Document Level Subjectivity Classification Experiments in DEFT'09 Challenge

Cigdem Toprak and Iryna Gurevych

Ubiquitous Knowledge Processing Lab, Technische Universität Darmstadt
Hochschulstr. 10, 64289, Darmstadt, Germany
www.ukp.tu-darmstadt.de

Résumé – Abstract

Cet article présente nos expériences de classification supervisée pour la subjectivité au niveau des documents, pour l'anglais et pour le français, au cours du Défi DEFT'09 de fouille de textes. Nous avons testé des traits portant sur les *mots*, les *parties du discours* et sur des *vocabulaires spécialisés* pour faire fonctionner un classificateur SVM. Nos expériences sur les traits des *mots* examinent d'une part l'utilité de l'information contextuelle, et procèdent d'autre part à une comparaison, sur cette tâche, entre les représentations *binaires* et *tf*idf*. Nous montrons que des distributions différentes pour les classes privilégient des représentations différentes pour les traits. Puis, sur l'anglais, nous comparons trois vocabulaires spécialisés dans l'expression des opinions, deux d'entre eux étant bien connus. Ce sont les indices de subjectivité de (Wiebe et Riloff, 2005; Wilson et al., 2005), *SentiWordNet* (Esuli et Sebastiani, 2006), et une liste de verbes compilée à partir de (Santini, 2007 ; Biber et al., 1999). Malgré sa faible couverture, ce lexique de 156 verbes donne d'assez bons résultats pour l'anglais.

In this paper, we present our supervised document level subjectivity classification experiments for English and French at the DEFT'09 Text Mining Challenge. We experiment with the *word*, *POS*, and *lexicon-based* features using an SVM classifier. Our *word* feature experiments *(i)* investigate the utility of the context information, and *(ii)* compare the *binary* and *tf*idf* feature representations in this task. We show that different class distributions favor different feature representations. Furthermore, on the English collection, we compare three, two of which are well-known, opinion lexicons at this task: the subjectivity clues from (Wiebe and Riloff, 2005; Wilson et al., 2005), *SentiWordNet* (Esuli and Sebastiani, 2006), and a list of verbs compiled from (Santini, 2007; Biber et al., 1999)[1]. We show that, despite its limited coverage, the verb lexicon, consisting of 156 verbs, establishes relatively good results in English.

Mots-clefs – Keywords

classification de textes, analyse supervisée de la subjectivité
text classification, supervised subjectivity analysis

1 Introduction

Distinguishing factual information from opinions plays a crucial role for many natural language processing applications in deciding which information to extract or retrieve or how to organize and present different types of information. For instance, an information retrieval system can aim at retrieving articles containing opinions in favor of a particular policy or decision, an information extraction system may need to extract only factual information, and a review aggregation system may require aggregating positive or negative opinions about a topic.

Subjectivity and sentiment analysis, a.k.a. opinion mining, are recent research directions focusing on the computational treatment of subjectivity, sentiments and opinions in text. Subjectivity analysis aims at classifying the content as objective vs. subjective. Sentiment analysis, on the other hand, involves several additional sub-tasks,

[1]Both English and French versions of these verbs can be found at: `http://www.ukp.tu-darmstadt.de/data/sentiment-analysis`

such as: *(i)* determining the emotional orientation (polarity) of the subjective content, *(ii)* determining the strength of the polarity, *(iii)* determining the targets of the opinions in text, and *(iv)* determining the holders of the opinions in text.

Two of the DEFT'09 Text Mining Challenge tasks this year have focused on subjectivity analysis:

- **Task-1** is a document level subjectivity classification task which required binary classification of the newspaper articles as *subjective* or *objective*.
- **Task-2** is detecting the subjective parts of each individual document.

Our team participated in the first task in English and French. We used Support Vector Machine (SVM) classifiers (Joachims, 1998; Forman, 2003) as SVMs are shown to be among the top performing classifiers for high dimensional feature spaces as in the case of document level text classification. We utilized *word*, *part-of-speech (POS)*, and *lexicon-based* features in different configurations. *Lexicon-based* features were created using SentiWordNet (Esuli and Sebastiani, 2006), a list of subjectivity clues from previous works (Wiebe and Riloff, 2005; Wilson et al., 2005), and a list of verbs from (Santini, 2007; Biber et al., 1999). For our official submissions, we adopted a "kitchen sink" approach combining a variety of features. In this paper, besides our preliminary implementation employed for submissions, we report on additional experiments on the training and test corpora that investigate the contribution of various feature classes separately.

