Comparison of Feature Selection Techniques for Multi-label Text Classification against a New Semantic-based Method

Wael Alkhatib, Steffen Schnitzer, Wei Ding, Peter Jiang, Yassin ALkhalili, and Christoph Rensing

Technische Universität Darmstadt, Fachgebiet Multimedia Kommunikation, S3/20,
Rundeturmstr. 10, 64283 Darmstadt, Germany
{wael.alkhatib,steffen.schnitzer,christoph.rensing}@kom.tu-darmstadt.de

Abstract. The under-explored research area of multi-label text classification has led to substantial amount of research in adapting feature selection techniques to handle multi-label data directly. A wide range of statistical techniques have been proposed for weighting and selecting features in order to reduce the high dimensionality of feature space. Those techniques suffer from losing semantic regularities of concepts as features and ignoring the dependencies and ordering between adjacent words. In this work, we undertake a comparative study across a set of statistical and semantic-based techniques for feature selection. Moreover, we propose a novel approach incorporating the text semantics in feature selection using typed dependencies. Our intensive experiments, using the EUR-lex dataset, showed that incorporating text semantics in feature selection can significantly improve the performance of multi-label classifiers. Moreover, it drastically decrease the computation costs by reducing the feature space. The experiments approved that our method applied to a combination of typed dependencies outperformed the state-of-the-art techniques for feature selection in terms of F1-measure.

Keywords: semantics; statistics; feature selection; dimensionality reduction; text classification; typed dependencies.

1 Introduction

Text classification has become a widespread problem in natural language processing and information retrieval as a result of the tremendous growth of data, most of which are unstructured [1]. Classification problems deal with the task of assigning a number of classes C out of a predefined set of classes L to an input. Such problems can either be binary, multi-class or multi label[**?**]. Binary classification is the problem of assigning one out of two labels meaning that $|C| = 1$ and $|L| = 2$. A problem where the task is to assign exactly one class C out of $|L|$ mutually exclusive classes to an input is called multi-class, while a classification problem is called a multi-label classification problem when the task is to classify the input into $m = |C|$ out of the set of classes L where $m \leq |L|$.

The classification strategies that deal with multi-label problems fall into two groups namely, transformation and adaptation methods. Transformation methods transform a multi-label learning problem into one or more single-label problems. Adaptation methods adapt or extend single-label classifiers to cope with multi-label data. Essentially for both categories, text representation is an essential preprocessing step where documents are transformed into a format consumable by machine learning models. This involves representing each document as a vector of words as features, where each dimension corresponds to the relevance of a word to the document [2]. Relevance can for example be computed using weighting schema i.e. TF-IDF. In general this method produces high dimensional, sparse vectors which are extremely challenging for learning algorithms. To increase the manageability of the problem, machine learning techniques apply a process called dimensionality reduction which aims at reducing redundancy and noise in the data set by mapping it into a lower dimensional space using a wide range of feature selection and extraction techniques.

This work is an extension of our previous work [3] on incorporating semantic knowledge into feature selection for dimensionality reduction with the novelty of better capability of identifying relations between candidate features even with more natural text. Using linguistic filters we extract all noun phrases to provide a terminology of basic and extended concepts. Then we extract semantic relations between the noun phrases based on the typed dependencies in order to build an undirected graph as a basic shallow ontology between the concepts. Relying on the shallow ontology of syntactic relations can drastically lower the computation costs for feature selection with regard to the statistical techniques. Using the undirected graph of concepts, we propose new method to select the features based on a combination of relationship between concepts from the typed dependencies. The empirical evaluation results showed that selecting the features based on the typed dependencies outperforms the state-of-the-art techniques for feature selection using statistics and semantic-based techniques.

The paper is organized as follow: An overview of related works in feature selection for text classification is provided in Sect. 2. We introduce our concept for the semantic-based feature selection in Sect. 3. Section 4 presents the evaluation objectives and demonstrates the comparative analysis of the proposed method against a wide range on feature selection techniques. Finally, Sect. 6 summarizes the paper and discusses future work.

