# Scary Films Good, Scary Flights Bad

## Topic Driven Feature Selection For Classification Of Sentiment

Scott Nowson
Appen Pty. Ltd
1 Railway Street, Chatswood
Sydney, Australia
snowson@appen.com.au

## ABSTRACT

This paper describes preliminary work on feature selection for classification of review text by both sentiment rating and topic. The premise stems from the notion that one size does not fit all; that feature sets for sentiment analysis should be tailored to the topic of a text. Thus it naturally follows that for this to be effective it is also necessary to first determine the topic of a text. Following successful work on classification of texts by author demographics, a corpus of review texts labelled with attributed rating, topic area, and user demographics has been compiled. This collection was divided for this work into different topic groups in order to automatically classify between both text topic and subjective rating. By using a single supervised statistical approach to feature selection, it is shown that improvements can be made to classification accuracy using topic tuned features sets over more generic features.

**Categories and Subject Descriptors:** J.4 [Social and Behavioural Sciences]: Miscellaneous

**General Terms:** Human Factors; Performance.

**Keywords:** Sentiment Analysis, Topic, Feature Selection, Text Classification.

## 1. INTRODUCTION

As is clear from dedicated workshops such as this and increasing attention within broader social media venues, there is considerable interest in affective language processing. In general, affective computing varies from short-lived feelings, through emotions, to moods, and ultimately to long-lived, slowly-changing personality characteristics [22]. Within datamining the majority of this work is focused on the analysis of sentiment and opinion as directed at specific topics or entities. By utilising natural language processing it is possible to explore the relationship between the language used in a text and the sentiment it is written to express.

Considering how people feel and the entities to which these feelings are directed has considerable applications. Such work has been explored on many levels from definite opinions [5, 19] to the more transient concept of mood [12, 27]. Studies have explored many fields from tracking and predicting consumer opinions [13] and question answering [26] to exploring political beliefs [8, 3] and looking for radicalization [2]. Data investigated includes strictly subjective review texts [19], blog posts [10] or the latest in consumer generated media - microdata on Twitter via an explosion of tools such as twendz[1], Twitter Sentiment[2] and twitrratr.[3]

Since the earliest binary classifications of sentiment – a thumbs up or down decision to determine if a movie review is positive of negative [20, 28] – work has been conducted on closed sets of data. However, the scope of real world applications is significantly larger with potential datasets expanding exponentially. So, in addition to the pursuit of accurately determining sentiment from any text, there is a desire to do this for increasingly vast numbers of texts. The ultimate solution to such a broad problem must be both accurate and efficient.

In order to address the first part of this problem, this paper explores the choice of feature set for classification. Just as not all authors speak with one voice – as Pang and Lee observe one person's four-star review is another's two-star [19] – so too do voices choose different words when talking on different topics. This study investigates a statistical approach to feature selection for sentiment classification dependent on topic area.

Of course, this would seem counterintuitive to the desire for efficiency in computation – to create multiple feature sets instead of one. However, the approach employed here allows feature sets of various sizes to be created depending upon the criteria applied. Thus it is possible to answer the second part of the problem posed above by exploring how the number of features effects accuracy. It also follows that in order to select a topic-specific feature set for classification, it is necessary to know the topic area of a text. This study therefore also applies the same feature selection technique to topic identification.

This study is part of a larger program of work which applies linguistic analyses within various forms of computer-mediated communication (CMC). We have previously successfully demonstrated the application of NLP approaches to the classification of text by author demographic and psychometric traits [18, 4]. While these approaches and applications continue to be improved, it is a natural extension of

[1] http://twendz.waggeneredstrom.com/
[2] http://twittersentiment.appspot.com/
[3] http://twitrratr.com/

*TSA'09,* November 6, 2009, Hong Kong, China.
Copyright 2009 ACM 978-1-60558-805-6/09/11 ...$10.00.

this work to explore sentiment. This paper will demonstrate the utility of employing topic-tuned feature sets to the classification of the sentiment rating of texts; it will also report the same approach applied to the classification of topic.

This paper is structured as follows: it begins by discussing the motivation behind exploring feature set creation, and looks at prior work in this area. It then describes the data used in this study, and outlines the statistical basis upon which feature sets will be defined. Experiments in sentiment rating classification are reported, followed by parallel work on topic classification. In discussing the results, planned avenues for future analysis are laid out.

