Classifying with Co-Stems
A New Representation for Information Filtering

Nedim Lipka and Benno Stein

Bauhaus-Universität Weimar, 99421 Weimar, Germany
nedim.lipka@uni-weimar.de benno.stein@uni-weimar.de

Abstract. Besides the content the writing style is an important discriminator in information filtering tasks. Ideally, the solution of a filtering task employs a text representation that models both kinds of characteristics. In this respect word stems are clearly content capturing, whereas word suffixes qualify as writing style indicators. Though the latter feature type is used for part of speech tagging, it has not yet been employed for information filtering in general. We propose a text representation that combines both the output of a stemming algorithm (stems) and the stem-reduced words (co-stems). A co-stem can be a prefix, an infix, a suffix, or a concatenation of prefixes, infixes, or suffixes. Using accepted standard corpora, we analyze the discriminative power of this representation for a broad range of information filtering tasks to provide new insights into the adequacy and task-specificity of text representation models. Altogether we observe that co-stem-based representations outperform the classical bag of words model for several filtering tasks.

1 Introduction

Identifying relevant, interesting, high quality, or humorous texts in wikis, emails, and blogs is the tedious job of every searcher. Algorithmic information filtering [4] simplifies this process by finding those texts in a stream or a collection that fulfill a given information interest. Current information filtering technology mostly relies on text classification where the classes describe the information interests. Usually the text representations are content-based, although various filtering tasks are characterized by their intricate combination of content and style.

In this paper we evaluate whether the untapped potential of a style representation in fact is substantial. We propose a model that encodes both (1) text content and (2) text style in the form of word stems and word co-stems respectively. To draw a clear and comprehensive picture of the underlying effects and their importance we resort to a straightforward vector representation. We consider the computational simplicity of this representation as a useful contribution, and to the best of our knowledge the co-stem representation has not been proposed or investigated in this respect. Also the number and heterogeneity of information filtering tasks that are compared in this paper goes beyond existing evaluations. In particular, we analyze the tasks in Table 1 that are marked with ticks (✓) and refer to the relevant literature. In this table, d denotes a plain text document, extracted from an email, a wiki page, a blog entry, or a web page, depending on the task in question.

Table 1. Overview of information filtering tasks. Those tasks which are analyzed in this paper are tagged with the ✓-symbol.

Task	Description	Reference	
Age Group Detection	Determine the age of the author who wrote d.	[17]	
Authorship Attribution	Determine the author of d, given a set of authors.	[18]	✓
Authorship Verification	Determine if d is written by more than one author.	[5]	
Gender Detection	Determine the gender of the author who wrote d.	[17]	✓
Genre Analysis	Determine the genre of d, given a set of genres.	[19]	✓
Information Quality Assessment	Determine whether d is of high quality.	[8]	✓
Language Identification	Determine the language of d.	[3]	
Sarcasm Detection	Determine whether d is sarcastic.	[20]	
Sentiment Analysis	Determine the sentiment expressed in d.	[13]	✓
Spam Detection (email, web page)	Determine whether d is spam or non-spam.	[1, 10]	✓
Topic Detection	Determine the topic of d.	[7]	✓
Vandalism Detection	Determine whether d is vandalized.	[15]	

2 Co-stems

This section introduces the construction of co-stems as the following operation: given a word its stem is computed first, and then the residuals of the word without its stem are taken as co-stems. For example, consider the words "timeless" and "timelessly" along with the application of different stemming algorithms, shown in Table 2.

Table 2. Different co-stems for the words "timelessly" and "timeless" resulting from different stemming algorithms.

Stemming Algorithm	Co-stem	Stem	Co-stem	Co-stem	Stem	Co-stem	Reference
Porter, Lancaster, Krovetz	-	timeless	ly	-	timeless	-	[14, 11, 6]
Lovins	-	time	lessly	-	tim	eless	[9]
Truncation(3)	-	tim	elessly	-	tim	eless	
rev. Truncation(3)	-	timeles	sly	-	timel	ess	

Stems are the output of a stemming algorithm, which is "[. . .] a computational procedure which reduces all words with the same root (or, if prefixes are left untouched, the same stem) to a common form, usually by stripping each word of its derivational and inflectional suffixes" [9]. A root is the base form of a word and cannot be reduced without losing its identity. An inflectional suffix changes the grammatical role of a word in a sentence, it indicates gender, number, tense, etc. A derivational suffix is used for word-formation. For example, the word "timelessly" has the inflectional suffix "ly" and the derivational suffix "less".

