Expert Systems with Applications 42 (2015) 1083–1093

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A multi-label classification based approach for sentiment classification

Shuhua Monica Liu *, Jiun-Hung Chen [1]

Department of Public Administration, Fudan University, Shanghai 200433, China

ARTICLE INFO

Article history:
Available online 17 September 2014

Keywords:
Sentiment analysis
Multi-label classification
Microblogs

ABSTRACT

A multi-label classification based approach for sentiment analysis is proposed in this paper. To the best of our knowledge, this work is the first to propose to use multi-label classification for sentiment classification of microblogs. The proposed prototype has three main components, text segmentation, feature extraction, and multi-label classification.

Raw segmented words and sentiment features based on the three different sentiment dictionaries, Dalian University of Technology Sentiment Dictionary, National Taiwan University Sentiment Dictionary and HowNet Dictionary, are the features and the bag of words is the feature representation.

A detailed empirical study of different multi-label classification methods on sentiment classification is conducted to compare their classification performances. Specifically, total 11 state of the art multi-label classification methods are compared on two microblog datasets and 8 evaluation metrics are used. The effects of the three sentiment dictionaries for multi-label classification are empirically studied and compared, which, to the best of our knowledge, have not been performed. The performed empirical comparisons show that Dalian University of Technology Sentiment Dictionary has the best performance among the three different sentiment dictionaries.

1. Introduction

Sentiment analysis which extracts emotions from natural language texts is an active research area and has been extensively applied to different applications such as understanding customer feedbacks and public opinions, monitoring real-world events and financial prediction (Bollen, Mao, & Zeng, 2011; Brown et al., 2011; Liu & Zhang, 2012; O'Connor, Balasubramanyan, Routledge, & Smith, 2010; Pang & Lee, 2008; Tang, Tan, & Cheng, 2009).

Lexicon-based approaches (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011) and machine learning based approaches (Pang, Lee, & Vaithyanathan, 2002) are two common approaches for sentiment classification. In contrast with most machine learning based methods which rely on either binary classification algorithms (Go, Bhayani, & Huang, 2009; Pang et al., 2002) or multiclass classification algorithms (Alm, Roth, & Sproat, 2005; Davidov, Tsur, & Rappoport, 2010; Purver & Battersby, 2012), a multi-label classification approach is proposed in this paper for sentiment classification of microblogs because some of the microblogs used in this paper have multiple emotional states. Multi-label classification has been extensively studied and used in many applications (Tsoumakas, Katakis, & Vlahavas, 2010; Zhang & Zhou, 2013). For example, while Huang, Peng, Li, and Lee (2013) proposed a multi-task multi-label classification model for the sentiment and topic classification of tweets jointly, the sentiment classification of the tweets they used is a multiclass problem and not a multi-label problem. Specifically, each tweet in their work has exactly one of the three labels, positive, negative and neutral while in this paper the microblogs used can allow to have multiple emotion states because the sentiment classification problem of microblogs are is framed as a multi-label problem. Bhowmick, Basu, and Mitra (2009) used multi-label classification for sentiment classification of news sentences extracted from Times of India news paper archive. However, to the best of our knowledge, we have not seen that multi-label classifications has been used for sentiment classification of microblogs.

Our contributions are summarized below. First, to the best of our knowledge, this work is the first to propose to use multi-label classification for sentiment classification of microblogs. Second, a detailed empirical study of different multi-label classification methods on sentiment classification is conducted to compare their classification performances. Specifically, total 11 state of the art multi-label classification methods are compared on two microblog datasets and 8 evaluation metrics are used. In addition, to the best of our knowledge, this work is the first to study and compare the

* Corresponding author.
E-mail addresses: Shuhua.Monica.Liu@gmail.com (S.M. Liu), JiunHung.Chen@gmail.com (J.-H. Chen).

[1] The work was done when this author visited the first author's research center at Fudan University.

http://dx.doi.org/10.1016/j.eswa.2014.08.036
0957-4174/

effects of the three sentiment dictionaries, Dalian University of Technology Sentiment Dictionary (DUTSD) (Xu, Lin, Pan, Ren, & Chen, 2008), National Taiwan University Sentiment Dictionary (NTUSD) (NTUSD, 2006), HowNet Dictionary (HD) (HowNet, 2007) for multi-label classification.

The paper is organized as follows. Reviews on sentiment analysis and multi-label classification are given in Section 2. The proposed system is presented in Section 3. Section 4 presents experimental results and discussions. The conclusions and future work are in Section 5.

2. Review

2.1. Sentiment classification

Sentiment analysis is a very active and important research area and has been extensively used to study microblogs and product reviews (Tang et al., 2009). Polarity classification which classifies the polarity of a given document is a basic task in sentiment analysis and there are three main types of approaches for polarity classification. First, lexicon-based approaches (Taboada et al., 2011; Turney, 2002) calculate the polarity of a document from the polarities of words or phrases in the document. Second, machine learning based approaches (Pang et al., 2002) build classifiers to determine the polarity of a document (Pang et al., 2002). The third approach is a hybrid approach which combines the first two approaches together.

Sentiment classification or emotion classification which classifies a document into emotional states such as "happy" and "angry" is a generalized form of polarity classification and the above three types of approaches can be easily adapted for solving sentiment classification.

In addition to common textual features, hashtags, smileys and emoticons are used as features and labels for sentiment analysis (Davidov et al., 2010; Go et al., 2009). Linguistic features such as negation (Wiegand, Balahur, Roth, Klakow, & Montoyo, 2010) have been used for sentiment analysis.

Because microblogs used in this paper can have multiple emotional states, the sentiment classification of microblogs is framed as a multi-label classification problem and not a multi-class nor a binary classification problem. To the best of our knowledge, two contributions in the area of sentiment classification of microblogs are made in the paper, which has not been done by previous studies. First, this work is the first to propose to use multi-label classification for sentiment classification of microblogs. Second, this work is the first to study and compare the effects of the three sentiment dictionaries, DUTSD, NTUSD, HD for multi-label classification to perform multi-label sentiment classification of microblogs.

2.2. Multi-label classification

In contrast with traditional single-label classification which is to classify an instance into one of labels, multi-label classification is to classify an instance into a set of labels. There are two main types of methods for multi-label classification, problem transformation methods and algorithm adaptation methods (Tsoumakas et al., 2010; Zhang & Zhou, 2013).

Problem transformation methods transform a multi-label classification problem into one or multiple single-label problems. In other words, at training time, transform the multi-label training data to single-label data and then learn a single-label classifier from the transformed data. At testing time, given a test instance, use the learned classifier to make a single-label prediction and then translate the prediction to a multi-label prediction. Binary relevance (BR) (Boutell, Luo, Shen, & Brown, 2004) transforms a multi-label learning problem to multiple binary classification problems with one-vs-all strategy where each binary classification problem handles whether an instance belongs to a particular label in the label space or not. In contrast, label powerset (LP) (Tsoumakas, Katakis, & Vlahavas, 2011a) considers each element in the power set of the label set in as a class and hence transforms a multi-label classification problem to a multi-class classification problem. Random k-Labelsets (Tsoumakas et al., 2011a) (RAkEL) constructs an ensemble of LP classifiers where each LP classifier is trained on a different random subset of the set of labels and combines the output from the ensemble via a voting scheme to make a final multi-label prediction. Homer (Tsoumakas, Katakis, & Vlahavas, 2008) builds a hierarchy of multi-label classifiers and gains the advantages that each multi-label classifier in the hierarchy handles a much smaller set of labels and has a more balanced example distribution. The balanced clustering problem in HOMER is solved with balanced k means.

To model interdependencies between labels, classifier chains (CC) (Read, Pfahringer, Holmes, & Frank, 2011) transform a multi-label classification problem into a chain of binary classification problems where the size of the chain is equal to the number of labels. Each classifier in the chain handles a binary relevance problem associated with a unique label and its feature space is augmented with all the predictions of its preceding classifiers in the chain. The ordering of labels in the chain can have a strong effect of the performances of classifier chains. To deal with the effect of ordering, ensemble of classifier chains (ECC) builds multiple CCs with different orderings of labels and a threshold is used to select most popular labels to form the final predicted labels. Probabilistic classifier chains (Cheng, Hüllermeier, & Dembczynski, 2010) require each classier in the chain to output a probability that a corresponding label is true and use the output probabilities as extra features instead of the binary predictions as extra features.