## 2. MOTIVATION AND BACKGROUND

Following Melville, Gryc and Lawrence [10], traditionally there have been two broad approaches to feature set creation for sentiment classification: knowledge-based and learning-based. The most straightforward knowledge-based approaches consist of human created lexicons of subjectively oriented words. There are human annotated subjectivity lexicons available for use in English [30], and there has been work investigating the possibility of using a lexicon in one language to create one in another [11].

There have been some semi-automatic approaches to generating a lexicon. Kim and Hovy [6] began with a small seed list of verbs and adjectives, and used the synonym and antonym relationships from WordNet to grow the list. Rather than throw out ambiguous terms, they used statistical measures to approximate the degree of polarity. Similarly, Gamon and Aue [5] started with an even smaller set of seed terms and tried two approaches to growing their list. The first was based on Turney's [28] concept of semantic orientation, selecting terms with strong pointwise mutual information (PMI) associations with the seed terms. The second method involved looking for co-occurrence of new and known terms in unlabeled sentiment data. They found that an iterative combination of the two approaches worked best. It should be noted that Turney's original approach can be considered near automatic, since he began with only two seed terms.

Learning-based approaches tend to consider sentiment analysis framed as a text classification problem. Pang, Lee and Vaithyanathan [20] used a standard bag-of-words technique and found it superior to a hand-crafted lexicon. Along with word-unigrams, they tried adding bigrams and higher level part-of-speech information, but found that the performance improved little for the increased complexity.

Melville et al. [10] used a combination of both prior knowledge and a learning-based approach. They began with a non-topic specific subjectivity lexicon and refined it based on training examples drawn from different domains. They found that performance improved more, and with fewer training examples, than either approach in isolation.

The intention of the methodology to be used in this paper is similar in the sense that it is intended to be a generalised framework for producing feature sets for different domains or topics, or indeed different tasks. The approach is different in the sense that a lexicon of terms is produced by means of a purely learning-based non-seeded strategy, in order to drive a subsequent knowledge-based classification process.

Mei et al. [9] proposed a similar combination of topic and statistical feature extraction for sentiment determination. However, their approach was to combine subjective language models drawn from different topics in order to best generalise to unseen topics, as opposed to keep them distinct. The generalised nature of a feature set is attractive when considering classifying large volumes of undirected blog data, as they were proposing to do.

However, as further motivation for considering distinct feature sets for delineated categories, consider the alternative subjective interpretations of adjectives in different contexts. The adjective *scary* could be considered positive in reviews of films or books, particularly those of the horror genre. However, it is unlikely that it could be a good word when reviewing a hotel or airline, or perhaps even a film aimed at children. Similarly, though it may have seemed counterintuitive at first, Melville et al. [10] found the word *truth* to be down-weighted with respect to politically motivated texts.

Before describing the approach to feature set creation to be used in this study, it is first necessary to report on the nature of the data upon which it is to be deployed.

## 3. DATA COLLECTION

In order to carry out future work relating sentiment to author traits (see section 7) it was necessary that data meet a number of requirements. First, texts must be labelled in some manner for sentiment or opinion; and second, texts must be from authors about whom demographic traits are known or can be determined. There are a number of available corpora, which satisfy the first criterion (for example [29]); and from previous studies of individual differences [25, 21, 14] there are corpora to meet the second. However, since there are few (if any) text collections which answer both requirements it was necessary to create one.

Review texts were collected from a online review site within which members have a user profile page. From the information on these pages it was possible to extract each author's geographical location and determine their gender. Review texts were collected for each author of known gender and are labelled in a number of ways: a star rating out of five (1-5, zero is not an option); a binary 'recommended' marker; product and category information; and an indication of the 'helpfulness' of the review as judged by others.

Overall, over 150,000 reviews were collected. However, there was a definite bias toward the positive in the collection: over 70% have ratings of four or five. Balancing the corpus to the smallest set – the reviews with a score of one – results in 50K documents. For these preliminary studies, a further subset was created – balanced for both gender and review score – of 10K texts.

### 3.1 Review Topics

Though the experimental corpus was balanced for work still in progress on gender and sentiment, it also contains considerable information on the topic area of each review. The product or service of each review is given a hierarchical category listing such as 'Home > Electronics > DVD Players'. This allows for the study of different granularities of topic distinction. Note that in this paper topic is used synonymously with category, as opposed to referring to the direct subject of a text or sentence.

