Hierarchical versus Flat Classification of Emotions in Text

Diman Ghazi [(a)], Diana Inkpen [(a)], Stan Szpakowicz [(a, b)]

[(a)] School of Information Technology and Engineering, University of Ottawa
[(b)] Institute of Computer Science, Polish Academy of Sciences
{dghaz038,diana,szpak}@site.uottawa.ca

Abstract

We explore the task of automatic classification of texts by the emotions expressed. Our novel method arranges neutrality, polarity and emotions hierarchically. We test the method on two datasets and show that it outperforms the corresponding "flat" approach, which does not take into account the hierarchical information. The highly imbalanced structure of most of the datasets in this area, particularly the two datasets with which we worked, has a dramatic effect on the performance of classification. The hierarchical approach helps alleviate the effect.

1 Introduction

Computational approaches to emotion analysis have focused on various emotion modalities, but there was only limited effort in the direction of automatic recognition of emotion in text (Aman, 2007).

Oleveres et al.(1998), as one of the first works in emotion detection in text, uses a simple Natural Language Parser for keyword spotting, phrase length measurement and emoticon identification.
They apply a rule-based expert system to construct emotion scores based on the parsed text and contextual information. However their simple word-level analysis system is not sufficient when the emotion is expressed by more complicated phrases and sentences.
More advanced systems for textual emotion recognition performed sentence-level analysis. Liu et al. (2003), proposed an approach aimed at understanding the underlying semantics of language using large-scale real-world commonsense knowledge to classify sentences into "basic" emotion categories. They developed a commonsense affect model enabling the analysis of the affective qualities of text in a robust way.
In SemEval 2007, one of the tasks was carried out in an unsupervised setting and the emphasis was on the study of emotion in lexical semantics (Strapparava and Mihalcea, 2008; Chaumartin, 2007; Kozareva *et al*., 2007; Katz *et al*., 2007). Neviarouskaya et al.(2009) applied a rule-based approach to affect recognition from a blog text. However, statistical and machine learning approaches have became a method of choice for constructing a wide variety of NLP applications (Wiebe et al., 2005).

There has been previous work using statistical methods and supervised machine learning, including (Aman, 2007; Katz *et al*., 2007; Alm, 2008; Wilson et al., 2009). Most of that research concentrated on feature selections and applying lexical semantics rather than on different learning schemes. In particular, only *flat* classification has been considered.

According to Kiritchenko *et al*. (2006), "Hierarchical categorization deals with categorization problems where categories are organized in hierarchies". Hierarchical text categorization places new items into a collection with a predefined hierarchical structure. The categories are partially ordered, usually from more generic to more specific.
Koller and Sahami (1997) carried out the first proper study of a hierarchical text categorization problem in 1997. More work in hierarchical text categorization has been reported later. Keshtkar and Inkpen (2009) applied a hierarchical approach to mood classification: classifying blog posts into 132 moods. The connection with our work is only indirect, because – even though moods and emotions may seem similar – their hierarchy structure and the classification task are quite different. The work reported in (Kiritchenko *et al*., 2006) is more general. It explores two main aspects of hierarchic-

Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pages 140–146,
Los Angeles, California, June 2010.

al text categorization: learning algorithms and performance evaluation.

In this paper, we extend our preliminary work (Ghazi *et al.*, 2010) on hierarchical classification. Hierarchical classification is a new approach to emotional analysis, which considers the relation between neutrality, polarity and emotion of a text. The main idea is to arrange these categories and their interconnections into a hierarchy and leverage it in the classification process.

We categorize sentences into six basic emotion classes; there also may, naturally, be no emotion in a sentence. The emotions are *happiness*, *sadness*, *fear*, *anger*, *disgust*, and *surprise* (Ekman, 1992). In one of the datasets we applied, we did consider the class *non-emotional*.

For these categories, we have considered two forms of hierarchy for classification, with two or three levels. In the two-level method, we explore the effect of neutral instances on one dataset and the effect of polarity on the other dataset. In the three-level hierarchy, we consider neutrality and polarity together.

Our experiments on data annotated with emotions show performance which exceeds that of the corresponding flat approach.

Section 2 of this paper gives an overview of the datasets and feature sets. Section 3 describes both hierarchical classification methods and their evaluation with respect to flat classification results. Section 4 discusses future work and presents a few conclusions.

2 Data and Feature Sets

2.1 Datasets

The statistical methods typically require training and test corpora, manually annotated with respect to each language-processing task to be learned (Wiebe *et al.*, 2005). One of the datasets in our experiments is a corpus of blog sentences annotated with Ekman's emotion labels (Aman, 2007). The second dataset is a sentence-annotated corpus resource divided into three parts for large-scale exploration of affect in children's stories (Alm, 2008).

In the first dataset, each sentence is tagged by a dominant emotion in the sentence, or labelled as non-emotional. The dataset contains 173 weblog posts annotated by two judges. Table 1 shows the details of the dataset.

In the second dataset, two annotators have annotated 176 stories. The affects considered are the same as Ekman's six emotions, except that the *surprise* class is subdivided into *positive surprise* and *negative surprise*. We run our experiments on only sentences with high agreement- sentences with the same affective labels annotated by both annotators. That is the version of the dataset which merged *angry* and *disgusted* instances and combined the positive and negative *surprise* classes. The resulting dataset, therefore, has only five classes (Alm, 2008). Table 1 presents more details about the datasets, including the range of frequencies for the class distribution (Min is the proportion of sentences with the most infrequent class, Max is the proportion for sentences with the most frequent class.) The proportion of the most frequent class also gives us a baseline for the accuracies of our classifiers (since the poorest baseline classifier could always choose the most frequent class).

Table 1. Datasets specifications.

	Domain	*Size*	*# classes*	*Min-Max%*
Aman's Data set	Weblogs	2090	7	6-38 %
Alm's Data set	Stories	1207	5	9-36%

2.2 Feature sets

In (Ghazi *et al.*, 2010), three sets of features – one corpus-based and two lexically-based – are compared on Aman's datasets. The first experiment is a corpus-based classification which uses unigrams (bag-of-words). In the second experiment, classification was based on features derived from the *Prior-Polarity* lexicon[1] (Wilson *et al.* 2009); the features were the tokens common between the prior-polarity lexicon and the chosen dataset. In the last experiment, we used a combination of the emotional lists of words from *Roget's Thesaurus*[2] (Aman and Szpakowicz, 2008) and *WordNet Affect*[3] (Strapparava and Valitutti, 2004); we call it the *polarity feature set*.

[1] www.cs.pitt.edu/mpqa
[2] The 1987 Penguin's Roget's Thesaurus was used.
[3] www.cse.unt.edu/~rada/affective text/data/WordNetAffectEmotioLists.tar.gz

Based on the results and the discussion in (Ghazi *et al.,* 2010), we decided to use the polarity feature set in our experiments. This feature set has certain advantages. It is quite a bit smaller than the unigram features, and we have observed that they appear to be more meaningful. For example, the unigram features include (inevitably non-emotional) names of people and countries. It is also possible to have misspelled tokens in unigrams, while the prior-polarity lexicon features are well-defined words usually considered as polar. Besides, lexical features are known to be more domain- and corpus-independent. Last but not least, our chosen feature set significantly outperforms the third set.

2.3 Classification

As a classification algorithm, we use the support vector machines (SVM) algorithm with tenfold cross-validation as a testing option. It is shown that SVM obtains good performance in text classification: it scales well to the large numbers of features (Kennedy and Inkpen, 2006; Aman, 2007).

We apply the same settings at each level of the hierarchy for our hierarchical approach classification.

In hierarchical categorization, categories are organized into levels (Kiritchenko *et al*., 2006). We use the hierarchical categories to put more knowledge into our classification method as the category hierarchies are carefully composed manually to represent our knowledge of the subject. We will achieve that in two forms of hierarchy. A two-level hierarchy represents the relation of emotion and neutrality in text, as well as the relation of positive and negative polarity. These two relations are examined in two different experiments, each on a separate dataset.

A three-level hierarchy is concerned with the relation between polarity and emotions along with the relation between neutrality and emotion. We assume that, of Ekman's six emotions, *happiness* belongs to the positive polarity class, while the other five emotions have negative polarity. This is quite similar to the three-level hierarchy of affect labels used by Alm (2008). In her diagram, she considers happiness and positive surprise as positive, and the rest as negative emotions. She has not, however, used this model in the classification approach: classification experiments were only run at three separate affect levels. She also considers positive and negative surprise as one *Surprise* class.

For each level of our proposed hierarchy, we run two sets of experiments. In the first set, we assume that all the instances are correctly classified at the preceding levels, so we only need to be concerned with local mistakes. Because we do not have to deal with instances misclassified at the previous level, we call these results *reference results*.

In the second set of experiments, the methodology is different than in (Ghazi *et al*. 2010). In that work both training and testing of subsequent levels is based on the results of preceding levels. A question arises, however: once we have good data available, why train on incorrect data which result from mistakes at the preceding level? That is why we decided to train on correctly-labelled data and when testing, to compute global results by cumulating the mistakes from all the levels of the hierarchical classification. In other words, classification mistakes at one level of the hierarchy carry on as mistakes at the next levels. Therefore, we talk of *global results* because we compute the accuracy, precision, recall and F-measure globally, based on the results at all levels. These results characterize the hierarchical classification approach when testing on new sentences: the classifiers are applied in a pipeline order: level 1, then level 2 on the results of the previous level (then level 3 if we are in the three-level setting).

In the next section, we show the experiments and results on our chosen datasets.

3 Results and discussions

3.1 Two-level classification

This section has two parts. The main goal of the first part is to find out how the presence of neutral instances affects the performance of features for distinguishing between emotional classes in Aman's dataset. This was motivated by a similar work in polarity classification (Wilson *et al*., 2009).

In the second part, we discuss the effect of considering positive and negative polarity of emotions for five affect classes in Alm's dataset.

Table 2. Two-level emotional classification on Aman's dataset (the highest precision, recall, and F-measure values for each class are shown in bold). The results of the flat classification are repeated for convenience.

		Two-level classification			Flat classification		
		Precision	Recall	F-measure	Precision	Recall	F-measure
1st level	*emo*	0.88	0.85	0.86	--	--	--
	non-emo	**0.88**	0.81	**0.84**	0.54	**0.87**	0.67
2nd level reference results	*happiness*	0.59	**0.95**	**0.71**	**0.74**	0.60	0.66
	sadness	**0.77**	**0.49**	**0.60**	0.69	0.42	0.52
	fear	**0.91**	**0.49**	**0.63**	0.82	**0.49**	0.62
	surprise	**0.75**	**0.32**	**0.45**	0.64	0.27	0.38
	disgust	0.66	**0.35**	**0.45**	**0.68**	0.31	0.43
	anger	**0.72**	**0.33**	**0.46**	0.67	0.26	0.38
Accuracy			**68.32%**			61.67%	
2-level experiment global results	*non-emo*	**0.88**	0.81	**0.84**	0.54	**0.87**	0.67
	happiness	0.56	**0.86**	**0.68**	**0.74**	0.60	0.66
	sadness	0.64	**0.42**	0.51	**0.69**	**0.42**	**0.52**
	fear	0.75	0.43	0.55	**0.82**	**0.49**	**0.62**
	surprise	0.56	**0.29**	**0.38**	**0.64**	0.27	**0.38**
	disgust	0.52	0.29	0.37	**0.68**	**0.31**	**0.43**
	anger	0.55	**0.27**	0.36	**0.67**	0.26	**0.38**
Accuracy			**65.50%**			61.67%	

3.1.1 Neutral-Emotional

At the first level, emotional versus non-emotional classification tries to determine whether an instance is neutral or emotional. The second step takes all instances which level 1 classified as emotional, and tries to classify them into one of Ekman's six emotions. Table 2 presents the result of experiments and, for comparison, the flat classification results. A comparison of the results in both experiments with flat classification shows that in both cases the accuracy of two-level approach is significantly better than the accuracy of flat classification.

One of the results worth discussing further is the precision of the non-emotional class: it increases while recall decreases. We will see the same pattern in further experiments. This happens to the classes which used to dominate in flat classification but they no longer dominate in hierarchical classification. Classifiers tends to give priority to a dominant class, so more instances are placed in this class; thus, classification achieves low precision and high recall. Hierarchical methods tend to produce higher precision.

The difference between precision and recall of the *happiness* class in the flat approach and the two-level approach cannot be ignored. It can be explained as follows: at the second level there are no more non-emotional instances, so the happiness class dominates, with 42% of all the instances. As explained before, this gives high recall and low precision for the happiness class. We hope to address this big gap between precision and recall of the happiness class in the next experiments, three-level classification. It separates happiness from the other five emotions, so it makes the number of instances of each level more balanced.

Our main focus is comparing hierarchical and flat classification, assuming all the other parameters are fixed. We mention, however, the best previous results achieved by Aman (2007) on the same dataset. Her best result was obtained by combining corpus-based unigrams, features derived from emotional lists of words from *Roget's Thesaurus* (explained in 2.2) and common words between the dataset and *WordNetAffect.* She also applied SVM with tenfold cross validation. The results appear in Table 3.

Table 3. Aman's best results on her data set.

	Precision	Recall	F-Measure
happiness	0.813	0.698	0.751
sadness	0.605	0.416	0.493
fear	0.868	0.513	0.645
surprise	0.723	0.409	0.522
disgust	0.672	0.488	0.566
anger	0.650	0.436	0.522
non-emo	0.587	0.625	0.605

Table 4. Two-level emotional classification on Alm's dataset (the highest precision, recall, and F-measure values for each class are shown in bold).

		Two-level classification			Flat classification		
		Precision	Recall	F-measure	Precision	Recall	F-measure
1st level	*neg*	0.81	0.93	0.87	--	--	--
	pos	**0.84**	0.64	**0.72**	0.56	**0.86**	0.68
2nd level reference results	*sadness*	0.65	**0.68**	**0.66**	**0.67**	0.53	0.59
	fear	**0.59**	**0.40**	**0.47**	**0.59**	0.38	0.46
	surprise	**0.45**	**0.21**	**0.29**	0.35	0.10	0.16
	anger	0.49	**0.73**	**0.59**	**0.54**	0.43	0.48
Accuracy			59.07%			57.41%	
2-level experiment global results	*happiness*	**0.84**	0.64	**0.72**	0.56	**0.86**	0.68
	sadness	0.55	**0.61**	0.58	**0.67**	0.53	**0.59**
	fear	0.45	**0.39**	0.42	**0.59**	0.38	**0.46**
	surprise	0.27	**0.21**	**0.19**	**0.35**	0.10	0.16
	anger	0.43	**0.68**	**0.53**	**0.54**	0.43	0.48
Accuracy			56.57%			57.41%	

By comparing the reference results in Table 2 with Aman's result shown in Table 3, our results on two classes, *non-emo* and sadness are significantly better. Even though recall of our experiments is higher for *happiness* class, the precision makes the F-measure to be lower. The reason behind the difference between the precisions is the same as their difference between in our hierarchical and flat comparisons. As it was also mentioned there we hope to address this problem in three-level classification. Both precision and recall of the *sadness* in our experiments is higher than Aman's results. We have a higher precision for *fear*, but recall is slightly lower. For the last three classes our precision is higher while recall is significantly lower.

The size of these three classes, which are the smallest classes in the dataset, appears to be the reason. It is possible that the small set of features that we are using will recall fewer instances of these classes comparing to the bigger feature sets used by Aman (2007).

3.1.2 Negative-Positive polarity

These experiments have been run on Alm's dataset with five emotion classes. This part is based on the assumption that the *happiness* class is positive and the remaining four classes are negative.

At the first level, positive versus negative classification tries to determine whether an instance bears a positive emotion. The second step takes all instances which level 1 classified as negative, and tries to classify them into one of the four negative classes, namely *sadness*, *fear*, *surprise* and *anger-disgust*. The results show a higher accuracy in *reference results* while it is slightly lower for global results. In terms of precision and recall, however, there is a high increase in precision of positive (*happiness*) class while the recall decreases.

The results show a higher accuracy in *reference results* while it is slightly lower for global results. In terms of precision and recall, however, there is a high increase in precision of positive (*happiness*) class while the recall decreases.

We also see a higher F-measure for all classes in the reference results. That confirms the consistency between the result in Table 2 and Table 4.

In the global measurements, recall is higher for all the classes at the second level, but the F-measure is higher only for three classes.

Here we cannot compare our results with the best previous results achieved by Alm (2008), because the datasets and the experiments are not the same. She reports the accuracy of the classification results for three sub-corpora separately. She randomly selected neutral instances from the annotated data and added them to the dataset, which makes it

different than the data set we used in our experiments.

3.2 Three-level classification

In this approach, we go even further: we break the seven-class classification task into three levels. The first level defines whether the instance is emotional. At the second level the instances defined as emotional by the first level will be classified on their polarity. At the third level, we assume that the instances of *happiness* class have positive polarity and the other five emotions negative polarity. That is why we take the negative instances from the second level and classify them into the five negative emotion classes. Table 5 presents the results of this classification. The results show that the accuracy of both reference results and global results are higher than flat classification, but the accuracy of the global results is not significantly better.

At the first and second level, the F-measure of *no-emotion* and *happiness* classes is significantly better. At the third level, except in the class *disgust*, we see an increase in the F-measure of all classes in comparison with both the two-level and flat classification.

Table 5. Three-level emotional classification on Aman's dataset (the highest precision, recall, and F-measure values for each class are shown in bold)

		Three-level Classification		
		Precision	Recall	F
1st level	*emo*	0.88	0.85	0.86
	non-emo	**0.88**	0.81	**0.84**
2nd level reference results	*positive*	**0.89**	**0.65**	**0.75**
	negative	0.79	0.94	0.86
3rd level reference results	*sadness*	0.63	**0.54**	**0.59**
	fear	**0.88**	**0.52**	**0.65**
	surprise	**0.79**	**0.37**	**0.50**
	disgust	0.42	**0.38**	0.40
	anger	0.38	**0.71**	**0.49**
Accuracy			**65.5%**	
3-level experiment global results	*non-emo*	**0.88**	0.81	**0.84**
	happiness	**0.77**	**0.62**	**0.69**
	sadness	0.43	**0.49**	0.46
	fear	0.52	0.4	0.45
	surprise	0.46	**0.32**	**0.38**
	disgust	0.31	**0.31**	0.31
	anger	0.35	**0.55**	**0.43**
Accuracy			**62.2%**	

Also, as shown by the two-level experiments, the results of the second level of the reference results approach an increase in the precision of the *happiness* class. That makes the instances defined as *happiness* more precise.

By comparing the results with Table 3, which is the best previous results, we see an increase in the precision of *happiness* class and its F-measure consequently; therefore in these results we get a higher F-measure for three classes, *non-emo*, *sadness* and *fear*. We get the same F-measure for *happiness* and slightly lower F-measure for *surprise* but we still have a lower F-measure for the other two classes, namely, *disgust* and *anger*. The other difference is the high increase in the recall value for *fear*.

4 Conclusions and Future Work

The focus of this study was a comparison of the hierarchical and flat classification approaches to emotional analysis and classification. In the emotional classification we noticed that having a dominant class in the dataset degrades the results significantly. A classifier trained on imbalanced data gives biased results for the classes with more instances. Our results, based on a novel method, shows that the hierarchical classification approach is better at dealing with the highly imbalanced data. We also saw a considerable improvement in the classification results when we did not deal with the errors from previous steps and slightly better results when we evaluated the results globally.

In the future, we will consider different levels of our hierarchy as different tasks which could be handled differently. Each of the tasks has its own specification. We can, therefore, definitely benefit from analyzing each task separately and defining different sets of features and classification methods for each task rather than using the same method for every task.

References

Alm, C.: “Affect in text and speech”, PhD dissertation, University of Illinois at Urbana-Champaign, Department of Linguistics (2008)

Aman, S.: “Identifying Expressions of Emotion in Text”, Master's thesis, University of Ottawa, Ottawa, Canada (2007)

Aman, S., Szpakowicz, S.: “Using Roget’s Thesaurus for Fine-grained Emotion Recognition”. Proc. Conf. on Natural Language Processing (IJCNLP), Hyderabad, India, 296-302 (2008)

Chaumartin. F.: “Upar7: A knowledge-based system for headline sentiment tagging”, Proc. SemEval-2007, Prague, Czech Republic, June (2007)

Ekman, P.: “An Argument for Basic Emotions”, Cognition and Emotion, 6, 169-200 (1992)

Ghazi, D., Inkpen, D., Szpakowicz, S.: “Hierarchical approach to emotion recognition and classification in texts”, A. Farzindar and V. Keselj (eds.), Proc. 23rd Canadian Conference on Artificial Intelligence, Ottawa, ON. Lecture Notes in Artificial Intelligence 6085, Springer Verlag, 40–50 (2010)

Katz, P., Singleton, M., Wicentowski, R.: “Swatmp: the semeval-2007 systems for task 5 and task 14”, Proc. SemEval-2007, Prague, Czech Republic, June (2007)

Kennedy, A., Inkpen, D.: “Sentiment classification of movie reviews using contextual valence shifter”, Computational Intelligence 22. 110-125 (2006)

Keshtkar, F., Inkpen, D.: “Using Sentiment Orientation Features for Mood Classification in Blog Corpus”, IEEE International Conf. on NLP and KE, Dalian, China, Sep. 24-27 (2009)

Kiritchenko, S., Matwin, S., Nock, R., Famili, F.: “Learning and Evaluation in the Presence of Class Hierarchies: Application to Text Categorization”, Lecture Notes in Artificial Intelligence 4013, Springer, 395-406 (2006)

Koller, D., Sahami, M.: “Hierarchically Classifying Documents Using Very Few Words”. Proc. International Conference on Machine Learning, 170-178 (1997)

Kozareva, Z., Navarro, B., Vazquez, S., Montoyo, A.: “UA-ZBSA: A headline emotion classification through web information”, Proc. SemEval-2007, Prague, Czech Republic, June (2007)

Liu, H., Lieberman, H., Selker, T.: “A Model of Textual Affect Sensing using Real-World Knowledge”. In Proc. IUI 2003, 125-132 (2003)

Neviarouskaya, A., Prendinger, H., and Ishizuka, M.: “Compositionality Principle in Recognition of Fine-Grained Emotions from Text”, In: Proceedings of Third International Conference on Weblogs and Social Media (ICWSM’09), AAAI, San Jose, California, US, 278-281 (2009)

Olveres, J., Billinghurst, M., Savage, J., Holden, A.: “Intelligent, Expressive Avatars”. In Proc. of the WECC’98, 47-55 (1998)

Strapparava, C., Mihalcea, R.: “SemEval-2007 Task 14: Affective Text” (2007)

Strapparava, C., Mihalcea, R.: “Learning to Identify Emotions in Text”, Proc. ACM Symposium on Applied computing, Fortaleza, Brazil, 1556-1560 (2008)

Wilson, T., Wiebe, J., Hoffmann, P.: “Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis”, Computational Linguistics 35(3), 399-433 (2009)

Wiebe, J., Wilson, T., Cardie, C.: “Annotating Expressions of Opinions and Emotions in Language”, Language Resources and Evaluation 39, 165-210 (2005)