Several approaches based on stacked aggregation are proposed for multi-label classification (Godbole & Sarawagi, 2004; Montañés et al., 2014; Montañés, Quevedo, & del Coz, 2011). In Godbole and Sarawagi (2004), in the first stage, binary relevance is used to the labels for an input example. All the binary classification outputs (i.e., whether the input has a particular label or not for all labels in the label space) in the first stage are stacked with the original feature space of the input example. The stacked output then serve as an input to all meta classifiers in the second stage where the number of meta classifiers is the same as the number of labels and a meta classifier is responsible for deciding a unique label. The binary classification outputs in the second stage are the final results for the input example. Dependent binary relevance models (Montañés et al., 2014) are the same as (Godbole & Sarawagi, 2004) except that during the training phase, the ground truth labels are used instead of predicted labels. In contrast with Godbole and Sarawagi (2004), a different way to stack features for a meta classifier, using ground truth labels during the training phase and using the results from both stages to form final results (e.g., using or rule) are proposed in (Montañés, Quevedo, & del Coz, 2011).

Calibrated label ranking transforms a multi-label classification problem into a label ranking problem with pairwise comparisons and artificial calibration labels which separate the relevant labels from the irrelevant ones (Fürnkranz, Hüllermeier, Mencía, & Brinker, 2008).

Algorithm adaptation methods generalize existing classification algorithms to handle multi-label data. For example, multi-label *k*-nearest neighbor algorithm (MLkNN) (Zhang & Zhou, 2007) modifies *k*-nearest neighbor (kNN) algorithm to handle multi-label data and uses maximum a posteriori rule to make multi-label prediction. BRkNN (Spyromitros, Tsoumakas, & Vlahavas, 2008) is adapted kNN algorithm for multi-label classification with the concept of BR. BRKNN-a and BRKNN-b are two extensions of the basic

BRkNN with different calculations of confidence scores for each label obtained from the basic BRkNN. Multi-label decision tree algorithm (Clare & King, 2001) modifies C4.5 decision tree algorithm to deal with multi-label data and defines a new entropy function for multi-label data. Rank support vector machine (SVM) (Elisseeff & Weston, 2001) extends SVMs to handle multi-label data and aims to minimizes a ranking loss with a large margin. Predictive clustering trees (PCT) (Blockeel, De Raedt, & Ramon, 1998) are decision trees like algorithms where the root corresponds to the largest cluster which contains all data and a parent's cluster is recursively partitioned to smaller clusters its children correspond to and can handle different types of structured outputs such as multi-labels and time series. Random forests of predictive clustering trees (RF-PCT) (Kocev, Vens, Struyf, & Džeroski, 2007) use random forests and majority voting to combine multiple PCTs to form a final decision.

It is recommend to use RF-PCT, HOMER, BR, CC as benchmark methods for multi-label learning because of their overall best performances in a recent extensive experimental comparison of multi-label classification methods (Madjarov, Kocev, Gjorgjevikj, & Džeroski, 2012) but none of the text datasets is a microblog dataset. One contribution in the area of multi-label classification made in this work is to conduct an empirical study of different multi-label classification methods on multi-label sentiment classification and to compare their classification performances, which has not been done by previous studies.

3. Prototype

The proposed prototype is consisted of three main components, text segmentation, features, and multi-label classification.

3.1. Text segmentation

Text segmentation is the process of break a text into meaningful units such as words or sentences. A python package, Jieba (Sun, 2012), is used for text segmentation in this paper. Its algorithm is to identify the maximum tangential points based on word frequency combination. To handle unknown words in Jieba, a hidden Markov model based approach is used and is solved with the Viterbi algorithm (Bishop & Nasrabadi, 2006).

3.2. Feature extraction and representation

Two types of features are used in this paper. Given a segmented word, the first type of features is sentiment features which are the combination of the sentiment, the sentiment strength and sentiment polarity of the word obtained from a sentiment dictionary if any. Three sentiment dictionaries, DUTSD, NTUSD, HD[2] are used for this purpose.[3] However, because the three sentiment dictionaries do not cover all possible sentiment words, raw segmented words are also used as the second type of features. The common bag of words representation is used to represent these features.

3.3. Multi-label classification

Because a microblog used in this paper can potentially have multiple sentiments, the classification problem is framed as a multi-label classification problem. Due to the lacking of previous work on multi-label classification performances for sentiment classification of microblogs, one research question of interest is how different multi-label classification methods perform for sentiment classification of microblogs and to find the one with the best performance. In order to answer this research question, an empirical study of different multi-label classification methods on sentiment classification is conducted to compare their multi-label classification performances.

Eight out of twelwe state of the art multi-label classification methods compared in Madjarov et al. (2012), BR, CC, CLR, HOMER, RAkEL, ECC, MLkNN, and RF-PCT, are selected including the suggested ones with best performances, RF-PCT, HOMER, BR and CC. BRkNN, BRkNN-a, BRkNN-b are also added for comparisons. Hence, total 11 multi-classification methods are compared in the study. The selection of these 11 algorithms covers various algorithm adaptation methods and problem transformation methods for multi-label learning.

Table 1
The numbers of parts associated with different numbers of sentiments on HR.

Number of sentiments	Number of parts
1	365
2	17
3	3

4. Experiments and discussions

4.1. Data sets

To understand public sentiments from microblogs about major incidents, microblogs about two major incidents in China, "2013 Huangpu River dead pigs incident"[4] (HR) and "2013 Influenza A virus subtype H7N9"[5] (IA) are collected in this paper. The reasons why they are selected are that they gained national attentions and a large amount of microblogs were discussed about them. For HR, 385 parts of microblogs are collected and manually encoded with a set of sentiments which are selected from these five sentiments, worry, anger, helplessness, disgust and surprise. For IA, 762 parts of microblogs are collected and manually encoded with a set of sentiments which are selected from these 10 sentiments, worry, anger, fear, surprise, hope, rejoicing, helplessness, indifference, blessing and despair. Tables 1 and 2 show the number of parts associated with different numbers of sentiments while Tables 3 and 4 show the numbers of parts associated with different sentiments. Each dataset is partitioned into two disjoint sets of equal size, one for training and one for testing. In other words, the numbers of training and testing examples are 192 and 193 on HR respectively while both of them are 381 in IA.

Two examples of text segmentation of microblogs are shown below. "一死几千头大家都懂得" (Several thousands of pigs died once and all knew what happened.) ⇒ "(一, m) (死, v) (几千头, m) (,x) (大家, n) (都, d) (懂得, v)". "好恐怖！最近不要碰家禽类！" (It is very horrible. Do not eat poultry recently.) ⇒ "(好, a) (恐怖, a) (！, x) (最近, f) (不要, df) (碰, v) (家禽类, n) (！, x)". The first microblog is encoded with worry while the second one is encoded with worry and fear.

4.2. Evaluation metrics

Eight evaluation metrics are selected from the sixteen ones used in a recent extensive experimental comparison of multi-label classification methods (Madjarov et al., 2012). The 8 evaluation

[2] Only the positive and negative sentiment keywords are used.
[3] Note that NTUSD and HD only have sentiment polarity while DUTSD have sentiment, sentiment strength and sentiment polarity.
[4] More details about this incident can be found via this wiki page, http://en.wikipedia.org/wiki/2013_Huangpu_River_dead_pigs_incident.
[5] More details about this incident can be found via this wiki page, http://en.wikipedia.org/wiki/Influenza_A_virus_subtype_H7N9#Reported_cases_in_2013.

Table 2
The numbers of parts associated with different numbers of sentiments on IA.

Number of sentiments	Number of parts
1	632
2	122
3	7
4	1

Table 3
The numbers of parts associated with different sentiments on HR.

Sentiment	Number of parts
担忧 (worry)	194
愤怒 (anger)	126
无奈 (helplessness)	38
厌恶 (disgust)	23
惊讶 (surprise)	27

Table 4
The numbers of parts associated with different sentiments on IA.

Sentiment	Number of parts
担忧 (worry)	354
愤怒 (anger)	47
害怕 (fear)	184
惊讶 (surprise)	71
期望 (hope)	58
庆幸 (rejoicing)	25
无奈 (helplessness)	68
无所谓 (indifference)	32
祝福 (blessing)	41
失望 (despair)	21

metrics can be grouped into three types, example based metrics, label-based metrics and ranking based metrics and the implementations of these metrics in Mulan (Tsoumakas, Spyromitros-Xioufis, Vilcek, & Vlahavas, 2011b) are used. Three example based metrics, Hamming loss, subset accuracy, and example based F1, two label based metrics, micro F1 and macro F1 and three ranking based metrics, average precision, coverage and one error are used.[6] A brief description about these metrics are given below and refer to Madjarov et al. (2012) and Zhang and Zhou (2013) for more details.

Hamming loss is the fraction of labels that are incorrectly predicted for a sample and the normalized Hamming loss which is normalized over samples is reported. Its value range is between 0 and 1. The smaller the value of Hamming loss, the better the performance is. Subset accuracy is defined as whether the set of labels predicted for a sample exactly matches the corresponding set of ground truth labels for the sample and the fraction of correctly classified samples based on the above definition of subset accuracy is reported. Its value range is between 0 and 1. The higher the value of subset accuracy, the better the performance is. Example based F1 is the average of the harmonic mean of example-precision and example-recall for every example. The example-precision is defined for an example as the size of the intersection of the set of its predicted labels and the set of its ground truth labels divided by the size of the set of its predicted labels. The example-recall is defined for an example as the size of the intersection of the set of its predicted labels and the set of its ground truth labels divided by the size of the set of its ground truth labels. Its value range is between 0 and 1. The higher the value of example based F1, the better the performance is.