2 Foundations and Related Works

Feature selection handles the problem of selecting a subset of features that is most effective for building a good predictor. This can be done by statistics or semantic-based measures [4]. In the following, we introduce a variety of methods which fall into these two categories and we relate them to our methodology.

2.1 Statistics-based Feature Selection

The more widely used feature selection methods are the statistics-based [5–7]. In this study, we consider four widely used techniques, namely information gain, information gain ratio, chi-squared statistic, and correlation.

Information Gain: Information gain is a well-established technique for term goodness criterion. It measure the level of impurity present in the information and filters out the variables or terms based on entropy. Let $\{c_1\}_{i=1}^{m}$ denotes the set of a label in the target space. The information gain of term t (words or phrase) is defined as:

$$\begin{array}{l} G(t) = -\sum_{i=1}^{m} P_r(c_i)\ logP_r(c_i) \\ +P_r(t) + \sum_{i=1}^{m} P_r(c_i|t)\ logP_r(c_i|t) \\ +P_r(\bar{t}) + \sum_{i=1}^{m} +P_r(c_i|\bar{t})\ logP_r(c_i|\bar{t}) \end{array} \tag{1}$$

This general form of information gain definition is used in order to measure the goodness of a term globally with the respect to all labels on average [8].

Information Gain Ratio: Information gain or mutual information is another measure for the goodness of a term. The estimated information gain between a term t and a label c is defined to be:

$$G(t,c) \approx log\frac{A \times N}{(A+C) \times (A+B)} \tag{2}$$

Where N is the number of documents, A is the co-occurrences of t and c, B is the occurrences of t without c, and C is the occurrences of c without.

Chi-squared Statistic: The Chi-squared statistic measures the independence between a term and a label. The term-goodness measure is defined to be:

$$X^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \tag{3}$$

Where A is the co-occurrences of t and c, B is the occurrences of c without t, D is the number of times neither c nor t occurs, and N is the number of documents. The main difference to Information gain ratio is that CHI-squared is normalized value, thus its values are comparable across terms for the same label.

Correlation: In correlation, the term-goodness is computed by measuring the correlation with the label. Pearson's correlation coefficient (r) is a measure of the strength of the association between a term and a label.

The major drawback of these statistics-based feature selection methods is ignoring textual features dependencies, structure and ordering.

2.2 Semantic-based Feature Selection

Incorporating text semantics can provide better performance with regard to the used feature selection techniques. Masuyama et al. [9] analyzed the impact of selecting terms as features based on their part-of-speech (POS) specifically nouns, verbs, adjectives and adverbs. By analyzing the different combinations of these four categories, they found out that a much smaller feature set of nouns is able to perform better than other POS combined. D.D. Lewis used all noun phrases that occurred at least twice as feature phrases in text categorization [10]. After applying clustering of phrases and words, he concluded that phrases produce less effective representation than single words. Y. Liu et al. showed that using bi-gram and tri-gram to leverage context information of word depending on previous or next words can improve the performance, however, word sequence of more than 3 decreases the performance [11, 12]. A. Khan et al [13] used frequent sequence (MSF) for extracting associated frequent sentences and co-occurring terms. Also, they used WordNet [14], a lexical database, as a domain ontology to convert these terms to concepts and update the SVM with new feature weights . Other researchers incorporate the ontological knowledge for training-less ontology-based text classification or to provide meta-information for feature selection [15–17].

Previously, researchers have incorporated text semantics in feature selection by selecting noun phrases or n-grams as features, others tried to leverage external lexical databases mainly WordNet to enhance the performance more by selecting relevant concepts. However, extracting ontological associations using external lexical resources or patterns has shortcomings due to the small coverage of concepts for particular domains and thus less ontological entities can be acquired. In our previous work, we have proposed new methods for incorporating semantic knowledge into feature selection for dimensionality reduction [3]. In those methods, noun phrases, which appear in a taxonomic relation are automatically extracted using Hearst six patterns [18] for taxonomic relations. The results showed that the proposed Concept-Document Frequency (C-DF) method significantly outperformed the Bag-of-Word (BOW) frequency based feature selection method with term frequency/inverse document frequency (TF-IDF) for features weighting. However, the applied patterns might work with significantly lower precision for more natural texts, also the number of discovered taxonomic relations will be much lower.

