Computational Linguistics in the Netherlands Journal 12 (2022) 165-181 Submitted 12/2022; Published 12/2022

Unsupervised Text Classification with Neural Word Embeddings

Andriy Kosar[*,**] ANDREW@TEXTGAIN.COM
Guy De Pauw[*] GUY@TEXTGAIN.COM
Walter Daelemans[**] WALTER.DAELEMANS@UANTWERPEN.BE

[*] *Textgain, Antwerp, Belgium*

[**] *University of Antwerp (CLiPS), Antwerp, Belgium*

Abstract

The paper presents the experiments and results of unsupervised multiclass text classification of news articles based on measuring semantic similarity between class labels and texts through neural word embeddings. The experiments have been conducted on the news texts of various lengths in English and Dutch with a wide range of pre-trained (word2vec, GloVe, fastText) and trained in-domain (word2vec and Doc2Vec) neural word embeddings. The paper demonstrates that distance-based multiclass text classification with neural word embedding can be improved through in-domain training (word2vec and Doc2Vec). Furthermore, we propose two techniques that enrich class label representation with adjacent words in the embedding space: substituting class label with class concept and augmenting class label with additional class label instances. We also argue that improved distance-based text classification with neural word embeddings can be employed for fast text classification in case of a lack of labeled data or frequent changes in class labels, since it is more computationally efficient than novel NLI approaches. Finally, we suggest that the aforementioned method is especially effective if applied to low-resource languages.

1. Introduction

Unsupervised text classification remains an important task in text classification because it avoids time-consuming, costly, and error-prone human data annotation. The task becomes even more significant in environments where textual data changes rapidly (social media, news etc.) and requires fast text classification without labeled data or re-classification of the existing labeled data due to changes in distribution of classes (i.e., merging or splitting existing classes) using predefined classes.

The task of unsupervised text classification can be completed with at least three main methods (Yin et al. 2019), mainly based on 1) frequency of occurrence of the class label in a text, 2) distance between the class label and text in a single vector space, or 3) natural language inference, whether a class label can be inferred from text with pre-trained classifiers.

Recently, with the development of transformer models, the latter method has attracted more attention in the NLP community due to its promising results (Yin et al. 2019, Ding et al. 2022), however with some challenges, such as difficulties to collect NLI data, lack of generalization and stability of classification results (Ma et al. 2021). Even though unsupervised text classification through natural language inference achieves outstanding performance, the downside of it is a high computational cost which prevents it from being applied in real-life situations that deal with large numbers of texts and classes (Reimers and Gurevych 2019)[1]. Meanwhile, more computationally efficient methods through distance-based text classification via neural word embeddings have not yet been systematically studied, a limitation this paper tries to remedy.

1. Reimers and Gurevych (2019) experimentally compared computational time for a semantic textual similarity (STS) task for two similar approaches: a distance-based method with SBERT embeddings and inference for sentence pair classification with BERT.

In our study, we conduct a comprehensive analysis of distance-based news text classification with neural word embeddings in a multilingual set up and describe techniques that enhance this method further. Mainly, we evaluate a wide range of pre-trained (word2vec - Mikolov et al. (2013a); GloVe - Pennington et al. (2014); fastText - Grave et al. (2018)) and custom trained (word2vec and Doc2Vec - Le and Mikolov (2014)) neural word embeddings for English and Dutch. We apply the aforementioned method to news texts of various lengths with a diverse list of simple and complex class labels to evaluate the method's effectiveness. Additionally, we propose two techniques that self-augment class label representation and improve classification results by a large margin. The results of our experiments demonstrate that enhanced distance-based text classification with pre-trained or trained in-domain neural word embeddings can provide a sufficient alternative for more complex unsupervised text classification methods, especially for low-resource languages.

The paper unfolds as follows: Section 2 outlines previous research on unsupervised text classification; Section 3 presents the main method and proposed extensions; Section 4 explains the experimental setup; Section 5 presents evaluation results.

2. Related work

Unsupervised text classification, also known as *dataless text classification* or *zero-shot text classification*, leverages semantic relatedness between class label and document to classify documents without the need for training data. Such a concept was first introduced by Chang et al. (2008) and was implemented by applying Explicit Semantic Analysis (ESA) and Wikipedia as an external knowledge base to encode class labels and document texts in a single semantic space and classify them based on their proximity. This method was further developed by Song and Roth (2014) for hierarchical text classification and by Song et al. (2016) for cross-lingual text classification.

With the emergence of neural word embeddings introduced by Mikolov et al. (2013a) and Mikolov et al. (2013b) word embeddings prevail as a representation of texts for various NLP tasks, and also for unsupervised text classification. Song and Roth (2014) evaluated several types of word embedding models, including word2vec, to compare their semantic representation to ESA. Sappadla et al. (2016) solely evaluated word2vec for multi-label text classification based on semantic similarity between label and document text. Haj-Yahia et al. (2019) employed pre-trained GloVe and trained in-domain word2vec to enrich class labels with similar words and classified documents based on similarity of class label and document encoded with Latent Semantic Analysis (LSA).

