Experimenting with Distant Supervision for Emotion Classification

Matthew Purver[*] and **Stuart Battersby**[†]
[*]Interaction Media and Communication Group [†]Chatterbox Analytics
School of Electronic Engineering and Computer Science
Queen Mary University of London
Mile End Road, London E1 4NS, UK
m.purver@qmul.ac.uk stuart@cbanalytics.co.uk

Abstract

We describe a set of experiments using automatically labelled data to train supervised classifiers for multi-class emotion detection in Twitter messages with no manual intervention. By cross-validating between models trained on different labellings for the same six basic emotion classes, and testing on manually labelled data, we conclude that the method is suitable for some emotions (happiness, sadness and anger) but less able to distinguish others; and that different labelling conventions are more suitable for some emotions than others.

1 Introduction

We present a set of experiments into classifying Twitter messages into the six basic emotion classes of (Ekman, 1972). The motivation behind this work is twofold: firstly, to investigate the possibility of detecting emotions of multiple classes (rather than purely positive or negative sentiment) in such short texts; and secondly, to investigate the use of *distant supervision* to quickly bootstrap large datasets and classifiers without the need for manual annotation.

Text classification according to *emotion* and *sentiment* is a well-established research area. In this and other areas of text analysis and classification, recent years have seen a rise in use of data from online sources and social media, as these provide very large, often freely available datasets (see e.g. (Eisenstein et al., 2010; Go et al., 2009; Pak and Paroubek, 2010) amongst many others). However, one of the challenges this poses is that of data annotation: given very large amounts of data, often consisting of very short texts, written in unconventional style and without accompanying metadata, audio/video signals or access to the author for disambiguation, how can we easily produce a gold-standard labelling for training and/or for evaluation and test? One possible solution that is becoming popular is crowd-sourcing the labelling task, as the easy access to very large numbers of annotators provided by tools such as Amazon's Mechanical Turk can help with the problem of dataset size; however, this has its own attendant problems of annotator reliability (see e.g. (Hsueh et al., 2009)), and cannot directly help with the inherent problem of ambiguity – using many annotators does not guarantee that they can understand or correctly assign the author's intended interpretation or emotional state.

In this paper, we investigate a different approach via *distant supervision* (see e.g. (Mintz et al., 2009)). By using conventional markers of emotional content within the texts themselves as a surrogate for explicit labels, we can quickly retrieve large subsets of (noisily) labelled data. This approach has the advantage of giving us direct access to the authors' own intended interpretation or emotional state, without relying on third-party annotators. Of course, the labels themselves may be noisy: ambiguous, vague or not having a direct correspondence with the desired classification. We therefore experiment with multiple such conventions with apparently similar meanings – here, emoticons (following (Read, 2005)) and Twitter hashtags – allowing us to examine the similarity of classifiers trained on independent labels but intended to detect the same underlying class. We also investigate the precision and correspondence of particular labels with the desired emotion classes by testing on a small set of man-

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 482–491,
Avignon, France, April 23 - 27 2012.

ually labelled data.

We show that the success of this approach depends on both the conventional markers chosen and the emotion classes themselves. Some emotions are both reliably marked by different conventions and distinguishable from other emotions; this seems particularly true for *happiness*, *sadness* and *anger*, indicating that this approach can provide not only the basic distinction required for sentiment analysis but some more finer-grained information. Others are either less distinguishable from short text messages, or less reliably marked.

2 Related Work

2.1 Emotion and Sentiment Classification

Much research in this area has concentrated on the related tasks of *subjectivity* classification (distinguishing objective from subjective texts – see e.g. (Wiebe and Riloff, 2005)); and *sentiment* classification (classifying subjective texts into those that convey positive, negative and neutral sentiment – see e.g. (Pang and Lee, 2008)). We are interested in *emotion* detection: classifying subjective texts according to a finer-grained classification of the emotions they convey, and thus providing richer and more informative data for social media analysis than simple positive/negative sentiment. In this study we confine ourselves to the six basic emotions identified by Ekman (1972) as being common across cultures; other finer-grained classifications are of course available.