Micro F1 is the harmonic mean of micro-precision and micro-recall where micro-precision and micro-recall are the precision and the recall which are averaged over all the example and label pairs. The value of micro F1 is between 1 and 0. The higher the value of micro F1, the better the performance is. Macro F1 is average of the harmonic mean of precision and recall over all labels and the precision and the recall are defined for each label. The value of macro F1 is between 1 and 0. The higher the value of macro F1, the better the performance is.

Average precision is the average fraction of labels which are ranked higher an actual label belonging to an example. The value of average precision is between 0 and 1. The higher the value of average precision, the better the performance is. Coverage is the average number of steps to move down from the top of the list of the ranked labels to cover all the actual labels for an example. The higher the value of coverage, the better the performance is. One error is the fraction of examples whose top-ranked label does not belong to the relevant label set for the example. The value of one error is between 0 and 1. The smaller the value of one error, the better the performance is.

4.3. Experimental setup

In order to provide a thorough and rigorous analysis, the experimental setup closely follows the one used in a recent extensive experimental comparison of multi-label classification methods (Madjarov et al., 2012).

Total 11 multi-classification methods are compared in the study, BR, CC, CLR, HOMER, RAkEL, ECC, MLkNN, RF-PCT, BRkNN, BRkNN-a and BRkNN-b where the first 8 ones are among the 12 state of the art multi-label classification methods compared in Madjarov et al. (2012) and RF-PCT, HOMER, BR, CC are the ones recommended to use as benchmark methods for multi-label learning because of their overall best performances (Madjarov et al., 2012). The implementation of RF-PCT and PCT in CLUS[7] and the implementations of BR, CC, CLR, HOMER, RAkEL, ECC, MLkNN, BRkNN, BRkNN-a, BRkNN-b in Mulan (Tsoumakas et al., 2011b) are used in the comparisons.

4.3.1. Base classifiers and model selection

While RF-PCTs use PCTs from CLUS as base classifiers, SVMs with radial basis kernels are used as base classifiers for BR, CC, CLR, HOMER, RAkEL and ECC and the SVM implementation in LibSVM (Chang & Lin, 2011) is used. In contrast, MLkNN, BRkNN, BRkNN-a and BRkNN-b all use kNN as a base classifier.

For a multi-label classification method and a dataset, model selection for selecting the key parameters of the method is performed by using 3-fold cross-validation with only training data and the metric to optimize in model selection is example-based F1. The parameters with the best metric value found in model selection are used to train the method with the training data from the dataset to obtain a multi-label classifier and the learned classifier is tested with the testing data from the dataset.

For BR, CC, CLR, RAkEL and ECC and each dataset, model selection for selecting the kernel parameter and the penalty is performed. The parameter value ranges, $[10^{-3}, 10^{-2}, \ldots, 10^{1}]$ and $[10^{-2}, 10^{-1}, \ldots, 10^{4}]$, similar to the ones used in Madjarov et al. (2012), are considered for the kernel parameter and the penalty, respectively. For HOMER, in addition to the two parameters for SVM, the number of clusters is also considered in model selection

[6] The output prediction files from CLUS are used in computing the metric values. However, the ranking based metrics can not be computed for CLUS due to the lacking of enough details in CLUS user's manual about the output prediction files which are required in computing the metric values.

[7] http://clus.sourceforge.net.

and the three parameters jointly are optimized during model selection for HOMER. The parameter range, [2,3,...,6], suggested in Madjarov et al. (2012), is considered for the number of clusters for HOMER. For MLkNN, BRkNN, BRkNN-a and BRkNN-b, the parameter range, [6,8,...,20], suggested in Madjarov et al. (2012), is considered for the number of neighbors in kNN during the model selection for them.

For RF-PCT, the significance level for the F-test used in the pre-pruning strategy of PCTs is automatically selected from $[0.001, 0.005, 0.01, 0.05, 0.1, 0.125]$ with a reserved prune-set which contains 25% of the training set and is not used for training.

4.3.2. Parameter setting

The following parameters of the methods are set as ones used in Madjarov et al. (2012), which also follows the recommendation from the previous literature. For RF-PCT, the number of models is 100 and the size of the feature subsets is $\lfloor 0.1 * D + 1 \rfloor$ where D is the dimensionality of the input space. The number of models used in ECC and RAkEL are 10 and $min(2 * Q, 100)$, respectively where Q is the number of labels in a dataset. For RAkEL, the size of the label-sets is set as half of the number of labels in a dataset.

4.4. Experiments

There are two research questions to be answered by the experiments. The first one is how different multi-label classification methods perform for sentiment classification of microblogs and what methods perform best with raw segmented words. The second one is what the impacts of different features on the performances of different multi-label classification methods and how different sentiment dictionaries perform for sentiment classification of microblogs. The two research questions will be answered in the following two subsections, respectively.

4.4.1. Performance comparisons with raw segmented words

In this experiment, only raw segmented words is used. The testing performances of different methods on HR and IA are shown in Tables 5 and 6, respectively. While RAkEL and ECC are clearly the best and second best methods on HR, respectively, CLR, HOMER and ECC performed best on IA. In contrast, BRkNN performed worst on both HR and IA.

The effect of different types of base classifiers are noticeable on IA. Method using kNN as base classifiers performed poorly in both example-based metrics and label-based metrics. However, no such patterns are observed on HR.

ECC performed better than CC across all the metrics in both datasets except macro F1 in IA. BRkNN-a and BRkNN-b, effectively improved almost all metrics in both datasets for BRkNN. All the best metric values except Hamming loss on HR are better than those on IA indicates that IA is a harder problem than HR.

Table 5
The testing performances of different methods on HR. For a metric, the best and the second best values are in bold and in italic, respectively. HL, SA, EF, MiF, MaF, AP, C and OE, are shorthand for Hamming loss, subset accuracy, example based F1, micro F1, macro F1, average precision, coverage and one error, respectively.

Method	HL	SA	EF	MiF	MaF	AP	C	OE
BR	0.212	0.466	0.480	0.478	0.131	0.615	1.767	0.513
CC	0.212	0.466	0.480	0.478	0.131	0.615	1.767	0.513
CLR	0.212	0.466	0.480	0.478	0.131	0.700	0.922	0.513
HOMER	0.188	0.508	*0.579*	0.568	0.388	0.687	1.399	0.425
RAkEL	**0.150**	*0.518*	*0.579*	**0.621**	**0.495**	*0.755*	1.073	**0.337**
ECC	*0.167*	**0.523**	**0.582**	*0.594*	*0.425*	**0.766**	**0.824**	*0.373*
MLkNN	0.198	0.399	0.413	0.459	0.131	0.700	*0.891*	0.518
BRkNN	0.204	0.026	0.026	0.048	0.027	0.562	1.611	0.648
BRkNN-a	0.196	0.497	0.518	0.519	0.226	0.702	0.886	0.508
BRkNN-b	0.217	0.446	0.466	0.468	0.202	0.702	0.886	0.508
RF-PCT	0.192	0.508	0.522	0.524	0.282	n.a	n.a	n.a

Table 6
The testing performances of different algorithms on IA. For a metric, the best and the second best values are in bold and in italic, respectively. HL, SA, EF, MiF, MaF, AP, C and OE, are shorthand for Hamming loss, subset accuracy, example based F1, micro F1, macro F1, average precision, coverage and one error, respectively.

Method	HL	SA	EF	MiF	MaF	AP	C	OE
BR	**0.103**	0.268	0.368	0.474	0.251	0.476	4.412	0.604
CC	0.106	0.278	0.394	0.483	**0.274**	0.495	4.278	0.580
CLR	*0.104*	0.265	0.372	0.478	*0.258*	**0.655**	**1.766**	*0.496*
HOMER	0.118	*0.339*	**0.494**	**0.519**	0.252	0.566	3.396	0.517
RAkEL	*0.104*	0.289	0.397	0.484	0.180	0.555	3.753	0.499
ECC	**0.103**	0.286	0.417	*0.509*	0.266	0.625	2.761	**0.480**
MLkNN	0.109	0.142	0.208	0.326	0.098	*0.627*	*2.068*	0.520
BRkNN	0.118	0.045	0.080	0.144	0.053	0.482	3.108	0.709
BRkNN-a	0.168	0.097	0.202	0.236	0.050	0.491	2.727	0.751
BRkNN-b	0.162	0.123	0.231	0.263	0.083	0.477	2.916	0.719
RF-PCT	0.118	**0.344**	*0.445*	0.458	0.131	n.a	n.a	n.a

4.4.2. Impacts of different features

To study the effects of different features on the classification performances of different algorithms, the performances of an algorithm with different features are compared in this experiment. There are 11 features, raw segmented words (R), the sentiment features based on DUTSD (SD), the sentiment features based on NTUSD (SN), the sentiment features based on HD (SH), and all combinations of SD, SN, SH with R, R + SD, R + SN, R + SH, R + SD + SN, R + SD + SH, R + SN + SH, R + SD + SN + SH. Tables 7–12 show the effects of different features on the testing performances of different algorithms.