Recent development and success of large pre-trained language models, started by Devlin et al. (2019), has shifted unsupervised text classification more into natural language inference tasks. Yin et al. (2019) conducted a comprehensive analysis of unsupervised text classification methods, mainly frequency-based, distance-based (ESA, word2vec) methods with a focus on an entailment approach based on fine-tuning a pre-trained BERT model on various entailment corpora. Ding et al. (2022) and Wang et al. (2022) further developed this approach by fine-tuning the model on Wikipedia categories (TE-Wiki), enhancing model architecture (S-BERT-CAM) correspondingly and also report on performance of word2vec.

3. Method

3.1 Main method

We formulate the problem of unsupervised text classification in the following way: given a set of predefined class labels, classify texts based on the semantic similarity between class label and text (Figure 1). For this purpose, we embed texts and class labels in the same embedding space with pre-trained neural word embeddings, mainly word2vec, GloVe, fastText and trained in-domain neural word embeddings such as Doc2Vec and word2vec. Text classification is conducted using cosine similarity between class labels and text vectors. To assign a class label to a text, we select a class

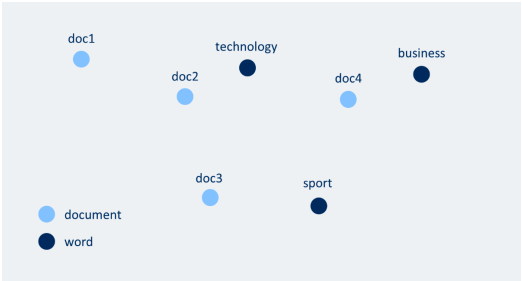

Figure 1: Classification based on the distance between class label and text.

label that is the closest to a given text in one semantic space (smallest cosine distance between class label and text).

3.2 Method improvement

We propose two alternative techniques to improve a class label representation and overcome the sensitivity of classification results to learned class label representations in the pre-trained models: 1) to substitute a class label vector representation with a latent class concept vector representation using words related to a class label, or 2) to augment a class label with additional class label instances based on words related to a class label.

We develop our proposed methods based on the assumption that a class label refers to a latent concept, and describes it explicitly or implicitly. However, there might be a related class label which describes the latent concept better. Sometimes a word or phrase might not be sufficient to convey a concept, therefore humans often use a group of related words or examples to describe it. Following the same logic, we hypothesize that it is possible to make a class label representation more salient by substituting it with a class concept or augmenting the class label with additional class label instances. Both techniques allow to obtain a more robust representation for the class label and consequently improve classification results. A similar principle of class label augmentation - but for other purposes - was applied by Haj-Yahia et al. (2019) with neural word embeddings and by Meng et al. (2020) through Masked Language Modeling (MLM).

Substituting the class label with a class concept.
To obtain a vector representation for a latent class concept we retrieve the N most related words to a class label in the neural word embedding space and construct a centroid vector by averaging vectors of the N most similar words and a class label (Figure 2). For example, to construct a concept vector for the concept sport which is denoted by the class label "sports", we retrieve the top 5 most similar words from the neural word embedding model (word2vec), such as "sport", "sporting", "athletics", "football", "soccer", and make a centroid vector by averaging the vectors for all the retrieved words and the class label itself. In text classification we substitute the class label representation with the concept representation and assign the class concept which is the most semantically similar to the text.

Augmenting the class label with additional class label instances.
To enrich a class label with additional class label instances, we retrieve the N most related words to a class label in the neural word embedding space, and use them as class label alternatives (Figure 3). For example, class label "sports" would have 5 alternatives: "sport", "sporting", "athletics", "football", "soccer". In text classification we assign the class label based on which class label instance or class label is most similar to a text.

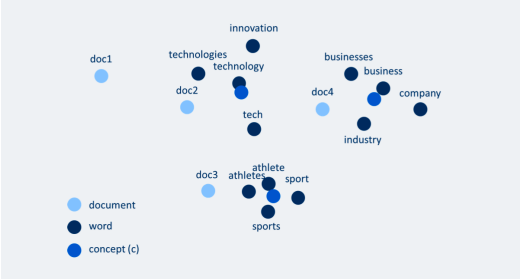

Figure 2: Classification based on the distance between class concept and text.

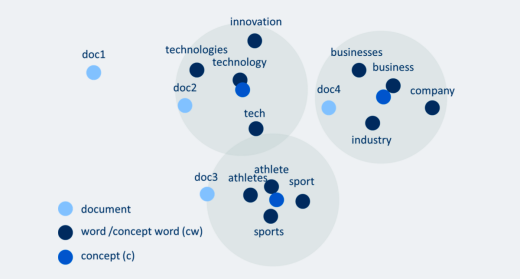

Figure 3: Classification based on the distance between class concept words and text.

4. Experiments

4.1 Datasets & Baseline

For our experiments we used English and Dutch news data. In addition to previously released news datasets, such as AG News (Zhang et al. 2015), BBC News (Greene and Cunningham 2006) and NOS (NL)[2], we constructed two new datasets from New York Times (NYT) and HLN Stories news texts to evaluate the proposed methods' effectiveness on a wider range of classes, more complex class labels and more versatile news texts. As class labels for NYT and HLN Stories we used the news section categories. Additionally, we constructed corresponding datasets of news headlines and abstracts to test the proposed method on various lengths of texts. AG News and HNL Stories lack abstracts. To compare the effect of increase in the number of class labels and impact of ambiguous class labels on classification results we subsetted datasets with simple class labels from NYT, NOS and HLN Stories (NYT6, NOS4, HLN6). The datasets statistics are demonstrated in Appendix A (Table A.1). List of class labels for all datasets is provided in Appendix A (Table A.2).