In this work, we extend and improve these semantic-based methods further with a new approach using typed dependencies to extract syntactic relations between concepts, aiming to achieve better performance even with more natural text and lower computation costs for feature selection. We chose the performance of C-DF from our previous work [3] as a baseline to compare with.

3 Proposed Modification of Semantic-based Method

In the proposed method, we incorporate text semantics by taking context information and dependencies of words in consideration to select features. This

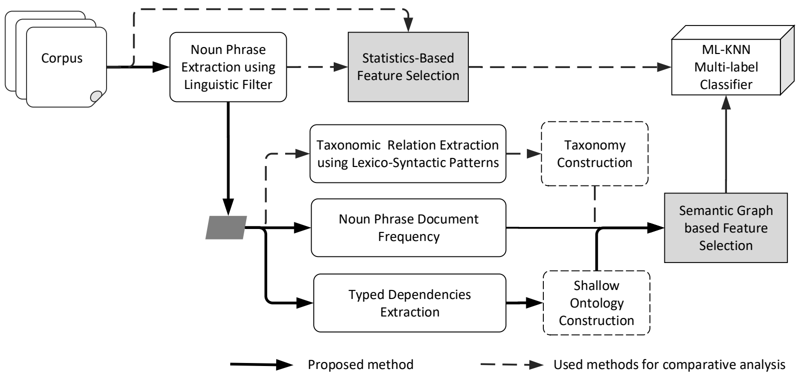
Corpus
Noun Phrase Extraction using Linguistic Filter
Statistics-Based Feature Selection
ML-KNN Multi-label Classifier
Taxonomic Relation Extraction using Lexico-Syntactic Patterns
Taxonomy Construction
Noun Phrase Document Frequency
Semantic Graph based Feature Selection
Typed Dependencies Extraction
Shallow Ontology Construction
Proposed method
Used methods for comparative analysis

Fig. 1. Block diagram of the proposed semantic-based feature selection method

process starts with selecting relevant concepts using a linguistic filter, then identifying semantic and syntactic relations between concept pairs based on the typed dependencies. By constructing a shallow ontology of the extracted relations between concepts as shown in Fig.2, different combinations of semantic relations between the candidate concepts can be used to select the features. Finally, we analyze the performance of the proposed technique and compare it against a wide range of statistics and semantics based feature selection techniques.

3.1 Linguistic Filter

In the first step we identify the domain terminology by extracting all noun phrases in order to form the basis for our semantic relation extraction phase. The role of the linguistic filter is to recognize essential concepts and filter out sequence of words that are unlikely to be concepts. In the linguistic component, the documents need to be preprocessed by a part-of-speech tagger for marking up the words in a text (corpus), based on their context, as corresponding to a particular part of speech i.e. noun, preposition, verb, etc. Multi-word NP like Supervised Machine Learning will be considered as one feature and concatenated as supervised_machine_learning. Then, words that are unlikely to be part of concepts are excluded using stop-words list. A combination of 3 linguistic filters is used to extract multi-word noun phrases NPs that can reflect essential concepts.

- $Noun\,Noun+$
- $Adj\ Noun+$
- $(Adj|\ Noun)+\ Noun$

3.2 Semantic Relation Extraction using Typed Dependencies

A triple based representation such as, *abbreviated relation name (governor, dependent)* is mentioned as the typed dependency relation between words from the same sentence [19]. The term *abbreviated relation name* represents the type of dependency relation between any two words in the sentence, "governor" and "dependent" simply represents the position of the words within the sentence. The parsing technique converts a sentence depending on the part-of-speech tagging of words and hierarchy of the typed dependencies into tree structure, then using this new representation, the syntactic relations on the sentence level can be identified to create the so called shallow ontology. Figure 2 illustrates a sample sentence and its corresponding typed dependencies graph.

