Cross-domain Polarity Classification using a Knowledge-enhanced Meta-classifier

Marc Franco-Salvador[a,*], Fermín L. Cruz[b], José A. Troyano[b], Paolo Rosso[a]

[a] *Pattern Recognition and Human Language Technology (PRHLT) Research Center*
Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain
[b] *Department of Languages and Computer Systems*
University of Seville, Av. Reina Mercedes s/n, 41012 Sevilla, Spain

Abstract

Current approaches to single and cross-domain polarity classification usually use bag of words, n-grams or lexical resource-based classifiers. In this paper, we propose the use of meta-learning to combine and enrich those approaches by adding also other knowledge-based features. In addition to the aforementioned classical approaches, our system uses the BabelNet multilingual semantic network to generate features derived from word sense disambiguation and vocabulary expansion. Experimental results show state-of-the-art performance on single and cross-domain polarity classification. Contrary to other approaches, ours is generic. These results were obtained without any domain adaptation technique. Moreover, the use of meta-learning allows our approach to obtain the most stable results across domains. Finally, our empirical analysis provides interesting insights on the use of semantic network-based features.

Keywords: Sentiment analysis, Cross-domain polarity classification, Meta-learning, Word sense disambiguation, Semantic network

1. Introduction

Text classification (also known as text categorization) is the task of assigning a category or categories to a text document from a set of predefined categories. Although at first this topic was approached from a knowledge

*Corresponding author.
Email address: mfranco@prhlt.upv.es (Marc Franco-Salvador)

engineering perspective (manually defining a set of rules encoding expert knowledge), in the 90s machine learning became the main approach, and so it stands today. A good survey on machine learning approaches to text classification can be found in (Sebastiani, 2002).

The nature of the predefined categories in text classification can be very heterogeneous. The most common task is that of topic-based classification, attemping to classify documents according to their subject matter (e.g. Sports vs. Politics vs. Economics). More recently, in the context of the Web 2.0 and social media, it emerged the task of deciding whether a subjective text (typically, a textual review of some product or a cultural or political issue) is positive or negative, depending on the overall sentiment detected. This particular task is known as polarity classification or sentiment classification (Turney, 2002; Pang et al., 2002). Although it can be defined in terms of text classification (being positive and negative the predefined categories) and tackled with similar approaches, polarity classification has been proved to be a more difficult task (Pang et al., 2002): while topics are often identifiable by keywords alone, sentiment can be expressed in a more subtle manner, and even more when for instance irony is employed (Reyes and Rosso, 2013). Therefore, solutions based only on bag-of-words representations of documents may not be enough.

In this work we are interested in single and cross-domain polarity classification. Since we are applying machine learning techniques, we start with a training set of documents to build some classifiers. In this context, single-domain classification is the aforementioned common text classification; it refers to training and testing classifiers on the same domain (e.g. movie reviews). Meanwhile, cross-domain classification refers to testing on a different domain (target domain) from that or those used in training (source domains), e.g. training on movie reviews and testing on books reviews. Because manually labeled documents are needed for training, the latter allows to work with domains where no labeled documents are available. The problem of cross-domain text classification was first tackled by Dai et al. (2007), and the first results on cross-domain polarity classification were reported by Blitzer et al. (2007).

In order to combine different approaches from the research literature and recent knowledge-based approaches, and also to measure the contributions of each one, we propose the use of a meta-learning scheme called Stacked Generalization (Wolpert, 1992). The set of base classifiers to be combined using that scheme include solutions used in the past as a TF-IDF bag-of-

words classifier, a TF-IDF word n-gram classifier, and a lexical resource for opinion mining-based classifier; but also two new proposals, a word sense disambiguation-based classifier and a vocabulary expansion-based classifier. The latter two classifiers are trained on the basis of knowledge graphs, a subset of a semantic network, i.e., BabelNet (Navigli and Ponzetto, 2012), focused on the concepts belonging to the text being classified.

The rest of the paper is structured as follows. In Section 2 we describe the related work on single and cross-domain polarity classification. In Section 3 we introduce our new knowledge-enhanced meta-classifier. In Section 4 we evaluate our approach in the tasks of single and cross-domain polarity classification, and compare it with other state-of-the-art approaches. In that section we evaluate also the performance of our different base classifiers. Finally, in Section 5 we draw the conclusions and mention directions for future work.

2. Related Work

The first experiments on single-domain polarity classification using machine learning techniques were performed by Pang et al. (2002). They used a movie review dataset extracted from IMDb.[1] They concluded that polarity classification achieves worse results than other text classification tasks when applying the standard machine learning techniques. Another interesting conclusion was that using unigram presence instead of unigram frequency leads to better results, contrary to observations in other works on text classification (McCallum and Nigam, 1998)

Recent works on polarity classification use the Multi-Domain Sentiment Dataset (Blitzer et al., 2007) for evaluation. In its last version, the resource is composed by Amazon product reviews of 25 product types, though most works report results on only the four domains used by Blitzer et al. (2007): Books, Electronics, DVDs and Kitchen appliances. Focused on single-domain polarity classification, Dredze et al. (2008) presented a new online learning method named confidence-weigthed learning. The method is based on measuring the confidence of each parameter of the classifier; less confident parameters are updated more aggressively than more confident ones. They performed experiments on standard datasets related to different text classification tasks, reporting very good results for the Multi-Domain Sentiment

[1] `http://www.cs.cornell.edu/people/pabo/movie-review-data/`

Dataset. Another approach, proposed by Li and Zong (2008), use n-grams combined with Binormal Separation (Forman, 2008), an alternative to TF-IDF to select the optimal set of features. They reported interesting results in single domain classification.

Cross-domain polarity classification has gained popularity thanks to the advances in domain adaptation (Daumé III, 2007; Blitzer et al., 2008; Ben-David et al., 2010). These techniques make use of labeled data from a source domain, and unlabeled data from source and target domains to train their classifiers. Using the different domains available in the Multi-Domain Sentiment Dataset, Blitzer et al. (2007) was also the first to report results on cross-domain classification proposing two algorithms: structural correspondence learning (SCL), and its variant using mutual informaton (SCL-MI). The SCL model selects pivot (unigram and bigram) features frequently appearing in both source and target domains. Then it learns to predict those pivot features in the unlabeled data from both domains. Later, a singular value descomposition is performed to reduce dimensions, and a binary classifer is trained to determine the polarity. Similarly, interesting results on cross-domain polarity classification have been reported by spectral feature alignment (SFA) (Pan et al., 2010). Using unigram and bigram features, the model exploits the mutual information between each feature and the domain label to differentiate domain-specific and domain-independent features. Next, a bipartite graph is constructed by dividing both types of features. An edge connects features from different types if there exists co-ocurrence. Finally, a spectral clustering is performed to generate feature clusters and a binary classifer is built for the polarity classification. More recently, Bollegala et al. (2011, 2013) used a cross-domain lexicon creation to generate a sentiment-sensitive thesaurus (SST) that groups different words expressing the same sentiment, using also unigram and bigram features as representation. This approach also obtained competitive results in single-domain polarity classification.

Note that all cross-domain approaches use domain adaptation techniques extracting relevant features from the source domains, in order to obtain important features to classify the target domain. In contrast, we do not use unlabeled data from the target domain. Our approach is focused on proposing new knowledge-based features which allows for training models using the source domains that are able to be directly applied to the target domain. In Section 4.4 we compare our approach in the task of single-domain polarity classification against SST and the state-of-the-art approaches proposed

by Dredze et al. (2008) and Li and Zong (2008). Next, in Section 4.5 we compare our approach in the task of cross-domain polarity classification against SCL-MI, SFA and SST models.

3. Knowledge-enhanced Meta-classifier

We propose the use of a meta-learning scheme for combining different classical approaches, i.e., bag of words, n-grams or lexical resource-based classifiers. Key to our approach is adding also other knowledge-based classifiers. By using a semantic network, we perform word sense disambiguation and generate new independent classifiers for the main part-of-speech tags: disambiguated adjectives, nouns, verbs and adverbs. Using the disambiguated terms, the semantic network allows us to obtain a vocabulary expansion-based classifier. In Section 3.1 we present the semantic network, and the word sense disambiguation and vocabulary expansion methods. Then, in Section 3.2 we describe the base classifiers that compose our system. Finally, in Section 3.3 we define the Stacked Generalization that we use to combine those classifiers.