4.2 Compared Neural Embeddings

We compared three types of pre-trained neural word embeddings, namely word2vec, GloVe and fastText for English and word2vec and fastText for Dutch. Moreover, we experimented with trained in-domain neural embeddings, such as word2vec (Skip-gram) and Doc2Vec (PV-DBOW)[3] to evaluate whether genre specific neural word embeddings can improve the classification results. To train English language models we used New York Times and for Dutch - De Morgen, Het Parool and HLN Stories. For Dutch language we additionally compared four released (Tulkens et al. 2016) pre-trained word2vec models (Wiki, SoNaR, Comb, COW)[4].

4.3 Experimental Setup

As a baseline for the unsupervised text classification evaluation we used frequency-based text classification. Mainly, by calculating the frequency of occurrence of class label or its parts (in case of more complex class label) in a text and assigning the class with highest frequency or randomly selecting a class label in case of ties. To obtain embeddings for compound class labels or texts we averaged word embeddings of constituent words of a text that are present in the models' vocabulary.

2. Available from: https://www.kaggle.com/datasets/maxscheijen/dutch-news-articles
3. We trained Doc2Vec together with word2vec with Gensim library (Řehůřek and Sojka 2010) using the following parameters: Doc2Vec (PV-DBOW), vector_size=300, window=15, min_count=50, epochs=100, hs=1, negative=0, dbow_words=1.
4. Available from: https://github.com/clips/dutchembeddings

Additionally, we experimented with an alternative method of obtaining text embeddings - Doc2Vec, through inferring text embeddings from custom pre-trained Doc2Vec models. For headlines, in case all of the constituent words are not present in a model, we randomly generated an embedding.

We run experiments from 0 up to 300 class concept and class instance sizes with the increment of 5. Where 0 size means the class label itself, and $n > 0$ the number of added adjacent words. For concepts that are constructed from adjacent words to class labels in word embedding space, we selected only those that are present in the corresponding corpus.

To evaluate classification results, we used F1 weighted average. Due to the large list of datasets and models used in our study, we based our conclusions based on general performance (average F1).

5. Results

5.1 Comparing pre-trained word embeddings in default set-up

The results of our experiments demonstrate that distance-based text classification with pre-trained word2vec and trained in-domain Doc2Vec neural word embeddings outperform the frequency-based method by a large margin for English (Table 1) and Dutch (Table 2) and for texts of various lengths. In contrast to the weak performance of frequency-based classification on short texts (headlines and abstracts), distance-based classification with the aforementioned embeddings achieves more than double the F-score. Among pre-trained word embeddings word2vec demonstrates better performance compared to GloVe and fastText. The improved performance of trained in-domain Doc2Vec compared to trained in the same set-up, word2vec indicates that the former is more suitable for encoding document semantics for the current task.

EN Corpora	**Frequency-based**	**Distance-based**				
		Pre-trained			**In-domain trained**	
		word2vec	GloVe	fastText	word2vec	Doc2Vec
NYT6 text body	0.40	0.63	0.32	0.13	0.48	**0.67**
NYT6 abstract	0.23	0.61	0.46	0.25	0.60	**0.65**
NYT6 headline	0.21	0.46	0.41	0.25	**0.56**	0.52
NYT17 text body	0.31	0.10	0.01	0.01	0.01	**0.49**
NYT17 abstract	0.15	0.32	0.05	0.04	0.03	**0.44**
NYT17 headline	0.10	0.28	0.12	0.08	**0.35**	**0.35**
AGNews4 text body	0.27	**0.58**	0.31	0.12	0.28	0.43
AGNews4 headline	0.28	0.48	0.38	0.35	0.35	**0.53**
BBC5 text body	0.35	**0.64**	0.07	0.07	0.07	0.63
BBC5 abstract	0.21	**0.65**	0.26	0.13	0.11	**0.65**
BBC5 headline	0.19	0.53	0.48	0.33	0.33	**0.54**
AVG	**0.25**	**0.48**	**0.26**	**0.16**	**0.29**	**0.54**

Table 1: Comparison of the results (F1 score) obtained from the frequency-based model and distance-based text classification with pre-trained and in-domain trained neural word embedding of English corpora.

The results of distance-based classification with various pre-trained word2vec models for the Dutch language indicate that word2vec models trained on the Wikipedia corpus show the best performance for our task (Appendix B, Table B.1)

5.2 Effect of renaming class labels

In our study, we also analyzed the effect of naming class labels on classification results. Our experiments demonstrate that selecting more explicit class labels yields better performance for both languages. For example, analysis of the distance-based text classification with trained in-domain Doc2Vec on NOS9 text body shows that using clearer class labels (e.g., politics/politiek,

NL Corpora	Frequency-based	Distance-based			
		Pre-trained		In-domain trained	
		word2vec	fastText	word2vec	Doc2Vec
NOS4 text body	0.31	0.49	0.16	0.16	**0.62**
NOS4 headline	0.24	0.51	0.37	0.44	**0.58**
NOS9 text body	0.16	0.21	0.03	0.03	**0.36**
NOS9 headline	0.12	0.28	0.20	0.20	**0.33**
HNL6 text body	0.28	0.30	0.20	0.07	**0.47**
HNL6 headline	0.20	0.44	0.27	0.32	**0.50**
HNL11 text body	0.15	0.06	0.02	0.02	**0.25**
HNL11 headline	0.11	0.20	0.10	0.02	**0.27**
AVG	**0.20**	**0.31**	**0.17**	**0.16**	**0.42**

Table 2: Comparison of the results (F1 score) obtained from the frequency-based model and distance-based text classification with pre-trained and custom trained neural word embedding of Dutch corpora. For Dutch word2vec we report results for the model trained on the Wikipedia corpus.