This paper is organized as follows: Section 2 introduces the related work in document level subjectivity classification. Section 3 explains our features. We discuss our experimental results in Section 4. Finally, we draw some conclusions in Section 5.

2 Related Work

Diverse features and classification algorithms have been investigated in document level subjectivity and sentiment classification tasks in previous works. Highest performance in document level subjectivity classification task for newspaper articles (F-measure 0.97) was established by Yu and Hatzivassiloglou (2003) using a Naive Bayes classifier with unigrams as features without stemming and stopword removal. Wiebe et al. (2004) present a detailed study for identifying *potential subjective elements*, i.e., subjective words and phrases, by clustering words according to their distributional similarity. They report accuracies up to 0.94 for document level subjectivity classification on a similar newspaper collection using the k-nearest neighbor algorithm based on the normalized counts of the *potential subjective elements* in each document. Similarly, we utilize *lexicon-based* features representing normalized counts of lexicon instances in a document.

Pang et al. (2002) compared three classification algorithms, i.e. Naive Bayes, maximum entropy and SVM, with different feature configurations in a document level sentiment classification task for movie reviews. They show that using words as binary features performs better than using word frequencies as features. Pang et al. (2002) perform their evaluations using accuracy. For French, we observe no difference between two representations in terms of accuracy. However, we receive a better recall at the cost of a lower precision with the *binary* representation. For English, frequency-based features (*tf*idf*) yield a better result than the *binary* features. Furthermore, they report that unigrams outperform bigrams in the same task. We confirm this finding for the English collection. They also show that SVM is the best performer although not by a significant margin.

Subjectivity classification has its roots in genre classification. Similar to genre, subjectivity of documents can be regarded as orthogonal to the topic, i.e., an objective or a subjective document may have the same topic. Finn and Kushmerick (2003) view document level subjectivity classification as a genre classification task and aim at building domain independent subjectivity classifiers. They investigate the utility of three different types of features (bag-of-words, POS statistics and text statistics) across three domains for subjectivity classification. They show that bag-of-words performs best in single topic domains and worst in the cross domain experiments indicating that there are keywords conveying subjectivity within each topic domain. POS statistics yields the best results in cross domain experiments as it allows a better abstraction over a topic dependent model. We explore *POS* features in isolation and in combination with domain independent *lexicon-based* features. As our bag-of-word approaches, i.e., *word* features, outperform our domain independent lexicon or POS combinations, we also confirm that keywords play a crucial role in this subjectivity classification task.

In a document level sentiment classification task, Généreux and Santini (2007) explore the effect of different feature weighting schemes and the utility of macro-features called *linguistic facets*, which were shown to be effective in the

Web genre classification by Santini (2007). *Linguistic facets* include features which can be functionally interpreted, e.g., high frequency of the first person pronouns indicate an argumentative style. We use some *linguistic facet* features introduced in (Généreux and Santini, 2007) like the communication and mental verbs from (Biber et al., 1999) among our *lexicon-based* features.

3 Approach

We used an SVM^{perf} classifier[2][3] with a linear kernel. SVMs are *large margin* classifiers which aim to find a hyperplane (for two class problems) for separating the document vectors in one class from those in the other while keeping the separation, i.e., the *margin*, as large as possible. Classifying new documents is done by determining which side of the hyperplane they fall into.

Typically, in text classification documents are represented as vectors of feature counts. A feature can be as simple as the occurrence of a certain word or represent complex phenomena which can be observed in the document. For instance, a feature may represent the co-occurrence of a modal verb and a first person pronoun in the same sentence. There are different ways to represent feature counts. One way is to use a binary representation which indicates the presence (1) or absence (0) of the feature in the document. Another common approach is to represent each feature with a function of its frequency in the document. We explored both binary and frequency based representations in our experiments. For the frequency based representation, we used *tf*idf* (term frequency multiplied by inverse document frequency) as shown in the formula:

$$tf * idf = (1 + log(tf_{i,j}))log\frac{N}{df_i} \quad (1)$$

where $tf_{i,j}$ is the number of occurrences of $word_i$ in $document_j$, N is the total number of documents, df_i is the number of documents which $word_i$ occurs in.