When it came to selecting topics, the obvious first step was to look at those which were most popular. The top five categories for review were 'Videos & DVDs', 'Books', 'Music', 'Toys', and 'Baby Care'. The presence of 'Baby Care'

**Table 1: Corpora sizes**

| | Total | Documents per | | |
|---|---|---|---|---|
| Corpora | Size | rating | topic | sub-topic |
| TwoTopic | **900** | 90 | 450 | – |
| TravelSent | **450** | 90 | – | – |
| TravelTopic | **300** | – | – | 100 |
| EntertSentTopic | **1500** | 100 | – | 500 |

led to further investigation which revealed that there were clear gender biases in review topic. Reviews of ‘Toys’ and ‘Baby Care’ products are significantly more likely to have been written by females (approximately 90% of the time); similarly men author the majority of the reviews of ‘Video Games’ and sports related products. The effects of gender and sentiment are outside the scope of the study reported here. Therefore, in order to best mitigate any dependent effects this might have, it was decided to use topics that were similarly authored by both genders.

For this study, two groups of three topic areas were selected: ‘Music’, ‘Books’, and ‘Videos & DVDs’ were grouped as ‘*Entertainment*’; while ‘Airlines’, ‘Hotels and Resorts’, and ‘Destinations’ were grouped as ‘*Travel*’. For the different experiments of this study sub-corpora were created, balanced variously for topic, sentiment rating or both. The total number of texts, including details of number of texts per topic or rating, as balanced to the smallest sub-group can be seen in table 1.

# 4. FEATURE SELECTION

The approach to feature selection to be used in this study is a supervised statistical approach designed to be flexible for language modeling in a number of domains and tasks in order to minimize the size of the feature set used.

## 4.1 Description of approach

Word n-gram approaches provide a potentially vast feature set with which to work. However, as the volume of potential data which needs to be processed in any practical application increases it is beneficial to computational tractability to reduce the feature space as much as possible. Therefore, following the lack of improvement found from looking beyond unigrams (as reported earlier, [19]) it is this single unit space in which this study works.

There are a number of approaches to selecting features which best fit a dataset - utilising information gain for example. The method used here is a supervised statistical approach which has previously been shown to be effective in language and affect studies such as classification of author personality [18] (following [17]), as well as having been applied to gender studies [24]. The technique is based on log-likelihood corpus comparison [23] using frequency information to produce a G2 value for each feature.

Consider two corpora whereby corpus 1 has total word count $c$ and corpus 2 has count $d$. Suppose for a single feature, such as a unigram, the frequencies in the respective corpora are $a$ and $b$. These observed values (O) and the word counts (N) can be used to determine the expected values (E) as in equation 1.

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i} \tag{1}$$

In the example here, $E_1 = \frac{c*(a+b)}{(c+d)}$ while $E_2 = \frac{d*(a+b)}{(c+d)}$. Normalization is not required, since the calculation of E takes into account the corpora sizes. The log-likelihood is then calculated according to equation 2.

$$-2\ \ln\lambda = 2\sum_i O_i\ \ln\left(\frac{O_i}{E_i}\right) \tag{2}$$

This equates to calculating the log-likelihood G2 value as: $G2 = 2*((a*\ln\frac{a}{E_1})+(b*\ln\frac{b}{E_2}))$. The higher the G2 value is the more significant the difference in frequencies of a feature between the two corpora. By taking all those scores which reach certain levels of significance, it is possible to determine which features most distinctly represent each corpora.

This technique has the advantage of evaluating corpora simultaneously, rather than in isolation. This is similar to approaches described previously that consider both the positivity and negativity of a word in combination. The difference here is that using just unigrams it is able to still compare ‘good’ which is a positive unigram with the negative bigram ‘not good’. If *good* is used with similar frequency in both the positive and negative corpora, then it will not show up as a significant feature for either. By extension, this will also cancel the effect of phrases such as ‘*it did not turn out quite as good as I had hoped*’ which in other approaches might require more complex language modeling to detect.