3.1. Word Sense Disambiguation and Vocabulary Expansion via a Semantic Network

A semantic network (Sowa, 2006) is a (un)directed graph consisting of vertices, which represent concepts, and edges, which represent semantic relations between them. Concepts are usually organized into a taxonomic hierarchy. Figure 1 shows a simple example of semantic network.

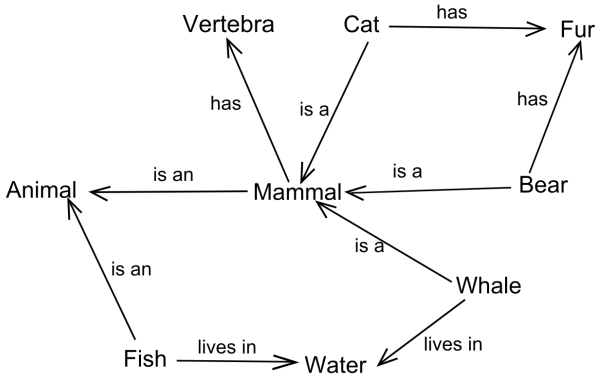

Figure 1: Semantic network example focused on the animal world.

In this work we use the semantic network graph to: (i) perform word sense disambiguation, and (ii) perform a vocabulary expansion using the disambiguated words. Despite having the WordNet Semantic Network (Fellbaum, 1998), which is an historical resource including 117,000 synsets[2] in English, in this work we are interested in employing a larger size wide-coverage lexical knowledge resource. Among those, we can find knowledge bases extracted automatically from Wikipedia such as DBPedia (Bizer et al., 2009) or YAGO (Hoffart et al., 2013). However, due to its WordNet-based internal structure combined with Wikipedia, the high amount of synsets included, and the lexicalizations of its concepts available in multiple languages,[3] we chose the BabelNet Multilingual Semantic Network.

3.1.1. BabelNet

BabelNet[4] 2.5 (Navigli and Ponzetto, 2012) is a multilingual semantic network whose concepts and relations are obtained from the automatic mapping onto Wordnet of Wikipedia,[5] OmegaWiki,[6] Wiktionary,[7] Wikidata,[8] and Open Multilingual WordNet.[9] BabelNet is therefore a multilingual "encyclopedic dictionary" that combines lexicographic information with wide-coverage encyclopedic knowledge. Concepts in BabelNet are represented similarly to WordNet, i.e., by grouping sets of synonyms in the different languages into multilingual synsets. Multilingual synsets contain lexicalizations from WordNet and Open Multilingual WordNet synsets, the corresponding Wikipedia pages, the OmegaWiki, Wiktionary and Wikidata entries, and additional translations by a statistical machine translation system. The relations between synsets are collected from WordNet, Open Multilingual WordNet, and from Wikipedia's hyperlinks between pages. The current version of BabelNet includes 9,348,287 synsets, covers 50 languages, and has a WordNet-Wikipedia mapping correctness of 91% (Navigli et al., 2013).

[2]Set of word synonyms.

[3]While this work is exclusively evaluated on English, this multilinguality allows us to perform at multilingual level.

[4]`http://babelnet.org`

[5]`http://wikipedia.org`

[6]`http://omegawiki.org`

[7]`http://wiktionary.org`

[8]`http://wikidata.org`

[9]`http://compling.hss.ntu.edu.sg/omw/`

3.1.2. Word Sense Disambiguation

Word sense disambiguation (WSD) (Navigli, 2009) is the process of identifying which sense (i.e., meaning) of a word is used in a sentence, when the word is polysemic. In general, the approaches for WSD can be classified into three types: (i) supervised, with a considerable effort for new languages and domains due to the huge amount of annotated data required (Shen et al., 2013; Pilehvar and Navigli, 2014); (ii) unsupervised approaches, which have to deal with data sparsity and an intrinsic difficulty with their evaluation (Agirre et al., 2006; Di Marco and Navigli, 2013); (iii) knowledge-based approaches, which exploit the knowledge available in structured knowledge bases (Ponzetto and Navigli, 2010; Navigli and Lapata, 2010; Agirre et al., 2014; Moro et al., 2014). Vocabulary expansion benefits from the WSD performed using a knowledge base by exploiting the relations in its network.

BabelNet has been used for WSD in several works, including some of the aforementioned publications and also as part of the Multilingual Word Sense Disambiguation Task of the SemEval Workshop (Navigli et al., 2013). Similarly to Navigli and Ponzetto (2012) and Franco-Salvador et al. (2013, 2014), we followed Navigli and Lapata (2010) to create knowledge graphs[10] in order to perform the WSD and the vocabulary expansion. The five-step method we used to perform the WSD is the following:

(i) Part-of-speech tagging and lemmatization. Initially we process a document d with tokenization, multi-word extraction, part-of-speech (POS) tagging and lemmatization[11] to obtain the list of tuples (lemma,tag) T. We are interested only in the POS tags available on BabelNet (adjectives, nouns, verbs and adverbs).

(ii) Populating the graph with initial concepts. Next, we create an initially-empty knowledge graph $G = (V, E)$, i.e., such that $V = E = \emptyset$. We populate the vertex set V with the set S_K of all the synsets in BabelNet which contain any tuple (lemma,tag) in T in the document language L, that is:

[10] A knowledge graph is a subset of the original semantic network focused on the concepts belonging to a text, and in the intermediate concepts and relations between them.

[11] For this purpose we used the Stanford Log-linear Part-Of-Speech Tagger: `http://nlp.stanford.edu/software/tagger.shtml`. For the multi-word extraction we implemented our own tool based on the matching of typical patterns.

$$S_K = \bigcup_{t \in T} \text{Synsets}_L(t), \tag{1}$$

where $\text{Synsets}_L(t)$ is the set of synsets which contains a tuple (lemma,tag) t in the language of interest L.

(iii) Creating the knowledge graph. We create the knowledge graph by searching on BabelNet to obtain the set of paths P connecting pairs of synsets in V. Formally, for each pair $\{v, v'\} \in V$ such that v and v' do not share any lexicalization[12] in T, for each path in BabelNet $v \to v_1 \to \cdots \to v_n \to v'$, we set: $V := V \cup \{v_1, \ldots, v_n\}$ and $E := E \cup \{(v, v_1), \ldots, (v_n, v')\}$. That is, we add all the path vertices and edges to G. Following Navigli and Ponzetto (2012), the path length is limited to maximum length of 3, in order to avoid an excessive semantic drift.

As a result of populating the graph with intermediate edges and vertices, we obtain a knowledge graph which models the semantic context of document d.

(iv) Knowledge graph weighting. The next step consists of weighting all the concepts and semantic relations of the knowledge graph G. For weighting relations we use the original weights from BabelNet, which provide the degree of relatedness between the synset end points of each edge.[13] For weighting concepts different methods, including the PageRank (Page et al., 1998) algorithm, have been tested in the past. In this work, we score each concept using its own outdegree, which has proved to obtain the best results.(Navigli and Ponzetto, 2012)

(v) Selecting the corresponding disambiguations. Finally, for each tuple (lemma,tag) $t \in T$, we collect from BabelNet the set of synsets S_t containing t, and we select as proper disambiguation t_{WSD} the synset with the highest score:

$$t_{WSD} = {}_{s \in S_t} \operatorname{argmax} \text{score}(s), \tag{2}$$

[12]This prevents different senses of the same term from being connected via a path in the resulting knowledge graph.

[13]At this point, we removed the edges below a certain threshold that represents a low semantic relationship.

