

Experiments with Mood Classification in Blog Posts

Gilad Mishne
Informatics Institute, University of Amsterdam
Kruislaan 403, 1098SJ Amsterdam, The Netherlands
gilad@science.uva.nl

ABSTRACT

We present preliminary work on classifying blog text according to the mood reported by its author during the writing. Our data consists of a large collection of blog posts – online diary entries – which include an indication of the writer’s mood. We obtain modest, but consistent improvements over a baseline; our results show that further increasing the amount of available training data will lead to an additional increase in accuracy. Additionally, we show that the classification accuracy, although low, is not substantially worse than human performance on the same task. Our main finding is that mood classification is a challenging task using current text analysis methods.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing

Keywords

Subjective content, blogs, moods

1. INTRODUCTION

With the increase in the web’s accessibility to the masses in the last years, the profile of web content is changing. More and more web pages are authored by non-professionals; part of this “publishing revolution” is the phenomenon of *blogs* (short for web-logs) – personal, highly opinionated journals publicly available on the internet. The *blogspace* – the collective term used for the collection of all blogs – consists of millions of users who maintain an online diary, containing frequently-updated views and personal remarks about a range of issues.

The growth in the amount of blogs is accompanied by increasing interest from the research community. Ongoing research in this domain includes a large amount of work on social network analysis, but also content-related work, e.g., [5, 7]. Some of this work is intended to develop technologies for

organising textual information not just in terms of topical content, but also in terms of metadata, subjective meaning, or stylometric aspects. Information needs change with the type of available information; the increase in the amount of blogs and similar data drives users to access textual information in new ways – for example, analyzing consumers’ attitudes for marketing purposes [16].

In this paper, we address the task of classifying blog posts by mood. That is, given a blog post, we want to predict the most likely state of mind with which the post was written: whether the author was depressed, cheerful, bored, and so on. As in the vast majority of text classification tasks, we take a machine learning approach, identifying a set of features to be used for the learning process.

Mood classification is useful for various applications, such as assisting behavioral scientists and improving doctor-patient interaction [8]. In the particular case of blog posts (and other large amounts of subjective data), it can also enable new textual access approaches, e.g., filtering search results by mood, identifying communities, clustering, and so on. It is a particularly interesting task because it also offers a number of scientific challenges: first, the large variety in blog authors creates a myriad of different styles and definitions of moods; locating features that are consistent across authors is a complex task. Additionally, the short length of typical blog entries poses a challenge to classification methods relying on statistics from a large body of text (these are often used for text classification). Finally, the large amounts of data require a scalable, robust method.

The main research questions addressed in this paper are:

- In what way does mood classification in blogs differ from mood classification in other domains? Many types of features seem to be good indicators of mood – which ones are effective in blogs? How complex is the task to begin with?
- How much data is required for reliable training, and how many features are required for each instance of the data? It has been observed that, for NLP tasks, continuously increasing the training set size improves results consistently [1]; does this hold in our domain?

Put differently, the paper is largely exploratory in nature, taking a large collection of blog posts, broad sets of features, and varying the amount of training data exploited by the

machine learner, evaluating the effect on the classification accuracy.

The remainder of the paper is organized as follows. In Section 2 we survey existing work in affect analysis and related fields. In Section 3 we describe the collection of blog posts we use for our experiments. Section 4 follows with details regarding the features we used for the classification process, dividing them into sets of related features. Our experiments and results are reported in Section 5, and we conclude in Section 6.

2. RELATED WORK

Most work on text classification is focused on identifying the topic of the text, rather than detecting stylistic features [28]. However, stylometric research – in particular, research regarding emotion and mood analysis in text – is becoming more common recently, in part due to the availability of new sources of subjective information on the web. Read [23] describes a system for identifying affect in short fiction stories, using the statistical association level between words in the text and a set of keywords; the experiments show limited success rates, but indicate that the method is better than a naive baseline. We incorporate similar statistical association measures in our experiments as one of the feature sets used (see Section 4). Rubin *et al.* investigated discriminating terms for emotion detection in short texts [24]; the corpus used in this case is a small-scale collection of online reviews. Holzman and Pottenger report high accuracy in classifying emotions in online chat conversations by using the phonemes extracted from a voice-reconstruction of the conversations [9]; however, the corpus they use is small and may be biased. Liu *et al.* present an effective system for affect classification based on large-scale “common-sense” knowledge bases [17].