2.1.1 Emotion Classification

The task of emotion classification is by nature a multi-class problem, and classification experiments have therefore achieved lower accuracies than seen in the binary problems of sentiment and subjectivity classification. Danisman and Alpkocak (2008) used vector space models for the same six-way emotion classification we examine here, and achieved F-measures around 32%; Seol et al. (2008) used neural networks for an 8-way classification (hope, love, thank, neutral, happy, sad, fear, anger) and achieved per-class accuracies of 45% to 65%. Chuang and Wu (2004) used supervised classifiers (SVMs) and manually defined keyword features over a seven-way classification consisting of the same six-class taxonomy plus a *neutral* category, and achieved an average accuracy of 65.5%, varying from 56% for *disgust* to 74% for *anger*. However, they achieved significant improvements using acoustic features available in their speech data, improving accuracies up to a maximum of 81.5%.

2.2 Conventions

As we are using text data, such intonational and prosodic cues are unavailable, as are the other rich sources of emotional cues we obtain from gesture, posture and facial expression in face-to-face communication. However, the prevalence of online text-based communication has led to the emergence of textual conventions understood by the users to perform some of the same functions as these acoustic and non-verbal cues. The most familiar of these is the use of emoticons, either Western-style (e.g. `:)`, `:-(` etc.) or Eastern-style (e.g. `(^_^)`, `(>_<)` etc.). Other conventions have emerged more recently for particular interfaces or domains; in Twitter data, one common convention is the use of *hashtags* to add or emphasise emotional content – see (1).

(1) a. Best day in ages! #Happy :)

b. Gets so #angry when tutors don't email back... Do you job idiots!

Linguistic and social research into the use of such conventions suggests that their function is generally to emphasise or strengthen the emotion or sentiment conveyed by a message, rather than to add emotional content which would not otherwise be present. Walther and D'Addario (2001) found that the contribution of emoticons towards the sentiment of a message was outweighed by the verbal content, although negative ones tended to shift interpretation towards the negative. Ip (2002) experimented with emoticons in instant messaging, with the results suggesting that emoticons do not add positivity or negativity but rather increase valence (making positive messages more positive and vice versa). Similarly Derks et al. (2008a; 2008b) found that emoticons are used in strengthening the intensity of a verbal message (although they serve other functions such as expressing humour), and hypothesized that they serve similar functions to actual non-verbal behavior; Provine et al. (2007) also found that emoticons are used to "punctuate" messages rather than replace lexical content, appearing in similar grammatical locations to verbal laughter and preserving phrase structure.

2.3 Distant Supervision

These findings suggest, of course, that emoticons and related conventional markers are likely to be useful features for sentiment and emotion classification. They also suggest, though, that they might be used as surrogates for manual emotion class labels: if their function is often to complement the verbal content available in messages, they should give us a way to automatically label messages according to emotional class, while leaving us with messages with enough verbal content to achieve reasonable classification.

This approach has been exploited in several ways in recent work; Tanaka et al. (2005) used Japanese-style emoticons as classification labels, and Go et al. (2009) and Pak and Paroubek (2010) used Western-style emoticons to label and classify Twitter messages according to positive and negative sentiment, using traditional supervised classification methods. The highest accuracies appear to have been achieved by Go et al. (2009), who used various combinations of features (unigrams, bigrams, part-of-speech tags) and classifiers (Naïve Bayes, maximum entropy, and SVMs), achieving their best accuracy of 83.0% with unigram and bigram features and a maximum entropy; using only unigrams with a SVM classifier achieved only slightly lower accuracy at 82.2%. Ansari (2010) then provides an initial investigation into applying the same methods to six-way emotion classification, treating each emotion independently as a binary classification problem and showing that accuracy varied with emotion class as well as with dataset size. The highest accuracies achieved were up to 81%, but these were on very small datasets (e.g. 81.0% accuracy on *fear*, but with only around 200 positive and negative data instances).

We view this approach as having several advantages; apart from the ease of data collection it allows by avoiding manual annotation, it gives us access to the author's own intended interpetations, as the markers are of course added by the authors themselves at time of writing. In some cases such as the examples of (1) above, the emotion conveyed may be clear to a third-party annotator; but in others it may not be clear at all without the marker – see (2):

(2) a. Still trying to recover from seeing the #bluewaffle on my TL #disgusted #sick

b. Leftover ToeJams with Kettle Salt and Vinegar chips. #stress #sadness #comfort #letsturnthisfrownupsidedown

3 Methodology

We used a collection of Twitter messages, all marked with emoticons or hashtags corresponding to one of Ekman (1972)'s six emotion classes. For emoticons, we used Ansari (2010)'s taxonomy, taken from the Yahoo messenger classification. For hashtags, we used emotion names themselves together with the main related adjective – both are used commonly on Twitter in slightly different ways as shown in (3); note that emotion names are often used as marked verbs as well as nouns. Details of the classes and markers are given in Table 1.