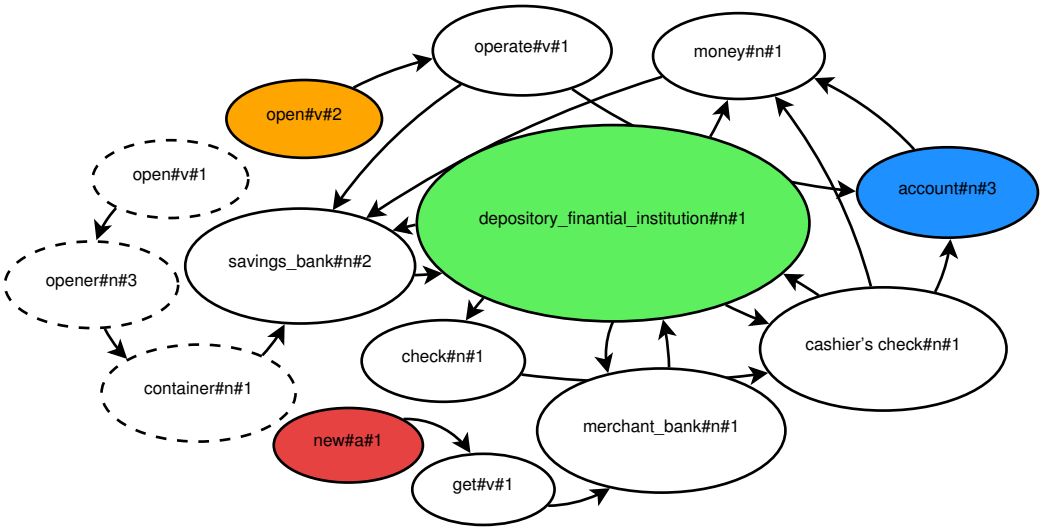

Figure 2: Simplified knowledge graph created from the sentence "I opened a new bank account". Colored nodes are the resulting disambiguations while white nodes are expanded concepts. Dashed noded will not be included in the vocabulary expansion set.

3.1.3. Vocabulary Expansion

Once we have disambiguated the words of a document d, to enrich and increase the available context, we perform an automatic vocabulary expansion (Ehrlich, 1995; Ehrlich and Rapaport, 1997) using the BabelNet graph topology. A simple vocabulary expansion can be done using directly any connected concept to a disambiguated one, up to a certain distance in the graph. However, to preserve as much context as possible and to avoid introducing noise, we include only intermediate concepts between pairs of disambiguated words. Formally, using the knowledge graph G created in Section 3.1.2, we obtain a vocabulary expansion as follows:

(i) Collecting the disambiguation senses. We first use the process described in the previous section to obtain the set S_{WSD}. This set is composed by the disambiguation synsets of the original words of document d.

(ii) Removing alternative senses. We create a path set P' by removing from the path set P all the paths between synsets which are not in S_{WSD}. This step removes noise by creating a knowledge graph focused on the disambiguated concepts.

(iii) Obtaining the expanded concepts. We obtain the vocabulary expansion by creating a set S_{exp} including the intermediate concepts in the paths of P'. We remove the source and target concepts from paths to evaluate the performance of the vocabulary expansion without the original words (see Section 4.3).[14]

Figure 2 provides an example[15] of disambiguation and vocabulary expansion using knowledge graphs.

3.2. Base Classifiers

We can now define the base classifiers that compose our system. We first include a TF-IDF bag-of-words classifier, a TF-IDF word n-gram classifier and a lexical resource for the opinion mining-based classifier. The choice of these components has been motivated by the good results that they achieved in the past. In addition, in this work we want to investigate the impact of knowledge-based classifiers; therefore we include an independent classifier to study the contribution of WSD for each POS tag employed (adjectives, nouns, verbs and adverbs). Finally, under the assumption that semantically-related concepts have a common near relative, we want to exploit this possible relateness between concepts including a vocabulary expansion-based classifier. Next we explain in more detail our eight base classifiers:

(i) Bag-of-words classifier. This approach transforms a document d into a traditional vector representation. Following the literature, we selected the most widely used representation for real-valued feature vectors, commonly used as baseline: the Term Frequency-Inverse Document Frequency (TF-IDF) weighting (Salton et al., 1983; Salton and McGill, 1986).

$$\text{tf-idf}(w) = \text{tf}(w)N/n(w). \tag{3}$$

where $tf(w)$ is the number of times a term w occurs in document d, N is the total number of documents in the collection and $n(w)$ is the number of documents that contain w. We removed stopwords from documents for all the base classifiers.

[14]This last part is optional, although it helps to focus on the vocabulary expanded concepts.

[15]Weights and nodes representing alternative senses or intermediate concepts are removed for simplicity.

As classifier, we selected Support Vector Machines (SVM) (Chang and Lin, 2011), with a linear kernel function,[16] given its good performance for text classification (Joachims, 1998) using TF-IDF weighting.

(ii) Word n-gram classifier. The use of word n-grams has been proposed several times (Cavnar, 1995; Mayfield and McNamee, 1999; Li and Zong, 2008) as a better alternative to single word vector representation due to the additional information that it provides. Using n-grams is a plus for a complex classification task like polarity classification: while topics are often identifiable by keywords alone, sentiment can be expressed in a more subtle manner (Pang et al., 2002). For example, the keyword *like* may be correlated with positive sentences (e.g. "I *like* this paper a lot.") or with negative sentences (e.g. "I do *not like* this paper at all."). Using n-grams also allows us to learn frequent, opinion-bearing multiword expressions (e.g. "*you will love* (this story)").

This n-gram representation is processed with a TF-IDF weighting and an SVM classifier. Since larger n-grams will not be frequent, we included only a combination (Li and Zong, 2008) of 1, 2, and 3-grams.

(iii) Lexical resource-based classifier. The use of lexical resources for opinion mining was strongly popularized by the release of SentiWordNet (Esuli and Sebastiani, 2006; Baccianella et al., 2010). This resource assigns to each synset of WordNet three sentiment scores: positivity, negativity, objectivity. It has been sucessfully applied to polarity classification in the past (Ohana and Tierney, 2009; Hamouda and Rohaim, 2011).

We selected as lexical resource ML-SentiCon (Cruz et al., 2014), which proved to make several improvements with respect to the original SentiWordnet 3.0, with a significative better positivity, negativity and objectivity estimation, reflecting those results on their evaluation.

For this base classifier, we decided to use the tree-based C4.5 (Quinlan, 1996) model, which infers a hierarchy of rules as a function of different feature values to determine the final class, and provides good performance for

[16] We use the linear kernel function for all the SVM base classifiers.

[17] As we can see, we take advantage of WSD to remove noise (unrelared synsets).

[18] We refer to the disambiguations of the original words of the document.

[19] Since the format of ML-SentiCon is the same as SentiWordNet, and BabelNet has a synset for each WordNet synset, we can map directly our disambiguated words to that lexical resource.

Model features
Number of words in document d.
Number of disambiguated synsets[17] in the knowledge graph G (see Section 3.1.2).
Number of directly connected disambiguated synsets in G[18].
Number of adjetives in d.
Number of nouns in d.
Number of verbs in d.
Number of adverbs in d.
Average positivity of the disambiguated words of d[19].
Average negativity of the disambiguated words of d.
Average objetivity of the disambiguated words of d

Table 1: List of features selected for the lexical resource-based classifier.

polarity classification (Jia et al., 2009). Its use is also motivated by the different types of features that we selected for this classifier (see Table 1): some of them are discrete and unbounded. In addition, considering that there are only 10 features, using SVM did not pose any additional advantage with regard to a simpler C4.5 tree-based classifier.

(iv-vii) Word sense disambiguation-based classifiers. As we stated at the beginning of this section, to study the impact of WSD on polarity classification, we generate an independent classifier for each POS tag available on BabelNet (adjectives, nouns, verbs and adverbs) on the basis of the method explained in Section 3.1.2.

During the prototyping process, we realized that due to the use of independent classifiers for each POS tag, and the error introduced by wrong disambiguations, the TF-IDF weighting provided an imprecise representation of documents, and worse results than using only binary TF (presence or not of the word w in the document). Since the use of this technique has been studied in the past with good results (Pang et al., 2002), for the WSD-based models we decided to use binary TF as weighting and SVM as classifier.