Two important points differentiating our work from existing work on affect analysis are the domain type and its size. As far as we are aware there is no published work on computational analysis of affect in blogs; this is an interesting and challenging domain due to its rising importance and accessibility in recent years, and the properties that make it different from other domains (e.g., highly personal, subjective writing style and the use of non-content features such as emoticons – see Section 4).

Closely related areas to mood classification are the fields of authorship attribution [19, 15] and gender classification [14], both of which are well-studied. Since these tasks are focused on identifying attributes that do not change over time and across different contexts, useful features typically employed are non-content features (such as the usage of stopwords or pronouns). In contrast, moods are dynamic and can change – for the same author – in a relatively short span. This causes both the features used for mood classification to be more content-based features, and the documents used for classification to be different: while authorship attribution and gender detection work well on long documents such as journal articles and even books, mood classification should be focused on short, time-limited documents.

Finally, a large body of work exists in the field of Sentiment Analysis. This field addresses the problem of iden-

tifying the semantic polarity (positive vs. negative orientation) of words and longer texts, and has been addressed both using corpus statistics [31, 6], linguistic tools such as WordNet [11], and “common-sense” knowledge bases [17]. Typically, methods for sentiment analysis produce lists of words with polarity values assigned to each of them. These values can later be aggregated for determining the orientation of longer texts, and have been successfully employed for applications such as product review analysis and opinion mining [3, 30, 21, 20, 2, 4].

3. A BLOG CORPUS

We now describe the collection of blog entries we used for our experiments.

We obtained a corpus of 815494 blog posts from Livejournal,¹ a free weblog service with a large community (several millions of users; considered the largest online blogging community). The web interface used by Livejournal, allowing users to update their blog, includes – in addition to the input fields for the post text and date – an optional field indicating the “current mood.” The user can either select a mood from a predefined list of 132 common moods such as “amused”, “angry” and so on, or enter free-text. If a mood is chosen while adding a blog entry, the phrase “current mood: X” will appear at the bottom of the entry, where X is the mood chosen by the user.

One obvious drawback of the mood “annotation” in this corpus is that it is not provided in a consistent manner; the blog writers differ greatly from each other, and their definitions of moods differ accordingly. What may seem to one person as a frustrated state of mind might appear to another as a different emotional state – anger, depression, and so on. Of course, this is also an advantage in a way, since unlike other corpora, in this case we have direct access to the writer’s opinion about her state of mind at the time of writing (rather than an external annotator).

The blog corpus was obtained as follows. First, for each one of the 132 common moods given by Livejournal as predefined moods, we used the Yahoo API [32] to get a list of 1000 web pages containing a Livejournal blog post with that mood. Since the Livejournal web pages contain multiple blog posts (up to 20), some of the web pages overlapped; in total, our list contained 122624 distinct pages, from 37009 different blogs. We proceeded to download the posts in these pages, getting in total the 815494 posts mentioned above – 22 posts per blog, on average. Of these posts, 624905 (77%) included an indication of the mood; we disregarded all other posts.

As expected, the distribution of different moods within the posts follows a power law. The number of unique moods in the corpus is 54487, but 46558 of them appear only once, and an additional 4279 appear only twice; such moods are inserted by users in the free-text field rather than chosen from the predefined list. Table 3 shows the distribution of the most popular moods in our corpus (percentages are calculated from the total number of posts with moods, rather than from the total number of posts altogether).