(3) a. Gets so #angry when tutors don't email back... Do you job idiots!

b. I'm going to say it, Paranormal Activity 2 scared me and I didn't sleep well last night because of it. #fear #demons

c. Girls that sleep w guys without even fully getting to know them #disgust me

Messages with multiple conventions (see (4)) were collected and used in the experiments, ensuring that the marker being used as a label in a particular experiment was not available as a feature in that experiment. Messages with no markers were not collected. While this prevents us from experimenting with the classification of neutral or objective messages, it would require manual annotation to distinguish these from emotion-carrying messages which are not marked. We assume that any implementation of the techniques we investigate here would be able to use a preliminary stage of subjectivity and/or sentiment detection to identify these messages, and leave this aside here.

(4) a. just because people are celebs they dont reply to your tweets! NOT FAIR #Angry :(I wish They would reply! #Please

Data was collected from Twitter's Streaming API service.[1] This provides a 1-2% random sample of all tweets with no constraints on language

[1] See `http://dev.twitter.com/docs/streaming-api`.

Table 1: Conventional markers used for emotion classes.

happy	:-) :) ;-) :D :P 8) 8-| <@o
sad	:-(:(;-(:-< :'(
anger	:-@ :@
fear	:| :-o :-O
surprise	:s :S
disgust	:$ +o(
happy	#happy #happiness
sad	#sad #sadness
anger	#angry #anger
fear	#scared #fear
surprise	#surprised #surprise
disgust	#disgusted #disgust

or location. These are collected in near real time and stored in a local database. An English language selection filter was applied; scripts collecting each conventional marker set were alternated throughout different times of day and days of the week to avoid any bias associated with e.g. weekends or mornings. The numbers of messages collected varied with the popularity of the markers themselves: for emoticons, we obtained a maximum of 837,849 (for happy) and a minimum of 10,539 for anger; for hashtags, a maximum of 10,219 for happy and a minimum of 536 for disgust. [2]

Classification in all experiments was using support vector machines (SVMs) (Vapnik, 1995) via the LIBSVM implementation of Chang and Lin (2001) with a linear kernel and unigram features. Unigram features included all words and hashtags (other than those used as labels in relevant experiments) after removal of URLs and Twitter usernames. Some improvement in performance might be available using more advanced features (e.g. n-grams), other classification methods (e.g. maximum entropy, as lexical features are unlikely to be independent) and/or feature weightings (e.g. the variant of TFIDF used for sentiment classification by Martineau (2009)). Here, our interest is more in the difference between the emotion and convention marker classes - we leave investigation of absolute performance for future work.

[2]One possible way to increase dataset sizes for the rarer markers might be to include synonyms in the hashtag names used; however, people's use and understanding of hashtags is not straightforwardly predictable from lexical form. Instead, we intend to run a longer-term data gathering exercise.

4 Experiments

Throughout, the markers (emoticons and/or hashtags) used as labels in any experiment were removed before feature extraction in that experiment – labels were not used as features.

4.1 Experiment 1: Emotion detection

To simulate the task of detecting emotion classes from a general stream of messages, we first built for each convention type C and each emotion class E a dataset D_E^C of size N containing (a) as positive instances, $N/2$ messages containing markers of the emotion class E and no other markers of type C, and (b) as negative instances, $N/2$ messages containing markers of type C of *any other* emotion class. For example, the positive instance set for emoticon-marked anger was based on those tweets which contained :-@ or :@, but none of the emoticons from the happy, sad, surprise, disgust or fear classes; any hashtags were allowed, including those associated with emotion classes. The negative instance set contained a representative sample of the same number of instances, with each having at least one of the happy, sad, surprise, disgust or fear emoticons but not containing :-@ or :@.

This of course excludes messages with no emotional markers; for this to act as an approximation of the general task therefore requires a assumption that unmarked messages reflect the same distribution over emotion classes as marked messages. For emotion-carrying but unmarked messages, this does seem intuitively likely, but requires investigation. For neutral objective messages it is clearly false, but as stated above we assume a preliminary stage of subjectivity detection in any practical application.