(viii) Vocabulary expansion-based classifier. The last base classifier uses the vocabulary expansion explained in Section 3.1.3 to represent each document d as a binary TF of synsets, which are related to the original disambiguated ones of d. The classification is performed using SVM. Since we are removing the original concepts of the documents from the vocabulary expansion, a document containing the concepts "Michael Jordan" and "NBA" will be

Base classifier ID	Description	Weighting	Classifier	Avg. # feat.
BOW	Bag-of-words representation	TF-IDF	SVM	19,976
(1+2+3)-grams	Combine $\{1, 2, 3\}$-grams to represent documents	TF-IDF	SVM	58,636
ML-SentiCon	Use a lexical resource to extract different polarity-related features	-	C4.5	10
Noun WSD	Represent documents by its set of disambiguated nouns	Binary TF	SVM	13,139
Adjective WSD	Represent documents by its set of disambiguated adjectives	Binary TF	SVM	3,241
Verb WSD	Represent documents by its set of disambiguated verbs	Binary TF	SVM	2,138
Adverb WSD	Represent documents by its set of disambiguated adverbs	Binary TF	SVM	689
Vocab. Exp.	Use a vocabulary expansion to represent the documents	Binary TF	SVM	59,372

Table 2: Summary of base classifiers.

represented by concepts as "Basketball" and "Sport", but not by the original concepts. As previously stated, the original concept removal was performed because we are interested in evaluating the performance of the vocabulary expansion without the original words.

Table 2 provides a summary[20] of all the base classifiers.

3.3. Stacked Generalization

We combine the base classifiers with one of the most popular combination methods in meta-learning: stacking. It has been used successfully in Natural Language processing (NLP) tasks (Van Halteren et al., 1998; Enríquez et al., 2013) in the past. This method follows the original Stacked Generalization method (Wolpert, 1992) to project documents onto a new dimensional space, which is composed by the annotations of a first-level base classifiers set. This combination is able to exploit additional information from a corpus by processing it with different classifiers. A second-level classifier uses all of the annotations of the first level to obtain a final decision, with the advantage of recognizing and classifying correctly patterns in which the correct class tag is in inferiority. In this work, instead of representing the results of the first level as a vector of class tags, we represent them as a vector of class probabilities, which proved to obtain better results using SVM (Martín-Valdivia et al., 2013).

We can see the Stacked Generalization method detailed in Algorithm 1. Lines 1–3 correspond to the first level of the classifier, which makes the transformation of the training corpus. The second level of the classfier is

[20]Column "Avg. # feat." shows the average number of potential features of the classifier across domains before applying their respective thresholds (see Section 4.2).

Algorithm 1 Stacking Generalization algorithm.

Require: a tagged training corpus T and a untagged test corpus t.

Ensure: a tagged test corpus t''.

1: Split T into K parts to obtain $T_{1,...,K}$ partitions.
2: Tag $T_{1,...,K}$ using cross-validation with the $C_{1,...,N}$ base classifiers to obtain $T'_{1,...,K}$ partitions containing the transformed samples of T.
3: Using $T_{1,...,K}$ for training, classify t with $C_{1,...,N}$ to obtain the transformed corpus t'.
4: Use $T'_{1,...,K}$ as a single partition to train the second-level classifier $C_{comb.}$.
5: Classify t' with $C_{comb.}$ to obtain the tagged test corpus t''.

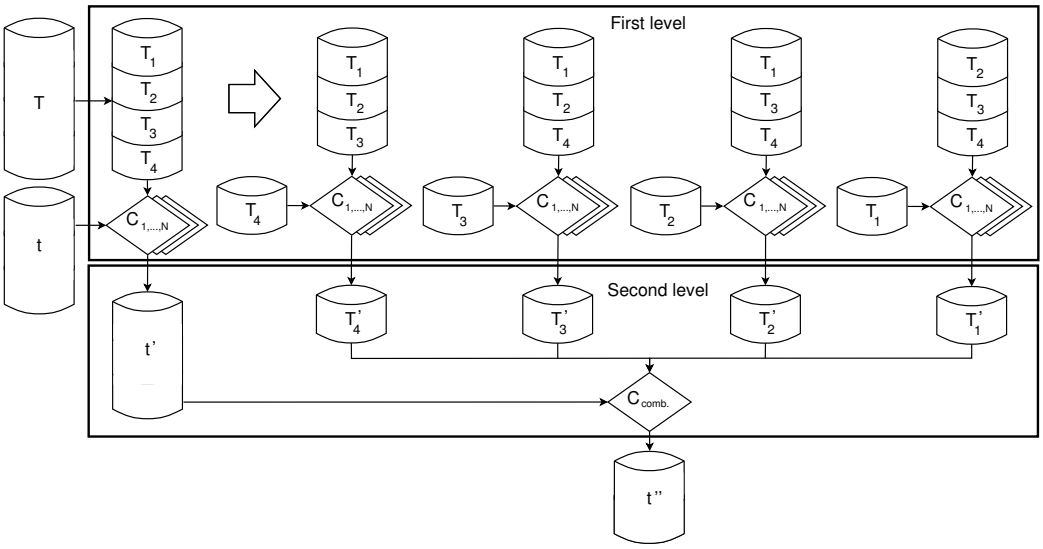

Figure 3: Stacked Generalization scheme. Training and test partitions are projected into a new dimensional space which is composed by the first-level classifier class probabilities. The second-level classifier uses those probabilities to obtain the final decision.

explained in Lines 4–5, which obtains the final classification of the test corpus. A complete scheme of the model is shown in Figure 3.

4. Evaluation

In this section we evaluate the base classifiers of our Knowledge-enhanced Meta-classifier (KE-Meta), and compare our approach with state-of-the-art models on single and cross-domain polarity classification.

4.1. Dataset

To evaluate our system we chose a classical state-of-the-art dataset, the Multi-Domain Sentiment Dataset (version 2.0)[21] (Blitzer et al., 2007), which has been used for the evaluation of several research works on sentiment analysis (Dredze et al., 2008; Li and Zong, 2008; Blitzer et al., 2007; Bollegala et al., 2013). The dataset is composed by Amazon product reviews of 25 product types. Each review contains metadata including a rating of 0-5 stars, the reviewer name and location, the product name, the review date and title, and the review text. In addition, for research purposes, a subset of the reviews with rating < 3 were originally labeled as negative, and with rating > 3 as positive. Following the literature, in this work we use the Books, Electronics, DVDs, and Kitchen appliances reviews, with 1,000 positive and 1,000 negative documents per domain, having a total of 8,000 reviews. With this setup, we can compare our results on single and cross-domain polarity classification directly with the state of the art.

4.2. Methodology

The evaluation of our approach in single-domain polarity classification is performed using a stratified 10-fold cross-validation setup for each domain. In cross-domain, we followed the same 10-fold cross-validation setup,[22] in this case, training always with all domains available and excluding the target domain to classify, e.g. we train with Books, Electronics, and DVDs, and we classify Kitchen reviews. We selected as the evaluation metric the accuracy of the classifiers, which is the proportion of correctly classified reviews among the test dataset. We detail the models compared with our approach on its respective evaluation sections. Note that the number of dimensions of all our base classifiers is limited to a maximum number of 20,000. However, similar results were obtained with sizes ranging between 15,000 and 25,000 during the prototyping step.

4.3. Evaluation of Base Classifiers

To evaluate the eight base classifiers that compose our approach (cfr Section 3.2) summarized in Table 2, we first employ a traditional measure

[21] `http://www.cs.jhu.edu/~mdredze/datasets/sentiment/`

[22] The cross-validation here is used only to train our KE-Meta classifier, which needs a splitting of the data to obtain training and testing partitions to generate the final second-level classifier.

of information theory (Hall and Smith, 1998): the information gain ratio (IGR) (Quinlan, 1986; Raileanu and Stoffel, 2004). Once analyzed the IGR, we will continue with the study of the accuracy of classification of each base classifier.

