Event Based Emotion Classification for News Articles

Minglei Li, Da Wang, Qin Lu, Yunfei Long
Computing Department, The Hong Kong Polytechnic University,
Hung Hom, Hong Kong
{csmli, csluqin, csylong}@comp.polyu.edu.hk
danwang.km@connect.polyu.hk

Abstract

Reading of news articles can trigger emotional reactions from its readers. But comparing to other genre of text, news articles that are mainly used to report events, lack emotion linked words and other features for emotion classification. In this paper, we propose an event anchor based method for emotion classification for news articles. Firstly, we build an emotion linked news corpus through crowdsourcing. Then we propose a CRF based event anchor extraction method to identify event related anchor words that can potentially trigger emotions. These anchor words are then used as features to train a classifier for emotion classification. Experiment shows that our proposed anchor word based method achieves comparable performance to bag-of-word based method and it also performs better than emotion lexicon features. Combining anchor words with bag-of-words can increase the performance by 7.0% under weighted F-score. Evaluation on the SemEval 2007 news headlines task shows that our method outperforms most of other methods.

1 Introduction

Emotion classification from text, an extension of sentiment analysis, aims at assigning emotional labels to a given text. It has wide applications such as customer review (Pang et al., 2002), emotion based recommendation (Cambria et al., 2011), emotional human-computer interaction (Hollinger et al., 2006), eLearning (Rodriguez et al., 2012), etc. It is also important to understand reader's emotion reactions for reading news articles as they may trigger emotionally charged reactions which may lead to serious social and political consequences. However, news articles are normally used to describe recent events. To maintain objectivity, writers normally avoid using subjectivity and emotion-linked words. Thus, current works on emotion analysis, which use more social media type of text, would not work well for news text.

Generally speaking, emotion classification can be done either at document level or at sentence level. In this paper, we focus on document level emotion classification for news articles. Due to the nature of new articles, we need to address two main issues: 1) How to obtain sufficiently high quality labeled news corpus for training and prediction; and 2) How to identify suitable features for this genre of text. To address the first issue, we make use of the crowdsourcing method to obtain labeled data for a set of news articles provided in ACE 2005 (Walker et al., 2006) and through appropriate filtering, to obtain a reasonably good emotion-labeled corpus. To address the second issue, we first investigate the commonly used features for emotion prediction, including N-gram, Part-Of-Speech (POS), and emotion lexicons (Lin et al., 2007). However, these features, suited for sentence level classification, seem to be noisy for document level classification. Since news articles mainly describe a specific event and based on psychological studies that event can trigger emotions (Cacioppo and Gardner, 1999), we further explore event related features for emotion classification. Our hypothesis is that for news articles, a specific set of event linked anchor words can trigger emotions of readers and are therefore more important than most of the other words which may not have relations to emotions. Here the anchor word means the keyword of an event, such as *"die", "accident", "bomb"*, etc. Our proposed approach identifies event anchor words and use them as features for emotion classification. The main steps involved in event anchor word extraction involves three steps: First, we make use of the ACE 2005 data as our raw source corpus where event information was already annotated and crowdsourcing is used to obtain emotion linked labels for the news articles. Second, we use the annotated event information to train a CRF model for event anchor words extraction. Last, the extracted event anchor words can then be used as features to train a classifier for emotion prediction. This is different from lexicon based method because lexicon based method relies on externally prepared knowledge. In contrast, anchor words are automatically extracted from training data to be used as features. The main contributions of this work include:

1. The construction of an important annotated resource for event based emotion analysis based on ACE 2005 English news articles which can be made available to the research community.
2. The identification of more suitable features for document level emotion classification of news articles without emotion lexicon, and a feasible feature extraction method, which can also be used by other event based applications. The proposed features are more effective than emotion lexicon features and can improve the performance when combined with the bag-of-word features.

The rest of the paper is organized as follows: Section 2 discusses related works for emotion classification. Section 3 introduces the construction of the annotated corpus as a training data resource. Section 4 presents our event anchor word extraction and emotion analysis framework. Section 5 gives performance evaluation. The conclusion and future work are summarized in section 6.