Class labels	Precision	Recall	F1-score	Support
Binnenland	0.22	0.10	0.13	500
Buitenland	0.14	0.11	0.12	500
Cultuur Media	0.27	0.37	0.31	500
Economie	0.51	0.47	0.49	500
Koningshuis	0.53	0.93	0.68	500
Opmerkelijk	0.28	0.08	0.13	500
Politiek	0.43	0.55	0.49	500
Regionaal nieuws	0.42	0.27	0.33	500
Technologie	0.45	0.68	0.55	500
AVG	**0.36**	**0.40**	**0.36**	**4500**

Table 3: Per class results (F1 score) of distance-based text classification with a trained in-domain Doc2Vec model for NOS9 corpus (text body).

economy/economie, technology/technologie) provide higher F1-score compared to class labels that require additional contextual information (e.g. inland/binnenland, abroad/buitenland, regional news/regionaal nieuws (Table 3).

To elaborate further, we renamed vague class labels with clearer terms for the English corpora, AGNews4 and NYT17. A list of original and renamed class labels with corresponding F1 scores can be found in Appendix B (Table B.2). In most cases, we observe that more specific class labels improve per-class and consequently overall classification results (Table 4).

EN Corpora	Original class labels	Renamed class labels
AGNews4 text body	0.58	0.67
AGNews4 headline	0.48	0.59
NYT17 text body	0.10	0.40
NYT17 abstract	0.32	0.45
NYT17 headline	0.28	0.33
AVG	**0.35**	**0.49**

Table 4: Results (F1 score) of distance-based text classification with pre-trained word2vec model with original and renamed class labels for English corpora.

Class labels	Top 5 most similar words to class label	NYT17 text body	NYT17 abstract	NYT17 headline	AVG
Books	book, tomes, novels, booklist, textbooks	0.07	0.47	0.39	**0.31**
Business Day	businesses, week, morning, days, month	0.13	0.20	0.19	**0.17**
Education	educational, educations, curriculum, vocational, postsecondary	0.07	0.49	0.53	**0.36**
Fashion Style	styles, fashions, couture, flair, chic	0.13	0.12	0.17	**0.14**
Food	foods, foodstuffs, meals, nutritious, meal	0.35	0.41	0.36	**0.37**

Table 5: Table 8. Per class results (F1 score) of distance-based text classification with class labels for English corpus using pre-trained word2vec model. Additionally, the top 5 most similar words to selected class labels in the embedding space are reported.

5.3 Class label selection

In our study, we investigated possible objective criteria for selecting optimal class labels, since more clear class labels demonstrate better performance in classification. We observe that weak class labels tend to have fewer related adjacent words in the embedding space compared to more clear class labels. To illustrate this, we present classification results with class labels using a pre-trained word2vec model for English corpora, NYT17 (Table 5). Overall, F1 score is lower for weak (ambiguous) class labels (such as Business Day, Fashion & Style) compared to more clear class labels (such as Food, Education).

Furthermore, we evaluated the possible relationship between classification performance per class label (F1 score) and coherence of adjacent words for each class label. Coherence measures are a common metric to evaluate topic models(Röder et al. 2015) and can be applied for evaluating the coherence of words related to class labels. Our experiments indicate that there is a weak to moderate correlation between classification performance (F1 score) and coherence score[5] of class label adjacent words (Appendix B, Tables B.3, B.4). Therefore, we hypothesize that manual examination of a class label's representation in the embedding space through review of adjacent words may help to determine more optimal class labels.

5.4 Classifying with class concept and class label instances

Our experiments with substituting class labels with class concept and class label instances demonstrate that both techniques improve the classification results for English and Dutch corpora. While for various corpora optimal concept size varies (see Appendix C) and we lack a method that can determine it, we observe that in general the concept size from 5 to 20 adjacent words (Tables 6, 7; Appendix B, Tables B.5, B.6) yields better results compared to the main method (classification based on class labels). Additionally, we report the maximum achievable results (max) for each method to demonstrate that it might be possible to improve results even further by developing a method for determining optimal concept size.

Regarding the embedding methods, we observe a larger increase in F1 score for trained in-domain Doc2Vec compared to the pre-trained word2vec model for both English and Dutch corpora. Also, there is a larger improvement in the results for English corpora compared to Dutch corpora. We attribute such a difference between English and Dutch corpora to the weak (ambiguous) class labels in Dutch corpora, especially in the case of the NOS9 corpus. Mainly, we hypothesize that enrichment of vague labels with adjacent words dilutes the class label semantics.

5. We calculated c_v coherence score with Gensim library (Řehůřek and Sojka 2010) using NYT and HLN as reference corpora.