Having a training set T and its set of attributes $Attr$, the IGR measure provides a normalized estimation (between 0 and 1) of the amount of information that an attribute $a \in Attr$ provides to determine the class attribute.[23] The IGR of an attribute a is calculated as the ratio between the information gain (IG) and the intrinsic value (IV):

$$\mathrm{IGR}(T, a) = \frac{\mathrm{IG}(T, a)}{\mathrm{IV}(T, a)} \tag{4}$$

$$\mathrm{IG}(T, a) = \mathrm{H}(T) - \sum_{v \in \mathrm{values}(a)} \left(\frac{|\{x \in T | \, \mathrm{value}(x, a) = v\}|}{|T|} \cdot H(\{x \in T | \, \mathrm{value}(x, a) = v\}) \right) \tag{5}$$

where we substract to the total entropy H of the train set T the sum of the relative entropies of the different values of a in T. For each of the attributes, if a unique classification can be made for the result attribute, the information gain is equal to the total entropy of a. The IV is a normalization factor estimated as a function of the substracted entropies of $\mathrm{H}(T)$ in IG.

$$\mathrm{IV}(T, a) = - \sum_{v \in \mathrm{values}(a)} \frac{|\{x \in T | \, \mathrm{value}(x, a) = v\}|}{|T|} \cdot \log_2 \left(\frac{|\{x \in T | \, \mathrm{value}(x, a) = v\}|}{|T|} \right) \tag{6}$$

To obtain the IGR of our base classifiers, we estimated the IGR on each tested domain and we calculated the harmonic mean[24] of those results. This test was performed on single and cross-domain polarity classification. We show the results in Figure 4. As expected, the IGR in cross-domain is lower than working on single domain for almost all of the base classifiers. This is not the case of the model using ML-SentiCon, which, despite getting a low IGR, is able to preserve all its gain when performing at cross-domain level. These results put forward the advantage of knowledge bases to model the information in a domain-independent way. We can see that BOW and

[23] Note that each attribute $a \in Attr$ corresponds to a base classifier in our approach.

[24] The harmonic mean is the most adequate measure to average percentages of different domains.

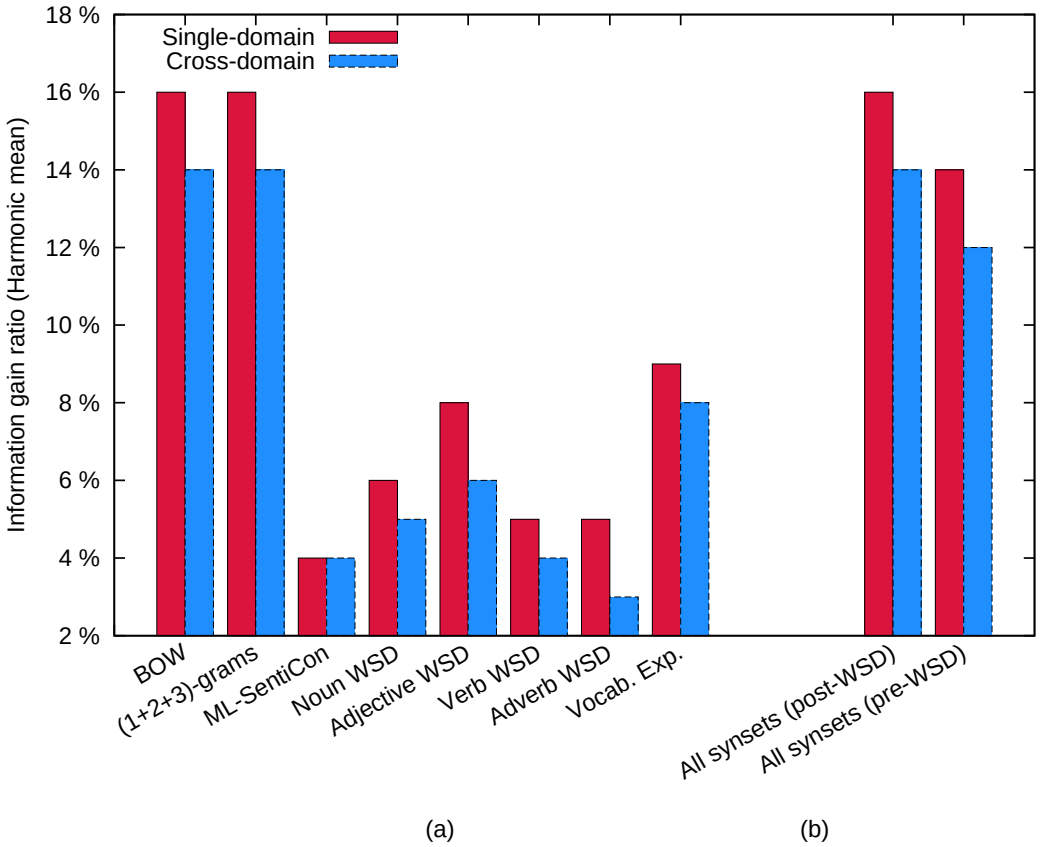

Figure 4: Information gain ratio of the eight base classifiers in single and cross-domain polarity classification. We show the harmonic mean of the IGR of each feature among the different tested domains. (a) Base classifiers; (b) other classifiers.

(1+2+3)-grams classifiers obtained the highest information gain ratios, with almost identical values. The results prove that these models are a good choice as base classifiers to be complemented with other classifiers. The vocabulary expansion, which does not include the original words of the documents, is able to obtain comparable results. Models disambiguating different POS tags obtained considerably low IGR. Adjective WSD was the most informative classifier. This is unsurprising if we consider that often, the polarity of a text could be given by adjectives. This is followed by the classifier for nouns, verbs, and finally adverbs. These last two with identical results on single-domain. Since WSD has been divided into four models, it is difficult to evaluate its contribution. For this reason, we included also the results of two

additional classifiers: All synsets (Post-WSD) and All synsets (Pre-WSD). They represent the IGR of a binary TF[25] classifier trained using SVM with: (i) all the disambiguated words together (All synsets (Post-WSD) classifier), and (ii) all the possible senses of the words together before disambiguation (All synsets (Pre-WSD) classifer). As we can see, the performance of All synsets (Post-WSD) significantly outperforms the Pre-WSD model, and obtained similar result to BOW and n-grams based approaches. This highlights the capability of WSD to remove noisy senses, leaving only the appropiate one.

Base classifiers	**Books**	**Electronics**	**DVDs**	**Kitchen**
BOW	0.788	0.803	0.804	0.821
(1+2+3)-grams	0.805	0.817	0.803	0.819
ML-SentiCon	0.612	0.644	0.644	0.651
Noun WSD	0.684	0.655	0.679	0.677
Adjective WSD	0.683	0.695	0.729	0.712
Verb WSD	0.669	0.670	0.633	0.675
Adverb WSD	0.651	0.638	0626	0.649
Vocab. Exp.	0.718	0.700	0.709	0.704
Other classifiers				
All synsets (Post-WSD)	0.775	0.782	0.785	0.806
All synsets (Pre-WSD)	0.758	0.765	0.784	0.800

Table 3: Base classifiers accuracy per domain in single-domain polarity classification.

Once evaluated the IGR of the base classifiers, the next step is to evaluate them separately in the polarity classification task. Following the setup of Section 4.2, we can see the results on single-domain in Table 3. The results are in line with those obtained for IGR: (1+2+3)-grams obtained the highest results, followed by BOW. The vocabulary expansion achieved averaged results followed by Adjective WSD and the rest of WSD-based classifiers. Finally, ML-SentiCon was the model with the lowest accuracy. Looking at the results on cross-domain in Table 4, we can see a similar trend. Despite there is a general decrease in the results, as we stated while analyzing its IGR, ML-

[25]Similarly to the other WSD-based classifiers, binary TF is preferred to TF-IDF to smooth the error in case of a wrong disambiguation.

Base classifiers	**Books**	**Electronics**	**DVDs**	**Kitchen**
BOW	0.756	0.804	0.791	0.809
(1+2+3)-grams	0.744	0.798	0.771	0.769
ML-SentiCon	0.643	0.652	0.639	0.673
Noun WSD	0.626	0.625	0.644	0.649
Adjective WSD	0.665	0.687	0.699	0.686
Verb WSD	0.584	0.619	0.590	0.605
Adverb WSD	0.617	0.661	0.622	0.646
Vocab. Exp.	0.666	0.695	0.694	0.695
Other classifiers				
All synsets (Post-WSD)	0.745	0.765	0.776	0.775
All synsets (Pre-WSD)	0.726	0.757	0.765	0.769

Table 4: Base classifiers accuracy per domain in cross-domain polarity classification.