We performed lemmatization, but applied no stop word removal. The documents are preprocessed with the TreeTagger[4] POS tagger (Schmid, 1994) and the Standford Named Entity Recognizer[5] (Finkel et al., 2005).

All of the features we used in our experiments can be grouped under three major classes as: *word features*, *POS features*, and *lexicon-based features*. Table1 illustrates all features used in our experiments. Next we explain each feature class in detail.

3.1 Word features

This feature class represents each word as a feature. We investigated the contribution of context information as well as the effect of unigrams and bigrams in our different experiments. The context information is represented with the *word_window* feature, which encodes the previous and the next token of the current token. Feature *lemma_tfidf* represents the tf*idf values of lemmas as features. Similar to the *word_window* feature, *lemma_tfidf_window* represents the context of the lemma, but uses tf*idf counts instead of the binary representation.

3.2 Lexicon-based features

Lexicon-based features are built based on three resources: the subjectivity clue lexicons from previous works (Wiebe and Riloff, 2005; Wilson et al., 2005), hereafter referred to as the *Wilson lexicon*, the lexical semantic resource *SentiWordNet* created by (Esuli and Sebastiani, 2006), and a list of verbs taken from (Santini, 2007) originating from (Biber et al., 1999), hereafter referred to as *C-M verb lexicon*.

Wilson lexicon consists of three lists of subjectivity clues: *(i)* the prior polarity lexicon, *(ii)* the intensifier lexicon, and *(iii)* the valence shifter lexicon. All three lexicons contain unigram as well as n-gram entries with *POS* and *stemming* attributes. The *POS* attribute indicates the POS of the subjectivity term. The *stemming* attribute indicates

[2] `http://svmlight.joachims.org/svm_perf.html`

[3] Classifier configuration: -c=1 -l=2 for English, -c=5 -l=2 for French. -c parameter represents the trade-off between training error and margin. -l parameter represents the loss function to use. We used the error rate, i.e., the percentage of errors in prediction vector as the loss function.

[4] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[5] http://nlp.stanford.edu/software/CRF-NER.shtml

whether the look-up should be performed with lemmas or tokens. For instance, the look-up for the lexicon entry (word1=abuse pos1=verb stemmed1=y) should be performed with lemmas and match all the verb instances of the entry like "abused" (verb), "abusing" (verb), but not "abuse" (noun) or "abuses" (noun). Entries of the prior polarity lexicon also have the *prior polarity* and *reliability* attributes. *Prior polarity* represents the polarity of an entry out of the context with the possible values of *positive*, *negative*, *both* or *neutral*. The *reliability* attribute indicates whether the entry has a subjective usage most of the time *(strongsubj)*, or whether it has only certain subjective usages *(weaksubj)*. The intensifier lexicon contains a list of intensifier words such as "fierce, enormous, more, most". The valence shifter lexicon contains entries which shift the polarity of an existing opinion towards negative or positive including negation words.

In order to increase the coverage of the original lexicon described above, we looked up the verbs in the prior polarity lexicon in WordNet to check if they also existed as nouns. Eventually, we added 61 nouns with positive and 192 nouns with negative polarities to the original lexicon. Binary features generated from the Wilson lexicon contain the *reliability* and the *polarity* information. Real-valued features *Wilson_clue_count* and *Wilson_strongSubj_clue_ratio* represent the normalized count of lexicon instances in a document and the percentage of the *strongsubj* instances among all existing clues in a document respectively.

SentiWordNet (Esuli and Sebastiani, 2006) is a lexical resource which assigns a triple of numerical scores for positivity *(PosScore)*, negativity *(NegScore)* and objectivity as *(1-(PosScore+NegScore))* to each synset in WordNet. Similar to the Wilson lexicon, *SentiWordNet* contains unigram as well as n-gram entries with the POS information besides the polarity scores. We used the *PosScore* and the *NegScore* of the first sense of the lexicon item as real-valued features. Similar to *Wilson_clue_count*, *SentiWN_count* represents the normalized count of the lexicon instances which have a non-zero *PosScore* or *NegScore* score for the first sense of the lexicon instance.

Communication and mental verbs (C-M verb lexicon): communication verbs include terms like *say, claim, accuse etc.* which often occur in reported speech and communication. Mental verbs include verbs conveying cognitive and emotional meaning such as *appreciate, love, judge etc.* C-M verbs have been taken from (Santini, 2007) who investigated them in the Web genre classification.