EN Corpora	word2vec									
	Class label	Class concepts					Class label instances				
		5	10	15	20	max	5	10	15	20	max
NYT6 text body	0.63	0.68	0.69	0.66	0.66	0.69	0.68	0.70	0.71	0.72	0.73
NYT6 abstract	0.61	0.65	0.66	0.65	0.65	0.66	0.65	0.66	0.67	0.68	0.70
NYT6 headline	0.46	0.47	0.49	0.50	0.50	0.51	0.48	0.49	0.50	0.51	0.53
NYT17 text body	0.10	0.15	0.16	0.25	0.18	0.25	0.17	0.16	0.07	0.02	0.17
NYT17 abstract	0.32	0.37	0.38	0.41	0.39	0.41	0.37	0.37	0.31	0.26	0.37
NYT17 headline	0.28	0.31	0.33	0.34	0.35	0.35	0.31	0.30	0.29	0.27	0.31
AGNews4 text body	0.58	0.60	0.57	0.59	0.60	0.60	0.67	0.67	0.66	0.56	0.67
AGNews4 headline	0.48	0.52	0.49	0.50	0.52	0.52	0.55	0.56	0.54	0.51	0.56
BBC5 text body	0.64	0.57	0.66	0.72	0.70	0.75	0.70	0.69	0.71	0.74	0.75
BBC5 abstract	0.65	0.61	0.68	0.72	0.71	0.75	0.64	0.66	0.71	0.73	0.73
BBC5 headline	0.53	0.51	0.54	0.57	0.58	0.60	0.51	0.49	0.55	0.57	0.58
AVG	**0.48**	**0.50**	**0.51**	**0.54**	**0.53**	**0.55**	**0.52**	**0.52**	**0.52**	**0.51**	**0.55**

Table 6: Results (F1 score) of distance-based text classification with class labels, class concepts (with size of 5-20 adjacent words) and class label instances (with size of 5-20 adjacent words) for English corpora with pre-trained word2vec.

NL Corpora	word2vec									
	Class label	Class concepts					Class label instances				
		5	10	15	20	max	5	10	15	20	max
NOS4 text body	0.49	0.62	0.41	0.35	0.39	0.62	0.35	0.40	0.42	0.54	0.66
NOS4 headline	0.51	0.54	0.52	0.49	0.51	0.57	0.46	0.51	0.48	0.55	0.54
NOS9 text body	0.21	0.20	0.15	0.16	0.14	0.21	0.19	0.15	0.17	0.13	0.34
NOS9 headline	0.28	0.27	0.26	0.26	0.26	0.28	0.26	0.27	0.26	0.26	0.31
HNL6 text body	0.30	0.34	0.38	0.35	0.37	0.53	0.36	0.33	0.33	0.35	0.70
HNL6 headline	0.44	0.44	0.46	0.45	0.46	0.55	0.45	0.42	0.43	0.45	0.60
HNL11 text body	0.06	0.12	0.17	0.19	0.20	0.23	0.09	0.09	0.12	0.12	0.38
HNL11 headline	0.20	0.22	0.23	0.24	0.24	0.26	0.22	0.22	0.23	0.23	0.32
AVG	**0.31**	**0.34**	**0.32**	**0.31**	**0.32**	**0.41**	**0.30**	**0.30**	**0.30**	**0.33**	**0.48**

Table 7: Results (F1 score) of distance-based text classification with class labels, class concepts (with size of 5-20 adjacent words) and class label instances (with size of 5-20 adjacent words) for Dutch corpora with pre-trained word2vec (Wiki).

5.5 Effect of filtering out of vocabulary words

While studying pre-trained word embeddings for distance-based text classification, we discovered that adjacent words to class labels in embedding space often consist of words that were absent from the target classification corpus and such words usually are difficult to interpret. For example, the word 'sports' in the pre-trained word2vec model has 'DeVillers_reports' and 'al_Sunaidy' in the top 10 adjacent words. We excluded such words from building class concept or class label instances to improve the interpretability of class concept and class concept instance. Our experiments show that such filtering not only improves the interpretability but also has a positive impact on classification results (Table 8; Appendix B, Table B.7).

6. Conclusion & Future Work

In this paper, we analyzed the potential of neural word embeddings in unsupervised multiclass text classification and described the effectiveness of various types of pre-trained and trained in-domain word embedding models and document embedding methods on a wide variety of English and Dutch news texts with simple and complex class labels. We conclude that distance-based classification

EN Corpora	word2vec				Doc2Vec			
	Class concept		Class label instances		Class concept		Class label instances	
	no filtering	filtering	no filtering	filtering	no filtering	filtering	no filtering	filtering
NYT6 text body	0.65	0.68	0.07	0.68	0.74	0.74	0.78	0.78
NYT6 abstract	0.66	0.65	0.36	0.65	0.70	0.70	0.73	0.73
NYT6 headline	0.51	0.47	0.36	0.48	0.57	0.57	0.60	0.60
NYT17 text body	0.06	0.15	0.06	0.17	0.52	0.52	0.52	0.52
NYT17 abstract	0.29	0.37	0.20	0.37	0.46	0.46	0.47	0.47
NYT17 headline	0.29	0.31	0.24	0.31	0.37	0.37	0.38	0.38
AVG	**0.41**	**0.44**	**0.21**	**0.44**	**0.56**	**0.56**	**0.58**	**0.58**

Table 8: Results (F1 score) of distance-based text classification with class concepts classes (with size of 5 adjacent words) and class label instances (with size of 5 adjacent words) for English corpora using word2vec and Doc2Vec without and with filtering out of vocabulary words.

with neural word embeddings significantly outperforms frequency based approach especially for short texts.