Performance was evaluated using 10-fold cross-validation. Results are shown as the **bold** figures in Table 2; despite the small dataset sizes in some cases, a χ^2 test shows all to be significantly different from chance. The best-performing classes show accuracies very similar to those achieved by Go et al. (2009) for their binary positive/negative classification, as might be expected; for emoticon markers, the best classes are happy, sad and anger; interestingly the best classes for hashtag markers are not the same

Table 2: Experiment 1: Within-class results. Same-convention (**bold**) figures are accuracies over 10-fold cross-validation; cross-convention (*italic*) figures are accuracies over full sets.

		Train	
Convention	Test	emoticon	hashtag
emoticon	happy	**79.8%**	*63.5%*
emoticon	sad	**79.9%**	*65.5%*
emoticon	anger	**80.1%**	*62.9%*
emoticon	fear	**76.2%**	*58.5%*
emoticon	surprise	**77.4%**	*48.2%*
emoticon	disgust	**75.2%**	*54.6%*
hashtag	happy	*67.7%*	**82.5%**
hashtag	sad	*67.1%*	**74.6%**
hashtag	anger	*62.8%*	**74.7%**
hashtag	fear	*60.6%*	**77.2%**
hashtag	surprise	*51.9%*	**67.4%**
hashtag	disgust	*64.6%*	**78.3%**

– `happy` performs best, but `disgust` and `fear` outperform `sad` and `anger`, and `surprise` performs particularly badly. For `sad`, one reason may be a dual meaning of the tag `#sad` (one emotional and one expressing ridicule); for `anger` one possibility is the popularity on Twitter of the game "Angry Birds"; for `surprise`, the data seems split between two rather distinct usages, ones expressing the author's emotion, but one expressing an intended effect on the audience (see (5)). However, deeper analysis is needed to establish the exact causes.

(5) a. broke 100 followers. #surprised im glad that the HOFF is one of them.

b. Who's excited for the Big Game? We know we are AND we have a #surprise for you!

To investigate whether the different convention types actually convey similar properties (and hence are used to mark similar messages) we then compared these accuracies to those obtained by training classifiers on the dataset for a different convention: in other words, for each emotion class E, train a classifier on dataset D_E^{C1} and test on D_E^{C2}. As the training and testing sets are different, we now test on the entire dataset rather than using cross-validation. Results are shown as the *italic* figures in Table 2; a χ^2 test shows all to be significantly different from the bold same-convention results. Accuracies are lower overall, but the highest figures (between 63% and 68%) are achieved for `happy`, `sad` and `anger`; here perhaps we can have some confidence that not only are the markers acting as predictable labels themselves, but also seem to be labelling the same thing (and therefore perhaps are actually labelling the emotion we are hoping to label).

4.2 Experiment 2: Emotion discrimination

To investigate whether these independent classifiers can be used in multi-class classification (distinguishing emotion classes from each other rather than just distinguishing one class from a general "other" set), we next cross-tested the classifiers between emotion classes: training models on one emotion and testing on the others – for each convention type C and each emotion class $E1$, train a classifier on dataset D_{E1}^C and test on D_{E2}^C, D_{E3}^C etc. The datasets in Experiment 1 had an uneven balance of emotion classes (including a high proportion of `happy` instances) which could bias results; for this experiment, therefore, we created datasets with an even balance of emotions among the negative instances. For each convention type C and each emotion class $E1$, we built a dataset D_{E1}^C of size N containing (a) as positive instances, $N/2$ messages containing markers of the emotion class $E1$ and no other markers of type C, and (b) as negative instances, $N/2$ messages consisting of $N/10$ messages containing only markers of class $E2$, $N/10$ messages containing only markers of class $E3$ etc. Results were then generated as in Experiment 1.

Within-class results are shown in Table 3 and are similar to those obtained in Experiment 1; again, differences between bold/italic results are statistically significant. Cross-class results are shown in Table 4. The `happy` class was well distinguished from other emotion classes for both convention types (i.e. cross-class classification accuracy is low compared to the within-class figures in italics and parentheses). The `sad` class also seems well distinguished when using hashtags as labels, although less so when using emoticons. However, other emotion classes show a surprisingly high cross-class performance in many cases – in other words, they are producing disappointingly similar classifiers.

This poor discrimination for negative emotion classes may be due to ambiguity or vagueness in the label, similarity of the verbal content associ-

Table 4: Experiment 2: Cross-class results. Same-class figures from 10-fold cross-validation are shown in *(italics)* for comparison; all other figures are accuracies over full sets.