2 Related Works

Any method on emotion analysis must rely on an emotion model to provide a framework for emotion classification. Emotion models can be characterized as discrete models and coordinate based models. Discrete models include the most commonly used six emotions (Ekman, 1993), the eight emotion model (Plutchik, 1980), Ortonys 22 emotions (Ortony, 1990), and Xu's seven emotion models (Xu and Lin, 2008), etc. Dimension based models include evaluation-activation based models (Whissell, 1989) and valence-arousal based models (Russell, 1980; Mehrabian, 1996; Wang et al., 2015), which is widely used for emotion classification. When using dimension based models, emotion prediction becomes a regression problem to predict the values in the two axis (Wu et al., 2013). When using discrete models, emotion prediction becomes a multi-class classification problem, which is the most commonly used methods in literature. This two representation methods are well compared in (Calvo and Mac Kim, 2013).

One of the problems that hinder emotion analysis is the lack of training data. Annotated emotion corpus is relative scarce compared to other NLP tasks. Manually labeled emotion corpora include SemEval 2007 for news headlines (Strapparava and Mihalcea, 2007), RenCECps for blogs in both sentence and document levels (Quan and Ren, 2009), and NLP&CC 2013 for Chinese microblogs. Because of the rapid development of social networks, many studies also try to automatically construct emotion corpus from the web using reader added emotion tags as labels. Lin crawled a news article corpus based on the emotion related tags by readers from Yahoo!s news (Lin et al., 2007). Hashtags, emoticons, and emoji characters are also used as naturally annotated labels to construct large emotion corpus from social media (Bandhakavi et al., 2014; Mohammad et al., 2013; Wang et al., 2012). However, these naturally annotated labels often contain noise. A recent trend is to make use of crowdsourcing to obtain annotated data. Crowdsourcing can be reliable if some control strategies are properly used. Example of resources obtained by crowdsourcing include lexicons constructed by Hutto (2013) and Mohammad (2013).

Emotion classification can be categorized into 1) rule based methods and machine learning based methods. As an example of rule based systems, the work by Chaumartin (2007) uses a set of hand crafted rules based on common knowledge to analyze the emotions of news headlines. In another system, Strapparava (2008) represents each emotion and text using latent semantic analysis (LSA) and analyzes the corresponding emotion based on the similarity between the text and the corresponding emotions. Machine learning based methods, on the other hand, heavily rely on the availability of training data as well as good feature selection methods. Mohammad shows that the combined using of emotion lexicon and N-gram features is more effective than N-gram feature only (Mohammad, 2012). Quan makes use of emotional words based features and tries to apply them to different classifiers such as SVM, Naive Bayes, and decision trees for sentence level blog emotion classification (Quan and Ren, 2009). Based on word embedding, Chen uses a sentence vector in combination with ML-KNN for microblog data (Chen et al., 2014). Inspired by Poria (2014) that uses dependency features and CRF model, a segment-based method is proposed to extract sentence segments using dependency trees and the semi-CRF model is used to label emotions of all the segments, and then the log linear model is used to infer the final emotion of the whole sentence (Wang, 2014). Similar idea is adopted by Wen (2014) where the data mining technique class sequential rules (CSR) mining is used to analyze the emotion of the whole microblog containing several sentences.

Based on psychological studies that events can trigger emotions (Cacioppo and Gardner, 1999), many studies propose event based emotion prediction methods. Tokuhisa extracts events that can trigger emotions from the web based on emotion words and uses k-NN to predict new text for dialogs (Tokuhisa et al., 2008). Extending from Tokuhisa (2008), Vu constructs an event corpus by first defining a set of seed events and then extends it using boot-strapping (Vu et al., 2014). Lee builds an emotion linked event corpus from Chinese stories (Lee et al., 2014). Li proposes a system to detect and extract the cause event in microblogs, and uses these events as features to train a classifier for emotion prediction in microblogs (Li et al., 2014).

3 Corpus Construction

In order to serve the objective of our work for event based emotion prediction on news articles, we need to first prepare an appropriate training data which is currently not available. With consideration of resources, we choose to use the crowsourcing platform to annotate the data.