We demonstrated that with trained in-domain Doc2Vec document representation, it is possible to further improve distance-based classification results compared to pre-trained or trained in domain word2vec. More importantly, we proposed two alternative techniques of self-augmenting class label representation with the most similar words in the embedding space that improves the results of text classification with pre-trained neural word embeddings. Finally, based on experiments with Dutch corpora we demonstrated that distance-based classification with neural word embeddings approach can also be used for low resource languages.

Our findings show that improving class label and document representation with a relatively simple and computationally effective method of distance-based text classification can yield better classification results. The promising directions of further research include the mitigation of the shortcomings of class label (ambiguity) and document (textual redundancy) representation through augmenting class labels with examples and encoding class labels and document texts with contextual embeddings.

7. Acknowledgment

This research was funded by Flanders Innovation & Entrepreneurship (VLAIO), grant HBC.2021.0222.

References

Chang, Ming-Wei, Lev Ratinov, Dan Roth, and Vivek Srikumar (2008), Importance of semantic representation: Dataless classification, *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, AAAI Press, p. 830–835.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019), BERT: Pre-training of deep bidirectional transformers for language understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. https://aclanthology.org/N19-1423.

Ding, Hantian, Jinrui Yang, Yuqian Deng, Hongming Zhang, and Dan Roth (2022), Towards open-domain topic classification, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, Association for Computational Linguistics, Hybrid: Seattle, Washington + Online, pp. 90–98. https://aclanthology.org/2022.naacl-demo.10.

Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov (2018), Learning word vectors for 157 languages, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), Miyazaki, Japan. https://aclanthology.org/L18-1550.

Greene, Derek and Pádraig Cunningham (2006), Practical solutions to the problem of diagonal dominance in kernel document clustering, *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, Association for Computing Machinery, New York, NY, USA, p. 377–384. https://doi.org/10.1145/1143844.1143892.

Haj-Yahia, Zied, Adrien Sieg, and Léa A. Deleris (2019), Towards unsupervised text classification leveraging experts and word embeddings, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, pp. 371–379. https://aclanthology.org/P19-1036.

Le, Quoc and Tomas Mikolov (2014), Distributed representations of sentences and documents, *in* Xing, Eric P. and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, Vol. 32 of *Proceedings of Machine Learning Research*, PMLR, Bejing, China, pp. 1188–1196. https://proceedings.mlr.press/v32/le14.html.

Ma, Tingting, Jin-Ge Yao, Chin-Yew Lin, and Tiejun Zhao (2021), Issues with entailment-based zero-shot text classification, *ACL-IJCNLP 2021*, Association for Computational Linguistics (ACL), pp. 786–796. https://www.microsoft.com/en-us/research/publication/issues-with-entailment-based-zero-shot-text-classification/.

Meng, Yu, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han (2020), Text classification using label names only: A language model self-training approach, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Mikolov, Tomás, Kai Chen, Greg Corrado, and Jeffrey Dean (2013a), Efficient estimation of word representations in vector space, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. http://arxiv.org/abs/1301.3781.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig (2013b), Linguistic regularities in continuous space word representations, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Atlanta, Georgia, pp. 746–751. https://aclanthology.org/N13-1090.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014), GloVe: Global vectors for word representation, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Doha, Qatar, pp. 1532–1543. https://aclanthology.org/D14-1162.

Řehůřek, Radim and Petr Sojka (2010), Software Framework for Topic Modelling with Large Corpora, *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, ELRA, Valletta, Malta, pp. 45–50.

Reimers, Nils and Iryna Gurevych (2019), Sentence-bert: Sentence embeddings using siamese bert-networks, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics. https://arxiv.org/abs/1908.10084.

Röder, Michael, Andreas Both, and Alexander Hinneburg (2015), Exploring the space of topic coherence measures, *Proceedings of the Eighth ACM International Conference on Web Search*

and Data Mining, WSDM '15, Association for Computing Machinery, New York, NY, USA, p. 399–408. https://doi.org/10.1145/2684822.2685324.

Sappadla, Prateek Veeranna, Jinseok Nam, Eneldo Loza Mencía, and Johannes Fürnkranz (2016), Using semantic similarity for multi-label zero-shot classification of text documents, *24th European Symposium on Artificial Neural Networks, ESANN 2016, Bruges, Belgium, April 27-29, 2016*. https://www.esann.org/sites/default/files/proceedings/legacy/es2016-174.pdf.

Song, Yangqiu and Dan Roth (2014), On dataless hierarchical text classification, *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, AAAI Press, p. 1579–1585.

Song, Yangqiu, Shyam Upadhyay, Haoruo Peng, and Dan Roth (2016), Cross-lingual dataless classification for many languages, *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, AAAI Press, p. 2901–2907.

Tulkens, Stephan, Chris Emmery, and Walter Daelemans (2016), Evaluating unsupervised dutch word embeddings as a linguistic resource, *in* Chair), Nicoletta Calzolari (Conference, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, European Language Resources Association (ELRA), Paris, France.

