter tweets), the difference in comment structure may affect the accuracy of sentiment analysis techniques. From our VoiceYourView trial, we collected over 2000 individual short comments on the topic of library refurbishment, many of which are transcribed spoken comments. We have shown success in determining the theme of comments by looking for the first noun and using a semantic tag set to categorise this noun and hence the comment for short comments. Sentiment is a measure of how positive or negative a comment is, and the actionability metric is a measure of how actionable the comment is, i.e. how useful it is. This paper looks towards applying methods from the literature to our dataset with the aim of evaluating methods of automatic sentiment and actionability analysis for our VoiceYourView application data and has relevance to data from other applications, e.g. those from the social media. With many social media commentary applications moving to add speech platforms, VoiceYourView data may be representative of the type of free-form spoken text input to be expected in such platforms.

Keywords: sentiment analysis; actionability analysis; text classification
VoiceYourView to be applied to the data analysis of social media, however we have no control over this data and the structure of the text, characters and abbreviations used make it slightly more complex to use. Methods to translate the typical abbreviations and language used in social media do exist, so could be used to make use of this data. Social media sites such as Twitter and Facebook now have speech inputs built by external developers, allowing people to comment using free form speech. The spoken library trial data may be representative of the type of comment received when the public are asked to comment in free-form conversational speech – the type of comment that will be perhaps left by spoken social media comments. One point of note is that the library user demographic is perhaps the inverse of the twitter demographic – being either young children or those of retirement age. This will perhaps be reflected in the type comments received and language used.

In this paper, four automatic methods of analysis are discussed for sentiment analysis including a tagger based approach, a naïve Bayesian classifier, the ReadMe tool [6] and a lexicon and rule method [12]. The naïve Bayesian classifier and ReadMe tool were further applied to test actionability. As will be shown, without any additional preprocessing of the data it was found that the lexicon and rule method gave the most accurate estimation of sentiment and the naïve Bayesian classifier gave the most accurate estimation of actionability category.

The remainder of this paper is organized as follows: Section II describes and summarizes the literature of sentiment and actionability analysis. Section III describes the methodology of the tests, including the operation of the three analysis types chosen. Section IV summarizes the results of these tests and Section VI contains the concluding discussion.

II. LITERATURE SUMMARY

A. Sentiment Analysis

Sentiment analysis has recently come to the fore as markets and governments alike have realized the potentials of web-based opinion mining [8]. There are many definitions of sentiment analysis, in our context we define it as “the task of identifying positive and negative opinions, emotions, and evaluations” [18].

Nasukawa and Yi [9] also recognize the importance of online opinion mining for marketing analysis purposes and ‘rumors detection’. There are some similarities between their system and the VoiceYourView system, for example, their model combines semantic analysis with a syntactic parser and the analysis is performed at statement (phrase) level to capture opposite sentiments in the same expressions. For example, the expression "Product A is good, but expensive" is split in two statements of a different sentiment orientation each, positive and negative respectively. Their system also includes ‘a sentiment lexicon’, that is a set of manually defined sentiment expressions. However, the objective of their research is different from ours: whereas they focus on capturing opinion trends over time and performing competitive analysis between ‘set’ topics, our research focuses on pairing sentiments to automatically extracted topics (or themes). Indeed, their system achieves high precision (75-95%, depending on the data) in extracting sentiments within Web pages and news articles. From our perspective, it would be interesting to further explore, for example, the role and impact of a bespoke sentiment lexicon or ‘sentiment markers’ in our analysis.

On this account, there is much research on the identification and use of these ‘markers’. Pang and Lee [10], for example propose ‘a subjectivity detector’ as a mean of improving the efficiency of sentiment analysis at document level. For this they develop a system that can identify and remove ‘objective sentences’ (e.g. plots in a movie review) and then apply a standard machine-learning classifier - such as support vector machines (SVMs) and Naive Bayes (NB) - to the resulting extracts or summaries. Their experiment shows that ‘subjectivity detection can compress reviews into much shorter extracts that still retain polarity information at a level comparable to that of the full review’.