Wilson_1.person and *C-M_1.person* features assess the co-occurrence of the lexicon instances and first person pronouns in the same sentence for the *Wilson* and the *C-M verb lexicons* respectively. Similarly, *Wilson_NE* and *C-M_NE* represents the co-occurrence of the lexicon instances and named entities in the same sentence.

Finally, the *C-M verb lexicon* is manually translated to French, and it is the only lexicon used in the experiments for French.

Feature Class	Feature Name	Feature Type	Description
Word	word_lemma	binary	lemma of tokens
	word_window	binary	lemma of the previous and next 2 lemmas
	bigrams	binary	lemmas of the bigrams
	lemma_tfidf	real-valued	tf*idf value of the lemmas
	lemma_tfidf_window	real-valued	tf*idf of the previous and next lemma
POS	POS	binary	POS of the tokens
	POS_window	binary	POS of the previous and next token
	preceded_by	binary	whether token is preceded by an adj. or adv.
	POS_statistics	real-valued	number of pronouns, adj., adv. in each sent.
	modal_in_sentence	binary	existence of a modal verb in each sent.
Lexicon-based	Wilson	binary	existence of Wilson lexicon instances
	Wilson_NE	binary	Wilson word and named entity in the same sent.
	Wilson_1.person	binary	Wilson word and 1st person pr. in the same sent.
	Wilson_clue_count	real-valued	number of lexicon instances pro document
	Wilson_strongSubj_clue_ ratio	real-valued	ratio of strongSubj instances over all clues
	C-M	binary	existence of C-M lexicon instances
	C-M_NE	binary	C-M verb and named entity in the same sent.
	C-M_1.person	binary	C-M verb and 1st person pr. in the same sent.
	SentiWN_Scores	real-valued	positive and negative scores of the first sense
	SentiWN_count	real-valued	number of lexicon instances per document

Table 1: An overview of features used in our experiments

3.3 POS features

The *POS* feature represents the POS of each token, and the *POS_window* feature represents the POS of the previous and the next token as binary features. The *preceded_by* feature encodes whether a token is preceded by an adjective or an adverb. The *modal_in_sentence* feature looks for the existence of modal verbs in each sentence. Finally, the *POS_statistics* feature assesses the number of pronouns, adjectives, and adverbs in each document.

4 Experiments

After the official submissions, we revised some parts of our system, performed additional experiments and evaluated them over the test collections. In this section, we present additional experiments as well as our submissions. Table 2 shows the number of documents in the training and test sets for both languages. Table 3 shows the statistics about the length of the documents in both collections. The English collection had a more balanced class distribution (56% subjective vs. 44% objective) compared to the French collection (17% subjective vs. 83% objective). Documents were labeled as *subjective* or *objective* based on the sections they appeared in within the newspapers. Newspaper articles from the opinionated sections such as *letter from the editor*, *debates* and *analyses* were labeled as *subjective* and articles from the sections reporting facts such as *news in local and foreign politics and economy* were labeled as *objective*.

Language	Subjective	Objective	Total
English train	4426 (56%)	3440 (44%)	7866
English test	2977	2268	5245
French train	4338 (17%)	20838 (83%)	25176
French test	2894	13894	16788

Table 2: Document distribution for the training and test sets

Document length	English	French
Minimum	43	1
Maximum	14316	91126
Average	517	454
St. deviation	374	842

Table 3: Document length in words

Table 4 and Table 6 show our experimental results for the English and French collections respectively. Feature combinations used in our submissions for both languages are presented in Table 5. Precision for each class label is calculated as $P_i = \frac{correctly\ classified\ instances\ for\ class_i}{all\ instances\ classified\ as\ class_i}$ and recall for each class label is calculated as $R_i = \frac{correctly\ classified\ instances\ for\ class_i}{number\ of\ class_i\ instances\ in\ gold\ standards}$ where $class_i \in \{sub, obj\}$. Overall precision and recall are calculated as the arithmetic mean of the precisions and recalls for both class labels. *F-score* is calculated as $F = \frac{2*Precision*Recall}{Precision+Recall}$ using the overall precision and recall. Additionally, we report the accuracy for each experiment to enable a comparison with a majority class baseline.

We experimented with each feature class in isolation to understand their contribution to the specific classification task at hand. Next, we discuss our results for both collections.