In its original application, this approach is used to determine the most distinct features of two corpora. However, following the previous work in personality [17, 18] the approach has been adapted to compare features across three corpora. Three two-way comparisons are run, and the results for each feature combined. By comparing the significance of the three comparisons for each feature it is possible to determine if it is a significant indicator of any of the corpora. Similarly, it is also possible to determine which features are indicative of two of the corpora but not the third.

In previous studies with small corpora [18], it has been observed that significant features can be unduly influenced by particularly verbose texts or authors. For example, a number of particularly long positive reviews of a specific product could result in the name of the item being included as a positive word. In order to reduce the effect of this, a filter can be applied to only include features which are used in a particular percentage of the texts in that group. By varying the two thresholds of significance and breadth of use it is possible to create smaller yet more directly relevant feature sets.

## 4.2 Deriving feature sets

Feature sets for this study will consist of all features for each sub-group within a selection. That is to say, for studying binary sentiment classification, the approach above can be employed to select significantly used words in both the positive and negative corpora, and these words will be combined into a single feature set. Similarly, either a binary or three-way comparison will be used to determine all features for classifying between topics.

There are multiple possible combinations of thresholding possible, but there will be two levels of strictness applied to feature selection for this study. The thresholds to be applied – on the statistical significance and breadth of usage within a corpora – are described in the table below. Note that significance is described in terms of the percentage level and probability of the feature occurring non-significantly, along with the percentile to which the item will therefore belong.

| Label | Significance | | | Usage |
|---|---|---|---|---|
| <50 | 0.01% lvl | $p < 0.00001$ | 99.99th perc | 50% |
| <10 | 5% lvl | $p < 0.05$ | 95th perc | 10% |

Feature sets are to be created for two tasks – sentiment and topic classification. For sentiment analysis, just as manually created lexicons tend to only consist of positive and negatively loaded terms, so the two most extreme ratings are used – feature sets are a combination of distinct unigrams for the 1- and 5-star rated texts. In order to evaluate the effect of tailoring feature sets to data, this is done for the two major topic areas both separately (Entertainment, Travel) and combined (labelled here as 'TwoTopic'). For topic analysis the same groupings are used, however the feature sets are generated by dividing the corpora along topic lines - Entertainment against Travel, or books against DVDs against music.

In order to compare the performance of the generated sentiment feature sets, a more traditional manually compiled lexicon is used. This is drawn from the subjectivity lexicon of Wilson et al. [30] and includes all words rated either positive or negative (regardless of strength) for prior polarity. The size of all feature sets of this study can be seen in table 2. Unless otherwise stated, feature selection process was binary: rating one versus five for sentiment driven features; and based on category for topic driven features.

What is clear from the sizes is that even the least restrictive feature sets are significantly smaller than the manual lexicon. The more selective sets are generally a factor smaller again. These sets lend themselves well to the notion of computational efficiency. However, the risk is that they do not generalise well nor even perform well upon the data from which they were created.

It is worth noting that while the significant unigrams selected were based on their frequency within each corpora, frequency is not the measure used in the tasks reported below. For all experiments there were three statistical measures generated for each feature: frequency, relative frequency, and ranked frequency – whereby the most frequent term for each document is assigned a rank value of 1, the second most frequent a value of 2 and so on, with any term with equal frequency given equal rank value. Across the board using ranked frequency was more accurate than the other two measures. Therefore, all features in any set used are based on the ranked frequency.

**Table 2: Number of features per feature set**

| Sentiment | | | |
|---|---|---|---|
| Polarity | | 5190 | |
| | | *(<50)* | *(<10)* |
| TwoTopic | | 31 | 250 |
| Entertainment | | 36 | 318 |
| Travel | | 27 | 302 |
| Topic | | | |
| | | *(<50)* | *(<10)* |
| TwoTopic | | 50 | 502 |
| Entertainment | (3) | 337 | 1226 |
| Travel | (3) | 103 | 740 |

**Table 3: Example words**

| Topic | Negative | Positive |
|---|---|---|
| Entert | ***book***, *really, instead*<br>*kids, plot, talent*<br>*waste, annoying* | ***finally***, *song, rock*<br>*fast, story, brilliant*<br>*favourite* |
| Both | *i, me, my, worse*<br>*worst, problem* | *is, are, you, most*<br>*great, best* |
| Travel | ***finally***, *time, flight*<br>*airline, room, never* | ***book***, *island, beaches*<br>*historic, recommend* |

### 4.3 Feature sets derived

This approach assumes that the feature sets created will be different. Therefore, in order to illustrate the results, a sample of words drawn from the Entertainment and Travel feature sets are presented in table 3. The table contains examples of words that are significantly positive or negative in either or both corpora. Note that words highlighted in **bold** are those which appear with opposite polarity between the topics.