SentiCon has even improved its results on cross-domain, taking advantage of all the other domains to train a domain independent model which is able to outperform the noun, verb and adverb WSD-based approaches. Note that, as we can see in both tables, All synsets (Post-WSD) classifier outperforms the Pre-WSD model, and gets similar results to the best base classifiers.

Looking at all the previous results, due to the different type of classifers selected, each one of them should provide different information when combined in a meta-classifier. The next experiment studies the improvement in the accuracy when adding base classifiers one by one to our KE-Meta approach. We can see the single-domain results in Figure 5. As expected, considering the harmonic mean, there is an improvement when each new base classifier is added. As one classifier might provide information included by others, the improvements were shown to be greater at the beginning. The results on cross-domain are shown in Figure 6. Also in this case there is a clear improvement compared to the first base classifier included, being BOW, ML-SentiCon and Adjective WSD, the models with higher contribution. However, the vocabulary expansion seems to have a negative contribution in this cross-domain combination. We assume that expanding vocabulary from different domains and combining all the documents together, contributes to obtaining a noisy base classifier with several clusters of vocabulary of concepts related to each training domain. In the next cross-domain experiments we will show also the results without the vocabulary expansion base classifier.

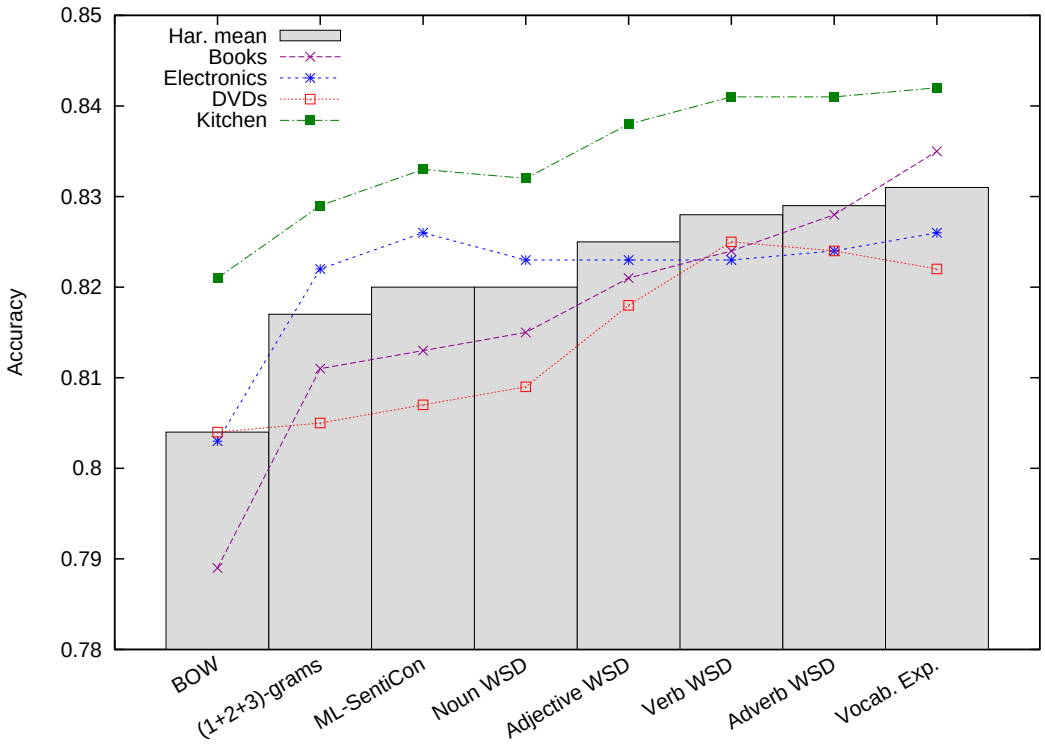

Figure 5: KE-Meta classifier improvement across domains when incrementally adding new base classifiers to single-domain polarity classification. Each column represents the accuracy of the model when we combine that base classifier with the classifiers at its left.

4.4. Single-domain Polarity Classification

We compared our knowledge-enhanced meta classifier against the state-of-the-art SST model, and those proposed by Dredze et al. (2008) and Li and Zong (2008)[26] (cfr Section 2). In addition we included the results of our BOW and (1+2+3)-grams classifiers as baselines.

As we can see from Table 5,[27] thanks to the additional information included when combining groups of words as single feature, (1+2+3)-grams obtained better results than BOW. However, all of the compared models outperformed these baselines. Dredze et al.'s approach obtained interesting re-

[26]Results of compared approaches are taken from their original works: Bollegala et al. (2013), Dredze et al. (2008) and Li and Zong (2008).

[27]In this work, statistically significant results according to a χ^2 test are highlighted in bold.

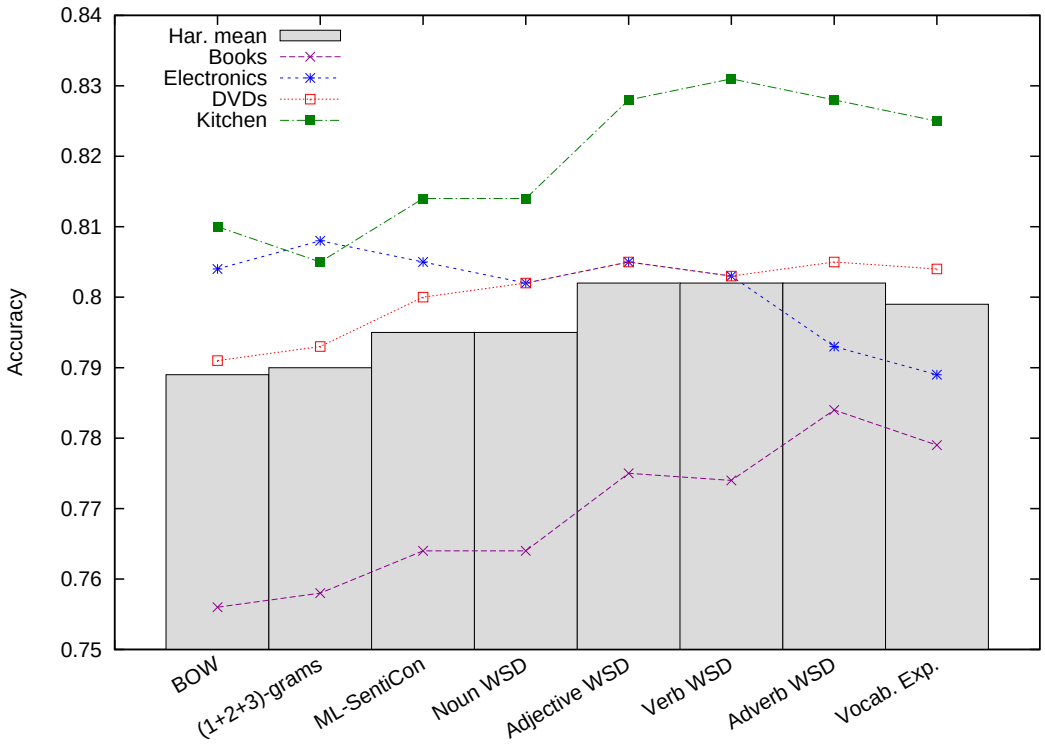

Figure 6: KE-Meta classifier improvement across domains when incrementally adding new base classifiers to cross-domain polarity classification. Each column represents the accuracy of the model when we combine that base classifier with the classifiers at its left.