Wiebe et al’s research [16] also focuses on sentiment markers and identifies three types of ‘similarity clues’ that are then used to detect whether an expression is subjective or objective. Interestingly, one of these clues include hapax legomena, the set of words that appear just once in the corpus ‘Apparently, people are creative when they are being opinionated’. Anecdotal evidence from the VYV trial in the Lancaster library would confirm this. However, Wiebe remarks on the importance of ‘context’ and the need for an NLP system to be able to disambiguate these objective/subjective expressions in context.

Wilson et al’s research [18][17] further develops Wiebe et al’s work by addressing the importance of the context in automatic sentiment analysis. It does so by taking a two step-approach: first assigning each ‘subjectivity clue’ a prior polarity, and then disambiguating the contextual-polarity of the clues marked as polar. Prior-polarity is the polarity manually assigned to a word in a lexicon (e.g. MPQA), whereas the contextual polarity ‘is the polarity of the expression in which a word appears’. In particular, Wilson’s research focuses on contextual polarity that is ‘disambiguating the contextual polarity of words with positive or negative prior polarity’. Differently from Wilson et al [18][17], Piao et al. [12] take clauses, not phrases, as the basic syntactic units for the sentiment analysis, suggesting that clauses provide an independent unit of sentiment orientation for complex sentences.

In summary, much research has been done in the field of sentiment analysis but the approaches and the sentiment analysis levels (e.g. document [10], clause [12], or phrase [17] level) are so different that very few have attempted a systematic comparison between tools that use such techniques and that are now available ‘off the shelf’.

B. Actionability analysis

As mentioned previously, the concept of actionability is a novel one in the context of automated text analysis and opinion mining. In our case we consider as actionable statements that provide a clear suggestion on how a product or a service (or a feature of both) could be improved. For example, the statement ‘To keep the library looking as good as new, food and drink
should be banned’ contains actionable knowledge, in that it suggests what actions - ‘banning food and drink’ - should be taken to keep the library looking like new. In this paper we argue that actionable knowledge can be quantified, thus we define ‘actionability’ as the measure of actionable knowledge in a statement.

Actionable knowledge is a term that has been used in Knowledge Management, Alavi and Leidner [1], for example suggest that, in order to be productive, information resources should be converted to actionable knowledge. In contrast, we argue that actionable knowledge can be found, but it needs to be recognized and extracted.

In the field of Knowledge Discovery and Data Mining (KDD) Cao et al. [3] link the concept of actionability to ‘business interestingness’ and describe actionable knowledge discovery as ‘one of the Grand Challenges’. In this context, much research has been done towards the development of a framework for actionable knowledge extraction [3][4]. However, their data-mining approach seems to be confined to interpretation of patterns extracted from structured data (e.g. Australian stock market data) and not from unstructured data such as spontaneous comments and narratives like in our case.

Lee et al. [8] highlight the importance of ‘automatically extracting actionable knowledge from customer feedback data on the Web’ for opinion mining, however they solely focus on the extraction of the sentiment associated with a product or with its features but not on the identification of customers suggestions on how a product could be improved.

Our challenge is to design a system that can automatically identify sentences as actionable. For example, a number of ‘actionability markers’ could be used to identify and measure the ‘actionability’ of a sentence. For this, research in pragmatics could potentially help our research in this direction. There are examples such as research on the use of ‘so’ as a marker of incipient actions [2] or on the use of modal verbs [14] and on a broad range of other constructs [5].

III. METHODOLOGY

This section reports on the methods used to perform the sentiment and actionability analysis. Manual analysis is performed, then, an automatic rule-based tagger approach and a lexicon/rule based method to sentiment estimation. Two machine-learning classifiers, a Naive Bayes machine-learning classifier and the ReadMe classifier are used to estimate sentiment and actionability. These four methods were chosen as they are straightforward to use and required little to no knowledge of internal operation to produce estimation of category. Other methods from the literature require detailed knowledge of NLP methods and careful preparation of data, selection of variables etc. before we can use them for text categorization.