3.1 Data Source

The raw data comes from the Automatic Content Extraction 2005 (ACE 2005), a complete set of multi-language training data for the ACE 2005 evaluation (Walker et al., 2006). In this work, we only use the English collection of Automatic Content Extraction 2005 (ACE 2005)[1] which contains 754 English texts collected from newswire (18.57%), broadcast news (39.79%), broadcast conversation (9.15%), web log (18.30%), UseNet newsgroups/discussion forum (8.22%), and conversational telephone speech (5.97%). Though not all of these texts are news articles, they are all descriptions about events, consistent for our event based emotion analysis for news articles. So we simply name this set of the data as the news articles in the rest of this paper.

The news articles dataset was originally created by the Linguistic Data Consortium (LDC) prepared for event identification. The news articles dataset already have annotated event related information such as the anchor words for events, event types and event subtypes. For example, in the below sentence, *114 people were wounded in Tues-day's southern Philippines airport*, it contains an event about *injury*. The event anchor word is *"wounded"* while the event type and subtype are *"Life"* and *"Injure"* correspondingly. This dataset is quite appropriate to serve as the training data for our work because it is highly related with event description. However, there is no emotion related annotation that can be used directly as training data for emotion analysis.

3.2 Data Annotation

In order to use this collection as training data, we need to identify the emotion associated with each article. For annotation, we choose the most commonly used emotion model (Ekman, 1993), which includes six discrete emotion labels: *anger, disgust, fear, joy, sadness, and surprise*, respectively. We also add the category, *neutral*, to be used for those articles which may not trigger any emotion. This label is partic-

[1] http://www.itl.nist.gov/iad/mig/tests/ace/2005/

ularly suitable for news articles. Naturally, different people may have different emotions even when reading the same text. Therefore, to eliminate bias by a single crowdsourcing contributor, we request each article to be annotated by 5 contributors. Obviously, there may be different labels given by different annotators. The principle is to use the majority as the label for each article. If the result has no majority, the filtering process after crowdsourcing is initiated. The annotation platform used is CrowdFlower[2]. To ensure quality of annotation, a quality control (QC) mechanism is included in CrowdFlower to prevent people from randomly labeling the text (and also possibly eliminate people who have low English proficiency). Inspired by Hutto (2014), the QC process is conducted through a four step process as described below:

1. A subset H of eight articles are labeled by the research team as the ground truth for QC purpose. We asked 3 persons in the research team to serve as experts to annotate a selected set of articles independently. The articles which received the same labels by all three people are used as the ground truth for future usage.

2. We choose a subset H_t (6 articles) from H as quality test data and the contributors who correctly labelled 80% in H_t are qualified for future batches of work. Here correctness means the label given by the contributor is the same as the ground truth.

3. In the real annotation tasks, we randomly pick one instance from H in every batch of 6 articles to test whether they are doing random labeling. Results by those who wrongly labeled the test instances are discarded and the person will not be given new tasks. Even though this added redundancy costs more for annotation, it gives us more assurance of the quality of acquired data.

4. After completing the annotation for each article, the annotators are asked to briefly give rationales for their choice of label. We randomly check the written responses to ensure that the contributors are not making random choices.

Type	**Pattern**	**%(Number)**	Non-Neutral
1	5,0,0,0,0	3.18%(24)	1.59%(12)
2	4,1,0,0,0	11.14%(84)	6.23%(47)
3	3,2,0,0,0	9.81%(74)	6.63%(50)
4	3,1,1,0,0	22.68%(171)	15.78%(119)
5	2,2,1,0,0	22.81%(172)	10.08%(76)
6	2,1,1,1,0	26.13%(197)	20.42%(154)
7	1,1,1,1,1	4.24%(32)	-
Sum		100.00%(754)	60.74%(458)
Fleiss Kappa		0.212	

Table 1: Crowdsourcing based annotation result

Table 1 shows the distribution of the seven types of patterns in the annotation result (including the articles in H. Different annotators may give the same article different labels. The first pattern (5,0,0,0,0) means that all five annotators gives an identical emotion label. The second pattern (4,1,0,0,0) means 4 people give the same label whereas one person gives a different label. The pattern (1,1,1,1,1) means that every annotator give it a different emotion label. The labels for text in H are also refined through the labels given by the contributors. The 3rd column in **Table 1** shows the distribution percentage (%) and total articles number (Number) and the last column shows the percentage for data that have one major label and falls into the non-neutral categories. Out of the 7 possible patterns, only 3.18% of data falls into Type 1, the best scenario where the same label is given by all annotators. In Type 7, everyone gives a different label and 4.24%(32 articles) of data falls into this category.