A word can have at most three co-stems, namely the part before, after, and inside the stem. Depending on the used stemming algorithm, a co-stem can be a single affix or a combination of affixes. Note that most stemming algorithms are language-dependent, and that some stemming algorithms regard a stem as one or more root morphemes plus a derivational suffix (the Lovins stemmer does not).

3 Evaluation

The general setting in our evaluation is as follows: Given a task, the Lovins stemming algorithm [9] computes the stems of an extracted plain text. The algorithm uses a list of 297 suffixes and strips the longest suffix from a word; hence the resulting co-stems in this study are suffixes. Since the goal is to capture the writing style of a text, we enhance the set of co-stems with stopwords and punctuation. A plain text d is represented by a vector $\mathbf{x}$, where each dimension specifies the frequency of its associated token. A token can be a word, a stem, or a co-stem. We apply as classification technologies a generative approach as well as a discriminative approach, namely Naïve Bayes and linear support vector machines (SVM).

Performance Comparison. Table 3 compares the classification performances of words, stems, co-stems, and stems combined with co-stems. The symbols ○ and ● indicate statistically significant improvement and degradation respectively, compared to the bag of words model in a paired T-test with 0.05 significance. For each precision value, recall value, F-Measure value, and area under ROC curve value (**P**, **R**, **F**, **Auc**) the average is given, weighted by the class distribution. The best solution of a task in terms of the F-Measure is shown bold. The performance scores are averaged over ten repetitions of a 10-fold cross validation. The table also shows details of the used corpora, whereby each corpus is specific for its respective field and accepted as a comparable standard. Since we consider only binary classification tasks, we randomly select two categories for those corpora that cover more than two categories.

Co-stems are effective in Gender Detection, while the combination of co-stems and stems leads to the best classification result. The combination leads also to the best results in Information Quality Assessment and Authorship Attribution, and it is able to compete in Topic Detection. For Genre Analysis (e-shop vs private home page) and Genre Analysis (course vs non course) the performances of the representation based on stems is comparable with the best solution. Finally, the standard bag of words model performs best in Spam Detection and Sentiment Analysis.

Influence of co-stems. To understand of the influence of stems and co-stems in information filtering, Figure 1 illustrates the feature importance characteristics for each task. They show the 10-fold cross validation accuracy scores of the SVM and Naïve Bayes classifiers when the top k features (stems and co-stems) are used. The top k features are computed by the information gain criteria on the training split in each fold. The striped bars below the figures illustrate the preference between stems (white) and co-stems (black) according to the information gain criterion among the top 200 features. The frequent occurrence of co-stems in all tasks among the top 200 features emphasize the discriminative power of co-stems. Each task has its own characteristics that are shown by the classification accuracy and the color distribution. A dark left area indicates a superior impact of co-stems with respect to the classification performance in the specific task, which can be observed in particular for Gender Detection and Information Quality Assessment. Co-stems are valuable within tasks where the texts to be filtered typically originate from a specific writer or group of uniform writers. Examples are Authorship Attribution and Information Quality Assessment, where a high quality Wikipedia article is edited by a group of writers who are likely to share style elements.

Table 3. Classification performance.