Several key findings can be found from different comparisons. The comparisons among SD, SN and SH show that SD is the overall best single feature. The better performance for SD can be attributed to that DUTSD has much more sentiment keywords, sentiments and sentiment information (e.g, sentiment strength) than the others and hence finds more discriminant information than the others. In fact, BRkNN, BRkNN-a and BRkNN-b performed best with SD among all 11 features for all metrics and all data sets except BRkNN with SN performed best for SA on IA. However, there is a noticeable exception that MLkNN with SN performed best for 4 metrics on both datasets.

The comparisons among R + SD, R + SN, R + SH show that R + SD is the overall best. There are some exceptions for both datasets. For example, MLkNN with R + SN and BRkNN-a with R + SN performed better for most metrics on HR while BR with R + SH and RAkEL with R + SN or R + SH together performed better for many metrics on IA.

For kNN based multi-label classifiers, SD is a better feature than R + SD on HR. However, this observation hold true for BRkNN and its variants on IA but not for MLkNN. In contrast, R + SD is a overall better feature than SD for the rest of the multi-label classifiers. There are no performance changes when BR, CLR, CC and ECC use SN or R + SN on HR. In contrast, HOMER, RAkEL, RP-PCT performed better with R + SN than with SN on HR. While BRkNN and its variants preferred SN than R + SN on IA, the rest of the methods overall preferred R + SN than SN on IA. On HR, there are no performance changes for when BR, CC and CLR use SH or R + SH on HR. In contrast, HOMER, RAkEL, ECC and RP-PCT performed better with R + SH than with SH on HR. On IA, R + SH is a overall better feature than SH for most methods except BRkNN-a and BRkNN-b for most metrics. From the above comparisons, BRkNN and its variants in many cases may not be able to use additional discriminative information when R is combined with SD, SN or SH and in contrast, most of the rest methods can effectively do so.

Table 7
The testing performances of different algorithms with different features on HR. The table is divided into three sections. The first section is for Hamming loss, the second is for subset accuracy and the third is for example based F1. For each row, the best and the second best values are in bold and in italic, respectively.

Method	R	SD	SN	SH	R + SD	R + SN	R + SH	R + SD + SN	R + SD + SH	R + SN + SH	R + SD + SN + SH
BR	0.212	0.165	0.212	0.212	0.153	0.212	0.212	0.152	*0.150*	0.212	**0.147**
CC	0.212	0.165	0.212	0.212	0.150	0.212	0.212	**0.141**	0.149	0.212	*0.148*
CLR	0.212	0.166	0.212	0.212	0.153	0.212	0.212	0.153	*0.149*	0.212	**0.148**
HOMER	0.188	0.167	0.212	0.212	**0.160**	0.187	0.179	0.167	*0.162*	0.183	0.178
RAkEL	0.150	0.171	0.212	0.212	*0.149*	0.158	0.161	**0.147**	0.152	0.159	*0.149*
ECC	0.167	0.169	0.212	0.212	**0.145**	0.212	0.173	*0.146*	0.148	0.212	0.152
MLkNN	0.198	**0.191**	0.208	0.207	0.196	0.198	0.204	0.194	*0.192*	0.209	0.216
BRkNN	0.204	**0.195**	*0.197*	0.211	0.201	0.206	0.204	0.201	0.200	0.206	0.200
BRkNN-a	**0.196**	**0.185**	0.203	0.215	0.215	0.202	0.231	0.215	0.212	0.215	0.212
BRkNN-b	0.217	**0.188**	0.206	0.212	*0.200*	0.223	0.225	*0.200*	0.206	0.231	0.206
RF-PCT	0.192	0.196	0.212	0.212	0.181	0.187	0.187	**0.172**	*0.175*	0.196	**0.172**
BR	0.466	**0.544**	0.466	0.466	0.482	0.466	0.466	0.482	0.508	0.466	*0.513*
CC	0.466	*0.580*	0.466	0.466	0.554	0.466	0.466	0.565	0.570	0.466	**0.585**
CLR	0.466	**0.544**	0.466	0.466	0.477	0.466	0.466	0.477	*0.503*	0.466	*0.503*
HOMER	0.508	*0.570*	0.466	0.466	**0.575**	0.508	0.518	0.560	0.565	0.503	0.497
RAkEL	0.518	**0.560**	0.466	0.466	0.549	0.508	0.492	*0.554*	*0.554*	0.513	*0.554*
ECC	0.523	0.560	0.466	0.466	0.575	0.466	0.518	*0.580*	**0.585**	0.466	0.560
MLkNN	0.399	0.368	0.425	0.104	0.212	0.394	0.415	0.223	**0.440**	0.420	*0.425*
BRkNN	0.026	**0.487**	0.332	*0.466*	0.036	0.021	0.026	0.036	0.041	0.021	0.041
BRkNN-a	*0.497*	**0.534**	0.477	0.461	0.461	0.482	0.409	0.461	0.461	0.451	0.466
BRkNN-b	0.446	**0.528**	0.477	0.461	*0.497*	0.430	0.430	*0.497*	0.482	0.415	0.482
RF-PCT	0.508	0.461	0.466	0.466	0.513	0.513	0.492	**0.539**	0.513	0.461	*0.534*
BR	0.480	**0.573**	0.480	0.480	0.513	0.480	0.480	0.523	0.535	0.480	*0.542*
CC	0.480	0.598	0.480	0.480	0.582	0.480	0.480	*0.599*	0.598	0.480	**0.603**
CLR	0.480	**0.573**	0.480	0.480	0.518	0.480	0.480	0.523	*0.544*	0.480	0.541
HOMER	0.579	0.607	0.480	0.480	**0.620**	0.575	0.579	0.611	*0.616*	0.561	0.569
RAkEL	0.579	0.577	0.480	0.480	0.594	0.546	0.551	**0.609**	0.592	0.565	*0.608*
ECC	0.582	0.591	0.480	0.480	0.617	0.480	0.566	**0.625**	*0.623*	0.480	0.601
MLkNN	0.413	0.385	*0.453*	0.111	0.230	0.408	0.432	0.237	0.468	0.440	**0.456**
BRkNN	0.026	**0.501**	0.359	*0.480*	0.036	0.021	0.026	0.036	0.041	0.021	0.041
BRkNN-a	*0.518*	**0.547**	0.508	0.475	0.475	0.503	0.430	0.475	0.478	0.472	0.480
BRkNN-b	0.466	**0.542**	0.494	0.478	*0.511*	0.451	0.447	*0.511*	0.496	0.432	0.496
RF-PCT	0.522	0.475	0.480	0.480	0.530	0.527	0.506	**0.556**	0.527	0.475	*0.547*

Table 8
The testing performances of different algorithms with different features on HR. The table is divided into two sections. The first section is for micro F1 and the second is for macro F1. For each row, the best and the second best values are in bold and in italic, respectively.