We use Fleiss Kappa value to evaluate the consistence between different contributors and the value of 0.212 indicates a fair agreement. The relatively low value of Fleiss Kappa indicates the difficulty in emotion annotation because emotions is very subjective depends a lot on the annotators. This is particularly true for event linked emotions as they can be dependent on the annotators' background and preferences such as religions, political stands, etc. Since emotion classification is naturally multilabeled and personal variations are also natural, we consider the data quality is reasonably good. However, for training purpose, we further filter the data and only retain those which have a major shared emotion. We

[2] http://www.crowdflower.com/

consider five of the patterns to have a major shared emotion including Type 1 to Type 4 and type 6. If the major label is the neutral label, however, the data will be removed. In other words, only the articles that has non-neutral labels are used as training data. In fact, in our data, 22.28% percent of annotated data falls into the neutral class which is only natural for news type of text. Obviously, this is very different for text from social media. The Type 5 pattern indicates two major emotion labels with equal numbers. Only if one of the major labels is neutral, the data is retained. Finally, we obtain 458 (60.74% of 754) articles with an improved Fleiss Kappa value of 0.214 (fair agreement), which is slightly better than the original 754. The distribution of the 6 emotion classes for the 458 articles are listed in **Table 2**. Note that the ratio of the largest set to the smallest set is about 3.2. Compare to other genre of text, the training data is not so skewed as the emotion labels in social media based corpus (Chen et al., 2014). To support research in emotion analysis. We make the annotated data available.[3]

Major emotion	**Number**	**Percentage%**
Fear	41	9.0
Sadness	114	24.9
Disgust	36	7.9
Surprise	115	25.1
Anger	61	13.3
Joy	91	19.9
Sum	458	100.00

Table 2: Emotion distribution of obtained data

4 Our Proposed Method

Our method for classification consists of two parts: the first part is anchor word extraction and the second part is the appropriate classification method for emotion classification.

4.1 Anchor Word Extraction

Many emotion prediction methods use NLP related features such as N-gram, POS tags, and position information of lexical sequences because they can be easily extracted to train classifiers. The ACE 2005 data contains many annotated latent information can potentially be useful for event identification. The annotated data in ACE 2005 at the summarization level includes topic, event type, event subtype, and event anchor word, etc. However, without appropriate method to extract latent information for testing data, they cannot be used. Because of this reason, other than lexical features which we can extract using NLP tools, most of the annotation information in ACE 2005 are not used as it is difficult to automatically infer event related summary information. We choose to focus our attention on extract event anchor words (anchors for short) as our features because they are easier to extract. Generally speaking we can consider event anchor word extraction as a kind of keyword extraction. The only difference is that the keyword here is linked to certain event (indicated by actions), and thus events provide the cues to identify the corresponding anchors.

Problem definition: Given a text sequence, $X = \{x_1, x_2, \ldots, x_n\}$ where x_i is a corresponding word and n is the number of words in X. Our goal is to find one or more words $x_j, \ldots, x_k$ used to describe the event in X. This problem can be converted into a sequential labeling problem. The objective of sequential labeling is to find the corresponding label sequence $Y = \{y_1, y_2, \ldots, y_n\}$ where $y_i \in \{0, 1\}$; 0 means not an anchor word; 1 otherwise. As this is a typical sequential labeling problem, the Conditional Random Field (CRF) algorithm can be used for anchor word extraction, the same method used by Zhang for keyword extraction (Zhang, 2008). The most important performance issue for CRF is feature construction. In our algorithm, we consider a context window of 2 on both sides of an anchor. Since we can easily use NLP tools to identify POS tags, features considered for anchor words include both the context words and their POS tags. The CRF model is trained using our 754 news articles which already contain the anchor annotation.