4.1 Results

For both languages, we observe that *word* class features perform superior compared to the other classes. For a fairly balanced corpus of English, we observe that all groups, in isolation or combined, outperform a majority class baseline (56%) in terms of accuracy. The French collection, on the other hand, has a highly-skewed class distribution in favor of the objective class, i.e., a majority class baseline would already establish a high accuracy of 0.83. For the French collection, only the *word* class features significantly exceed such a baseline in terms of accuracy.

Word class experiments: The experiments performed with *word* class features give us some insights about *(i)* how two different feature representations, i.e. *binary* vs. *tf*idf*, behave, and *(ii)* whether providing context information,

i.e., using *bigrams* or *word_window* as opposed to using *unigrams* as features, would aid the document level subjectivity classification task.

*Binary vs. tf*idf:* For the English collection, we observe that the *tf*idf* representation (W4 in Table 4) outperforms the *binary* representation (W1 in Table 4) for all precision and recall values. On both collections, the *tf*idf* representation increases the subjective precision (Psub), however, on the French collection at the cost of a dramatic decrease in the subjective recall (Rsub) (F-W1 vs. F-W3 in Table 6). In other words, on the French collection which has a highly-skewed class distribution in favor of the objective class, the *binary* representation identifies more instances of the minority class (high subjective recall), whereas the *tf*idf* representation is more precise. Frequency-based representations such as *tf*idf* are known to be effective for classical topic categorization tasks as they determine the content words, i.e., keywords. However, we hypothesize that for subjectivity classification on an uneven class distribution, the binary representation may determine the non-content words constituting evidence for the minority class (for instance, subjective terms), thus, increasing the recall for the minority class. However, this observation requires more investigation before drawing definite conclusions.

Context vs. no context: The binary features *word_window* and *bigrams* provide context information to the classifier. The *word_window* feature represents the previous and the next lemma as features. The *bigram* feature represents two consecutive lemmas as features. The experiments using *word_window* (W2 in Table 4) and bigrams (W3 in Table 4) are outperformed by the experiments using unigrams as features (W1). However, for the French collection, using *word_window* features (F-W2 in Table 6) increases the subjective precision and the objective recall compared to unigrams as features (F-W1). For the French collection, context information enables the classifier to make more precise decisions without damaging the recall too much. However, for the English collection, the classifier does not benefit from the context information.

Lexicon-based class experiments: For the English collection, the experiments using the *lexicon-based* class features in isolation (L1-L7 in Table 4) aim at comparing three domain independent subjectivity lexicons. In other words, the lexicons contain no knowledge of the training or the test collections. The *Wilson lexicon* contains about 6850 unique entries from different POS classes, out of which 990 are multi-word expressions. The *C-M verb lexicon* has 156 verb entries. *SentiWordNet* assigns positivity and negativity scores to all synsets in WordNet Version 2.0, out of which around 9420 unigrams have non-zero subjectivity scores. As a result, *SentiWordNet* constitutes the largest resource, followed by the *Wilson lexicon*. In terms of accuracy, all lexicons perform well above the majority class baseline proving their value for modeling subjectivity regardless of the domain. We see that the *Wilson lexicon* alone (L1) performs better than *SentiWordNet* alone (L5) and the *C-M verb lexicon* alone (L3). Adding complex features, which represent the co-occurrence of the named entities/first person pronouns and a lexicon item in the same sentence, improves the performance for the *C-M verb lexicon* (L3 vs. L4 in Table 4), but it does not contribute to the performance of the *Wilson lexicon* (L1 vs. L2 in Table 4). L3 and L4, the results obtained from the *C-M verb lexicon* restricted to verbs only reveal the potential of using verbs from certain semantic verb categories in subjectivity classification. Finally, the best performance for the *lexicon-based* experiments is obtained by combining all lexicons (L7 in Table 4).

For the French collection, we utilized the manual translations of the *C-M verb lexicon* (FL in Table 6). They proved to be insufficient for identifying subjective documents, establishing a low subjective recall.

POS class experiments: On both collections, the experiments with the *POS* class features (P1 in Table 4 and FP in Table 6) deliver similar results to the experiments using *C-M verb lexicon* (L4 in Table 4 and FL in Table 6). While for the English corpus, using *POS* class features alone establishes an accuracy significantly better than the majority class baseline, for the unbalanced corpus of French, *POS* class shows performance similar to a majority class baseline.