¹<http://www.livejournal.com>

| Mood | Occurrences | | Mood | Occurrences | | Mood | Occurrences | |
|--------------|-------------|--------|---------------|-------------|--------|------------|-------------|--------|
| amused | 24857 | (4.0%) | contemplative | 10724 | (1.7%) | anxious | 7052 | (1.1%) |
| tired | 20299 | (3.2%) | awake | 10121 | (1.6%) | exhausted | 6943 | (1.1%) |
| happy | 16471 | (2.6%) | calm | 10052 | (1.6%) | crazy | 6433 | (1.0%) |
| cheerful | 12979 | (2.1%) | bouncy | 10040 | (1.6%) | depressed | 6386 | (1.0%) |
| bored | 12757 | (2.0%) | chipper | 9538 | (1.5%) | curious | 6330 | (1.0%) |
| accomplished | 12200 | (1.9%) | annoyed | 8277 | (1.3%) | drained | 6260 | (1.0%) |
| sleepy | 11565 | (1.8%) | confused | 8160 | (1.3%) | sad | 6128 | (1.0%) |
| content | 11180 | (1.8%) | busy | 7956 | (1.3%) | aggravated | 5967 | (1.0%) |
| excited | 11099 | (1.8%) | sick | 7848 | (1.3%) | ecstatic | 5965 | (1.0%) |

Table 1: Frequently occurring moods in our corpus

To ensure a minimal amount of training data for each mood we attempt to classify, we use only posts for which the mood is one of the top 40 occurring moods in the entire corpus. This leaves us with 345014 posts, the total size of which is 366MB (after cleanup and markup removal). The number of words in the corpus is 69149217 (average of 200 words per post), while the unique number of words is 596638.

An additional point important to note about our corpus is that while it contains a large amount of different authors, it does not constitute a representative sample of adult writers. In fact, many of the blog maintainers are not even adults: according to Livejournal, the median age of blog authors is about 18, so half of the writers are actually teenagers.

4. FEATURE SET

When designing a classification experiment, the most important decision – more important than the choice of the learning algorithm itself – is the selection of features to be used for training the learner. In the case of text classification, several feature sets such as word counts are commonly used; in the blog domain, additional sets of features seem beneficial. In this section we list the features we used in our experiments, grouped by “feature family”.

First, we employ “classic” features in text analysis – features which are used in various types of classification tasks, both style-related and topic related.

Frequency Counts

Perhaps the most common set of features used for text classification tasks is information regarding the occurrence of words, or word n-grams, in the text. The absolute majority of text classification systems treat documents as simple “bag-of-words” and use the word counts as features [28]. Other measures commonly used as features in text classifiers are frequencies of Part-of-Speech (POS) tags in the text. In our experiments, we used both the word counts and the POS tag counts as features; an additional feature that we used was the frequencies of word lemmas. Both the POS tags and the lemmas were acquired with TreeTagger [27].

We have experimented both with single word/POS/lemma features and with higher-order n-grams; due to time and space constraints, in this paper we report only on unigram features.

Length-related

Four features are used to represent the length of a blog post: the total length in bytes, the number of words in the post, the average length of a sentence in bytes, and the average number of words in a sentence. A naive method was used for sentence splitting, taking standard punctuation marks as sentence delimiters.

Next, we make use of features that are related to the subjective nature of text in blogs – the fact that they tend to contain a larger amount of personal, opinionated text than other domains.

Semantic Orientation Features

Semantic orientation seems like a particularly useful feature for mood prediction: some moods are clearly “negative” (annoyed, frustrated) and some are clearly “positive” (cheerful, loved); it is anticipated that positive blogs posts will have, on average, a more positive orientation than negative ones.

In our experiments, we use both the total orientation of a blog post and the average word orientation in the blog as features. Since the estimation of word semantic orientation is highly dependent on the method used for calculating it, we use two different sources for the word-level orientation estimation.

The first source is a list of 21885 verbs and nouns, each assigned with either a positive, negative, or neutral orientation. The method used for creating this list is described by Kim and Hovy in [12]. In a nutshell, the method uses the WordNet distances of a word from a small set of manually-classified keywords. For calculating the total and average orientation of a post, we assign a value of +1 to every positive word and -1 to every negative one, summing (or averaging) the words.

The second source we use is a similar list of 1718 adjectives with their corresponding real-numbered polarity values, either positive or negative. This list was constructed using Turney and Littman’s method described in [30]; their method is based on measuring the co-occurrence of a word with a small set of manually-classified keywords on the web.