Method	R	SD	SN	SH	R + SD	R + SN	R + SH	R + SD + SN	R + SD + SH	R + SN + SH	R + SD + SN + SH
BR	0.478	0.589	0.478	0.478	0.582	0.478	0.478	0.591	*0.594*	0.478	**0.603**
CC	0.478	0.595	0.478	0.478	0.615	0.478	0.478	**0.636**	0.621	0.478	*0.623*
CLR	0.478	0.588	0.478	0.478	0.587	0.478	0.478	0.591	*0.602*	0.478	**0.604**
HOMER	0.568	0.600	0.478	0.478	**0.617**	0.567	0.579	0.604	*0.614*	0.565	0.576
RAkEL	0.621	0.578	0.478	0.478	0.625	0.589	0.593	**0.636**	0.616	0.598	*0.631*
ECC	0.594	0.589	0.478	0.478	*0.637*	0.478	0.577	**0.639**	0.634	0.478	0.620
MLkNN	0.459	0.452	**0.472**	0.180	0.327	0.456	0.463	0.335	0.501	0.463	*0.467*
BRkNN	0.048	**0.510**	0.431	*0.480*	0.067	0.039	0.048	0.067	0.077	0.039	0.077
BRkNN-a	*0.519*	**0.545**	0.508	0.473	0.473	0.504	0.433	0.473	0.478	0.473	0.478
BRkNN-b	0.468	**0.539**	0.494	0.478	*0.509*	0.453	0.448	*0.509*	0.494	0.433	0.494
RF-PCT	0.524	0.496	0.478	0.478	0.543	0.534	0.524	**0.568**	0.549	0.496	*0.563*
BR	0.131	0.504	0.131	0.131	0.498	0.131	0.131	*0.512*	0.505	0.131	**0.523**
CC	0.131	0.495	0.131	0.131	0.546	0.131	0.131	0.529	*0.547*	0.131	**0.550**
CLR	0.131	0.503	0.131	0.131	0.500	0.131	0.131	*0.512*	0.509	0.131	**0.523**
HOMER	0.388	0.476	0.131	0.131	**0.494**	0.399	0.410	0.479	*0.488*	0.403	0.464
RAkEL	0.495	0.398	0.131	0.131	0.513	0.404	0.451	*0.529*	0.511	0.442	**0.533**
ECC	0.425	0.407	0.131	0.131	0.509	0.131	0.401	0.507	**0.523**	0.131	*0.519*
MLkNN	0.131	0.150	0.150	0.064	*0.148*	0.131	0.131	0.150	**0.178**	0.134	0.139
BRkNN	0.027	**0.252**	*0.156*	0.131	0.037	0.022	0.027	0.037	0.042	0.022	0.042
BRkNN-a	*0.226*	**0.329**	0.197	0.156	0.157	0.220	0.186	0.157	0.151	0.207	0.155
BRkNN-b	*0.202*	**0.324**	0.190	0.142	0.184	0.196	0.193	0.184	0.170	0.187	0.170
RF-PCT	0.282	0.138	0.131	0.131	0.302	0.308	0.303	*0.333*	0.310	0.168	**0.336**

The performances of R + SD + SN on both datasets are overall better than those of R + SN, which clearly indicates that some key information captured in SD but not in R + SN plays an important role in the discrimination. A similar explanation can also explain why the performance of R + SD + SH on HR is overall better than that of R + SH on HR. In contrast, no similar observations are found in the comparisons between R + SD + SN and R + SD and the comparisons between R + SD + SH and R + SD and the comparisons among R + SN, R + SH and R + SN + SH. In fact, most methods preferred R + SD than R + SD + SH on IA.

The comparisons between R + SN + SH and R + SD + SN + SH show that R + SD + SN + SH overall performed better than

Table 9
The testing performances of different algorithms with different features on HR. The table is divided into three sections. The first section is for average precision, the second is for coverage and the third is for one error. For each row, the best and the second best values are in bold and in italic, respectively. Note that RF-PCT is not listed because of no metric values computed.

Method	R	SD	SN	SH	R + SD	R + SN	R + SH	R + SD + SN	R + SD + SH	R + SN + SH	R + SD + SN + SH
BR	0.615	*0.678*	0.615	0.615	0.644	0.615	0.615	0.647	0.661	0.615	**0.669**
CC	0.615	0.696	0.615	0.615	0.694	0.615	0.615	*0.708*	0.707	0.615	**0.711**
CLR	0.700	0.756	0.700	0.700	*0.794*	0.700	0.700	0.791	**0.796**	0.700	0.792
HOMER	0.687	0.703	0.615	0.615	**0.714**	0.688	0.688	0.706	*0.710*	0.674	0.671
RAkEL	**0.755**	0.689	0.615	0.615	*0.743*	0.718	0.733	0.741	*0.743*	0.728	0.742
ECC	0.766	0.723	0.615	0.643	0.790	0.615	0.759	**0.799**	0.794	0.615	*0.795*
MLkNN	0.700	**0.727**	0.703	0.699	0.717	*0.721*	0.699	0.718	0.716	0.705	0.699
BRkNN	0.562	**0.728**	*0.709*	0.702	0.610	0.559	0.566	0.606	0.607	0.562	0.606
BRkNN-a	0.702	**0.745**	0.709	0.693	*0.721*	0.699	0.692	*0.721*	0.711	0.689	0.711
BRkNN-b	0.702	**0.745**	0.709	0.693	*0.721*	0.699	0.692	*0.721*	0.711	0.689	0.711
BR	1.767	**1.508**	1.767	1.767	1.601	1.767	1.767	1.565	1.523	1.767	**1.503**
CC	1.767	1.435	1.767	1.767	1.430	1.767	1.767	**1.321**	*1.352*	1.767	*1.352*
CLR	0.922	0.725	0.922	0.922	*0.591*	0.922	0.922	0.622	**0.580**	0.922	0.601
HOMER	1.399	1.358	1.767	1.767	**1.311**	1.389	1.394	1.342	*1.316*	1.451	1.435
RAkEL	**1.073**	1.446	1.767	1.767	1.145	1.244	1.176	1.140	1.140	1.244	*1.135*
ECC	0.824	1.171	1.767	1.492	0.725	1.767	0.798	0.684	*0.679*	1.767	**0.653**
MLkNN	0.891	**0.762**	0.938	0.927	0.860	0.850	0.907	0.860	*0.829*	0.902	0.896
BRkNN	1.611	**0.813**	*0.886*	0.896	1.316	1.627	1.611	1.321	1.326	1.611	1.321
BRkNN-a	0.886	**0.756**	0.886	0.933	*0.855*	0.891	0.912	*0.855*	0.881	0.917	0.881
BRkNN-b	0.886	**0.756**	0.886	0.933	*0.855*	0.891	0.912	*0.855*	0.881	0.917	0.881
BR	0.513	**0.425**	0.513	0.513	0.477	0.513	0.513	0.477	0.456	0.513	*0.440*
CC	0.513	0.399	0.513	0.513	0.399	0.513	0.513	*0.389*	*0.389*	0.513	**0.378**
CLR	0.513	0.415	0.513	0.513	0.358	0.513	0.513	**0.352**	*0.358*	0.513	*0.358*
HOMER	0.425	0.404	0.513	0.513	**0.383**	0.425	0.425	*0.394*	*0.394*	0.446	0.456
RAkEL	**0.337**	0.415	0.513	0.513	*0.352*	0.389	0.368	0.358	0.358	0.363	0.358
ECC	0.373	0.404	0.513	0.513	*0.342*	0.513	0.399	**0.332**	*0.342*	0.513	0.347
MLkNN	0.518	*0.482*	0.503	0.513	*0.482*	**0.477**	0.518	*0.482*	0.487	0.503	0.518
BRkNN	0.648	**0.466**	*0.497*	0.513	0.627	0.653	0.637	0.637	0.632	0.648	0.637
BRkNN-a	0.508	**0.440**	0.497	0.523	*0.472*	0.513	0.528	*0.472*	0.492	0.534	0.492
BRkNN-b	0.508	**0.440**	0.497	0.523	*0.472*	0.513	0.528	*0.472*	0.492	0.534	0.492

Table 10
The testing performances of different algorithms with different features on IA. The table is divided into three sections. The first section is for Hamming loss, the second is for subset accuracy and the third is for example based F1. For each row, the best and the second best values are in bold and in italic, respectively.