As one might expect, a number of the words can be considered explicitly subjective, such as 'brilliant' and 'problem'. Differences in the use of such words suggest that merely being a subjective word does not automatically suggest efficacy for sentiment classification. There are also a number of topic-specific words which show an association with polarity despite being seemingly objective, for example 'historic', 'song' and 'flight'. The utility of the approach reported in this paper is further demonstrated by the words which appear in both features set, yet indicate opposing polarities. In the entertainment corpora 'book' is used in a negative context, yet it is used positively in travel. These conflicting uses result in 'book' not appearing in the feature set drawn from both topics – its obvious usefulness for classification lost.

Though it is outside the scope of this work to consider in detail the implications of these feature sets, even this small sample lends itself to such considerations. For example, that 'plot' and 'story' appear with different polarities suggests that in a critical review context they are not used as synonymously as might be expected. Also, the use of self-reference appears to be related to polarity. Within this corpora, it appears that negative reviews are written from a first person perspective, reflected by the significant use of first-person pronouns; positive reviews, on the other hand, appear to be written with the reader in mind.

## 5. CLASSIFYING SENTIMENT

In this study, the task of identifying sentiment rating from text will be a binary one: low ratings versus high ratings. The most traditional form of this task is simply to make a decision between the highest rating – the texts that should exhibit the most positive polarity – and the lowest – the most negative texts. Texts with these ratings are the most distinct in terms of opinion expressed, and it follows that they should show the greatest linguistic distance linguistically.

However, it is not only the extreme levels of subjectivity that are of interest in this field. To only consider the far ends of the scale – the 1- and 5-star ratings – is to ignore a significant proportion of data. It also significantly limits our ability to considered change in sentiment, since there are only two points on the scale. So, for this study the definition

**Table 4: Classification accuracy of rating comparisons**

| CLASS | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| 1 | 59.9 | 71.1 | 79.2 | 83.3 |
| 2 | | 70.0 | 69.9 | 77.0 |
| 3 | | | 58.8 | 70.4 |
| 4 | | | | 58.6 |

of positive and negative ratings is not just 1 vs 5, but also grouping 1 & 2 vs 4 & 5. Note that the two tasks will be labelled for the remainder of the paper as '1-2' and '12-45' respectively.

There are three corpora balanced for rating that are to be used to assess the feature set. Along with the polarity feature set, each corpus has been used to create a feature set using the method described in section 4. It is expected that while the polarity set with broad focus will perform well, the more tightly focused topic specific feature sets will perform best on the topic corpora from which they were drawn.

In all classification experiments in this study support vector machines as implemented in LIBSVM [1] are employed, along with five fold cross validation. In the experimentation process, a wide range of parameter values were tried. Though the specific settings are not discussed here, only the best performance is reported. However, before reporting results, an explanation is given for why – particularly in light of prior work into using the neutral case – this study is solely concerned with binary classification.

### 5.1 Why so binary?

There have been a number of studies which have shown the utility of using the neutral case in automatic classification of polarity (eg. [7]). However, the corpus of this paper in its original form was heavily biased in favour of the existence of positive reviews. This suggests, perhaps, that users may have a general preference for being positive (similar to prior findings that personal weblog authors tend to have high Openness personalities [15, 16]. This effect can be controlled – at least in terms of external labelling – by the construction of corpora where each class is equally represented.

However, this does not account for internal consistency *between* ratings – that there is not also a similarly positive bias. Since there is no zero option on the data collected a rating of three is the theoretically neutral score. Inspired by previous work on multi-class classification (most notably [19]) in order to determine if this is the case a comparison of all rating classifications against each other is performed. This is conducted on the whole of the 10K document corpus, using the subjectivity lexicon words as features. The best results for each classification pair can be seen in table 4.