Of the 2000 comments in the database, we randomly sampled 300 comments for analysis. The 300 comment sample size was chosen in another study for manual analysis to be representative at confidence interval 5 and confidence level 95% of the whole dataset. Here we do not use any type of ‘subjectivity detector’, instead choosing to run every comment through the systems. For this study the data set was split into 200 for use as training data with the two classifier methods, and 100 to use as test cases for all automated methods.

A. Manual Analysis

Six researchers independently manually rated each of the 300 comments for sentiment and actionability, then calibrated their ratings in a discussion session. The calibration was performed to reduce the possibility of biasing in the results. In future it would be possible to use the k coefficient to calculate inter-rater agreement scores (Wilson et al. 2009) [17], however in this study the calibration procedure we adopted involved a discussion to arrive at an agreed score. This manual rating created a baseline to compare automatic methods against.

Comments were rated as very negative if it contained multiple negative words or strong negative inference (e.g. “Very poor stock choice is worse than a small branch library”). A comment was negative if it had negative inference (e.g. “It’s not very welcoming”). A neutral comment was if there was no sentiment apparent (e.g. “What happened to the wooden shelves?”). A comment was positive if it had positive tone (e.g. “A good renovation”) and very positive if there were multiple positives in the comment (e.g. “Lovely library very helpful staff”). The resolution of this analysis was reduced for this study to a 3-point scale with very positive and positive comments placed in positive category, neutral staying as neutral and negative and very negative comments placed in a negative category.

We rated a comment as actionable if it provided specific information about what is good or bad about the topic and, at the same time, provided a clear suggestion on how it could be realistically improved (e.g., “Not enough chairs in the music room, more seats please!”). We rated a comment not actionable if there was no content that could specifically be acted upon to solve an issue (e.g., “I am very disappointed in the new library.”). There were also comments that pointed out a problem but did not offer a specific solution; these were rated somewhat actionable (e.g., “Noise echo throughout the library.”).

B. Tagger Analysis

We first used a simple tagger approach to act as a baseline comparison against other methods – simple methods can often perform only marginally worse than complex methods but are often easier to implement and understand; we wanted to evaluate whether this holds in this case. The tagger used in the process of theme extraction was Wmatrix [13]. This tool tags each word with a Part Of Speech (POS) and a semantic category (SEM) taken from the USAS semantic tag set [11]. The SEM tag includes an indication of the sentiment of individual words, e.g. the tag G2.1 has G2.1+ for “Lawful” and G2.1- for “Crime”. A number of rules were written in an attempt to deal with the incompleteness of the tag set for our purposes – e.g. tag Z6 “Negative” is for negative words which do not fit other categories, but does not have a “+.”. Using this you can imagine it would be possible to build up the sentiment of a complete sentence by adding the + and – of every word to score the sentence. However this system does not deal with double negatives without creating further rules. The rules were expanded in an attempt to account for situations where a word such as “very” influences the sentiment. E.g. “the food is very
good” would be rated as more positive than simply “the food is good”. Likewise, “the food is very bad” would be more negative than “the food is bad”. Other more complex situations are dealt with, the double negative such as “the food is not bad” would be rated positive.

If it were possible to build a rule set to define all usage of the English language then it might be feasible to create a very accurate tagger. However language (especially spoken language where tone and facial expression contribute to sentiment) has many foibles and it would be very complex to document with a rule based approach.

This method does not provide scope for rating comments for actionability at the moment, so the procedure was run only for sentiment data. The results can be seen in Figure 1.

C. ReadMe Analysis

Readme is a tool developed at Harvard University by Hopkins and King [6]. The tool has been the subject of a commercial spin out company and has had much exposure in the media. ReadMe is not specific to sentiment analysis – i.e., it just “reacts” to the oracle used in the training data. This makes it more general than the tagger and many other sentiment analysis methods and allows us to use it for the planned actionability analysis. Hopkins and King indicate that most social science applications are interested in the aggregate proportion of all documents which fall into each defined category, hence they are not interested in individual document categorization.

The method requires a set of “training data” be prepared. This involves manually coding a number of documents with the chosen categories. This is then prepared into an input text document along with the (normally much larger) test data set and preprocessed.