4.2 Classifier for Emotion Classification

Popularly used supervised machine learning methods for classification include Nave Bayes, k-NN, SVM, random forest, etc.. Study by (Fernandez-Delgado et al., 2014) shows that, among the reviewed 179 classifiers, random forest achieves the best result, closely followed by SVM. Since this work focuses on the effectiveness of our proposed

[3] https://github.com/MingleiLI/ACE2005_emotion_corpus

features rather than a classifier, we simply choose SVM as our classifier because it is widely used for multiclass classification. We further adopt the one-vs-all strategy for multiclass classification.

4.3 Features Used for Emotion Classification

To train an emotion classifier, we investigate the following features which are potentially useful. All features are considered using a context window of 5. In addition to an anchor word, the 2 words on each side of the anchor are included.

1. **F1: Frequency of anchors** - Occurrences of an identified anchor word in an article.
2. **F2: Word similarity** - Similarity between anchors and all the other words in the article. The motivation is if more words with similar meaning occur, the emotion tendency is more apparent. Similarity calculation is based on Lin's similarity module of WordNet, which is based on information content (Pedersen et al., 2004).
3. **F3: Frequency of POS tag of anchors** - Occurrences of POS tags of anchors.
4. **F4: Frequency of POS tag of context** - Occurrences of POS tag of context words. context words are not used because our training dataset is small.

Based on the above features, we form different feature sets to evaluate the effectiveness of these features and select the best one.

5 Performance Evaluation

Evaluations are conducted for both the event anchor word extraction and the selection of features in emotion prediction.

5.1 Evaluation on Anchor Word Extraction

Since anchor words are used to identify events, anchor word extraction can use all the 754 news articles as training data. The Stanford POS tagger is used for POS tagging[4]. CRF++[5] is used for event anchor extraction. As the training data is relatively small, 10-fold, 5-fold and 3-fold cross validation are conducted to see the effect of data size to anchor extraction performance. Results are shown in **Table 3**:

Fold num	precision	recall	accuracy	F-score
10	82.17	63.50	97.30	71.64
5	82.07	62.17	97.24	70.75
3	81.96	59.86	97.14	69.19

Table 3: Event anchor word extraction result

Table 3 shows that the size of training dataset does affect performance. However, the difference is mostly on recall. From 3-fold to 10-fold, the increase in F-score is only little over 2% when the training data size is increased by about 35%. Close examination found that the extracted anchors are very stable. In other words, they are very similar under similar event types and similar topics. Thus, they are very good representatives. For example, if a news text is about *injury*, it is highly likely that the text would contain anchor words such as *wounded* or *injured*. Anchor extraction is a special kind of keyword extraction, yet its performance is much better compared to the state-of-art keyword extraction (Hasan and Ng, 2014) which has F-score of 31.7% on news articles. In addition, event anchor extraction is different because general keyword extraction focuses on extracting only a few keywords for the whole article while anchor extraction focuses on on extracting keyword more at the sentence level where the event is described. Ultimately, we care if the extracted anchors do serve as good features for emotion classification.

5.2 Evaluation on Emotion Classification

We use LibSVM (Chang and Lin, 2011) as the SVM tool. Three sets of experiments are conducted. The first set tests the performance of different anchor based feature groups in emotion classification. The second set tests the effectiveness of our selected features compared to features used by other methods for emotion classification. We conduct the third set of experiments by applying our method to the publicly available dataset used in the SemEval 2007 task for emotion classification. This dataset is on news headlines which should also be qualified as event-based data (Strapparava and Mihalcea, 2007). In the **first** set of experiments, we use the 458 news articles as training and a 10-fold cross validation is used for testing. Out of the 4 feature presented in Section 4.1

[4] http://nlp.stanford.edu/software/tagger.shtml

[5] http://taku910.github.io/crfpp/

(F1 to F4), our test plan tries four feature groups to explore the best feature combination as shown in **Table 5**. In the first feature group (FG1), only anchor words are used. The other three groups use the basic anchor words to be combined with an additional single feature. **Table 5** gives the F-score of the 4 feature groups with details on the performance of each emotion type. The Weighted F-scores in the last row is the micro average of individual F-scores.