		Train					
Convention	Test	happy	sad	anger	fear	surprise	disgust
emoticon	happy	*(78.1%)*	17.3%	39.6%	26.7%	28.3%	42.8%
emoticon	sad	16.5%	*(78.9%)*	59.1%	71.9%	69.9%	55.5%
emoticon	anger	29.8%	67.0%	*(79.7%)*	74.2%	76.4%	67.5%
emoticon	fear	27.0%	69.9%	64.4%	*(75.3%)*	74.0%	61.2%
emoticon	surprise	25.4%	69.9%	67.7%	76.3%	*(78.1%)*	66.4%
emoticon	disgust	42.2%	54.4%	61.1%	64.2%	64.1%	*(73.9%)*
hashtag	happy	*(81.1%)*	10.7%	45.3%	47.8%	52.7%	43.4%
hashtag	sad	13.8%	*(77.9%)*	47.7%	49.7%	46.5%	54.2%
hashtag	anger	44.6%	45.2%	*(74.3%)*	72.0%	63.0%	62.9%
hashtag	fear	45.0%	50.4%	68.6%	*(74.7%)*	63.9%	60.7%
hashtag	surprise	51.5%	45.7%	67.4%	70.7%	*(70.2%)*	64.2%
hashtag	disgust	40.4%	53.5%	74.7%	71.8%	70.8%	*(74.2%)*

Table 3: Experiment 2: Within-class results. Same-convention (**bold**) figures are accuracies over 10-fold cross-validation; cross-convention (*italic*) figures are accuracies over full sets.

		Train	
Convention	Test	emoticon	hashtag
emoticon	happy	**78.1%**	*61.2%*
emoticon	sad	**78.9%**	*60.2%*
emoticon	anger	**79.7%**	*63.7%*
emoticon	fear	**75.3%**	*55.9%*
emoticon	surprise	**78.1%**	*53.1%*
emoticon	disgust	**73.9%**	*51.5%*
hashtag	happy	*68.7%*	**81.1%**
hashtag	sad	*65.4%*	**77.9%**
hashtag	anger	*63.9%*	**74.3%**
hashtag	fear	*58.9%*	**74.7%**
hashtag	surprise	*51.8%*	**70.2%**
hashtag	disgust	*55.4%*	**74.2%**

ated with the emotions, or of genuine frequent co-presence of the emotions. Given the close lexical specification of emotions in hashtag labels, the latter reasons seem more likely; however, with emoticon labels, we suspect that the emoticons themselves are often used in ambiguous or vague ways.

As one way of investigating this directly, we tested classifiers across labelling conventions as well as across emotion classes, to determine whether the (lack of) cross-class discrimination holds across convention marker types. In the case of ambiguity or vagueness of emoticons, we would expect emoticon-trained models to fail to discriminate hashtag-labelled test sets, but hashtag-trained models to discriminate emoticon-labelled test sets well; if on the other hand the cause lies in the overlap of verbal content or the emotions themselves, the effect should be similar in either direction. This experiment also helps determine in more detail whether the labels used label similar underlying properties.

Table 5 shows the results. For the three classes `happy`, `sad` and perhaps `anger`, models trained using emoticon labels do a reasonable job of distinguishing classes in hashtag-labelled data, and vice versa. However, for the other classes, discrimination is worse. Emoticon-trained models appear to give (undesirably) higher performance across emotion classes in hashtag-labelled data (for the problematic non-`happy` classes). Hashtag-trained models perform around the random 50% level on emoticon-labelled data for those classes, even when tested on nominally the same emotion as they are trained on. For both label types, then, the lower within-class and higher cross-class performance with these negative classes (`fear`, `surprise`, `disgust`) suggests that these emotion classes are genuinely hard to tell apart (they are all negative emotions, and may use similar words), or are simply often expressed simultaneously. The higher performance of emoticon-trained classifiers compared to hashtag-trained classifiers, though, also suggests vagueness or ambiguity in emoticons: data labelled with emoticons nominally thought to be

Table 5: Experiment 2: Cross-class, cross-convention results (train on hashtags, test on emoticons and vice versa). All figures are accuracies over full sets. Accuracies over 60% are shown in **bold**.