Our submissions: We made three submissions for English and two submissions for French. The first submission for English (S1 for English in Table 5) combines one of the best performing *word* class features, *lemma_tfidf*, with the *lexicon-based* class features. The second submission (S2 for English in Table 5) combines the *lexicon-based* class and the *POS* class features. The third submission is the best performing *word* class feature *lemma_tfidf_window*. For English, based on our revised system, S3 delivers the best results (SR_3 in Table 4). SR_1 in Table 4 (first submission) shows that the lexicons have almost no effect at all when combined with *lemma_tfidf*. The results from the second submission (SR_2), which combines the *lexicon-based* class and the *POS* class, show that adding *POS* features damages the performance of the *lexicon-based* class.

The first submission for French combines *lemma_tfidf* and the *lexicon-based* class features (S1 for French in Table 5). The *lexicon-based* features increase the subjective recall (F-SR_1 vs. F-W3 in Table 5). For the second submission, we combined the *lexicon-based* and *POS* feature classes, which performs better than each feature class in isolation.

Feature Class	Features	Psub	Rsub	Pobj	Robj	P	R	F	Acc
Baseline	assigning majority class	-	-	-	-	-	-	-	0.56
Word	W1: word_lemma	0.878	0.840	0.801	0.847	0.840	0.843	0.841	0.843
	W2: word_window	0.850	0.835	0.789	0.806	0.819	0.821	0.820	0.823
	W3: bigrams	0.876	0.831	0.792	0.845	0.834	0.838	0.835	0.837
	W4: lemma_tfidf	**0.896**	0.842	0.808	**0.872**	0.852	0.857	0.853	0.855
	W5: lemma_tfidf_window	0.893	**0.850**	**0.815**	0.866	**0.854**	**0.858**	**0.855**	0.857
Lexicon-based	L1: Wilson	0.848	0.801	0.757	**0.812**	0.802	0.806	0.803	0.806
	L2: L1, Wilson_NE, Wilson_1.person	0.835	0.814	0.764	0.790	0.800	0.802	0.801	0.804
	L3: C-M	0.738	0.722	0.645	0.663	0.691	0.693	0.692	0.697
	L4: L3, C-M_NE, C-M_1.person	0.748	0.737	0.661	0.675	0.705	0.706	0.705	0.710
	L5: SentiWN_scores	0.836	0.773	0.729	0.802	0.783	0.787	0.784	0.785
	L6: L2, L4	0.844	0.816	0.769	0.803	0.807	0.809	0.808	0.810
	L7: L5, L6	**0.849**	**0.829**	**0.783**	0.807	**0.816**	**0.818**	**0.817**	0.820
POS	P1: all POS group features	0.714	0.815	0.702	0.571	0.708	0.693	0.695	0.710
Submissions official	SO_1	0.836	0.863	0.812	0.778	0.824	0.821	0.822	-
	SO_2	0.791	0.819	0.751	0.716	0.771	0.767	0.769	-
	SO_3	0.783	0.931	0.880	0.662	0.832	0.796	0.814	-
Submissions revised	SR_1	0.876	0.841	0.801	0.843	0.838	0.842	0.840	0.842
	SR_2	0.839	0.824	0.775	0.792	0.807	0.808	0.807	0.810
	SR_3	0.893	0.850	0.815	0.866	0.854	0.858	0.855	0.857

Table 4: Experimental results for the English collection

Submission	English	French
S1	lemma_tfidf Wilson, Wilson_NE, Wilson_1.person C-M, C-M_NE, C-M_1.person SentiWN_Scores, SentiWN_count	lemma_tfidf C-M, C-M_NE, C-M_1.person
S2	Wilson_clue_count Wilson_strongSubj_clue_ ratio Wilson, Wilson_NE, Wilson_1.person C-M, C-M_NE, C-M_1.person SentiWN_Scores, SentiWN_count preceded_by, POS_statistics modal_in_sentence	C-M, C-M_NE, C-M_1.person preceded_by, POS_window modal_in_sentence
S3	lemma_tfidf_window	

Table 5: Features included in the submissions

5 Conclusions

In this paper, we presented our approach and experiments for the document level subjectivity classification task at the DEFT'09 Challenge which required the classification of the newspaper articles as *subjective* or *objective*. We experimented with an SVM^{perf} classifier using features from three different classes including *word*, *POS*, and *lexicon-based* features. We investigate how each feature class in isolation and in combination with other classes performs at the subjectivity classification task on the DEFT collections, the English and French, which have quite disparate class distributions.