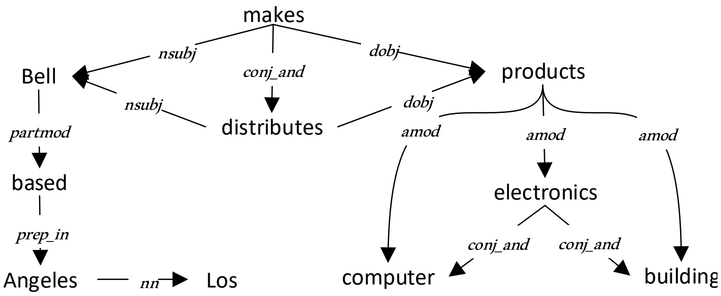

Fig. 2. Shallow ontology of dependencies for the sentence: *Bell, based in Los Angeles, makes and distributes electronic, computer and building products* [19].

There are different types of dependency systems available such as, Basic, Collapsed and Non-collapsed. We have considered the general collapsed dependencies, which represents the prepositions, conjunctions and other relative clause in a collapsed way to provide a direct relation between words. Using these dependencies, the syntactic features are defined and represented as triples of (governor, dependent, relations). A highlight in typed dependencies is meronomy. Meronymic relations are "part-whole" relations, where one entity is part or substance of another. Some dependencies like "including", "within", "involving", "inside", "containing" imply these kind of relations. *Berland & Charniak (1999)*[20] also tried using the dependency "of" to extract meronymic relations.

3.3 Semantic-Based Feature Selection

We propose a set of feature selection techniques based on the associations between the extracted concepts using the linguistic filter and the shallow ontology of typed dependencies.

- ***Concept Length***: Inspired by n-gram model, a noun phrase with multi-words may have different meaning or contain more specific information than when it is treated separately. For example, the length of noun phrase "French Financial Institutions" is 3. And "French Financial Institution" means a specific institution from a country, which is more specific and brings more information than "Institution".

- ***Typed Dependencies***: Typed dependencies between noun phrase represent the prepositional, conjunctional and verbal relations between words.

The shallow ontology provides the candidate features and their relations. Later, the semantic features can be selected based on two approaches, namely the Document Frequency (DF) and Concept Degree (CD). We define Concept Degree as the number of noun phrases connected to a specific noun phrase from the shallow ontology of typed dependencies, while Document Frequency represents the number of documents in which a term occurs. It is the simplest technique for feature selection. The basic heuristics behind using document frequency is that rare or non-frequent terms are non-informative for classification. Respectively, two weighting techniques, namely Binary weights and TF-IDF, will be used to weight the features with respect to the individual documents. TF-IDF is a global weighting method which reflects the importance of a word to a document in a corpus while Binary weighting is a local weighting method which reflects the presence of a word in a document.

4 Experimental Setup

In this section, we introduce the used classifier for multi-label classification, the evaluation metrics and the dataset.

4.1 ML-KNN Multi-label Classifier

Multi-label k Nearest Neighbors (ML-KNN) results from the modification of the k Nearest Neighbors (KNN) lazy learning algorithm using a Bayesian approach in order to deal with multi-label classification problems [21]. ML-KNN searches for the k nearest neighborhood of an input instance using KNN, then it calculates prior and posterior probabilities based on frequency counting of each label y in the set of labels L in order to determine the label set of the instance. This method has been selected to align the current experiments with our previous work.