		Train					
Convention	Test	happy	sad	anger	fear	surprise	disgust
emoticon	happy	**61.2%**	40.4%	44.1%	47.4%	52.0%	45.9%
emoticon	sad	38.3%	**60.2%**	55.1%	51.5%	47.1%	53.9%
emoticon	anger	47.0%	48.0%	**63.7%**	56.2%	50.9%	56.6%
emoticon	fear	39.8%	57.7%	57.1%	55.9%	50.8%	56.1%
emoticon	surprise	43.7%	55.2%	59.2%	58.4%	53.1%	54.0%
emoticon	disgust	51.5%	48.0%	53.5%	55.1%	53.1%	51.5%
hashtag	happy	**68.7%**	32.5%	43.6%	32.1%	35.4%	50.4%
hashtag	sad	33.8%	**65.4%**	53.2%	**65.0%**	**61.8%**	48.8%
hashtag	anger	43.9%	55.5%	**63.9%**	59.6%	**60.4%**	53.0%
hashtag	fear	44.3%	54.6%	56.1%	58.9%	**61.5%**	54.3%
hashtag	surprise	54.2%	45.3%	49.8%	49.9%	51.8%	52.3%
hashtag	disgust	41.5%	57.6%	**61.6%**	**62.2%**	59.3%	55.4%

associated with `surprise` produces classifiers which perform well on data labelled with many other hashtag classes, suggesting that those emotions were present in the training data. Conversely, the more specific hashtag labels produce classifiers which perform poorly on data labelled with emoticons and which thus contains a range of actual emotions.

4.3 Experiment 3: Manual labelling

To confirm whether either (or both) set of automatic (distant) labels do in fact label the underlying emotion class intended, we used human annotators via Amazon's Mechanical Turk to label a set of 1,000 instances. These instances were all labelled with emoticons (we did not use hashtag-labelled data: as hashtags are so lexically close to the names of the emotion classes being labelled, their presence may influence labellers unduly)[3] and were evenly distributed across the 6 classes, in so far as indicated by the emoticons. Labellers were asked to choose the primary emotion class (from the fixed set of six) associated with the message; they were also allowed to specify if any other classes were also present. Each data instance was labelled by three different annotators.

Agreement between labellers was poor overall. The three annotators unanimously agreed in only 47% of cases overall; although two of three agreed in 83% of cases. Agreement was worst for the three classes already seen to be problematic: `surprise`, `fear` and `disgust`. To create our dataset for this experiment, we therefore took only instances which were given the same primary label by all labellers – i.e. only those examples which we could take as reliably and unambiguously labelled. This gave an unbalanced dataset, with numbers varying from 266 instances for `happy` to only 12 instances for each of `surprise` and `fear`. Classifiers were trained using the datasets from Experiment 2. Performance is shown in Table 6; given the imbalance between class numbers in the test dataset, evaluation is given as recall, precision and F-score for the class in question rather than a simple accuracy figure (which is biased by the high proportion of `happy` examples).

Table 6: Experiment 3: Results on manual labels.

Train	Class	Precision	Recall	F-score
emoticon	happy	79.4%	75.6%	**77.5%**
emoticon	sad	43.5%	73.2%	**54.5%**
emoticon	anger	62.2%	37.3%	**46.7%**
emoticon	fear	6.8%	63.6%	12.3%
emoticon	surprise	15.0%	90.0%	25.7%
emoticon	disgust	8.3%	25.0%	12.5%
hashtag	happy	78.9%	51.9%	**62.6%**
hashtag	sad	47.9%	81.7%	**60.4%**
hashtag	anger	58.2%	76.0%	**65.9%**
hashtag	fear	10.1%	81.8%	18.0%
hashtag	surprise	5.9%	60.0%	10.7%
hashtag	disgust	6.7%	66.7%	11.8%

[3] Although, of course, one may argue that they do the same for their intended audience of readers – in which case, such an effect is legitimate.

Again, results for happy are good, and correspond fairly closely to the levels of accuracy reported by Go et al. (2009) and others for the binary positive/negative sentiment detection task. Emoticons give significantly better performance than hashtags here. Results for sad and anger are reasonable, and provide a baseline for further experiments with more advanced features and classification methods once more manually annotated data is available for these classes. In contrast, hashtags give much better performance with these classes than the (perhaps vague or ambiguous) emoticons.

The remaining emotion classes, however, show poor performance for both labelling conventions. The observed low precision and high recall can be adjusted using classifier parameters, but F-scores are not improved. Note that Experiment 1 shows that both emoticon and hashtag labels are to some extent predictable, even for these classes; however, Experiment 2 shows that they may not be reliably different to each other, and Experiment 3 tells us that they do not appear to coincide well with human annotator judgements of emotions. More reliable labels may therefore be required; although we do note that the low reliability of the human annotations for these classes, and the correspondingly small amount of annotated data used in this evaluation, means we hesitate to draw strong conclusions about fear, surprise and disgust. An approach which considers multiple classes to be associated with individual messages may also be beneficial: using majority-decision labels rather than unanimous labels improves F-scores for surprise to 23-35% by including many examples also labelled as happy (although this gives no improvements for other classes).