Examples of words with their values in both lists are given in Table 2, illustrating the occasional disagreement between the different sources.

| Word | Kim&Hovy | Turney&Littman |
|--------------|----------|----------------|
| pricey | Positive | -4.99 |
| repetitive | Positive | -1.63 |
| teenage | Negative | -1.45 |
| momentary | Negative | +0.01 |
| fair | Positive | +0.02 |
| earnest | Positive | +1.86 |
| unparalleled | Negative | +3.67 |
| fortunate | Positive | +5.72 |

Table 2: Semantic orientation values of words

Mood PMI-IR

The next set of features we use is based on Pointwise Mutual Information (PMI, [18]). PMI is a measure of the degree of association between two terms, and is defined as

$$PMI(t_1, t_2) = \log \frac{p(t_1 \& t_2)}{p(t_1)p(t_2)}$$

PMI-IR [29] uses Information Retrieval to estimate the probabilities needed for calculating the PMI using search engine hitcounts from a very large corpus, namely the web. The measure thus becomes

$$PMI-IR(t_1, t_2) = \log \frac{\text{hitcounts}(t_1 \& t_2)}{\text{hitcounts}(t_1) \cdot \text{hitcounts}(t_2)}$$

When estimating the total PMI of a text with a certain concept, it is common practice to sum the individual PMI values of all words in the text and the concept [30]. Since we are classifying text by mood, our “concepts” are all possible moods, and we would like to measure the association between words used in the blog entry and various moods. Thus, we pre-calculated the PMI-IR of the 2694 most frequently occurring words in the corpus with the top 40 occurring mood (for a total of $2694 \cdot 40 = 107760$ PMI-IR values). For the search engine hitcounts we used the Yahoo API; some example PMI-IR values are given in Table 3 (higher values depict greater association).

| Word | Mood | PMI-IR |
|-----------|---------|--------|
| nap | great | -15.51 |
| hugged | great | -25.61 |
| mirror | great | -40.23 |
| goodnight | sleepy | -22.88 |
| moving | sleepy | -26.58 |
| install | sleepy | -28.87 |
| homework | content | -29.24 |
| homework | annoyed | -26.04 |
| homework | bored | -25.52 |

Table 3: Example PMI-IR values of ⟨word,mood⟩ pairs

After calculating the PMI-IR values between the frequent words and the frequent moods, we used 80 additional features for each blog post: for each mood of the top 40 moods, we included two features representing the association of the post to that mood: the total PMI and the average PMI. The numerical values of the features are simply the sum of the

normalized PMI-IR values of words contained in the post (and included in the list of 2694 most frequent words for which PMI was pre-calculated), and the average of the PMI values. This approach is somewhat similar to the one used in [23].

Finally, we turn to features that are unique to online text such as blogs, as well as email and certain types of web pages.

Emphasized Words

Historically, written online text such as email was unformatted (that is, raw ASCII was used, without layout modifiers such as different font sizes, italic text and so on). This led to alternative methods of text emphasis, including using all-capitalized words (“I think that’s a GREAT idea”), and using asterisks or underscores attached to a word on both sides (“This is *not* what I had in mind”, “Did you bother .checking- it before sending??”).

While today most online text has extensive formatting options, usage of these emphasis methods is still popular, especially in cases where text is added through a standard text-box on a web page, containing no formatting options – the way many blog hosting services provide access to the blog maintainer.

We use as a feature the frequency of each emphasized word in a post, as well as the total number of stressed words per post. The intuition is that since these are words that the writer chose to emphasize, they may be important indicators of the written text.

Special Symbols

This set of features captures the usage of two types of special characters in the blog posts. The first type is punctuation characters such as ellipsis, exclamation marks, and so forth. The intuition behind modeling the frequencies of these symbols is that in some cases increased usage of them is beneficial for characterizing specific kinds of text [26]. Indeed, punctuation marks proved suitable in some text classification tasks, such as detecting email spam [25]. We use as features the frequencies of 15 common special symbols in each blog post; these include punctuation marks and some additional non-alphanumeric symbols such as asterisks and currency signs.