Our experiments with the *word* class features reveal that different class distributions favor different feature representations in the document level subjectivity classification task. The English collection, which is almost balanced, benefits consistently from the *tf*idf* representation for all precision and recall values. For the unbalanced French corpus, the *binary* representation yields better subjective recall and the *tf*idf* representation yields better subjective precision. Additionally, with the *word* class experiments we assess the utility of the context information in the subjectivity classification task. We observe that for the English collection, context information, which we model with the *bigram* and *word_window* features, does not contribute, whereby for the French collection, it increases the precision without damaging the F-measure.

The *lexicon-based* experiments investigate the utility of three domain-independent lexicons in the document level subjectivity classification task in English. The *Wilson lexicon* consists of the subjectivity clues from previous works (Wiebe and Riloff, 2005; Wilson et al., 2005). The lexical semantic resource *SentiWordNet* assigns a triplet

Feature Class	Features	Psub	Rsub	Pobj	Robj	P	R	F	Acc
Baseline	assigning majority class	-	-	-	-	-	-	-	0.83
Word	F-W1: word_lemma	0.787	**0.902**	**0.979**	0.949	0.883	**0.926**	**0.902**	0.941
	F-W2: word_window	0.861	0.816	0.962	0.972	0.911	0.894	**0.902**	0.945
	F-W3: lemma_tfidf	**0.920**	0.763	0.952	**0.986**	**0.936**	0.874	0.901	**0.947**
Lexicon	FL: C-M, C-M_NE, C-M_1st person	0.633	0.248	0.861	0.970	0.747	0.609	0.634	0.845
POS	FP: all POS group features	0.706	0.128	0.844	0.988	0.772	0.550	0.550	0.840
Submissions official	F-SO_1	0.783	0.122	0.844	0.993	0.814	0.557	0.662	-
	F-SO_2	0.527	0.752	0.943	0.860	0.735	0.806	0.769	-
Submissions revised	F-SR_1	0.886	0.789	0.957	0.978	0.921	0.884	0.901	0.946
	F-SR_2	0.676	0.359	0.878	0.964	0.777	0.661	0.694	0.859

Table 6: Experimental results for the French collection

of numerical scores for positivity *(PosScore)*, negativity *(NegScore)* and objectivity as *(1-(PosScore+NegScore))* to each synset in WordNet. The *C-M verb lexicon* constitutes a list of communication and mental verbs introduced in (Biber et al., 1999; Santini, 2007). For English, all three lexicons model document level subjectivity better than a majority class baseline. Nevertheless, they lag behind the *word* class features due to: *(i)* the word sense ambiguity of lexicon terms, and *(ii)* the keywords which support the classification task and are incorporated by the *word* class features, but do not appear in the domain independent subjectivity lexicons.

We conclude that the lexicon-based approaches have a potential to provide domain independent subjectivity classifiers. However, such lexicons also call for high-performance word sense disambiguation to be useful in document level subjectivity classification. Furthermore, as subjectivity lexicons do not contain domain specific keywords loaded with subjective connotations, models based on subjectivity lexicons need to be enriched with domain specific information for a better performance.

Acknowledgements

This work was supported by the German Research Foundation (DFG) as part of the *Research Training Group on Feedback Based Quality Management in eLearning* under the grant 1223, and by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806.

References

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Pearson Education Limited.

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

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics*, pages 363–370, Ann Arbor, USA.

Finn, A. and Kushmerick, N. (2003). Learning to classify documents according to genre. In *Proceedings of the Workshop on Computational Approaches to Style Analysis and Synthesis at International Joint Conference on Artificial Intelligence*, Acapulco, Mexico.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305.

Généreux, M. and Santini, M. (2007). Exploring the use of linguistic features in sentiment analysis. In *Proceedings of Corpus Linguistics Conference*, Birmingham, UK.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of European Conference on Machine Learning (ECML)*, pages 137–142, Chemnitz, Germany.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, Philadelphia, PA, USA.

Santini, M. (January 2007). *Automatic Identification of Genre in Web Pages*. PhD thesis, University of Brighton (UK).

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Wiebe, J. and Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. In *CICLing 2005: Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 486–497, Mexico City, Mexico.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., and Martin, M. (2004). Learning subjective language. *Computational Linguistics*, 30(3):277–308.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT/EMNLP'05: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada.

Yu, H. and Hatzivassiloglou, V. (2003). Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 129–136, Sapporo, Japan.