Method	R	SD	SN	SH	R + SD	R + SN	R + SH	R + SD + SN	R + SD + SH	R + SN + SH	R + SD + SN + SH
BR	0.103	0.126	0.120	0.119	0.103	*0.102*	**0.101**	*0.102*	0.104	0.104	*0.102*
CC	0.106	0.150	0.178	0.188	**0.103**	0.106	0.105	0.105	**0.103**	0.106	0.104
CLR	0.104	0.127	0.118	0.119	0.103	0.102	*0.101*	**0.100**	0.105	0.107	0.103
HOMER	0.118	0.129	0.124	0.126	**0.112**	0.118	0.120	*0.113*	0.115	0.121	0.114
RAkEL	*0.104*	0.115	0.124	0.124	0.104	0.104	**0.103**	0.105	0.107	*0.104*	0.107
ECC	0.103	0.125	0.125	0.124	*0.102*	**0.100**	**0.100**	*0.102*	0.104	*0.102*	*0.102*
MLkNN	**0.109**	0.119	0.126	0.122	*0.110*	0.117	0.112	**0.109**	0.112	0.112	0.114
BRkNN	0.118	**0.112**	0.115	0.118	0.115	0.117	0.118	*0.115*	*0.115*	0.117	*0.115*
BRkNN-a	0.168	**0.122**	*0.124*	0.126	0.134	0.168	0.150	0.139	0.139	0.155	0.132
BRkNN-b	0.162	**0.120**	*0.124*	*0.124*	0.136	0.169	0.152	0.133	0.137	0.160	0.137
RF-PCT	0.118	**0.110**	0.120	0.120	0.115	0.115	0.117	0.114	0.115	0.117	0.116
BR	*0.268*	0.163	0.110	0.050	0.262	0.262	**0.278**	0.262	0.252	0.255	0.252
CC	*0.278*	0.228	0.152	0.113	0.276	0.276	**0.281**	0.270	0.268	0.276	0.265
CLR	*0.265*	0.165	0.110	0.045	0.255	0.260	**0.273**	0.262	0.247	0.247	0.252
HOMER	0.339	0.299	0.315	0.302	0.354	0.336	0.318	**0.360**	*0.354*	0.320	0.349
RAkEL	0.289	**0.325**	*0.315*	*0.315*	0.310	0.291	0.302	0.320	0.302	0.307	0.312
ECC	0.286	*0.312*	*0.312*	**0.315**	0.297	0.302	0.291	0.304	0.286	0.283	0.304
MLkNN	0.142	0.076	*0.168*	0.003	0.134	0.047	0.084	**0.178**	0.160	0.084	0.147
BRkNN	0.045	*0.110*	**0.115**	0.047	0.094	0.034	0.052	0.094	0.087	0.045	0.089
BRkNN-a	0.097	**0.331**	*0.315*	0.304	0.268	0.097	0.189	0.255	0.249	0.168	0.278
BRkNN-b	0.123	**0.328**	*0.315*	*0.315*	0.260	0.092	0.186	0.270	0.255	0.131	0.255
RF-PCT	0.344	0.094	0.000	0.000	0.325	*0.346*	0.325	0.344	**0.349**	0.331	0.341
BR	0.368	0.288	0.147	0.073	**0.388**	0.366	0.383	**0.388**	*0.384*	0.369	0.382
CC	**0.394**	0.338	0.190	0.149	**0.394**	0.384	*0.392*	0.386	0.391	0.383	0.387
CLR	0.372	0.281	0.147	0.072	*0.396*	0.370	0.387	**0.399**	0.387	0.365	0.390
HOMER	0.494	0.441	0.421	0.417	**0.507**	0.492	0.486	*0.499*	0.493	0.481	0.486
RAkEL	0.397	**0.447**	*0.421*	*0.421*	0.417	0.400	0.408	0.423	0.406	0.409	0.416
ECC	0.417	**0.441**	0.417	0.421	0.428	0.416	0.420	0.431	0.425	0.425	*0.431*
MLkNN	0.208	0.113	*0.247*	0.015	0.197	0.090	0.129	**0.251**	0.223	0.129	0.216
BRkNN	0.080	**0.158**	*0.145*	0.073	0.132	0.058	0.090	0.134	0.126	0.082	0.130
BRkNN-a	0.202	**0.445**	*0.421*	0.420	0.373	0.203	0.297	0.351	0.349	0.271	0.381
BRkNN-b	0.231	**0.448**	*0.421*	*0.421*	0.365	0.196	0.286	0.377	0.360	0.239	0.360
RF-PCT	0.445	0.149	0.000	0.000	0.432	0.446	0.425	0.457	**0.464**	0.435	*0.457*

Table 11
The testing performances of different algorithms with different features on IA. The table is divided into two sections. The first section is for micro F1 and the second is for macro F1. For each row, the best and the second best values are in bold and in italic, respectively.

Method	R	SD	SN	SH	R + SD	R + SN	R + SH	R + SD + SN	R + SD + SH	R + SN + SH	R + SD + SN + SH
BR	0.474	0.383	0.216	0.124	*0.496*	0.476	0.491	**0.498**	0.494	0.478	0.494
CC	0.483	0.367	0.191	0.155	**0.498**	0.478	0.488	0.486	0.495	0.481	*0.492*
CLR	0.478	0.377	0.218	0.134	*0.505*	0.482	0.497	**0.512**	0.497	0.472	0.499
HOMER	0.519	0.474	0.436	0.432	**0.534**	0.519	0.514	*0.526*	0.521	0.510	0.517
RAkEL	0.484	0.491	0.435	0.435	*0.492*	0.483	0.489	**0.493**	0.478	0.484	0.483
ECC	0.509	0.472	0.431	0.435	0.518	0.512	**0.520**	0.518	0.512	*0.519*	0.517
MLkNN	0.326	0.181	0.314	0.033	0.310	0.165	0.220	**0.370**	0.331	0.222	*0.341*
BRkNN	0.144	**0.252**	*0.217*	0.135	0.215	0.108	0.158	0.215	0.206	0.149	0.212
BRkNN-a	0.236	**0.461**	0.435	*0.436*	0.391	0.237	0.321	0.368	0.368	0.296	0.399
BRkNN-b	0.263	**0.464**	*0.435*	*0.435*	0.382	0.229	0.308	0.394	0.378	0.270	0.378
RF-PCT	0.458	0.265	0.000	0.000	0.479	0.470	0.449	**0.495**	*0.492*	0.456	0.487
BR	0.251	0.250	0.047	0.029	0.280	0.253	0.268	**0.296**	0.279	0.272	*0.285*
CC	0.274	0.265	0.080	0.072	0.278	0.270	0.277	0.276	*0.289*	0.274	**0.294**
CLR	0.258	0.238	0.047	0.034	*0.292*	0.251	0.285	**0.310**	0.284	0.268	0.291
HOMER	0.252	0.108	0.067	0.068	**0.283**	0.255	0.260	0.257	*0.277*	0.245	0.255
RAkEL	0.180	0.150	0.065	0.065	0.178	0.184	**0.203**	0.179	0.172	*0.191*	0.174
ECC	0.266	0.230	0.064	0.065	**0.277**	0.251	0.256	**0.277**	0.266	*0.275*	0.272
MLkNN	0.098	0.044	0.052	0.013	0.093	0.072	0.087	**0.135**	*0.118*	0.082	0.108
BRkNN	0.053	**0.114**	0.043	0.033	*0.075*	0.045	0.057	0.074	0.072	0.055	0.074
BRkNN-a	0.050	**0.175**	0.065	0.070	0.154	0.055	0.126	0.141	0.141	0.122	*0.158*
BRkNN-b	0.083	**0.168**	0.065	0.065	*0.151*	0.052	0.146	0.125	0.150	0.095	0.150
RF-PCT	0.131	0.073	0.000	0.000	0.164	0.144	0.107	*0.166*	**0.168**	0.129	0.161

Table 12
The testing performances of different algorithms with different features on IA. The table is divided into three sections. The first section is for average precision, the second is for coverage and the third is for one error. For each row, the best and the second best values are in bold and in italic, respectively. Note that RF-PCT is not listed because of no metric values computed.

Method	R	SD	SN	SH	R + SD	R + SN	R + SH	R + SD + SN	R + SD + SH	R + SN + SH	R + SD + SN + SH
BR	0.476	0.402	0.302	0.243	0.485	0.474	**0.488**	*0.485*	0.482	0.475	0.482
CC	**0.495**	0.430	0.316	0.278	0.490	0.488	*0.494*	0.483	0.484	0.486	0.483
CLR	0.655	0.607	0.587	0.603	0.664	0.649	0.661	*0.664*	*0.664*	0.654	**0.673**
HOMER	0.566	0.526	0.508	0.504	**0.581**	0.566	0.561	*0.575*	0.569	0.555	0.560
RAkEL	**0.555**	0.542	0.509	0.509	0.546	0.552	0.544	*0.554*	0.543	0.551	0.539
ECC	**0.625**	0.594	0.531	0.548	0.607	0.611	0.609	0.612	0.611	**0.617**	0.609
MLkNN	0.627	0.618	0.585	0.595	0.628	0.623	*0.635*	0.627	0.624	**0.642**	0.623
BRkNN	0.482	**0.631**	0.598	*0.603*	0.568	0.485	0.497	0.572	0.571	0.496	0.568
BRkNN-a	0.491	**0.615**	0.599	*0.603*	0.559	0.501	0.497	0.558	0.556	0.496	0.555
BRkNN-b	0.477	**0.631**	0.599	*0.610*	0.559	0.501	0.497	0.576	0.556	0.484	0.555
BR	4.412	4.913	5.879	6.283	**4.226**	4.415	4.289	*4.249*	4.255	4.378	4.289
CC	4.278	4.829	5.840	6.147	**4.181**	4.323	4.268	4.273	*4.252*	4.331	4.299
CLR	1.766	2.160	2.601	2.304	**1.651**	1.814	1.722	1.711	1.619	1.803	*1.627*
HOMER	*3.396*	3.735	4.283	4.294	**3.370**	3.428	3.436	3.462	3.507	3.459	3.583
RAkEL	**3.753**	**3.759**	4.299	4.299	3.874	3.806	3.877	3.803	3.919	**3.753**	3.890
ECC	**2.761**	2.900	3.900	3.693	2.961	2.921	3.047	2.898	2.950	*2.824*	2.945
MLkNN	*2.068*	2.186	2.430	2.362	2.163	2.110	2.105	2.126	2.160	**2.066**	2.139
BRkNN	3.108	**2.247**	2.593	*2.341*	2.727	2.898	2.982	2.696	2.756	2.979	2.748
BRkNN-a	2.727	**2.370**	2.609	2.341	2.924	*2.648*	2.982	2.929	2.929	2.979	2.937
BRkNN-b	2.916	**2.247**	2.596	*2.286*	2.924	2.648	2.982	2.672	2.929	2.890	2.937
BR	*0.604*	0.724	0.814	0.903	0.609	0.609	**0.593**	0.606	0.612	0.612	0.612
CC	**0.580**	0.675	0.790	0.848	0.598	0.591	*0.585*	0.606	0.606	0.596	0.606
CLR	0.496	0.556	0.538	0.535	0.496	0.504	*0.488*	*0.488*	0.501	0.496	**0.483**
HOMER	0.517	0.559	0.525	0.533	**0.488**	0.517	0.525	**0.488**	*0.499*	0.533	0.504
RAkEL	0.499	0.525	0.522	0.522	0.501	*0.496*	0.509	**0.491**	0.504	0.507	0.517
ECC	**0.480**	0.533	0.528	0.522	0.499	0.491	**0.480**	0.493	0.491	*0.488*	*0.488*
MLkNN	0.520	0.528	0.567	0.551	0.514	0.525	*0.499*	0.520	0.522	**0.496**	0.525
BRkNN	0.709	**0.499**	*0.522*	0.535	0.593	0.706	0.701	0.588	0.585	0.703	0.593
BRkNN-a	0.751	**0.520**	*0.522*	0.535	0.588	0.743	0.701	0.591	0.596	0.703	0.598
BRkNN-b	0.719	**0.499**	*0.522*	0.525	0.588	0.743	0.701	0.580	0.596	0.714	0.598