The second type of special symbols we use as feature are *emoticons* (emotional icons). Emoticons are sequences of printable characters which are intended to represent human emotions or attitudes; often, these are sideways textual representations of facial expressions. Examples of such emoticons are :) (representing a smile) and ;) (representing a wink) – both viewed sideways. Usage of emoticons originated in email messages and quickly spread to other forms of online content; it is currently very popular in many domains including blogs. Similarly to the first set of special symbols, we use the frequencies of 9 popular emoticons in the blog posts as features.

5. EXPERIMENTAL EVALUATION

In this section we describe the experiments we performed for classifying the mood of a blog post. We start with an overview of the classification environment, and follow with a description of the experiments performed and their results.

5.1 Classification

Setup

For our experiments, we use SVMlight, a support-vector machine package.² SVMs are popular in text classification tasks since they scale to the large amount of features often incurred in this domain [10]; also, they have been shown to significantly outperform other classifiers for this type of experiments [33].

Although SVMs can effectively handle large feature spaces, for efficiency reasons we chose to reduce the number of features related to word frequencies in the text. This is a common practice in text classification tasks, due to the large feature space; many text classifiers employ methods for Feature Selection – choosing only some of the features to actually be used in the learning process.

The intuition behind our feature selection method is that each mood has a set of words (and similarly, POS tags and lemmas) that are more characteristic of text associated with this mood than of other texts. We identify this set of features for each mood, then aggregate the separate sets to a combined feature set of all words which are characteristic of at least one mood. The identification of the characteristic set of features per mood is done using standard tests for comparing frequency distributions, where we compare the distribution of the words in texts associated with a mood with the distribution of the words in all other texts, and similarly for POS tags and word lemmas.

More formally, for each mood m we define two probability distributions, Θ_m and $\Theta_{\bar{m}}$, to be the distribution of all words in texts associated with m and in other texts, respectively.³ We rank all the words in Θ_m , according to their log likelihood measure [22], as compared with $\Theta_{\bar{m}}$. We then set as the “set of characteristic features for mood m ” the top N -ranked features. Once we have completed this process for all moods, we combine all the characteristic sets obtained to one feature set. In the experiments reported in this paper, we set N to 50.

Examples of characteristic word n-grams in our corpus for some moods are given in Table 4; characteristic POS and lemma features were calculated similarly.

Since the vocabulary of stressed words is substantially smaller than that of all words, we do not employ any mechanisms for reducing the amount of features, as we did with the frequencies of words.

Experiments

We performed two sets of experiments. The first set is intended to evaluate the effectiveness of identifying specific,

²<http://svmlight.joachims.org/>

³We describe the process for word features, but it is equivalent for POS tag and word lemma features.

| Mood | Top words | Top bigrams | Top trigrams |
|------------|--|---|--|
| hungry | hungry eat bread sauce | am hungry hungry and some food to eat | I am hungry is finally happened I am starving ask my mother |
| frustrated | n't frustrated frustrating do | am done can not problem is to fix | I am done am tired of stab stab stab I do not |
| loved | love me valentine her | I love love you love is valentines day | I love you my god oh i love him love you so |

Table 4: Most discriminating word n-grams for some moods

individual moods in a blog post, and to examine the effect of changes in the training set size on classification accuracy. For each mood we created a training set with randomly drawn instances from the set of posts associated with that mood as positive examples, and an equal amount of negative examples, randomly drawn from all other moods. The test set we used contained, similarly, an equal amount of random positive and negative instances, distinct from those used for training.

For the second set of experiments, we manually partitioned the moods into two “mood sets” according to some abstraction, such as “positive moods” vs. “negative moods”. We then repeated the training and testing phase as done for the individual mood classification, treating all moods in the same set as equivalent. The purpose of these experiments was to test whether combining closely-related moods improves performance, since many of the moods in our corpus are near-synonyms (e.g., “tired” and “sleepy”).

In the experiments reported in this paper, we use the entire list of features given above, rather than select subsets of it and experiment with them separately. This was done due to space constraints; our ongoing work includes evaluating the performance gain contributed by each feature subset.