4.2 Evaluation Metrics

A classifier can either be evaluated by examining each label separately and then averaging the results. Such schemes are called *label-based.* Another approach is by considering the average difference between the expected and the predicted sets of labels over all test examples, such metrics are called *example-based.*

For a number of classifier predictions, we have the number of *true positive(TP), false positive (FP), true negative (TN)* and *false negative (FN)* predictions respectively. From those numbers we can calculate the evaluation metrics mentioned below:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F - Measure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{6}$$

The total label-based evaluation measures for a multi-label problem where TP_j, FP_j, TN_j, FN_j are the predictions for the *j-th* label. A micro-averaged metric M_{micro} is defined as:

$$M_{micro} = M\big(\sum_{j=1}^{q} TP_j, \sum_{j=1}^{q} FP_j, \sum_{j=1}^{q} TN_j, \sum_{j=1}^{q} FN_j\big) \tag{7}$$

While macro-averaged metric M_{micro} is defined as:

$$M_{macro} = \frac{1}{q}\sum_{j=1}^{q} M\big(TP_j, FP_j, TN_j, FN_j\big) \tag{8}$$

In addition,the Hamming Loss which is the fraction of labels that are incorrectly predicted,is used in our evaluation:

$$HammingLoss(h, D) = \frac{1}{|D|}\sum_{i=1}^{|D|} \frac{xor(Y_i, Z_i)}{|L|} \tag{9}$$

D is the set of examples (x_i, Y_i) with $Y_i \subseteq L$ and Z_i is the predicted set of labels for x_i. Lower values of Hamming Loss indicate better performance.

Four performance metrics have been used in this work, namely Hamming loss, Macro/Micro-averaged F-Measure and Average Precision.

4.3 Dataset and Experimental Settings

In the context of our comparative analysis, the EUR-lex dataset has been used [22]. It is a text dataset containing European Union laws, treaties, international agreements, preparatory acts and other public documents. It contains 19.348 text documents, which are published in 24 official languages of the European Union. The EUR-Lex repository readily contains three different labeling schemes - *directory-codes, subject-matters* and *eurovoc-descriptors* - for its documents. However, for the evaluation we used only *subject-matters.* A detailed description of parsing and obtaining the documents, as well as the dataset properties can be found here [23]. Table 1 provides a summary of the characteristics of the subject-matters labeling scheme also Fig. 3 shows the labels distribution among the dataset documents.

Stanford CoreNLP toolkit [24] was used in this work for performing the different natural language processing tasks (POS, linguistic filter, taxonomic relations extraction and typed dependencies extraction). It combines machine

Table 1. Data-Set statistics

	Unique Labels	Label Cardinality	Label Density
Subject Matters	201	2.21	1.10

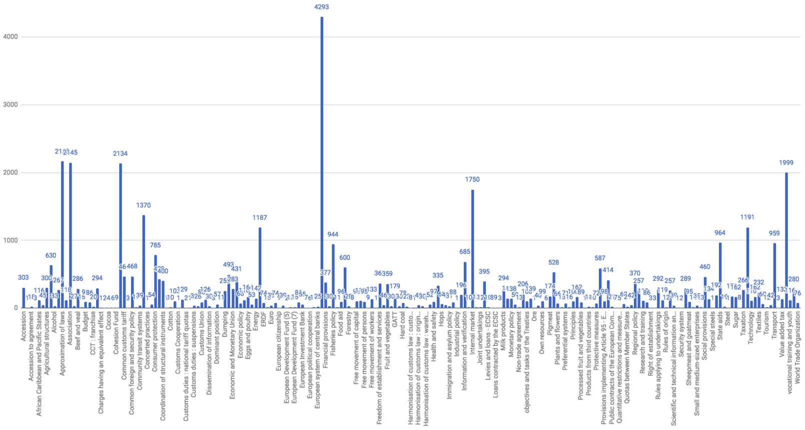

Fig. 3. shows the label distribution among the corpus

learning and probabilistic approaches to NLP with sophisticated, deep linguistic modeling techniques. The used linguistic filter to extract single and multi-word concepts resulted in 940685 distinct features.

The carried out experiments aimed to compare the effectiveness of using the typed dependencies for feature selection against a wide range of statistics and semantics based feature selection techniques, taking in consideration the feature extraction based on taxonomic relations as a baseline since it had provided the best performance in a previous work [3]. For multi-label classification we used ML-KNN with the number of nearest neighbors $K = 10$ as fixed parameter during the experiments. For the features weighting two techniques has been used, namely TF-IDF and binary weighting. In addition the number of features was fixed to 5000 features and Fold=5 for cross-validation evaluation for the comparative analysis with our previous work provided [3].