Since this is a balanced corpus, the baseline in each case is 50%, and a higher accuracy suggests a greater difference in the overall level of sentiment expressed between classes. For the most part results are consistent: a rating difference of one returns only 60% accuracy at distinguishing between classes; a difference of two is approximately 70%; three or more around 80%. The obvious exception is the accuracy between ratings of two and three - which performs more like a two star distinction than a one. This suggests a three star rating may not be neutral, but in fact tends toward the more positive side, embodied perhaps in a rating system more like 1 - 2 - 2.5 - 3 - 4 - 5 stars. Until this issue can be investigated further, it is considered for this study that there is no neutral group, and the three-star rated reviews are ignored.

**Table 6: Classification of topic using category driven feature sets**

| Feature<br>Set | Binary<br>TwoTop | Ternary<br>Entert | <br>Travel |
|---|---|---|---|
| Baseline | 50.0 | 33.3 | |
| TwoTop<50 | 92.6 | 73.5 | 51.7 |
| Entert<50 | **95.9** | **95.2** | 62.0 |
| Travel<50 | 91.2 | 54.9 | **90.7** |
| TwoTop<10 | **98.0** | 90.8 | 87.7 |
| Entert<10 | 97.4 | **96.6** | 74.3 |
| Travel<10 | 96.1 | 87.4 | **92.3** |

### 5.2 Results

The results of the best classification performance for each task/feature set combination can be seen in table 5. For ease of interpretation the best results for each task per group of derived features are in **bold**, and those that performed more accurately than the polarity set are marked with a *.

The human created Polarity lexicon performs as well as expected given the earlier identified limitations of using unigrams out of context. It does slightly less well in the Entertainment corpora than the others, but across the board it is less accurate when applied to the less distinct 12-45 division. In fact in almost all cases the performance in the more distinct 1-5 task is higher.

Similarly, it is clear that the performance of the <50 feature sets (see table 2) is lower, in many cases considerably so, than the <10 sets. This is understandable given the size of each set. However, considering the best of these (Travel<50 on Travel 1-5) performs at 77.2% from 27 features, this is not too great a degradation in performance compared with the polarity set with almost 200 times the number of features.

As expected the majority of the best results comes from the use of the topic-tuned feature sets on the appropriate topic corpora. It is unsurprising that the TwoTopic feature set does well in both sub-topic categories, since it is drawn from a combination of the two. However, it is interesting to note that the narrow topic sets also do reasonably well on the other tasks.

## 6. CLASSIFYING TOPIC

As with the sentiment classification task above, there are again three corpora to be used in this assessment, though here they are balanced for topic. Again, each of the corpora has been used to create a specific feature sets The main difference here is that the classification of topic *within* the two categories is a three class problem due to the three subcategories that make up each. Once again, it is hypothesised that feature sets will perform best on the topics upon which they were based.

### 6.1 Results

The results of the experiments can be seen in table 6. Once again, the best results are marked by **bold** font.

There are two main differences between the results here and those earlier for sentiment (table 5): the greater accuracy of the more restrictive (<50) task specific feature sets;

**Table 5: Classification of rating using sentiment driven feature sets**

| | Two topics | | Entertainment | | Travel | |
|---|---|---|---|---|---|---|
| | 1 v 5 | 12 v 45 | 1 v 5 | 12 v 45 | 1 v 5 | 12 v 45 |
| Polarity | 83.1 | 74.7 | 78.2 | 74.2 | 83.9 | 76.9 |
| TwoTop<50 | 68.9 | 63.9 | 61.3 | 60.0 | 75.6 | 70.0 |
| Entert<50 | 65.0 | 64.3 | **64.0** | **61.7** | 68.9 | 65.8 |
| Travel<50 | **73.1** | **67.1** | 58.2 | 58.2 | **77.2** | **73.1** |
| TwoTop<10 | **85.3**** | **76.4**** | 76.0 | 71.9 | 82.8 | **75.8** |
| Entert<10 | 81.1 | 72.1 | **82.3**** | **75.3**** | 82.2 | 71.2 |
| Travel<10 | 81.9 | 73.8 | 74.5 | 71.7 | **87.2**** | 73.1 |

and the comparably poorer performance of non-topic specific sets. For the three-way topic classification within the Entertainment and Topic domains, the targeted feature sets performed at over 90%, while the best performing non-targeted was less that 75%. It is, however, unsurprising that both targeted feature sets performed well on the binary task – though the words in each were selected to distinguish *within* the grouped category, it is understandable that they therefore perform well at distinguishing *between* the two groups.