For classifying individual moods, our training set size was limited to a maximum of a few thousand positive and a few thousand negative examples, since many moods did not have large amounts of associated blog posts (see Table 3). For classifying the mood sets, we used a larger amount of training material.

Since both our training and test sets contain the same number of positive and negative examples, the baseline to all our experiments is 50% accuracy (achieved by classifying all examples as positive or all examples as negative).

5.2 Results

Table 5 lists the results of the classification of individual moods. The test sets contained 400 instances; for the training sets we used varying amounts, up to 6400 instances; the table lists the results when training with 1600 instances and with 6400 instances. The results of the classification of two mood partitions – active/passive and positive/negative – are shown in Table 6.

| Mood | Correct | | Mood | Correct | |
|--|---------|--------|------------|---------|--------|
| | 1600 | 6400 | | 1600 | 6400 |
| confused | 56.00% | 65.75% | bored | 51.52% | 55.25% |
| curious | 60.25% | 63.25% | sleepy | 44.25% | 55.00% |
| depressed | 58.25% | 62.50% | crazy | 54.00% | 55.00% |
| happy | 54.50% | 60.75% | blank | 56.00% | 54.50% |
| amused | 57.75% | 60.75% | cheerful | 52.50% | 54.25% |
| sick | 54.75% | 60.25% | anxious | 51.75% | 54.25% |
| sad | 53.00% | 60.25% | aggravated | 52.75% | 54.25% |
| frustrated | 57.00% | 60.25% | content | 50.75% | 54.00% |
| excited | 55.50% | 59.75% | awake | 51.50% | 53.75% |
| ecstatic | 54.00% | 59.75% | busy | 50.75% | 53.50% |
| bouncy | 51.00% | 59.50% | cold | 50.25% | 53.25% |
| thoughtful | 52.75% | 59.00% | exhausted | 52.50% | 52.50% |
| annoyed | 57.00% | 59.00% | drained | 47.50% | 52.25% |
| loved | 57.00% | 57.75% | hungry | 51.50% | 50.75% |
| blah | 53.75% | 57.75% | good | 48.50% | 50.50% |
| hopeful | 51.50% | 57.50% | creative | 47.75% | 50.50% |
| cranky | 55.00% | 57.25% | okay | 46.75% | 49.00% |
| contemplative | 53.25% | 57.00% | calm | 44.75% | 49.00% |
| accomplished | 54.75% | 55.75% | | | |
| 400 test instances; 1600 and 6400 training instances | | | | | |

Table 5: Classification performance: individual moods

| Size of training set | Active/Passive | Positive/Negative |
|----------------------|----------------|-------------------|
| 800 | 50.51% | 48.03% |
| 1600 | 50.93% | 53.00% |
| 3200 | 51.50% | 51.72% |
| 6400 | 51.77% | 54.92% |
| 20000 | 53.53% | 54.65% |
| 40000 | 55.26% | 57.33% |
| 80000 | 57.08% | 59.67% |

Table 6: Classification performance: active vs. passive moods (size of test set: 65936) and positive vs. negative moods (size of test set: 55495)

5.3 Discussion

The classification performance on most moods is modest, with an average of 8% improvement over the 50% baseline (with 6400 training examples); a few moods exhibit substantially higher improvements, up to 15% improvement over the baseline, while a small number of moods are performing equivalently or worse than the baseline. Examining the better and worse performing moods, it seems that the better ones are slightly more concrete and focused than the worse ones, e.g., “depressed”, “happy” and “sick” compared to “okay” and “calm”. However, this is not consistent as some concrete moods show low accuracy (“hungry”) whereas some of the non-focused moods perform averagely (“blah”): the reasons for the different behavior on different moods need to be explored further.

Somewhat surprising, the classification of the aggregated sets does not seem to be an easier task than classifying a single mood, despite the substantial increase in the amount of training examples.