5 Evaluation

Five different evaluation scenarios were applied to deeply analyze the effect of incorporating the text semantic in feature selection.

Evaluation of Statistics-Based Methods: Figure 4 illustrates the different performance metrics for four statistics feature selection methods applied on the raw text documents after stemming and removing the stop words. While Fig.

5 demonstrates the performance by considering only the noun phrases as candidate features. Three different approaches for feature selection were evaluated, namely selecting equal number of features per label, proportionally to the label frequency and based on the average score for each feature over all labels. Selecting the feature based on labels distribution resulted in the best performance. Using the noun phrases during the feature selection instead of the raw data provides better performance. This is a very interesting result which proves the importance of embedding simple semantic through the multi-word terms to improve the performance and drastically reduce the computation costs by reducing the features space with a factor of almost 90%.

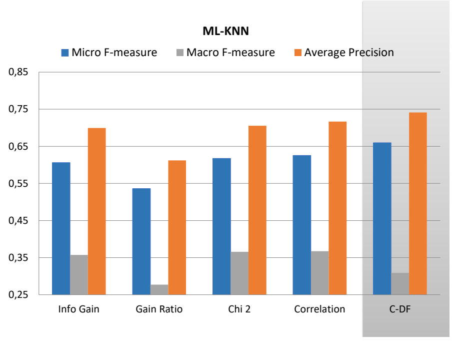

Fig. 4. ML-KNN performance with the different statistical feature selection techniques using terms as features.

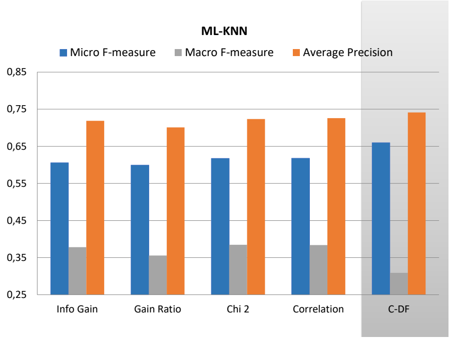

Fig. 5. ML-KNN performance with the different statistical feature selection techniques using noun phrases as features.

Evaluation of Noun Phrase Length: Table. 2 shows the effect of noun phrases length on selecting candidate features. We fixed the number of features to 5000 with TF-IDF weights. The comparison shows that choosing noun phrases with the length of 2 as features performs best, however, the performance is less than our baseline.

Table 2. Evaluation results for noun phrase lengths to select features

(# words)	Hamming Loss	Micro F1	Macro F1	Average Precision
1	0.0077±0.0001	0.5495±0.0101	0.3366±0.0087	0.6570±0.0024
2	**0.0073±0.0001**	**0.5789±0.0074**	**0.3515±0.0127**	**0.6614±0.0017**
3	0.0075±0.0001	0.5467±0.0091	0.3229±0.0112	0.6252±0.0047
4	0.0092±0.0001	0.3560±0.0134	0.2327±0.0128	0.5032±0.0069
5	0.0102±0.0000	0.1676±0.0105	0.1622±0.0151	0.3266±0.0037

Comparison between Document Frequency and Concept Degree: In the third experiments, we compare two techniques for selecting the features from the shallow ontology based on the document frequency or the concept degree. Using the most common relation “of” with binary weighting, we evaluated the performance impact of these two methods over a range of feature space. The Table 3 shows that DF results in better performance however the difference is small so both techniques can be used. Based on that, in following scenarios we have fixed the selection technique to Document Frequency.