R + SN + SH because of the addition of SD. The comparisons between R + SD + SN + SH and R on HR show that using sentiment features from all the three dictionaries together improved the performances for most methods except HOMER and BRkNN-a. The comparisons between R + SD + SN + SH and R on IA show that the performance gains due to using sentiment features from all the three dictionaries together exist for the example and label metrics for most methods but not for ranking based metrics.

From the above empirical comparisons, for kNN based multi-label classification methods, SD overall worked best on both datasets. For the rest of the methods, R + SD + SN, R + SD, R + SH and R are the overall the best features on HR while R + SD + SN + SH, R + SD + SN and R + SD are overall the best features on IA. In addition, DUTSD has the best multi-label classification performance among the three different sentiment dictionaries. The best performances for DUTSD can be attributed to that DUTSD has much more

sentiment keywords, sentiments and sentiment information (e.g, sentiment strength) than the others and hence finds more discriminant information than the others. On HR, ECC is overall the best while MLkNN and BRkNN was the worst. On IA, CLR and HOMER are better than the rest of the methods for most metrics but CLR has weak performances in subset accuracy and example based F1 and HOMER has weak performances in Hamming loss, average precision and coverage.

4.5. Discussions

The empirical comparisons clearly showed the importance of features on the multi-label classification performances for every method considered here and the performance gains can be significant. For example, the Hamming loss of ECC is reduced from 0.167 (with R) to 0.145 (R + SD) and while the subset accuracy is improved from 0.523 to 0.585. Their performance gains are 13.2% and 11.9% in this example, respectively. However, except for BRkNN and its variants, there does not exist a single feature which can help each method to achieve its best performance in every metric in the experiments. In most cases, different methods achieved their best performances in different metrics with different combinations of single features.

Although RF-PCT performed best in the benchmarks reported in Madjarov et al. (2012), it did not performed competitively in both experiments. This situation may be attributed to the following two hypotheses. First, the default parameter settings for RF-PCT suggested in Madjarov et al. (2012) may not necessarily work for the both datasets used in this paper. Second, in Madjarov et al. (2012), none of the text datasets is a microblog dataset. Hence, it is not guaranteed that RF-PCT can perform well in the experiments with microblogs.

Building and using intelligent tools in social media for sentiment analysis is very important and very relevant to the research community of expert systems with application (Tang et al., 2009; Tan & Zhang, 2008). To the best of our knowledge, using multi-label classification for sentiment classification of microblogs is proposed in this work, which can handle microblogs with multiple sentiment labels and has not been done by previous studies. The proposed multi-label sentiment classification of microblogs can be very helpful for building intelligent tools in social media for sentiment analysis, especially for microblogs with multiple sentiment labels because this work relaxes the current constraints of using either multiclass or binary classification for sentiment classification of microblogs and instead uses multi-label classification as a generalized solution. In addition, studying and comparing the effects of the three sentiment dictionaries, DUTSD, NTUSD, HD for multi-label sentiment classification of microblogs and conducting an empirical study and comparison of different multi-label classification methods on multi-label sentiment classification of microblog have not been done by previous studies. The experimental results clearly show that this work is a working prototype as an intelligent tool for multi-label classification for sentiment classification of microblogs. In addition, the empirical studies, comparisons and findings (e.g, DUTSD has the best performance among the three sentiment dictionaries) can serve as a good reference and provide practical guidelines like the contributions made in Tan and Zhang (2008) for the research community.

The microblogs collected in this paper are about two major incidents in China, "2013 Huangpu River dead pigs incident" and "2013 Influenza A virus subtype H7N9" which both gained national attentions. The collected datasets by themselves are valuable research resources for public administration and crisis management to understand public sentiments from microblogs about these major incidents. For example, Tables 3 and 4 clearly show that worry (50%) and anger (33%) are the two major sentiments about the first incident while worry (46%) and fears (24%) are two major sentiments about the second incident. To the best of our knowledge, these statistics and distributions from microblogs about these two incidents have not be reported in the research literature and they can be valuable for researchers and practitioners in public administration and crisis management. In addition, the datasets can be very useful for further analysis about different research questions such as how sentiments developed over time and what the effects of genders and ages on sentiments about these incidents are. Because public administration and crisis management are two key application areas for the research community of expert systems with application, the collected datasets are very relevant and useful for the research community.

5. Conclusions and future work

A multi-label classification based approach for sentiment analysis is proposed in this paper. The proposed prototype has three main components, text segmentation, feature extraction, and multi-label classification. The features used in this paper included raw segmented words and sentiment features based on the three different sentiment dictionaries, DUTSD, NTUSD and HD, and the bag of words is used for feature representation.

A detailed empirical study of different multi-label classification methods for multi-label sentiment classification of microblogs is conducted to compare their classification performances. Specifically, total 11 state of the art multi-label classification methods BR, CC, CLR, HOMER, RAkEL, ECC, MLkNN, and RF-PCT, BRkNN, BRkNN-a and BRkNN-b are compared on two microblog datasets. The selection not only covers various algorithm adaptation methods and problem transformation methods for multi-label classification but also includes eight of those compared and all the suggested ones with best performances in a recent extensive experimental comparison (Madjarov et al., 2012). In addition, 8 evaluation metrics are selected from the 16 ones used in Madjarov et al. (2012) and cover several example based metrics, label-based metrics and ranking based metrics.

The experiment with raw segmented words show that while RAkEL and ECC are clearly the best and second best methods on HR respectively, CLR, HOMER and ECC performed best on IA. In contrast, BRkNN performed worst on both HR and IA. Several key findings are observed in experiments with different features. For example, DUTSD has the best multi-label classification performance among the three different sentiment dictionaries. The better performance for SD can be attributed to that DUTSD has much more sentiment keywords, sentiments and sentiment information than the others. On HR, ECC is overall the best while MLkNN and BRkNN was the worst. On IA, CLR and HOMER are better than the rest of the methods for most metrics. However, CLR has weak performances in subset accuracy and example based F1 and HOMER has weak performances in Hamming loss, average precision and coverage. In addition, that RF-PCT did not perform competitively in the experiments potentially contradicted the finding that RF-PCT performed best in the study (Madjarov et al., 2012).

Three contributions are made in this paper. First, to the best of our knowledge, this work is the first to propose to use multi-label classification for sentiment classification of microblogs. The contribution is significant for microblogs with multiple sentiment labels because this work relaxes the current constraints of using either multiclass or binary classification for sentiment classification of microblogs and instead uses multi-label classification as a generalized solution. Second, an detailed study and performance comparisons of different multi-label classification methods for sentiment classification of microblogs were conducted in this work, which has not been done by previous studies. Third, an

empirical comparison of the effects of the three sentiment dictionaries, DUTSD, NTUSD, HD for multi-label classification was conducted in this paper, which has not been studied previously. The experimental findings found in this work can be helpful for the research community and practitioners of expert systems with application, especially for those building and using intelligent tools in social media for sentiment analysis (Tang et al., 2009; Tan & Zhang, 2008).