	Naïve Bayes				SVM			
Representation	**P**	**R**	**F**	**Auc**	**P**	**R**	**F**	**Auc**
Task: Gender Detection *Corpus*: 400/400 blog entries written by different male/female bloggers. *Source*: "The Blog Authorship Corpus" [17] with 681,288 blog entries from 19,320 bloggers on blogger.com.								
Bag of Words	0.72	0.71	0.71	0.78	0.70	0.70	0.70	0.74
Stems ∪ Co-Stems	0.82 ○	0.82 ○	0.82 ○	0.90 ○	0.87 ○	0.86 ○	**0.86** ○	0.91 ○
Stems	0.77 ○	0.77 ○	0.77 ○	0.84 ○	0.80 ○	0.80 ○	0.80 ○	0.85 ○
Co-Stems	0.83 ○	0.83 ○	**0.83** ○	0.90 ○	0.86 ○	0.85 ○	0.85 ○	0.91 ○
Task: Information Quality Assessment. *Corpus*: 255/255 "featured" (high quality) and "non-featured" articles. *Source*: The english version of Wikipedia.								
Bag of Words	0.79	0.78	0.78	0.87	0.84	0.83	0.83	0.87
Stems ∪ Co-Stems	0.80 ○	0.80 ○	**0.80** ○	0.88 ○	0.87 ○	0.87 ○	**0.87** ○	0.91 ○
Stems	0.80 ○	0.80 ○	**0.80** ○	0.86	0.81 ●	0.81 ●	0.81 ●	0.84 ●
Co-Stems	0.78	0.78	0.78	0.87	0.86 ○	0.85 ○	0.85 ○	0.91 ○
Task: Authorship Attribution *Corpus*: 357/481 blog entries from one author/from all other authors. *Source*: The engineering category from"The Blog Authorship Corpus" [17].								
Bag of Words	0.98	0.97	**0.97**	1.00	0.98	0.98	0.98	1.00
Stems ∪ Co-Stems	0.97	0.97	**0.97**	1.00	0.99 ○	0.99 ○	**0.99** ○	1.00
Stems	0.96 ●	0.96 ●	0.96 ●	1.00	0.98 ○	0.98 ○	0.98 ○	1.00
Co-Stems	0.96 ●	0.95 ●	0.95 ●	1.00	0.95 ●	0.94 ●	0.94 ●	0.99 ●
Task: Topic Detection. *Corpus*: 1,000/800 messages from the (top-level) newsgroups computer-related discussions/recreation and entertainment. *Source*: The well-known "20 Newsgroups" with 20,000 Usenet articles.								
Bag of Words	0.98	0.98	**0.98**	1.00	0.97	0.97	0.97	0.99
Stems ∪ Co-Stems	0.98	0.98	**0.98**	1.00	0.98 ○	0.98 ○	**0.98** ○	0.99
Stems	0.98	0.98	**0.98**	1.00	0.98 ○	0.98 ○	**0.98** ○	1.00
Co-Stems	0.83 ●	0.82 ●	0.82 ●	0.90 ●	0.88 ●	0.88 ●	0.88 ●	0.93 ●
Task: Genre Analysis (e-shop vs private home page). *Corpus*: 200/200 web pages from eshops/personal home pages. *Source*: The "7-web genre collection" [16] with 1,400 web pages.								
Bag of Words	0.96	0.96	0.96	0.99	0.94	0.94	**0.94**	0.98
Stems ∪ Co-Stems	0.95 ●	0.94 ●	0.94 ●	0.98 ●	0.94 ○	0.93	0.93	0.98
Stems	0.97 ○	0.97 ○	**0.97** ○	0.99	0.94	0.93	0.93	0.98
Co-Stems	0.87 ●	0.85 ●	0.85 ●	0.95 ●	0.88 ●	0.86 ●	0.86 ●	0.94 ●
Task: Genre Analysis (course vs non course) *Corpus*: 230/821 web pages about courses/non-courses. *Source*: The subset of "The 4 Universities Data Set" used in the Co-training Experiments [2].								
Bag of Words	0.94	0.94	**0.94**	0.98	0.93	0.91	0.91	0.98
Stems ∪ Co-Stems	0.92 ●	0.92 ●	0.92 ●	0.96 ●	0.91 ●	0.88 ●	0.89 ●	0.98
Stems	0.93 ●	0.93 ●	0.93 ●	0.97 ●	0.93 ○	**0.92** ○	0.92 ○	0.98
Co-Stems	0.88 ●	0.89 ●	0.88 ●	0.91 ●	0.91 ●	0.90	0.90 ●	0.95 ●
Task: Spam Detection *Corpus*: 160/320 spam/non-spam emails. *Source*: The "SpamAssassin public email corpus" with 1,397 spam and 2,500 non-spam emails. *http://spamassassin.apache.org*								
Bag of Words	0.92	0.92	**0.92**	0.97	0.95	0.94	**0.94**	0.98
Stems ∪ Co-Stems	0.92	0.91 ●	0.91 ●	0.96 ●	0.93 ●	0.91 ●	0.91	0.98 ●
Stems	0.92	0.91 ●	0.91 ●	0.96 ●	0.94 ●	0.92 ●	0.93	0.98 ●
Co-Stems	0.89 ●	0.89 ●	0.89 ●	0.95 ●	0.93 ●	0.93 ●	0.93 ●	0.96 ●
Task: Sentiment Analysis *Corpus*: 1,000/1,000 positve/negative movie reviews. *Source*: The "Cornell Movie Review Dataset" [12] with 1,000 positve and 1,000 negative reviews.								
Bag of Words	0.80	0.80	**0.80**	0.88	0.85	0.85	**0.85**	0.91
Stems ∪ Co-Stems	0.76 ●	0.76 ●	0.75 ●	0.84 ●	0.84 ●	0.83 ●	0.83 ●	0.91
Stems	0.81	0.80	**0.80**	0.89	0.82 ●	0.82 ●	0.82 ●	0.89 ●
Co-Stems	0.63 ●	0.62 ●	0.62 ●	0.68 ●	0.72 ●	0.72 ●	0.71 ●	0.79 ●