5 Survey

To further detemine whether emoticons used as emotion class labels are ambiguous or vague in meaning, we set up a web survey to examine whether people could reliably classify these emoticons.

5.1 Method

Our survey asked people to match up which of the six emotion classes (selected from a drop-down menu) best matched each emoticon. Each drop-down menu included a 'Not Sure' option. To avoid any effect of ordering, the order of the emoticon list and each drop-down menu was randomised every time the survey page was loaded. The survey was distributed via Twitter, Facebook and academic mailing lists. Respondents were not given the opportunity to give their own definitions or to provide finer-grained classifications, as we wanted to establish purely whether they would reliably associate labels with the six emotions in our taxonomy.

5.2 Results

The survey was completed by 492 individuals; full results are shown in Table 7. It demonstrated agreement with the predefined emoticons for sad and most of the emoticons for happy (people were unsure what 8-| and <@o meant). For all the emoticons listed as anger, surprise and disgust, the survey showed that people are reliably *un*sure as to what these mean. For the emoticon :-o there was a direct contrast between the defined meaning and the survey meaning; the definition of this emoticon following Ansari (2010) was fear, but the survey reliably assigned this to surprise.

Given the small scale of the survey, we hesitate to draw strong conclusions about the emoticon meanings themselves (in fact, recent conversations with schoolchildren – see below – have indicated very different interpretations from these adult survey respondents). However, we do conclude that for most emotions outside happy and sad, emoticons may indeed be an unreliable label; as hashtags also appear more reliable in the classification experiments, we expect these to be a more promising approach for fine-grained emotion discrimination in future.

6 Conclusions

The approach shows reasonable performance at individual emotion label prediction, for both emoticons and hashtags. For some emotions (happiness, sadness and anger), performance across label conventions (training on one, and testing on the other) is encouraging; for these classes, performance on those manually labelled examples where annotators agree is also reasonable. This gives us confidence not only that the approach produces reliable classifiers which can predict the labels, but that these classifiers are actually detecting the desired underlying emotional classes,

Table 7: Survey results showing the defined emotion, the most popular emotion from the survey, the percentage of votes this emotion received, and the χ^2 significance test for the distribution of votes. These are indexed by emoticon.

Emoticon	Defined Emotion	Survey Emotion	% of votes	Significance of votes distribution
:-)	Happy	Happy	94.9	$\chi^2 = 3051.7$ ($p < 0.001$)
:)	Happy	Happy	95.5	$\chi^2 = 3098.2$ ($p < 0.001$)
;-)	Happy	Happy	87.4	$\chi^2 = 2541$ ($p < 0.001$)
:D	Happy	Happy	85.7	$\chi^2 = 2427.2$ ($p < 0.001$)
:P	Happy	Happy	59.1	$\chi^2 = 1225.4$ ($p < 0.001$)
8)	Happy	Happy	61.9	$\chi^2 = 1297.4$ ($p < 0.001$)
8-|	Happy	Not Sure	52.2	$\chi^2 = 748.6$ ($p < 0.001$)
<@o	Happy	Not Sure	84.6	$\chi^2 = 2335.1$ ($p < 0.001$)
:-(	Sad	Sad	91.3	$\chi^2 = 2784.2$ ($p < 0.001$)
:(	Sad	Sad	89.0	$\chi^2 = 2632.1$ ($p < 0.001$)
;-(	Sad	Sad	67.9	$\chi^2 = 1504.9$ ($p < 0.001$)
:-<	Sad	Sad	56.1	$\chi^2 = 972.59$ ($p < 0.001$)
:'(	Sad	Sad	80.7	$\chi^2 = 2116$ ($p < 0.001$)
:-@	Anger	Not Sure	47.8	$\chi^2 = 642.47$ ($p < 0.001$)
:@	Anger	Not Sure	50.4	$\chi^2 = 691.6$ ($p < 0.001$)
:s	Surprise	Not Sure	52.2	$\chi^2 = 757.7$ ($p < 0.001$)
:$	Disgust	Not Sure	62.8	$\chi^2 = 1136$ ($p < 0.001$)
+o(	Disgust	Not Sure	64.2	$\chi^2 = 1298.1$ ($p < 0.001$)
:|	Fear	Not Sure	55.1	$\chi^2 = 803.41$ ($p < 0.001$)
:-o	Fear	Surprise	89.2	$\chi^2 = 2647.8$ ($p < 0.001$)

without requiring manual annotation. We therefore plan to pursue this approach with a view to improving performance by investigating training with combined mixed-convention datasets, and cross-training between classifiers trained on separate conventions.