Table 3. Comparison between feature selection techniques

	Document Frequency		Concept Degree	
# Features	Micro F1	Macro F1	Micro F1	Macro F1
250	**0.6272±0.0088**	**0.3551±0.0116**	0.6134±0.0094	0.3329±0.0118
500	**0.6415±0.0034**	**0.3654±0.0086**	0.6351±0.0098	0.3622±0.0112
1000	0.6567±0.0027	0.3822±0.0127	**0.6612±0.0067**	**0.3879±0.0114**
2000	**0.6632±0.0060**	0.3944±0.0141	0.6593±0.0051	**0.3956±0.0103**
2500	**0.6649±0.0050**	**0.4006±0.0133**	0.6620±0.0047	0.3973±0.0126
5000	**0.6643±0.0039**	**0.4042±0.0089**	0.6597±0.0057	0.4007±0.0112

Comparison between TF-IDF and Binary Weighting Techniques: The typed dependencies can form a shallow ontology using propositions, conjunctions and verbs as relation modifiers between the concepts. In this scenario we have investigated the quality of the extracted features based on two different weighting techniques, namely TF-IDF and Binary weights. Table 4 shows that using the binary weights significantly outperforms using TF-IDF as a weighting technique. This result is reasonable since the features are selected globally and not based on labels distribution. Also part-of verbs, reflecting verbal relations i.e. including, containing, part-of, etc.,in addition to the “of” relation, slightly outperforms the other combinations of propositions, conjunctions and verb typed dependencies. Moreover, it outperforms the baseline too.

Table 4. Comparison between feature weighting techniques

	Binary Weights		TF-IDF Weights	
Dependency	Micro F1	Macro F1	Micro F1	Macro F1
verb relations	0.6538±0.0050	0.3987±0.0168	0.5962±0.0111	0.3756±0.0130
meronomy relations	**0.6647±0.0039**	**0.4050±0.0095**	0.5940±0.0087	0.3707±0.0162
all relations	0.6586±0.0074	0.3915±0.0135	0.5851±0.0110	0.3692±0.0148

Comparison against the Baseline Table 5 shows the comparison between the baseline [3] and typed dependencies over a range of feature space. The results indicate that using only the typed dependencies has slightly better performance compared to embedding the taxonomic relations too with regard to Micro F1, however, it significantly outperforms the baseline and other techniques discussed in this paper with regard to the Macro F1. Considering the number of features, reducing the number to 2500 or 3000 can achieve nearly as good results as with 5000, in some cases, they even improve the performance. We can conclude that, classification works better for a lower number of features which agrees with other works in classification [23].

Table 5. Evaluation results against the baseline for the meronomic typed dependencies over different numbers of features

# features	Hamming Loss	Micro F1	Macro F1	Average Precision
250	0.0065±0.0000	0.6281±0.0086	0.3540±0.0092	0.7049±0.0064
500	0.0063±0.0000	0.6412±0.0034	0.3655±0.0089	0.7192±0.0065
1000	0.0062±0.0001	0.6554±0.0024	0.3824±0.0103	0.7356±0.0073
2000	0.0061±0.0000	0.6649±0.0059	0.3962±0.0121	0.7441±0.0078
2500	**0.0060±0.0000**	**0.6666±0.0050**	0.4008±0.0136	**0.7464±0.0068**
3000	**0.0060±0.0001**	0.6646±0.0043	0.3985±0.0115	0.7450±0.0075
5000	**0.0060±0.0001**	0.6639±0.0027	**0.4044±0.0070**	0.7436±0.0067
baseline	0.0061±0.0001	0.6642±0.0087	0.3162±0.0115	0.7425±0.0053

6 Conclusion and Future Work

In this work we proposed a new method to select semantic-based features using only the typed dependencies without relying on any external lexical databases, dictionaries or syntactic patterns. We improved on our previous work using taxonomic relations by relying on the typed dependencies which can identify these shallow relations even with more natural texts since it analyses syntactic relations on the sentence level. Our comprehensive evaluation against a wide range of feature selection techniques proved that taking in consideration syntactic relations between words can provide better performance with regard to the compared statistics and semantic-based approaches. In addition, it significantly reduces the computation costs for selecting the features by relaying on the shallow ontology for selecting and updating the features. In Future work, the proposed feature selection technique will be used for developing ontology-based training less classifier to overcome the limitation of selecting the number of features and considering their semantic similarity for labels assignment.