Feature Group	**F1**	**F2**	**F3**	**F4**
FG1	Y			
FG2	Y	Y		
FG3	Y		Y	
FG4	Y			Y

Table 4: Feature combinations

Emotion	**FG1**	**FG2**	**FG3**	**FG4**
Fear	12.9	**21.8**	13.8	12.4
Sadness	44.4	36.0	**46.0**	40.6
Disgust	**6.50**	3.2	5.3	0.0
Surprise	**36.4**	31.1	29.5	22.4
Anger	18.8	15.9	**19.9**	14.5
Joy	30.0	27.0	**31.8**	30.2
Weighted F-score	**30.3**	26.5	29.5	24.7

Table 5: Performance of different feature combinations

Table 5 shows that the best feature group is FG1 which takes only anchor text using frequency as the feature. The POS tags of context words (FG4) is the noisiest and produces the worst result. The use of the additional POS tags for anchor words (FG3) does not give overall better result. Yet, it gives better performance in 3 emotion types. Compared to the other two features, it is the least noisy because the performance degradation is less than 1%. This may be because frequency information is already used by FG1, and the frequency of POS tags of anchors are largely represented. It is interesting to see that context word does not give overall gain in performance except in the Fear emotion type. We can generally conclude that using POS tags do not translate into overall performance improvement. The F2 similarity feature degrades the performance maybe because this similarity is based on semantic similarity, not emotional similarity. In conclusion, anchor word as single feature achieves the best performance and thus we only use anchor words in the following experiments.

In the **second** set of experiments, we compare our event anchor based (EA) features to features used by other works for emotion classification using news articles. The features used by other methods include (1) lexicon feature based method (LF) which simply use a given emotion lexicon as features; (2) Bag of word (BOW) (Mohammad, 2012); (3) LF plus BOW based method (LF+BOW) that combines BOW and LF by increasing the weight of words that occur in the emotion lexicon, and (4) feature combination of event anchor words with BOW (EA+BOW) which does not use any external knowledge. The lexicon for LF is from WordNet-Affect (Strapparava and Valitutti, 2004). The parameters of SVM are the same for all SVMs used in this experiment. The result of F-score is shown in **Table 6**.

Table 6 shows that LF achieves the worst result which indicates that classification based only on an externally provided lexicon is not enough for document level emotion classification. BOW and EA use training data supplied information without any external knowledge and achieve much better result than LF as they are learning based methods. Our event anchor word based EA method achieves 75.1% better result than lexicon feature and slightly better performance than BOW based method. In this experiment, the size of the anchor words is 890, far smaller than the size of BOW at 13,793. This indicates that fewer effective features can actually achieve comparable result. As combined features, LF+BOW achieves better result than BOW (when LF in BOW is given more emphasis), which is consistent with the result of (Mohammad, 2012). In EA+BOW, anchor words in the bag of words are also given extra weight, and the performance is increased by 2.1% compared to BOW, which translates to 7.0% improvement over using BOW alone and also 4.9% improvement over using LF+BOW. The experiment shows that event anchors are more effective features than emotion lexicon and bag-of-words for news article emotion classification. This validates our assumption that news articles trigger emotions through specific set of event anchor words. Just a note that experiments show that increasing the frequency LF or EA in LF+BOW and EA+BOW by 3 achieves the best result.

Emotion	LF	BOW	EA	LF+ BOW	EA+ BOW
Fear	4.0	20.3	12.9	**20.4**	18.4
Sadness	17.5	41.5	44.4	39.5	**48.5**
Disgust	6.5	4.7	6.5	**7.3**	4.7
Surprise	20.8	28.8	30.4	28.8	**30.8**
Anger	11.1	24.7	18.8	24.6	**30.2**
Joy	27.1	35.9	30.1	**40.6**	32.5
Weighted F-score	17.3	30.2	30.3	30.8	**32.3**

Table 6: Results on Crowdsourcing Annotated Data

The **third** set of experiments are conducted on the SemEval 2007 data to test the usefulness of our proposed anchor feature. The SemEval 2007 affective task contains 1,000 annotated news headlines for testing and 250 annotated headlines as development data (though labelled, but too small to be used as training data). The dataset is similar in genre although has much less content. In this experiment, we directly use the event anchor words extracted in the second set of experiment as the feature set. To make easy comparison to other methods on the same dataset, our classifier is trained using the 250 validation dataset and test on the 1,000 test dataset for both the EA method and the EA+BOW method. We list the top three systems in SemEval 2007 labeled by SWAT, UA and UPAR7. We also compare with the DepecheMood (DM) method (Staiano and Guerini, 2014) which uses emotion lexicon as simple features. Their emotion lexicon contains aroun thirty seven thousand terms from 25.3K crowd-annotated news. The performance evaluation in terms of F-score is shown in **Table 7**