There are three important future directions and the research results obtained in these research directions can be highly useful for building and using intelligent tools for sentiment analysis in social media. First, Tables 7–12 clearly show that features play a meaningful role in the multi-label classification performances. Hence, exploring new features such as n-grams, different linguistic structures and dependency structures to identify useful linguistic and semantic relationships will be a key future direction. In addition, the sentiment dictionaries used in this paper are predefined and not customized for the datasets of interest. Instead of using predefined sentiment dictionaries, learning a customized sentiment dictionary from the datasets can be consequential.

Second, conducting an extensive experimental comparison like (Madjarov et al., 2012) for multi-label sentiment classification of microblogs will be investigated in the future. There are two reasons why this research is worth further investigation. The first reason is that none of the text datasets contains microblogs in Madjarov et al. (2012) and their recommendations may not be applicable to multi-label sentiment classification of microblogs. In fact, one empirical finding in the experiments that RF-PCT did not perform competitively potentially contradicted the finding that RF-PCT performed best in the study (Madjarov et al., 2012). The second reason is that to the best of our knowledge, no previous extensive experimental comparison for multi-label sentiment classification of microblogs has been done. Hence, several key research questions such as how different multi-label classification methods perform and what the best multi-label classification method is for multi-label sentiment classification of microblogs have not been completely and thoroughly answered although this work has provided some initial answers and has been one step closer to the answers. These answers can play a substantial role in building intelligent tools for multi-label sentiment analysis in social media.

Third, to collect more microblogs about different incidents with multi-label sentiments for further research is a significant future direction. For example, one major research question to answer is whether multi-label sentiment classification of microblogs is incident-dependent or incident-indecent. In other words, whether a learned multi-label sentiment classifier learned from a microblog dataset about an incident can be applied to solve the multi-label sentiment classification problem of a new microblog dataset about another incident. Another research question to answer is that whether there exist any correlations among different sentiments in microblogs and how to use them if any to help multi-label sentiment classification. It is not possible to answer these research questions without collecting more microblogs about different incidents with multi-label sentiments. In addition, the collected datasets by themselves can be valuable research resources for public administration and crisis management to understand public sentiments from microblogs about these major incidents. Furthermore, this direction is a precondition for conducting an extensive experimental comparisons for multi-label sentiment classification of microblogs. It is hence believed that this research direction is a significant future direction for the research community of expert systems with application, especially for researchers and practitioners who are interested in multi-label sentiment analysis in social media, public administration and crisis management.

Acknowledgement

We want to thank multiple funding agencies for their generous support that made this research possible: China National Social Science Foundation (Project ID: KRB3056068), China Ministry of Education (Project ID: JJH3056017), China Ministry of Science and Technology (Project ID: 201508016), Shanghai Pujiang Program (Project ID: KBH3056609), Shanghai Municipal Advisory Committee of Decision Support (Project ID: KEH3056089), and The Dr. Seaker Chan Center for Comparative Political Development at Fudan University (Project ID: CCPDS-FudanNDKT13035).

References

Alm, C. O., Roth, D., & Sproat, R. (2005). Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* (pp. 579–586). Association for Computational Linguistics.

Bhowmick, P. K., Basu, A., & Mitra, P. (2009). Reader perspective emotion analysis in text through ensemble based multi-label classification framework. *Computer and Information Science, 2*, 64–74.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 1). New York: Springer.

Blockeel, H., De Raedt, L., & Ramon, J. (1998). Top-down induction of clustering trees. In *Proceedings of the 15th international conference on machine learning* (pp. 55–63).

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*, 1–8.

Boutell, M. R., Luo, J., Shen, X., & Brown, C. M. (2004). Learning multi-label scene classification. *Pattern Recognition, 37*, 1757–1771.

Brown, C., Frazee, J., Beaver, D., Liu, X., Hoyt, F., & Hancock, J. (2011). Evolution of sentiment in the libyan revolution. White Paper at <https://webspace.utexas.edu/dib97/libya-report-10-30-11.pdf>.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*, 27:1–27:27<http://www.csie.ntu.edu.tw/simcjlin/libsvm>.

Cheng, W., Hüllermeier, E., & Dembczynski, K. J. (2010). Bayes optimal multilabel classification via probabilistic classifier chains. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (pp. 279–286).

Clare, A., & King, R. D. (2001). Knowledge discovery in multi-label phenotype data. In *Principles of data mining and knowledge discovery* (pp. 42–53). Springer.

Davidov, D., Tsur, O., & Rappoport, A. (2010). Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd international conference on computational linguistics: Posters* (pp. 241–249). Association for Computational Linguistics.

Elisseeff, A., & Weston, J. (2001). A kernel method for multi-labelled classification. In *Advances in neural information processing systems* (pp. 681–687).

Fürnkranz, J., Hüllermeier, E., Mencía, E. L., & Brinker, K. (2008). Multilabel classification via calibrated label ranking. *Machine Learning, 73*, 133–153.

Godbole, S., & Sarawagi, S. (2004). Discriminative methods for multi-labeled classification. In *Advances in knowledge discovery and data mining* (pp. 22–30). Springer.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford (pp. 1–12).

HowNet (2007). HowNet. <http://www.keenage.com/html/c_bulletin_2007.htm>.

Huang, S., Peng, W., Li, J., & Lee, D. (2013). Sentiment and topic analysis on social media: A multi-task multi-label classification approach. In *Proceedings of the 5th annual ACM web science conference* (pp. 172–181).

Kocev, D., Vens, C., Struyf, J., & Džeroski, S. (2007). Ensembles of multi-objective decision trees. In *Proceedings of the 18th European conference on machine learning* (pp. 624–631).

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data* (pp. 415–463). Springer.

Madjarov, G., Kocev, D., Gjorgjevikj, D., & Džeroski, S. (2012). An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition, 45*, 3084–3104.

Montañés, E., Quevedo, J. R., & del Coz, J. J. (2011). Aggregating independent and dependent models to learn multi-label classifiers. In *Machine learning and knowledge discovery in databases* (pp. 484–500). Springer.

Montañes, E., Senge, R., Barranquero, J., Ramón Quevedo, J., José del Coz, J., & Hüllermeier, E. (2014). Dependent binary relevance models for multi-label classification. *Pattern Recognition, 47*, 1494–1508.

NTUSD (2006). National Taiwan University Semantic Dictionary. <http://nlg18.csie.ntu.edu.tw:8080/lwku/pub1.html>.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM, 11*, 122–129.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on empirical methods in natural language processing* (Vol. 10, pp. 79–86).

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*, 1–135.

Purver, M., & Battersby, S. (2012). Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th conference of the European chapter of the association for computational linguistics* (pp. 482–491). Association for Computational Linguistics.

Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2011). Classifier chains for multi-label classification. *Machine Learning, 85*, 333–359.

Spyromitros, E., Tsoumakas, G., & Vlahavas, I. (2008). An empirical study of lazy multilabel classification algorithms. In *Proceedings of the 5th Hellenic conference on artificial intelligence: Theories models and applications* (pp. 401–406). Springer-Verlag.

Sun, J. (2012). Jieba. <https://github.com/fxsjy/jieba>.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics, 37*, 267–307.

Tang, H., Tan, S., & Cheng, X. (2009). A survey on sentiment detection of reviews. *Expert Systems with Applications, 36*, 10760–10773.

Tan, S., & Zhang, J. (2008). An empirical study of sentiment analysis for chinese documents. *Expert System with Applications, 34*, 2622–2629.

Tsoumakas, G., Katakis, I., & Vlahavas, I. (2008). Effective and efficient multilabel classification in domains with large number of labels. In *Proceedings of the ECML/PKDD 2008 workshop on mining multidimensional data (MMD'08)* (pp. 30–44).

Tsoumakas, G., Katakis, I., & Vlahavas, I. (2010). Mining multi-label data. In *Data mining and knowledge discovery handbook* (pp. 667–685). Springer.

Tsoumakas, G., Katakis, I., & Vlahavas, I. (2011a). Random k-labelsets for multilabel classification. *IEEE Transactions on Knowledge and Data Engineering, 23*, 1079–1089.

Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., & Vlahavas, I. (2011b). Mulan: A java library for multi-label learning. *Journal of Machine Learning Research, 12*, 2411–2414.

Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 417–424). Association for Computational Linguistics.

Wiegand, M., Balahur, A., Roth, B., Klakow, D., & Montoyo, A. (2010). A survey on the role of negation in sentiment analysis. In *Proceedings of the workshop on negation and speculation in natural language processing* (pp. 60–68).

Xu, L., Lin, H., Pan, Y., Ren, H., & Chen, J. (2008). Constructing the affective lexicon ontology. *Journal of The China Society For Scientific And Technical Information, 27*, 180–185<http://ir.dlut.edu.cn/EmotionOntologyDownload.aspx> .

Zhang, M.-L., & Zhou, Z.-H. (2007). Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition, 40*, 2038–2048.

Zhang, M.-L., & Zhou, Z.-H. (2013). A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*.