In general, it seems that the classification task is indeed a complex one, and that methods and features used for other stylistic analysis – even when augmented with a range of additional features – do not provide sufficient results. Disappointed by these results, we decided to let humans perform the individual mood classification task, and see if this yields substantially higher performance. For each one of the 40 most frequent moods, we randomly selected 10 posts annotated with that mood, and 10 posts annotated with a random other mood. We then presented these 20 posts to a human assessor without their accompanying moods; the assessor was told that exactly 10 out of the 20 posts are of mood X (the mood was explicitly given), and was asked to select which ones they are. This process simulates the same test data used with the machine learning experiments. The accuracy of the human over these 800 posts was 63%, and the assessor commented that in many of the cases, it seemed to him that much less than 10 posts were in fact related to the given mood, therefore driving him to choose randomly.

Some possible reasons for the low accuracy – both of the human and the machine – on this task are as follows.

- First, the subjective nature of the “annotation” in our corpus is problematic due to the large amount of widely-varying authors in it. Unlike lab-controlled experiments, where annotators follow guidelines and try to be consistent, our annotated corpus is fairly unstable.
- Additionally, the nature of the blog posts is problematic for text classification purposes. The average size of an entry is, as stated earlier, is 200 words; this is usually not enough to gather meaningful statistics, creating very sparse training data. Some posts hardly contain text at all – just a few links or pictures – and others yet are not even in English.

- Finally, the mood categories themselves – as defined by the Livejournal interface – are highly subjective; for any given mood there are lots of different situations that may bring this mood about, and correspondingly there could be many different types of blog entries labelled with the same mood.

One clear observation is the increasing the size of the training set affects favorably the performance in the vast majority of the cases, particularly for single-mood classification, and to a lesser extent also for mood-set classification. We believe this indicates that our results can still improve by simply further increasing the training data size.

6. CONCLUSIONS

We presented preliminary experiments in classifying moods of blog text, using as our corpus a large collection of blog posts containing the authors' indication of their state of mind at the time of writing. We use a variety of features for the classification process, including content and non-content features, and some features which are unique to online text such as blogs. Our results show a small, if consistent, improvement over a naive baseline; while the success rates are relatively low, human performance on this task is not substantially better.

Going back to our research questions, we witness that mood classification in blogs is a complex task—for humans as well as machines—and that the wealth of features available for the learning process does not ensure high performance. We do experience, however, a consistent improvement over a baseline for almost all given moods. Furthermore, our results indicate that increasing the amount of training data results in a clear improvement in effectiveness, and that our experiments did not reach a saturation point in the improvement – i.e., further improvement is expected with more training data.

In the future, we intend to thoroughly analyze which features are more beneficial for effective classification, and modify our feature set accordingly; preliminary investigations in this direction show that the mood PMIs are prominent features throughout all moods. In this context, we are examining the notion of “feature stability” [13] for identifying important features for style analysis in blogs. Additional directions we intend to explore are the relation between blog post length and the success in classifying it, and the reasons for the different performance of the classification process on different moods. Finally, to increase the level of “annotator agreement”—the consistency level regarding moods among bloggers—we intend to reduce the amount of different authors in the corpus, focusing on a relatively small amount of bloggers, with a large amount of posts each.

Acknowledgments

The author wishes to thank Maarten de Rijke for valuable comments and discussions, and Soo-Min Kim and Ed Hovy for providing their polarity-tagged lists. This work was supported by the Netherlands Organization for Scientific Research (NWO) under project number 220-80-001.