References

1. I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, 2016.

2. F. Sebastiani, “Text categorization,” in *Encyclopedia of Database Technologies and Applications*. IGI Global, 2005, pp. 683–687.
3. W. Alkhatib, C. Rensing, and J. Silberbauer, “Multi-label text classification using semantic features and dimensionality reduction with autoencoders,” in *International Conference on Language, Data and Knowledge*. Springer, 2017, pp. 380–394.
4. I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.
5. H. Ogura, H. Amano, and M. Kondo, “Feature selection with a measure of deviations from poisson in text categorization,” *Expert Systems with Applications*, vol. 36, no. 3, pp. 6826–6832, 2009.
6. P. Soucy and G. W. Mineau, “Beyond tfidf weighting for text categorization in the vector space model,” in *IJCAI*, vol. 5, 2005, pp. 1130–1135.
7. Y. Yang and J. O. Pedersen, “A comparative study on feature selection in text categorization,” in *Icml*, vol. 97, 1997, pp. 412–420.
8. Y. Yang and J. Pedersen, “A comparative study on feature selection in text categorization.” Morgan Kaufmann Publishers, 1997, pp. 412–420.
9. T. Masuyama and H. Nakagawa, “Cascaded feature selection in svms text categorization,” in *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2003, pp. 588–591.
10. D. D. Lewis, “Feature selection and feature extraction for text categorization,” in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 212–217.
11. Y. Liu, H. T. Loh, and W. F. Lu, “Deriving taxonomy from documents at sentence level,” *HAD Prado and E. Ferneda,” Emerging Technologies of Text Mining: Techniques and Applications”, Idea, Hershey, PA*, pp. 99–119, 2007.
12. J. Fürnkranz, “A study using n-gram features for text categorization,” *Austrian Research Institute for Artifical Intelligence*, vol. 3, no. 1998, pp. 1–10, 1998.
13. A. Khan, B. Baharudin, and K. Khan, “Semantic based features selection and weighting method for text classification,” in *Information Technology (ITSim), 2010 International Symposium in*, vol. 2. IEEE, 2010, pp. 850–855.
14. G. A. Miller, “Wordnet: a lexical database for english,” *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.
15. M. Janik and K. Kochut, “Training-less ontology-based text categorization,” in *Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR 2008) at the 30th European Conference on Information Retrieval, ECIR*, vol. 20, 2008.
16. Y.-H. Chang and H.-Y. Huang, “An automatic document classifier system based on naive bayes classifier and ontology,” in *Machine Learning and Cybernetics, 2008 International Conference on*, vol. 6. IEEE, 2008, pp. 3144–3149.
17. S. Chua and N. Kulathuramaiyer, “Feature selection based on semantics,” in *Innovations and Advanced Techniques in Systems, Computing Sciences and Software Engineering*. Springer, 2008, pp. 471–476.
18. M. A. Hearst, “Automatic acquisition of hyponyms from large text corpora,” in *Proceedings of the 14th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, 1992, pp. 539–545.
19. M.-C. De Marneffe and C. D. Manning, “Stanford typed dependencies manual,” Technical report, Stanford University, Tech. Rep., 2008.
20. M. Berland and E. Charniak, “Finding parts in very large corpora,” in *Proceedings of the 37th annual meeting of the Association for Computational*

Linguistics on Computational Linguistics. Association for Computational Linguistics, 1999, pp. 57–64.

21. M.-L. Zhang and Z.-H. Zhou, “Ml-knn: A lazy learning approach to multi-label learning,” *Pattern recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.
22. (2017, 01). [Online]. Available: http://www.ke.tu-darmstadt.de/resources/eurlex
23. E. L. Mencía and J. Fürnkranz, “Efficient multilabel classification algorithms for large-scale problems in the legal domain,” in *Semantic Processing of Legal Texts*. Springer, 2010, pp. 192–215.
24. C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, “The stanford corenlp natural language processing toolkit.” in *ACL (System Demonstrations)*, 2014, pp. 55–60.