	Method	Books	Electronics	DVDs	Kitchen
(a)	(Dredze et al., 2008)	**0.826**	**0.859**	0.809	0.857
	SST	0.804	0.844	0.824	**0.877**
	(Li and Zong, 2008)	0.790	**0.850**	**0.845**	0.845
(b)	(1+2+3)-grams	0.805	0.817	0.803	0.819
	BOW	0.788	0.803	0.804	0.821
(c)	KE-Meta	**0.835**	0.826	0.823	0.842

Table 5: Accuracy results in single-domain polarity classification. (a) State-of-the-art approaches; (b) baselines; (c) proposed approach.

sults, specially classifying electronics. This model benefited from confidence-weighted classification to create very precise linear frontiers among classes. The SST model, using its sentiment sensitive thesaurus, took advantage of

the type of reviews used in kitchen domain and obtained the best results, with good accuracy in the other domains. Li and Zong's approach, based on a optimized n-gram selection criteria, obtained the best results on DVD reviews. Our approach obtained the best results on Books domain and considerably high results on the rest. We hypothesize that when reviewers analyze books summarizing parts from the story of the book, our meta-classifier is able to distinguish this pattern by contrasting the probabilities of the base classifiers, and the polarity of the book summary has less influence in the final review classification. Note that our approach is the most stable, with no less than 82.3% of accuracy in all the tests. Using meta-classification, KE-Meta is able to determine which base classifier is better on each domain, maximizing its contribution in the combination. We highlight also that each state-of-the-art approach obtained specially low (or high) results in some domain. This may be produced by the writing style employed by reviewers when commenting on those products. At the end of Section 4.5, in Table 7 we analyze the vocabulary of domains to investigate these differences further.

4.5. Cross-domain Polarity Classification

	Method	**Books**	**Electronics**	**DVDs**	**Kitchen**
(a)	SST	0.763	**0.839**	0.783	**0.852**
	SFA	**0.777**	0.753	0.763	0.815
	SCL-MI	0.746	0.789	0.763	0.820
(b)	BOW	0.756	0.804	0.791	0.809
	(1+2+3)-grams	0.744	0.798	0.771	0.769
(c)	KE-Meta$_B$	**0.784**	0.793	**0.805**	0.828
	KE-Meta	**0.779**	0.789	**0.804**	0.825

Table 6: Accuracy results in cross-domain polarity classification. (a) State-of-the-art approaches; (b) baselines; (c) proposed approaches.

In this task we compared our KE-Meta approach against the state-of-the-art SFA, SCL-MI and SST approaches.[28] As we mentioned in Section 4.3, we included also the results of our approach without the vocabulary expansion-based base classifier: KE-Meta$_B$. The BOW and (1+2+3)-grams models are included as baselines.

[28]The results of the approaches compared are taken from Bollegala et al. (2013)

Table 6 shows the cross-domain polarity classification accuracy. The (1+2+3)-grams baseline achieved the lowest results. Training a cross-domain n-gram-based classifier using only three domains does not seem to be sufficient to obtain a good domain-independent n-gram inventory. Evidence of this observation are the close results obtained by SCL-MI and SFA, other two n-gram based approaches. SCL-MI excelled especially in the kitchen domain. We hypothesize that the singular value descomposition used to reduce dimensions worked better with the reduced size of the vocabulary in kitchen domain. The second domain with less vocabulary, electronics, excelled too. The bipartite graph constructed to differentiate domain-specific and independent n-grams helped SFA to obtain significant results on books domain. Precisely despite obtaining the lowest results in that domain, the BOW baseline outperformed SFA and SCL-MI on average. In contrast to n-gram-based approaches, the training data provided was sufficient to infer a vocabulary, which made this classifier more stable. The SST model proved to be a good option in cross-domain, with significative results on electronics and kitchen reviews. Bollegala et al. (2013) justified the low results on books because of the low number of unlabeled data available on that domain, which is necessary to create its sentiment sentitive thesaurus. Finally, our KE-Meta approach obtained the best results on books and DVD reviews, being again the most stable approach across domains, thanks to the combination of different base classifiers. KE-Meta$_B$, the classifier that does not consider the vocabulary expansion, obtained not significative better results in all domains. Since the use of this base classifer improved the results in single-domain, future work is needed in order to understand how to improve its performance also in cross-domain.

Experimental results of Tables 5 and 6 show that review polarity classification of evaluated approaches differ across domains. These differences could be due to the different language employed by reviewers when commenting on products of different domains. In Table 7 we can see some statistics of the corpus divided by domain. While kitchen appliance and electronic reviewers evaluate using short comments, reviews of book and DVD domains are longer, e.g. some of them include a summary of the story. Interesting also the reduced percentage of nouns in kitchen compared to the rest. It seems that kitchen appliance reviewers do not cite so often other products, and use more qualifying adjectives. This makes this domain the easiest to classify, probably also explained by its shorter length. In general, single-domain n-gram-based approaches obtained better results with the two domains with

Statistics	Books	Electronics	DVDs	Kitchen
Average document length	175	113	190	96
# different lemmas domain	26,108	13,947	28,757	11,095
Average # different lemmas per document	53.4	33.6	57.8	28.3
% nouns domain	66.5%	64.7%	67.4%	61.3%
% adjectives domain	16.7%	15.8%	15.5%	17.4%
% verbs domain	10.0%	11.9%	9.5%	14.1%
% adverbs domain	3.4%	3.7%	3.2%	4%
# different senses domain	17,523	8,809	18,487	8,416
Average # different senses per document	51.2	31.8	54.3	27.9
# different lemmas domain / # different senses domain	0.671	0.632	0.643	0.759
KE-Meta results (single-domain)	**0.835**	0.826	0.823	0.842
KE-Meta results (cross-domain)	**0.784**	0.793	**0.805**	0.828

Table 7: Corpus statistics per domain. Bold results indicate statistical significance.

shorter reviews. However, the same trend is not clearly appreciated for the BOW classifier.

We include in the table also statistics of the disambiguated senses. Note that the ratio between the number of different lemmas per domain and the different senses per domain is a measure of the polysemy employed[29] by reviewers. As we can see, the results of our KE-Meta approach are better when the percentage of polysemy is lower and, consequently, less WSD effort is required.

5. Conclusions

In this work we introduced a knowledge-enhanced meta-classifier for single and cross-domain polarity classification. The main contributions of this work are: (i) KE-Meta, a new generic approach that combines different types of classifiers to categorize documents according to their polarity; and (ii) the study of the impact of WSD and vocabulary expansion-based features as document representation.

In single and cross-domain polarity classification, KE-Meta has proven to perform at par or better than state-of-the-art when classifying Amazon product reviews. Thanks to the combination of different classifiers, our approach obtained the most stable results across domains, and was able to excel in domains such as books and DVDs, which often combine a review

[29] A value of 1.0 here highlights 0% of polysemy in the corpus.

and a summary of the product together. In contrast to the state-of-the-art, our meta-classifier does not perform any domain adaptation, which renders our approach generic. Moreover, the study of the information gain of our base classifiers concluded that WSD and vocabulary expansion-based features provide additional information not included in other BOW or n-gram-based classifiers.

Future work will investigate how it affects the inclusion of new base classifiers in KE-Meta. The use of other state-of-the-art approaches combined with our approach should provide better results. In addition, we will improve the current base classifiers, specially the vocabulary expansion-based one, to perform better both at single and cross-domain level. We will study also the performance of our classifier in other popular datasets like the well-known movie review dataset. Moreover, we will evaluate our polarity classification approach in other languages.[30] Finally, we will investigate how to apply multilingual semantic networks and knowledge graphs in other NLP tasks, from both, monolingual and multilingual perspectives.

Acknowledgments

This research has been carried out in the framework of the European Commission WIQ-EI IRSES (no. 269180) and DIANA-APPLICATIONS - Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01) projects. This research is partially funded by the national project ACOGEUS (TIN2012-38536-C03-02) and the regional project AORESCU (P11-TIC-7684 MO). We thank Juan M. Cotelo and Luis A. Leiva for their support and comments.

[30] As we stated in Section 3.1, our approach is multilingual. This is due to the use of BabelNet, which performs WSD, vocabulary expansion, and mapping of SentiWordNet with the disambiguated words.

References

Agirre, E., de Lacalle, O. L., Soroa, A., 2014. Random walks for knowledge-based word sense disambiguation. Computational Linguistics 40 (1), 57–84.

Agirre, E., Martínez, D., de Lacalle, O. L., Soroa, A., 2006. Two graph-based algorithms for state-of-the-art wsd. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 585–593.