Emotion	SWAT	UA	UPA R7	DM	BOW	EA	EA+ BOW
Fear	18.3	**20.1**	4.7	32	13.4	16.2	13.4
Sadness	17.4	1.8	30.4	40	29.9	**35.6**	27.9
Disgust	0	0	0	0	0	0	0
Surprise	11.8	**15.0**	2.3	16	6.9	10.6	14.6
Anger	7.1	**16.0**	3.0	0	0	12.4	12.6
Joy	14.9	4.2	11.9	30	34.5	16.8	**37.6**
Weighted F-score	14.5	8.6	11.9	**27.1**	22.0	18.7	**25.0**

Table 7: Results on News Headline Data

Table 7 shows that both of our methods (EA+BOW and EA) perform better than the top three performers of SemEval 2007 task with a large margin and BOW performs 17.6% better than EA news headlines are too short to identify anchor words. Secondly, the news sources of ACE is different from the news headlines in SemEval 2007. So the anchor word extracted from ACE may not cover the news headlines well. EA+BOW, however performs better (13.6% improvement) than BOW only, which indicates the usefulness of event anchors. However, comparing to the DM method, we are still behind by about 2.1%. This may be because we have only 890 anchor words extracted from only 754 news articles and the training data size of our method is only 250 news headlines, whereas the lexicon of DM comes from 25.3K documents and their lexicon size is 37K.

6 Conclusion and Future Work

In this paper, we propose a novel method to make use of event anchor words as features for emotion classification in news articles. The use of event anchor words is based on the intuition that a small set of semantically relevant features should be more useful than a large set of noisy features. Experimental results show that anchor words are indeed quite effective. Another contribution of this work is the establishment of an important annotated resource for event based emotion analysis based on the ACE 2005 English dataset. The first limitation of this work is that the dataset used to extracted event anchors is relatively small. The second limitation is anchor words associated with events may not be sufficient to represent an event as an emotionally linked event may also be related to other attributes such as who, when, where, and others. In the future, we can extend the method on a larger dataset and explore the use of topics, event types, and other information to further improve the performance.

Acknowledgments

This work is supported by HK Polytechnic University (PolyU RTVU and CERG PolyU 15211/14E).

References

Anil Bandhakavi, Nirmalie Wiratunga, Deepak P, and Stewart Massie. 2014. Generating a Word-Emotion Lexicon from #Emotional Tweets. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 12–21. 00000.

John T. Cacioppo and Wendi L. Gardner. 1999. Emotion. *Annual review of psychology*, 50(1):191–214. 01223.

Rafael A. Calvo and Sunghwan Mac Kim. 2013. Emotions in text: dimensional and categorical models. *Computational Intelligence*, 29(3):527–543.

Erik Cambria, Amir Hussain, and Chris Eckl. 2011. Taking refuge in your personal sentic corner. pages 35–43. Citeseer.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27. 21969.

Franois-Rgis Chaumartin. 2007. UPAR7: A knowledge-based system for headline sentiment tagging. pages 422–425. Association for Computational Linguistics.

Tao Chen, Ruifeng Xu, Qin Lu, Bin Liu, Jun Xu, Lin Yao, and Zhenyu He. 2014. A Sentence Vector Based Over-Sampling Method for Imbalanced Emotion Classification. In *Computational Linguistics and Intelligent Text Processing*, pages 62–72. Springer.

Paul Ekman. 1993. Facial expression and emotion. *American psychologist*, 48(4):384.

Manuel Fernandez-Delgado, Eva Cernadas, Senen Barro, and Dinani Amorim. 2014. Do we Need Hundreds of Classifiers to Solve Real World classification problems. *Journal of Machine Learning Research*, Microtome Publishing, No. 15, pp. 3133-3181 , 2014.:3133–3181.

Kazi Saidul Hasan and Vincent Ng. 2014. Automatic Keyphrase Extraction: A Survey of the State of the Art. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1:1262–1273. 00006.