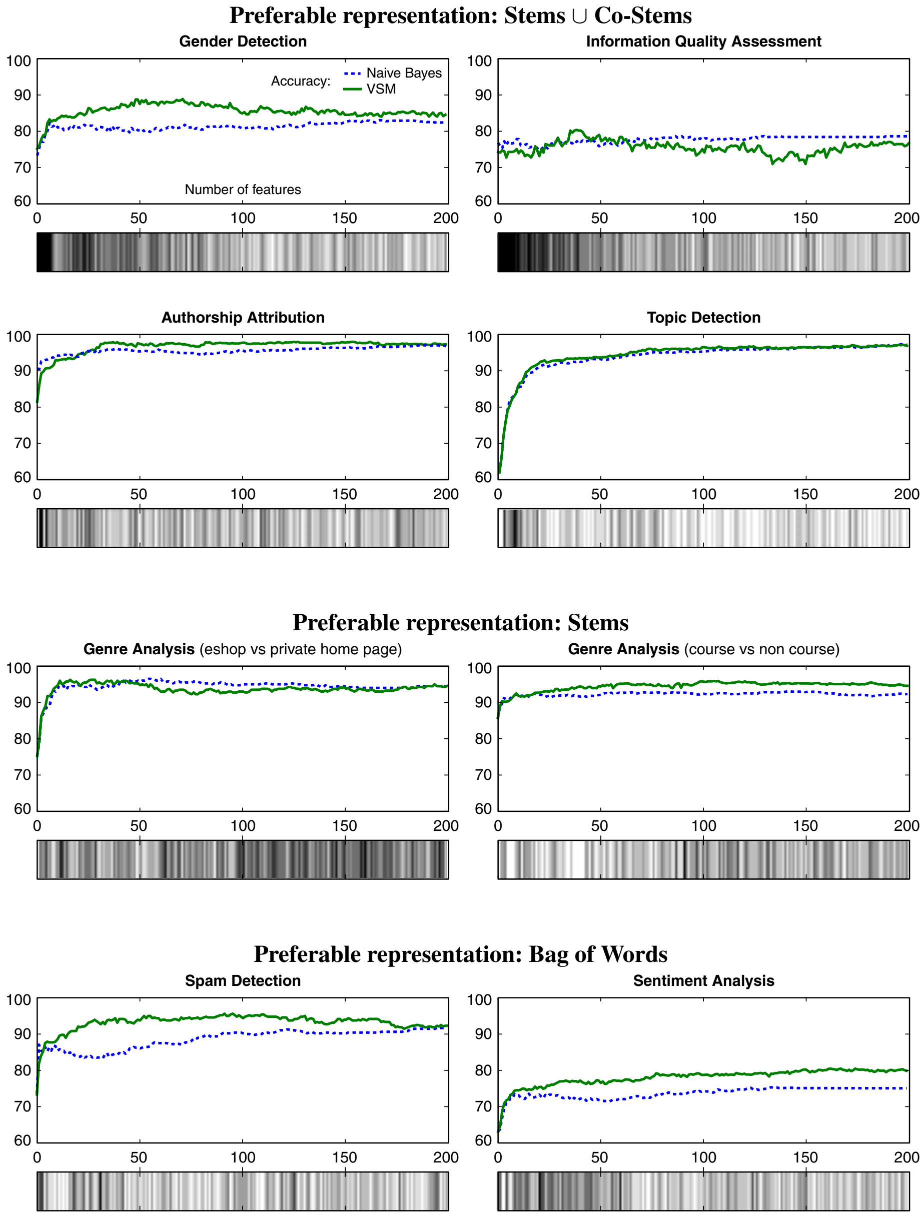

Fig. 1. Task-specific discrimination analysis of stems and co-stems. Each plot shows the classification accuracy (y-axis) over the number m of employed features (x-axis), $m \in [1, 200]$, for a given task. The two curves correspond to the Naïve Bayes classifier (dotted blue) and SVM (solid green) respectively. The striped bars below the plots illustrate whether a stem (white) or a co-stem (black) is chosen by the information gain criterion as the m-th feature: the results are obtained from a 10-fold cross validation, and the exact value of the average calculation is reflected by a gray-scale value. A dark left area indicates the superiority of co-stems over stems.

4 Conclusion

Each information filtering task has its own characteristics in terms of the importance of co-stems. For the tasks Gender Detection, Information Quality Assessment, and Authorship Attribution the combination of stems and co-stems leads to a statistically significant improvement compared to the bag of words model. We provide evidence for the discriminative power of co-stems by setting up experiments with accepted corpora, and by analyzing and illustrating the distribution of the top discriminating features.

References

1. Blanzieri, E., Bryl, A.: A survey of learning-based techniques of email spam filtering. Artificial Intelligence Review 29(1), 63–92 (2008)
2. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proc. of the Workshop on Computational Learning Theory. pp. 92–100 (1998)
3. Gottron, T., Lipka, N.: A comparison of language identification approaches on short, query-style texts. In: Proc. of ECIR. vol. 5993, pp. 611–614 (2010)
4. Hanani, U., Shapira, B., Shoval, P.: Information filtering: Overview of issues, research and systems. User Modeling and User-Adapted Interaction 11(3), 203–259 (2001)
5. Koppel, M., Schler, J.: Authorship verification as a one-class classification problem. In: Proc. of ICML. p. 62 (2004)