Baccianella, S., Esuli, A., Sebastiani, F., may 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). pp. 2200–2204.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J. W., 2010. A theory of learning from different domains. Machine learning 79 (1-2), 151–175.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S., 2009. Dbpedia - a crystallization point for the web of data. J. Web Sem. 7 (3), 154–165.

Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Wortman, J., 2008. Learning bounds for domain adaptation. In: Advances in neural information processing systems. pp. 129–136.

Blitzer, J., Dredze, M., Pereira, F., 2007. Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. pp. 187–205.

Bollegala, D., Weir, D., Carroll, J., 2011. Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, pp. 132–141.

Bollegala, D., Weir, D., Carroll, J., 2013. Cross-domain sentiment classification using a sentiment sensitive thesaurus. Knowledge and Data Engineering, IEEE Transactions on 25 (8), 1719–1731.

Cavnar, W., 1995. Using an n-gram-based document representation with a vector processing retrieval model. NIST SPECIAL PUBLICATION SP, 269–278.

Chang, C.-C., Lin, C.-J., 2011. Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2 (3), 27.

Cruz, F. L., Troyano, J. A., Pontes, B., Ortega, F. J., 2014. Building layered, multilingual sentiment lexicons at synset and lemma levels. Expert Systems with Applications 41 (13), 5984–5994.

Dai, W., Xue, G.-R., Yang, Q., Yu, Y., 2007. Co-clustering based classification for out-of-domain documents. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '07. ACM, New York, NY, USA, pp. 210–219.

Daumé III, H., 2007. Frustratingly easy domain adaptation. In: Proceedings of the annual meeting on association for computational linguistics (ACL 07). pp. 256–263.

Di Marco, A., Navigli, R., 2013. Clustering and Diversifying Web Search Results with Graph-Based Word Sense Induction. Computational Linguistics 39 (3), 709–754.

Dredze, M., Crammer, K., Pereira, F., 2008. Confidence-weighted linear classification. In: Proceedings of the 25th international conference on Machine learning. ACM, pp. 264–271.

Ehrlich, K., Rapaport, W. J., 1997. A computational theory of vocabulary expansion. Department of Computer Science, State University of New York at Buffalo.

Ehrlich, K. A., 1995. Automatic vocabulary expansion through narrative context. Ph.D. thesis, Buffalo, NY, USA, UMI Order No. GAX95-25550.

Enríquez, F., Cruz, F. L., Ortega, F. J., G Vallejo, C., Troyano, J. A., 2013. A comparative study of classifier combination applied to nlp tasks. Information Fusion 14 (3), 255–267.

Esuli, A., Sebastiani, F., 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In: In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06). pp. 417–422.

Fellbaum, C., 1998. WordNet: An electronic lexical database. Bradford Books.

Forman, G., 2008. Bns feature scaling: an improved representation over tf-idf for svm text classification. In: Proceedings of the 17th ACM conference on Information and knowledge management. ACM, pp. 263–270.

Franco-Salvador, M., Gupta, P., Rosso, P., 2013. Cross-language plagiarism detection using a multilingual semantic network. In: Proc. of the 35th European Conference on Information Retrieval (ECIR'13). LNCS(7814). Springer-Verlag, pp. 710–713.

Franco-Salvador, M., Rosso, P., Navigli, R., 2014. A knowledge-based representation for cross-language document retrieval and categorization. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Gothenburg, Sweden, pp. 414–423.

Hall, M. A., Smith, L. A., 1998. Practical feature subset selection for machine learning. In: Proceedings of Australian Computer Science Conference. pp. 181–191.

Hamouda, A., Rohaim, M., 2011. Reviews classification using sentiwordnet lexicon. In: World Congress on Computer Science and Information Technology.

Hoffart, J., Suchanek, F. M., Berberich, K., Weikum, G., 2013. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. Artificial Intelligence 194, 28–61.

Jia, L., Yu, C., Meng, W., 2009. The effect of negation on sentiment analysis and retrieval effectiveness. In: Proceedings of the 18th ACM conference on Information and knowledge management. ACM, pp. 1827–1830.

Joachims, T., 1998. Text categorization with support vector machines: Learning with many relevant features. Springer.

Li, S., Zong, C., 2008. Multi-domain sentiment classification. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers. Association for Computational Linguistics, pp. 257–260.

Martín-Valdivia, M.-T., Martínez-Cámara, E., Perea-Ortega, J.-M., Ureña-López, L. A., 2013. Sentiment polarity detection in spanish reviews combining supervised and unsupervised approaches. Expert Systems with Applications 40 (10), 3934 – 3942.

Mayfield, J., McNamee, P., 1999. Indexing using both n-grams and words. NIST SPECIAL PUBLICATION SP, 419–424.

McCallum, A., Nigam, K., 1998. A comparison of event models for naive bayes text classification. In: AAAI-98 workshop on learning for text categorization. Vol. 752. Citeseer, pp. 41–48.

Moro, A., Raganato, A., Navigli, R., 2014. Entity linking meets word sense disambiguation: A unified approach. Transactions of the Association for Computational Linguistics (TACL) 2, 231–244.

Navigli, R., 2009. Word Sense Disambiguation: a survey. ACM Computing Surveys 41 (2), 1–69.

Navigli, R., Jurgens, D., Vannella, D., 2013. Semeval-2013 task 12: Multilingual word sense disambiguation. In: Proceedings of the 7^{th} International Workshop on Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013). Atlanta, USA, pp. 222–231.

Navigli, R., Lapata, M., 2010. An experimental study of graph connectivity for unsupervised word sense disambiguation. IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (4), 678–692.

Navigli, R., Ponzetto, S. P., 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence 193, 217–250.

Ohana, B., Tierney, B., 2009. Sentiment classification of reviews using sentiwordnet. In: 9th. IT & T Conference. p. 13.

Page, L., Brin, S., Motwani, R., Winograd, T., 1998. The PageRank Citation Ranking: Bringing Order to the Web. Tech. rep., Stanford Digital Library Technologies Project.

Pan, S. J., Ni, X., Sun, J.-T., Yang, Q., Chen, Z., 2010. Cross-domain sentiment classification via spectral feature alignment. In: Proceedings of the 19th international conference on World wide web. ACM, pp. 751–760.

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-Volume 10. Association for Computational Linguistics, pp. 79–86.

Pilehvar, M. T., Navigli, R., 2014. A large-scale pseudoword-based evaluation framework for state-of-the-art word sense disambiguation. Computational Linguistics 1 (1).

Ponzetto, S. P., Navigli, R., 2010. Knowledge-rich Word Sense Disambiguation rivaling supervised system. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden, pp. 1522–1531.

Quinlan, J. R., 1986. Induction of decision trees. Machine learning 1 (1), 81–106.

Quinlan, J. R., 1996. Bagging, boosting, and c4. 5. In: AAAI/IAAI, Vol. 1. pp. 725–730.

Raileanu, L. E., Stoffel, K., 2004. Theoretical comparison between the gini index and information gain criteria. Annals of Mathematics and Artificial Intelligence 41 (1), 77–93.

Reyes, A., Rosso, P., 2013. On the difficulty of automatically detecting irony: beyond a simple case of negation. Knowledge and Information Systems, 1–20.

Salton, G., Fox, E. A., Wu, H., 1983. Extended boolean information retrieval. Communications of the ACM 26 (11), 1022–1036.

Salton, G., McGill, M. J., 1986. Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York, NY, USA.

Sebastiani, F., 2002. Machine learning in automated text categorization. ACM computing surveys (CSUR) 34 (1), 1–47.

Shen, H., Bunescu, R., Mihalcea, R., June 2013. Coarse to fine grained sense disambiguation in wikipedia. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity. Association for Computational Linguistics, Atlanta, Georgia, USA, pp. 22–31.

Sowa, J. F., 2006. Semantic networks. Encyclopedia of Cognitive Science.

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. Association for Computational Linguistics, pp. 417–424.

Van Halteren, H., Zavrel, J., Daelemans, W., 1998. Improving data driven wordclass tagging by system combination. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics, pp. 491–497.

Wolpert, D. H., 1992. Stacked generalization. Neural networks 5 (2), 241–259.