Geoffrey A Hollinger, Yavor Georgiev, Anthony Manfredi, Bruce A Maxwell, Zachary A Pezzementi, and Benjamin Mitchell. 2006. Design of a social mobile robot using emotion-based decision mechanisms. pages 3093–3098. IEEE.

C. J. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014.* 00010.

Sophia Lee, Shoushan Li, and Chu-Ren Huang. 2014. Annotating Events in an Emotion Corpus. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014).* 00000.

Chengxin Li, Huimin Wu, and Qin Jin. 2014. Emotion Classification of Chinese Microblog Text via Fusion of BoW and eVector Feature Representations. In *Natural Language Processing and Chinese Computing*, pages 217–228. Springer.

Kevin Hsin-Yih Lin, Changhua Yang, and Hsin-Hsi Chen. 2007. What emotions do news articles trigger in their readers? pages 733–734. ACM.

Albert Mehrabian. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a wordemotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.

Saif Mohammad. 2012. Portable features for classifying emotional text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 587–591. Association for Computational Linguistics.

Andrew Ortony. 1990. *The cognitive structure of emotions*. Cambridge university press.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. pages 79–86. Association for Computational Linguistics.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. WordNet:: Similarity: measuring the relatedness of concepts. In *Demonstration papers at HLT-NAACL 2004*, pages 38–41. Association for Computational Linguistics. 01145.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience*, 1(3):3–33.

Soujanya Poria, Erik Cambria, Grgoire Winterstein, and Guang-Bin Huang. 2014. Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems*, 69:45–63.

Changqin Quan and Fuji Ren. 2009. Construction of a blog emotion corpus for Chinese emotional expression analysis. pages 1446–1454. Association for Computational Linguistics.

Pilar Rodriguez, Alvaro Ortigosa, and Rosa M. Carro. 2012. Extracting emotions from texts in e-learning environments. In *Complex, Intelligent and Software Intensive Systems (CISIS), 2012 Sixth International Conference on*, pages 887–892. IEEE. 00021.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Jacopo Staiano and Marco Guerini. 2014. Depeche Mood: a Lexicon for Emotion Analysis from Crowd Annotated News. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2:427–433. 00000.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 70–74. Association for Computational Linguistics.

Carlo Strapparava and Rada Mihalcea. 2008. Learning to identify emotions in text. pages 1556–1560. ACM.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet Affect: an Affective Extension of WordNet. volume 4, pages 1083–1086.

Ryoko Tokuhisa, Kentaro Inui, and Yuji Matsumoto. 2008. Emotion classification using massive examples extracted from the web. pages 881–888. Association for Computational Linguistics.

Hoa Trong Vu, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Acquiring a Dictionary of Emotion-Provoking Events. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 128–132. 00000.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*. 00057.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2012. Harnessing twitter" big data" for automatic emotion identification. In *Privacy, Security, Risk and Trust (PASSAT),2012 International Confernece on Social Computing (SocialCom)*, pages 587–592. IEEE.

Jin Wang, K. Robert Lai, Liang-Chih Yu, and Xue-jie Zhang. 2015. A locally weighted method to improve linear regression for lexical-based valence-arousal prediction. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 415–420. IEEE.

Zengfu Wang. 2014. Segment-based Fine-grained Emotion Detection for Chinese Text. *CLP 2014*, page 52.

Shiyang Wen and Xiaojun Wan. 2014. Emotion Classification in Microblog Texts Using Class Sequential Rules. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Qubec City, Qubec, Canada.*, pages 187–193.

Cynthia Whissell. 1989. The dictionary of affect in language. *Emotion: Theory, research, and experience*, 4(113-131):94.

Bin Wu, Erheng Zhong, Derek Hao Hu, Andrew Horner, and Qiang Yang. 2013. Smart: Semi-supervised music emotion recognition with social tagging. In *SIAM International Conference on Data Mining*, pages 279–287. SIAM.

Linhong Xu and Hongfei Lin. 2008. Constructing the Affective Lexicon Ontology [J]. *Journal of the China Society for Scientific and Technical Information*, 2:006.

Chengzhi Zhang. 2008. Automatic keyword extraction from documents using conditional random fields. *Journal of Computational Information Systems*, 4(3):1169–1180. 00063.