7. REFERENCES

- [1] M. Banko and E. Brill. Scaling to very very large corpora for natural language disambiguation. In *Proceedings ACL 2001*, pages 26–33, 2001.
- [2] S. Das and M. Chen. Yahoo! for Amazon: Sentiment Parsing from Small Talk on the Web. In *EFA 2001*, 2001.
- [3] K. Dave, S. Lawrence, and D. Pennock. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *WWW2003: the 13th international conference on World Wide Web*, 2003.
- [4] G. Grefenstette, Y. Qu, J. Shanahan, and D. Evans. Coupling niche browsers and affect analysis. In *RIAO'2004*, 2004.
- [5] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *WWW2004: the 13th international conference on World Wide Web*, pages 491–501, New York, NY, USA, 2004. ACM Press.
- [6] V. Hatzivassiloglou and K. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings EACL 1997*, 1997.
- [7] S. Herring, L. Scheidt, S. Bonus, and E. Wright. Bridging the gap: A genre analysis of weblogs. In *HICSS '04: Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 4*, page 40101.2, Washington, DC, USA, 2004. IEEE Computer Society.
- [8] J. Hodgson, C. Shields, and S. Rousseau. Disengaging communication in later-life couples coping with breast cancer. *Families, Systems, & Health*, (21):145–163, 2003.
- [9] L. Holzman and W. Pottenger. Classification of emotions in internet chat: An application of machine learning using speech phonemes. Technical Report LU-CSE-03-002, Lehigh University, 2003.
- [10] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *ECML '98: Proceedings of the 10th European Conference on Machine Learning*, pages 137–142, London, UK, 1998. Springer-Verlag.
- [11] J. Kamps, M. Marx, R. Mokken, and M. de Rijke. Using WordNet to measure semantic orientations of adjectives. In *Proceedings LREC 2004*, 2004.
- [12] S.-M. Kim and E. Hovy. Determining the Sentiment of Opinions. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, 2004.
- [13] M. Koppel, N. Akiva, and I. Dagan. A corpus-independent feature set for style-based text categorization. In *IJCAI'03 Workshop On Computational Approaches And Synthesis*, 2003.

- [14] M. Koppel, S. Argamon, and A. Shimoni. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412, 2002.
- [15] M. Koppel and J. Schler. Authorship verification as a one-class classification problem. In *ICML '04: Twenty-first international conference on Machine learning*, New York, NY, USA, 2004. ACM Press.
- [16] B. Liu, M. Hu, and J. Cheng. Opinion observer: analyzing and comparing opinions on the web. In *WWW2005: the 14th international conference on World Wide Web*, pages 342–351, New York, NY, USA, 2005. ACM Press.
- [17] H. Liu, H. Lieberman, and T. Selker. A model of textual affect sensing using real-world knowledge. In *IUI '03: Proceedings of the 8th international conference on Intelligent user interfaces*, pages 125–132, New York, NY, USA, 2003. ACM Press.
- [18] C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999.
- [19] A. McEnery and M. Oakes. *Handbook of Natural Language Processing*, volume 2, chapter Authorship Studies / Textual Statistics. Marcel Dekker, 2000.
- [20] T. Nasukawa and J. Yi. Sentiment analysis: capturing favorability using natural language processing. In *Proceedings K-CAP'03: the international conference on Knowledge capture*, 2003.
- [21] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings EMNLP 2002*, 2002.
- [22] P. Rayson and R. Garside. Comparing corpora using frequency profiling. In *The workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000)*, 2000.
- [23] J. Read. Recognising affect in text using pointwise-mutual information. Master's thesis, University of Sussex, 2004.
- [24] V. Rubin, J. Stanton, and E. Liddy. Discerning emotions in texts. In *The AAAI Symposium on Exploring Attitude and Affect in Text (AAAI-EAAT)*, 2004.
- [25] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A Bayesian Approach to Filtering Junk E-Mail. In *Learning for Text Categorization: Papers from the 1998 Workshop*, Madison, Wisconsin, 1998. AAAI Technical Report WS-98-05.
- [26] B. Say and V. Akman. Current Approaches to Punctuation in Computational Linguistics. *Computers and the Humanities*, 30(6):457–469, 1996.
- [27] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester, UK, 1994.
- [28] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.
- [29] P. Turney. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *EMCL '01: Proceedings of the 12th European Conference on Machine Learning*, pages 491–502, London, UK, 2001. Springer-Verlag.
- [30] P. Turney and M. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4), 2003.
- [31] J. Wiebe. Learning subjective adjectives from corpora. In *Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence*. AAAI Press / The MIT Press, 2000.
- [32] Yahoo Development Network, URL: <http://developer.yahoo.net>.
- [33] Y. Yang and X. Liu. A re-examination of text categorization methods. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49, New York, NY, USA, 1999. ACM Press.