The performance of the larger feature sets is stronger, but not by a great deal. Increasing the size of the two topic feature set by a factor of ten has only produced a 2.1% increase in performance.

# 7. DISCUSSION

One of the most obvious conclusions one can draw from the two sets of experiments in this study is that topic is easier to classify than sentiment. It is perhaps more accurate, however, to say that despite perceived differences between texts at the extremes of sentiment, they are clearly less distinct than texts related to two different topics. This is evidenced in the size of the feature sets. These were created to be the terms which best distinguish between two or three corpora to a certain degree of significance. The number of strongly distinct terms is far greater for the topic driven sets.

A subsequent conclusion that this disparity in results suggests is that there is less variety in the language used to convey the topic of a text than is used to express the sentiment therein. Pang and Lee's observation concerning the different levels to which different authors can express the same self-perceived level of sentiment surely has no analogue in topic. It is perhaps obvious to say, but measuring sentiment – as one does when asked to quantify with a rating – is entirely subjective; the fact that the sentiment is being expressed about a movie or a hotel is entirely objective.

As was intended, this study has shown the utility of the methodology employed here, a form of divide and conquer. Employing feature selection to data stratified by topic has proven more effective that having all the data together. The approach to the subsequent feature selection has created reliable feature sets that lend themselves well to large scale computing. The feature sets used here have been particularly small and, to varying degrees depending on the task, have performed well – in some cases very well.

Of course, there are natural criticisms that can be leveled at the results. The most pertinent of these is that the results are too good, that they merely reflect over-fitting and the feature sets will fail to generalise. The very nature of the approach is to select features best suited to the task in hand. However, it is certainly true that for this preliminary study the features were created on the same data upon which they were then used to classify. Investigating beyond this specific subset of the collected data will be the first task following this study – not only drawing more data from the broader collection, but classifying on entirely unseen data.

## 7.1 Future Work

There are a number of directions that future work can follow, informed by the experiments, the data, and the long term goals of the work in which this study sits. The most obvious, as mentioned above, is to use larger corpora so as to include unseen data in experiments.

When it comes to determining topic the areas selected here were chosen as they were obvious categories that grouped well and statistics suggested they would show only minimal language effects due to author gender. For future work it is worth considering more closely the relationship between categories: how similar or different they are; and how this effects performance when distinguishing between them. It is also worth considering combining feature sets and classifiers in order to create a system capable of determining between multiple possible topics.

For sentiment classification, perhaps the most necessary step is to explore a finer grained classification than simply positive or negative. Being able to distinguish the degrees of strength of an opinion being expressed has significant applications in commercial and security domains. The first step here is to look more closely at the differences between the different rating levels: as described earlier, in this corpus there appears to be a great degree of positivity in the scale-implied 'neutral' 3-star rating level. It would also be interesting to investigate this phenomena in data drawn from alternative sources.

A related effort to this, which fits into to our broader interests is to explore the relationship between individual differences and the language used to express sentiment. As described, this corpus was particularly constructed so as to facilitate analysis of gender and sentiment, though the data also includes geographical information. Our previous work has looked also at age, level of education and psychometric traits of personality [18, 4]. It is intended that analysis developed in the review domain will be applied to a broader range of CMC text types, in a natural extension of our existing tools.

# 8. CONCLUSION

This paper has reported the first results of a program of sentiment analysis designed to complement a broader suite of work exploring language in CMC. It had two intentions: the first was to show that tuning a language model to a particular topic has greater accuracy on classifying the subjective rating of a text than a more generic set of features; the second was to demonstrate the utility of a single statistical technique for generating language models for the classification of a text by both sentiment rating and topic. The promising results presented here go some way to accomplishing these goals. In both situations the technique has created small, focused feature sets that have performed well. There is an obvious direction for future work to show that these results are not the result of over-fitting and that feature sets will generalise well to unseen data.

Classification of text by topic has proven a degree easier than by sentiment rating. However, it is no less interesting or important a task. Indeed, the approach to sentiment analysis used here hinges upon it – an approach that thus far seems quite effective.

## 9. ACKNOWLEDGMENTS

The author would like to thank all at Appen for their support, ideas and feedback, but particularly Julie Vonwiller and Stephen Norris. The author also thanks the reviewers for their comments and suggestions.