Wang, Yuqi, Wei Wang, Qi Chen, Kaizhu Huang, Anh Nguyen, and Suparna De (2022), Generalised zero-shot learning for entailment-based text classification with external knowledge, *2022 IEEE International Conference on Smart Computing (SMARTCOMP)*, pp. 19–25.

Yin, Wenpeng, Jamaal Hay, and Dan Roth (2019), Benchmarking Zero-shot Text Classification: Datasets, Evaluation, and Entailment Approach, *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://cogcomp.seas.upenn.edu/papers/YinHaRo19.pdf.

Zhang, Xiang, Junbo Zhao, and Yann LeCun (2015), Character-level convolutional networks for text classification, *in* Cortes, C., N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, Vol. 28, Curran Associates, Inc. https://proceedings.neurips.cc/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf.

Appendix A. Corpora

Corpora	Lang	Size	Classes	Mean tokens	Std tokens
NYT6 text body	EN	3000	6	959	562
NYT6 abstract	EN	3000	6	31	31
NYT6 headline	EN	3000	6	8	2
NYT17 text body	EN	8500	17	889	557
NYT17 abstract	EN	8500	17	32	34
NYT17 headline	EN	8500	17	8	2
AGNews4 text body	EN	2000	4	59	14
AGNews4 headline	EN	2000	4	6	2
BBC5 text body	EN	1000	5	354	263
BBC5 abstract	EN	1000	5	23	13
BBC5 headline	EN	1000	5	5	1
NOS4 text body	NL	2000	4	300	204
NOS4 headline	NL	2000	4	7	2
NOS9 text body	NL	4500	9	254	175
NOS9 headline	NL	4500	9	7	2
HNL6 text body	NL	3000	6	274	211
HNL6 headline	NL	3000	6	9	4
HNL11 text body	NL	5500	11	240	204
HNL11 headline	NL	5500	11	9	4

Table A.1: Corpora statistics.

English				Dutch			
NYT6	**NYT17**	**AG News4**	**BBC5**	**NOS4**	**NOS9**	**HLN Stories6**	**HLN Stories11**
Arts	Arts	Business	Business	Cultuur Media	Binnenland	Auto	Auto
Movies	Automobiles	Science Technology	Entertainment	Economie	Buitenland	Geld	Bizar
Science	Books	Sports	Politics	Politiek	Cultuur Media	Showbizz	De Krant
Sports	Business Day	World	Sport	Technologie	Economie	Sport	Geld
Technology	Education		Technology		Koningshuis	Wetenschap Planeet	In de Buurt
Travel	Fashion Style				Opmerkelijk	Woon	Nieuws
	Food				Politiek		Reizen
	Health				Regionaal Nieuws		Showbizz
	Home Garden				Technologie		Sport
	Movies						Wetenschap Planeet
	Real Estate						Woon
	Science						
	Sports						
	Technology						
	Theater						
	Travel						
	World						

Table A.2: Class labels for English and Dutch Corpora.

Appendix B. Results

NL Corpora	Pre-trained word2vec			
	Wiki	**COW**	**Comb**	**SoNaR**
NOS4 text body	0.49	0.29	0.23	0.13
NOS4 headline	0.51	0.43	0.38	0.31
NOS9 text body	0.21	0.13	0.08	0.03
NOS9 headline	0.28	0.26	0.24	0.18
HNL6 text body	0.30	0.36	0.25	0.10
HNL6 headline	0.44	0.44	0.32	0.21
HNL11 text body	0.06	0.02	0.02	0.02
HNL11 headline	0.20	0.05	0.08	0.07
AVG	**0.31**	**0.25**	**0.20**	**0.13**

Table B.1: Results (F1 score) of distance-based text classification with various pre-trained word2vec models for Dutch.

EN Corpora	Class labels		F1	
	Original	**Renamed**	**Original**	**Renamed**
AGNews4 text body	World	Politics	0.48	0.69
	Science Technology	Technology	0.44	0.49
AGNews4 headline	World	Politics	0.36	0.58
	Science Technology	Technology	0.42	0.54
NYT17 text body	Business Day	Business	0.13	0.52
	Fashion Style	Fashion	0.13	0.19
	Home Garden	Gardens	0.38	0.28
	World	Politics	0.00	0.54
NYT17 abstract	Business Day	Business	0.20	0.43
	Fashion Style	Fashion	0.12	0.17
	Home Garden	Gardens	0.29	0.25
	World	Politics	0.08	0.45
NYT17 headline	Business Day	Business	0.19	0.31
	Fashion Style	Fashion	0.17	0.15
	Home Garden	Gardens	0.21	0.16
	World	Politics	0.20	0.32

Table B.2: Per class results (F1 score) of distance-based text classification with original and renamed class labels (using pre-trained word2vec model) for English.

EN Corpora	Doc2Vec	word2vec
NYT6 text	0.54	0.77
NYT6 abstract	0.55	0.63
NYT6 heading	0.54	0.96
NYT17 text	0.89	-0.53
NYT17 abstract	0.86	0.16
NYT17 heading	0.74	0.04
AGNews4 text	0.27	0.35
AGNews4 heading	0.88	0.60
BBC5 text	0.66	0.03
BBC5 abstract	0.72	-0.07
BBC5 heading	0.62	0.10
AVG	**0.66**	**0.28**

Table B.3: Pearson correlation between F1 score and coherence score of top 5 adjacent words to class label in embedding space (using trained in-domain Doc2Vec and pre-trained word2vec) for English.