6. Krovetz, R.: Viewing morphology as an inference process. In: Proc. of SIGIR. pp. 191–202 (1993)
7. Lang, K.: Newsweeder: learning to filter netnews. In: Proc. of ICML. pp. 331–339 (1995)
8. Lipka, N., Stein, B.: Identifying Featured Articles in Wikipedia: Writing Style Matters. In: Proc. of WWW. pp. 1147–1148 (2010)
9. Lovins, J.B.: Development of a stemming algorithm. Mechanical Translation and Computational Linguistics 11, 22–31 (1968)
10. Ntoulas, A., Najork, M., Manasse, M., Fetterly, D.: Detecting spam web pages through content analysis. In: Proc. of WWW. pp. 83–92 (2006)
11. Paice, C.D.: Another Stemmer. SIGIR Forum 24(3), 56–61 (1990)
12. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proc. of ACL. pp. 271–278 (2004)
13. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proc. of the Conference on Empirical Methods in Natural Language Processing. pp. 79–86 (2002)
14. Porter, M.F.: An algorithm for suffix stripping. Program: Electronic Library & Information Systems 40(3), 211–218 (1980)
15. Priedhorsky, R., Chen, J., Lam, S.T.K., Panciera, K., Terveen, L., Riedl, J.: Creating, destroying, and restoring value in wikipedia. In: GROUP '07: Proc. of the International ACM Conference on Supporting Group Work. pp. 259–268 (2007)
16. Santini, M.: Common criteria for genre classification: Annotation and granularity. In: Third International Workshop on Text-Based Information Retrieval (2006)
17. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.: Effects of age and gender on blogging. In: Proc. of AAAI - Symposium on Computational Approaches for Analyzing Weblogs. pp. 191–197 (2006)
18. Stamatatos, E.: A survey of modern authorship attribution methods. JASIST 60, 538–556 (2009)
19. Stein, B., Eissen, S.M.z., Lipka, N.: Web genre analysis: Use cases, retrieval models, and implementation issues. In: Genres on the Web, vol. 42, pp. 167–189 (2011)
20. Tsur, O., Davidov, D., Rappoport, A.: A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Product Reviews. In: Proc. of AAAI - ICWSM (2010)