However, this cross-convention performance is much better for some emotions (happiness, sadness and anger) than others (fear, surprise and disgust). Indications are that the poor performance on these latter emotion classes is to a large degree an effect of ambiguity or vagueness of the emoticon and hashtag conventions we have used as labels here; we therefore intend to investigate other conventions with more specific and/or less ambiguous meanings, and the combination of multiple conventions to provide more accurately/specifically labelled data. Another possibility might be to investigate approaches to analyse emoticons semantically on the basis of their shape, or use features of such an analysis – see (Ptaszynski et al., 2010; Radulovic and Milikic, 2009) for some interesting recent work in this direction.

Acknowledgements

The authors are supported in part by the Engineering and Physical Sciences Research Council (grants EP/J010383/1 and EP/J501360/1) and the Technology Strategy Board (R&D grant 700081). We thank the reviewers for their comments.

References

Saad Ansari. 2010. Automatic emotion tone detection in twitter. Master's thesis, Queen Mary University of London.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for Support Vector Machines*. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Ze-Jing Chuang and Chung-Hsien Wu. 2004. Multimodal emotion recognition from speech and text. *Computational Linguistics and Chinese Language Processing*, 9(2):45–62, August.

Taner Danisman and Adil Alpkocak. 2008. Feeler: Emotion classification of text using vector space model. In *AISB 2008 Convention, Communication, Interaction and Social Intelligence*, volume 2, pages 53–59, Aberdeen.

Daantje Derks, Arjan Bos, and Jasper von Grumbkow. 2008a. Emoticons and online message interpretation. *Social Science Computer Review*, 26(3):379–388.

Daantje Derks, Arjan Bos, and Jasper von Grumbkow. 2008b. Emoticons in computer-mediated communication: Social motives and social context. *CyberPsychology & Behavior*, 11(1):99–101, February.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, Cambridge, MA, October. Association for Computational Linguistics.

Paul Ekman. 1972. Universals and cultural differences in facial expressions of emotion. In J. Cole, editor, *Nebraska Symposium on Motivation 1971*, volume 19. University of Nebraska Press.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Master's thesis, Stanford University.

Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. 2009. Data quality from crowdsourcing: A study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 27–35, Boulder, Colorado, June. Association for Computational Linguistics.

Amy Ip. 2002. The impact of emoticons on affect interpretation in instant messaging. Carnegie Mellon University.

Justin Martineau. 2009. Delta TFIDF: An improved feature space for sentiment analysis. *Artificial Intelligence*, 29:258–261.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL-IJCNLP 2009*.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the 7th conference on International Language Resources and Evaluation*.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Robert Provine, Robert Spencer, and Darcy Mandell. 2007. Emotional expression online: Emoticons punctuate website text messages. *Journal of Language and Social Psychology*, 26(3):299–307.

M. Ptaszynski, J. Maciejewski, P. Dybala, R. Rzepka, and K Araki. 2010. CAO: A fully automatic emoticon analysis system based on theory of kinesics. In *Proceedings of The 24th AAAI Conference on Artificial Intelligence (AAAI-10)*, pages 1026–1032, Atlanta, GA.

F. Radulovic and N. Milikic. 2009. Smiley ontology. In *Proceedings of The 1st International Workshop On Social Networks Interoperability*.

Jonathon Read. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the 43rd Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Young-Soo Seol, Dong-Joo Kim, and Han-Woo Kim. 2008. Emotion recognition from text using knowledge based ANN. In *Proceedings of ITC-CSCC*.

Y. Tanaka, H. Takamura, and M. Okumura. 2005. Extraction and classification of facemarks with kernel methods. In *Proceedings of IUI*.

Vladimir N. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer.

Joseph Walther and Kyle D'Addario. 2001. The impacts of emoticons on message interpretation in computer-mediated communication. *Social Science Computer Review*, 19(3):324–347.

J. Wiebe and E. Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts . In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-05)*, volume 3406 of *Springer LNCS*. Springer-Verlag.