## 10. REFERENCES

[1] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[2] H. Chen. Sentiment and affect analysis of dark web forums: Measuring radicalization on the internet. In J. Hajic and Y. Matsumoto, editors, *IEEE International Conference on Intelligence and Security Informatics*, pages 104–109, Taipei, June 2008. IEEE.

[3] K. T. Durant and M. D. Smith. Predicting the political sentiment of web log posts using supervised machine learning techniques coupled with feature selection. In *Advances in Web Mining and Web Usage Analysis*, pages 187–206. Springer, Berlin / Heidelberg, 2007.

[4] D. Estival, T. Gaustad, B. Hutchinson, S. B. Pham, and W. Radford. Author profiling for english emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING 2007)*, pages 263–272, Melbourne, Australia, 2007.

[5] M. Gamon and A. Aue. Automatic identification of sentiment vocabulary: exploiting low association with known sentiment terms. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, pages 57–64, Ann Arbor, MI, 2005.

[6] S.-M. Kim and E. Hovy. Determining the sentiment of opinions. In *Proceedings of Coling 2004*, pages 1367–1373, Geneva, Switzerland, Aug 23–Aug 27 2004.

[7] M. Koppel and J. Schler. The importance of neutral examples in learning sentiment. *Computational Intelligence*, 22(2):100–109, 2006.

[8] R. Malouf and T. Mullen. Taking sides: User classification for informal online political discourse. *Internet Research*, 18:177–190, 2008.

[9] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. Topic sentiment mixture: Modeling facets and opinions in weblogs. In *Proceedings of WWW 2007*, Banff, May 2007.

[10] P. Melville, W. Gryc, , and R. Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th Conference on Knowledge Discovery and Data Mining (KDD-09)*, Paris, June 2009.

[11] R. Mihalcea, C. Banea, and J. Weibe. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 976–983, Prague, June 2007.

[12] G. Mishne. Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access*, Salvador, Bahia, Brazil, August 2005.

[13] G. Mishne and N. Glance. Predicting movie sales from blogger sentiment. In *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, Palo Alto, CA, 2006.

[14] S. Nowson. *The Language of Weblogs: A study of genre and individual differences*. PhD thesis, University of Edinburgh, 2006.

[15] S. Nowson and J. Oberlander. The identity of bloggers: Openness and gender in personal weblogs. *AAAI Spring Symposium, Computational Approaches to Analysing Weblogs*, Stanford University., 2006.

[16] S. Nowson and J. Oberlander. Identifying more bloggers. In *Proceedings of the International Conference on Weblogs and Social Media*, Boulder, CO, 2007.

[17] J. Oberlander and A. J. Gill. Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes*, 42(3):239–270, 2006.

[18] J. Oberlander and S. Nowson. Whose thumb is it anyway? Classifying author personality from weblog text. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics and 21st International Conference on Computational Linguistics*, Sydney, Australia, 2006.

[19] B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 115–124, June 2005.

[20] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 79–86, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[21] J. W. Pennebaker and L. King. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77:1296–1312, 1999.

[22] R. W. Picard. *Affective Computing*. MIT Press, Cambridge, Ma., 1997.

[23] P. Rayson and R. Garside. Comparing corpora using frequency profiling. In J. Hajic and Y. Matsumoto, editors, *Workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000)*, pages 1–6, Hong Kong, October 2000.

[24] P. Rayson, G. Leech, and M. Hodges. Social differentiation in the use of english vocabulary: some analyses of the conversational component of the british national corpus. *International Journal of Corpus Linguistics*, 2(1):133–152, 1997.

[25] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker. Effects of age and gender on blogging. In *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, 2006.

[26] S. Somasundaran, T. Wilson, J. Wiebe, and V. Stoyanov. Qa with attitude: Exploiting opinion type analysis for improving question answering in on-line discussions and the news. In *Proceedings of the International Conference on Weblogs and Social Media*, Boulder, CO, 2007.

[27] S. O. Sood and L. Vasserman. Esse: Exploring mood on the web. In *Proceedings of International Conference on Weblogs and Social Media*, Seattle, WA, May 2009.

[28] P. D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 417–424, Philadelphia, July 2002.

[29] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2–3):165–210, 2005.

[30] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics.