CS-584 MACHINE LEARNING

REPORT

Hierarchical Approach to Emotion Recognition and Classification in Texts

Sahith Kumar Andel(A20329614)

Raviteja Chebrolu (A20326528)

Introduction

In this paper, we try to explore the task of classification of texts with respect to the sentiment expressed by them. Although, there is a type of classification called FLAT, we try to solve the problem by Hierarchical classification because Flat approach doesn't consider hierarchical information. We try to check how the existence of neutral instances affect the performance of classifying emotions.

In this paper, we use two-level classification where we basically look into the categorization of a sentence into two basic sentiments in the higher level namely positive and negative, whereas the lower level has a classification of three sentiments namely positive, negative and neutral where positive and negative are named as emotional and neutral as non-emotional. The lower level carries the hierarchical information to the higher level where the information is used to produce better results than when compared to Flat Classification. Hierarchical classification considers the relationship between polarity and emotion of a text.

We use Naive Bayes algorithm to classify the text into the sentiments. Naive Bayes allows to predict a class, given set of features using probability. In the below sections we explain the datasets used, algorithm implemented, and the evaluation of the results of both flat classification and hierarchical classification and also future work and conclusion.

Resources and Feature Sets

Resources:

These statistical methods generally require the training and the test data, which is manually annotated. The main priority in the finding the data is that the data should be rich in emotion expressions and also the dataset should contain all the emotion categories. Keeping these points in mind, we have use the data by sanders which is manually annotated twitter data. The data is contains about 5000 instances of positive, negative, neutral and Irrelevant categories.

Domain	SIze	#classes
Twitter	5114	4

Table1: Dataset Information

Cleaning and preprocessing:

The next step after collecting data was to clean the data and preprocess the data and make it ready for feature extraction and then classification. So we removed the irrelevant class and all the data related to that class.

Domain	Size	#classes
Twitter	3425	3

Table2: Dataset after removing irrelevant data

The next step was to preprocess the data to prepare if for feature extraction. The preprocessing for this dataset include some steps which are:

1. Convert the text to lowercase
2. Replace links with the string “URL”
3. Replace @...with ‘AT_USER’
4. Replace #word with the word
5. Remove stopwords(including url and user)
6. Tokenize the document into words(a list of words)

The above defined steps were used by us to preprocess the documents in the datasets. The steps 2,3,4 where implemented by using regular expression and the steps 5,6 where implemented by using an external package called nltk(natural language processing toolkit).

Feature Extraction:

Once the preprocessing is done, we extracted the features from training data and test data. We used Naive bayes for the classification. We used the naive bayes classifier inbuilt in the nltk itself to classify the data.

To perform the classification, we followed some steps which are:

1. Building the vocabulary (list of all the word in all the documents in the training data)
2. Representing each document with the presence/absence of these words in the document

   Example : {'the','worst','thing,'in','the','world'} -vocabulary

   {'the','worst','thing')- document

   Then the feature vector will be (1,1,1,0,0,0)- feature vector

3. Once all the training data is represented as explained, we used the nltk inbuilt naive bayes method to classify the data

The next sections will discuss the Flat based classification and Hierarchical Classification and the evaluation.

Hierarchical Classification

Hierarchical categorization deals with categorization problems in which categories are organized in hierarchies. For most text categorization tasks the category hierarchies have been carefully composed by humans and represent our knowledge of the subject. In this work, we use the hierarchical categories to impart an additional knowledge to our classification method.

In this paper, we tried to convey the information in one form of hierarchy. That is two-level hierarchy where we try represent the relation between emotion and the neural instances. In the remainder of this section, we will detailed explanation about the two level classification

In two-level classification, the first level is to classify between emotional and non-emotional where we try to determine whether the instance is emotional or neutral. The second step takes all the instances classified in the 1st level as emotional and tries to classify them into one of the two sentiments positive and negative respectively.

Hierarchical classification evaluation:

	Precision	Recall	F1-score
Emotional	0.51	0.90	0.65
Non-emotional	0.84	0.37	0.65
Positive	**0.76**	0.69	0.726
Negative	**0.84**	0.88	0.726

Table3: Hierarchical Evaluation

Flat Classification evaluation:

	Precision	Recall	F1-score
Positive	**0.12**	0.74	0.21
Negative	**0.34**	0.88	0.5
Neutral	0.99	0.73	0.84

Table4: Flat Evaluation

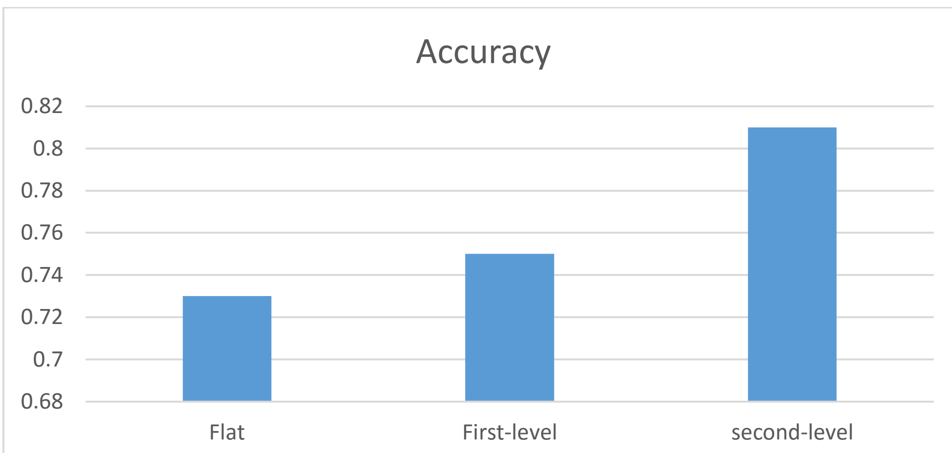
Graph1: Details of metrics

From the above tables we can understand how the neutral instances affect the classification of the sentiments, we can see in flat classification how low the values are due to presence of neutral instances whereas in the hierarchical classification we can see the scores are pretty good compared to flat classification. This shows that the usage of hierarchical information has a pretty good effect on the classification.

	Accuracy
Flat Classification	0.730656934307
Emotional/non-emotional	0.75985401459854018
Hierarchical	0.81465517241379315

Table 5: Accuracy of test data

The above shows that there is significant amount of effect of using hierarchical classification when compared to that of Flat classification.



Graph2: Depicting the accuracy of classifications

Conclusion

The focus of this study was an emotional analysis and classification of emotions in sentences. We used only one dataset with three different kinds of sentiments to classify them using flat and hierarchical and evaluate the performance of these methods on emotion analysis.

In this work, we noticed that having non-emotional instances in the dataset degrades the result significantly, which is why we implemented two-level classification which defines the non-emotional instances in one step and considers the rest as emotional. This process gave us better results in classifying.

The future work will be to collect more datasets which contains six types of emotions happiness, sadness, surprise, fear, disgust, and anger and to implements three-level classification of the hierarchy and to evaluate the performance of the classification methods. The three-level classification, has three different tasks, namely emotional versus non-emotional, polarity, and emotional classification

References

1) https://www.site.uottawa.ca/~diana/publications/DimanCanadianAI2010.pdf
2) Strapparava, C., Mihalcea, R.: Learning to Identify Emotions in Text. In: Proc. ACM Symposium on Applied computing, Fortaleza, Brazil, pp. 1556–1560 (2008)
3) Aman, S.: Identifying Expressions of Emotion in Text. Master's thesis, University of Ottawa, Ottawa, Canada (2007)
4) http://streamhacker.com/2010/05/10/text-classification-sentiment-analysis-naive-bayes-classifier/
5) 5. Alm, C.O., Roth, D., Sproat, R.: Emotions from text: machine learning for text- based emotion prediction. In: Proc. Joint Conference on Human Language Technology / Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), Vancouver, Canada, pp. 579–586 (2005)