First, text is preprocessed – converted to lowercase, removing punctuation and stemming words. E.g. the stem of “consistency”, “consistent” and “consisting” is “consist”. This pre-processing reduces the complexity of the text. The text is then summarized as dichotomous variables, indicating the presence or absence of each stem. A count is not made, then summarized as dichotomous variables, indicating the presence or absence of words from the training data. The ReadMe estimator is input with the manually tagged data set. The Naïve Bayes classifier implementation used here is uClassify [7], which is designed as a multi-purpose classifier for the classification of text sources. It is designed to be quick and simple to use – and offers no further control other than input texts. No manual pre-processing of the text was done before loading into the classifier system.

As before, 200 comments were used to train the classifier for positive, neutral and negative sentiment. The 100 individual test data comments were loaded into the classifier, and the category with the highest probability was returned and recorded. The exercise was repeated with another instance of the classifier for the actionability tagged data.

E. Lexicon and rule based method

We also tested the lexicon and rule based method described by Piao et al [12]. This method combines syntactic structure analysis, a scoring algorithm based on a subjectivity lexicon compiled by Wilson et al. (2005) and a set of rules. As mentioned earlier, this method takes the clauses as the basic units of analysis. So the main focus of this method is to detect the sentiment orientation of the clauses involved, and then aggregate the clause analysis results into the sentence sentiment score. Here the aggregation process does not simply sum up the scores, but applies a scheme to weight different types of clauses according to their relative importance in determining the sentence sentiment orientation. In our experiment, the syntactic analysis was carried out using OpenNLP parser (http://opennlp.sourceforge.net/). As there is no need for training, this method could directly be applied to the test data.

The sentiment analysis proceeds as follows. (1) The input text is split into sentences and parsed for identifying clauses. (2) The sentiment score is calculated for each clause. In the subjectivity lexicon, each word has a sentiment orientation value, such as +1 (strongly positive), +0.5 (weakly positive) etc. The subjective words in the clauses are detected using the lexicon and their sentiment scores are summed up. This sum is divided by the number of subjective words to produce the sentiment score for the clause. (3) The clause sentiment scores are aggregated and divided by the number of clauses to produce sentence sentiment score. In this step, a clause weighting scheme is used to assign different weights to
different types of clauses, e.g. weight 1 for main clauses and zero for concessive clauses (starting with although etc). The resultant sentiment scores range between [+1, -1], with positive, zero and negative scores indicate positive, neutral and negative sentiment respectively.

In the experiment, as was the case for other tools, the 100 test comments are processed with this method and the results are compared against the manual classification for evaluation.

IV. SENTIMENT ANALYSIS RESULTS

The results for the five methods applied to sentiment analysis are presented in Table 1. The classification of each individual comment was compared with the manual tag for that comment. The number of times the classifier got the categorisation correct is represented as a percentage in the 2\textsuperscript{nd} column. The 3\textsuperscript{rd} to 5\textsuperscript{th} columns represent an aggregated category estimation to allow comparison of the readme classifier. The number of comments tagged with each category is represented as a percentage.

<table>
<thead>
<tr>
<th>Analysis</th>
<th>% of correct individual classifications</th>
<th>Aggregated category estimation % of comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual</td>
<td>100</td>
<td>40 32 28</td>
</tr>
<tr>
<td>Tagger</td>
<td>51</td>
<td>19 24 57</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>53</td>
<td>32 50 18</td>
</tr>
<tr>
<td>ReadMe</td>
<td>-</td>
<td>14 57 29</td>
</tr>
<tr>
<td>Lexicon-Rule</td>
<td>55</td>
<td>14 36 50</td>
</tr>
</tbody>
</table>

As indicated by Table 1, none of the methods are highly accurate. The tagger method is the least accurate, and the aggregated scores indicate that it may have mis-categorised many negative comments as positive. The lexicon rule method is the most accurate, producing a correct classification 55\% of times. The naïve Bayes method produced the best fit to the aggregated categories. The ReadMe estimator misclassified the distribution, perhaps misestimating negative comments as neutral; however this table does not represent the direction, nor degree of error in the cases where the classification was incorrect.

V. ACTIONABILITY ANALYSIS RESULTS

The results for the two methods applied to actionability analysis are presented in Table 2. Again, the classification of each individual comment was compared with the manual tag for that comment. The number of times the classifier got the categorization correct is represented as a percentage in the 2\textsuperscript{nd} column. The 3\textsuperscript{rd} to 5\textsuperscript{th} columns represent an aggregated category estimation to allow comparison of the readme classifier. The number of comments tagged with each category is represented as a percentage.

<table>
<thead>
<tr>
<th>Analysis</th>
<th>% of correct individual classifications</th>
<th>Aggregated category estimation % of comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual</td>
<td>100</td>
<td>54 22 23</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>66</td>
<td>53 19 27</td>
</tr>
<tr>
<td>ReadMe</td>
<td>-</td>
<td>79 14 7</td>
</tr>
</tbody>
</table>

The results for the ReadMe classifier are disappointing given the published accuracy of the system. However it may be caused by the small size of the training data available in our evaluation. Worked examples used test and training data sets of the order of 1800+ documents [6]. Manually classifying 1800+ training examples entails a substantial amount of work, especially because each example needs to be classified by multiple people. In our evaluation, just 200 short comments were used as training data, so the result may not be fully representative of all of the language used by the different categories.

The Naïve Bayes classifier proved to be the most accurate approach for actionability analysis and second best for sentiment. There are a number of suggestions to improve the accuracy – it is expected that using stemmed words would have a positive effect. The Naïve Bayes approach proved more accurate for actionability estimation than sentiment analysis. This is perhaps due to very specific “actionable” words being present in actionable comments – words such as “should” and “need” appear frequently and only in actionable comments so are perhaps easier to classify. Sentiment words such as “like” appear in both positive and negative categories so would be much harder to categorise.

VI. CONCLUSIONS

Before performing this first review of analysis techniques on our VoiceYourView data, it was thought the tagger based approach to sentiment analysis would perform poorly. Our evaluation has revealed that in fact it is not so inaccurate in comparison with the methods tried. Our study highlights how complex it is to create a system for sentiment analysis. The results for the ReadMe classifier are disappoitting given the published accuracy of the system. However it may be caused by the small size of the training data available in our evaluation. Worked examples used test and training data sets of the order of 1800+ documents [6]. Manually classifying 1800+ training examples entails a substantial amount of work, especially because each example needs to be classified by multiple people. In our evaluation, just 200 short comments were used as training data, so the result may not be fully representative of all of the language used by the different categories.

The Naïve Bayes classifier proved to be the most accurate approach for actionability analysis and second best for sentiment. There are a number of suggestions to improve the accuracy – it is expected that using stemmed words would have a positive effect. The Naïve Bayes approach proved more accurate for actionability estimation than sentiment analysis. This is perhaps due to very specific “actionable” words being present in actionable comments – words such as “should” and “need” appear frequently and only in actionable comments so are perhaps easier to classify. Sentiment words such as “like” appear in both positive and negative categories so would be much harder to categorise.
This first look at applying classification techniques from the literature to our VoiceYourView data has highlighted the difficulties and complexity of classification systems. Just four approaches were tested on the data – there are many other classifiers including other machine learning classifiers, such as state vector machines which have not been evaluated. Indeed it is a major field of interest in computer science, certainly with no single approach appearing to be totally reliable for automatic sentiment analysis. Here we were unable to reproduce the ball park 80% accuracy cited by other researchers, and tailoring of algorithms to the specific data set may be required. Tailoring of algorithms to a specific data set requires expertise in how they work and in the domain of the data – if these techniques are to be used widely, their application needs to be economic and easy to use so tailoring may not be realistic. This tailoring of systems, however, may go against the ideology of VoiceYourView – which states that comments can be on any topic and contain anything that occurs to commenter. Similar results to those discussed here are expected when analyzing data from social media, especially if spoken comments are transcribed.

ACKNOWLEDGMENT

This research was funded by the RCUK Digital Economy Programme.

REFERENCES