NL Corpora	Doc2Vec	word2vec
NOS4 text	0.32	0.48
NOS4 heading	0.56	0.43
NOS9 text	0.51	0.81
NOS9 heading	0.52	0.75
HNL6 Stories text	-0.61	-0.38
HNL6 Stories heading	-0.66	-0.37
HNL11 Stories text	0.14	-0.05
HNL11 Stories heading	0.03	0.08
AVG	**0.10**	**0.22**

Table B.4: Pearson correlation between F1 score and coherence score of top 5 adjacent words to class label in embedding space (using trained in-domain Doc2Vec and pre-trained word2vec) for Dutch.

EN Corpora	Doc2Vec									
	Class label	Class concepts					Class label instances				
		5	10	15	20	max	5	10	15	20	max
NYT6 text body	0.67	0.74	0.75	0.76	0.75	0.76	0.78	0.79	0.79	0.78	0.80
NYT6 abstract	0.65	0.70	0.71	0.72	0.71	0.72	0.73	0.74	0.74	0.74	0.75
NYT6 headline	0.52	0.57	0.57	0.57	0.56	0.57	0.60	0.60	0.60	0.58	0.60
NYT17 text body	0.49	0.52	0.51	0.51	0.51	0.52	0.52	0.52	0.53	0.53	0.53
NYT17 abstract	0.44	0.46	0.46	0.46	0.46	0.46	0.47	0.46	0.47	0.47	0.47
NYT17 headline	0.35	0.37	0.37	0.37	0.37	0.37	0.38	0.37	0.37	0.36	0.38
AGNews4 text body	0.43	0.48	0.53	0.53	0.52	0.53	0.43	0.45	0.46	0.49	0.49
AGNews4 headline	0.53	0.55	0.55	0.57	0.57	0.57	0.54	0.53	0.54	0.55	0.56
BBC5 text body	0.63	0.63	0.65	0.67	0.64	0.67	0.65	0.65	0.67	0.66	0.71
BBC5 abstract	0.65	0.65	0.65	0.67	0.65	0.67	0.66	0.66	0.69	0.68	0.71
BBC5 headline	0.54	0.54	0.54	0.55	0.54	0.56	0.52	0.54	0.56	0.57	0.59
AVG	**0.54**	**0.57**	**0.57**	**0.58**	**0.57**	**0.58**	**0.57**	**0.57**	**0.58**	**0.58**	**0.60**

Table B.5: Results (F1 score) of distance-based text classification with class labels, class concepts (with size of 5-20 adjacent words) and class label instances (with size of 5-20 adjacent words) for English corpora with trained in-domain Doc2Vec.

NL Corpora	Doc2Vec									
	Class label	Class concepts					Class label instances				
		5	10	15	20	max	5	10	15	20	max
NOS4 text body	0.62	0.65	0.64	0.64	0.65	0.67	0.64	0.62	0.62	0.62	0.64
NOS4 headline	0.58	0.58	0.59	0.58	0.59	0.60	0.56	0.56	0.55	0.57	0.59
NOS9 text body	0.36	0.35	0.34	0.33	0.33	0.35	0.35	0.33	0.32	0.32	0.35
NOS9 headline	0.33	0.33	0.32	0.32	0.31	0.33	0.32	0.32	0.31	0.30	0.32
HNL6 text body	0.47	0.44	0.45	0.43	0.42	0.45	0.48	0.47	0.47	0.48	0.49
HNL6 headline	0.50	0.51	0.51	0.51	0.50	0.51	0.53	0.51	0.52	0.52	0.53
HNL11 text body	0.25	0.23	0.23	0.22	0.21	0.23	0.25	0.24	0.25	0.25	0.25
HNL11 headline	0.27	0.27	0.28	0.27	0.26	0.28	0.28	0.27	0.27	0.27	0.28
AVG	**0.42**	**0.42**	**0.42**	**0.41**	**0.41**	**0.43**	**0.43**	**0.42**	**0.41**	**0.42**	**0.43**

Table B.6: Results (F1 score) of distance-based text classification with class labels, class concepts (with size of 5-20 adjacent words) and class label instances (with size of 5-20 adjacent words) for Dutch corpora with trained in-domain Doc2Vec.

NL Corpora	word2vec				Doc2Vec			
	Class concept		Class label instances		Class concept		Class label instances	
	no filtering	filtering	no filtering	filtering	no filtering	filtering	no filtering	filtering
NOS4 text body	0.49	0.62	0.42	0.35	0.65	0.65	0.64	0.64
NOS4 headline	0.49	0.54	0.49	0.46	0.58	0.58	0.56	0.56
NOS9 text body	0.19	0.20	0.19	0.19	0.35	0.35	0.35	0.35
NOS9 headline	0.27	0.27	0.26	0.26	0.33	0.33	0.32	0.32
HNL6 text body	0.36	0.34	0.36	0.36	0.44	0.44	0.50	0.46
HNL6 headline	0.45	0.44	0.45	0.45	0.51	0.51	0.54	0.50
AVG	**0.38**	**0.40**	**0.36**	**0.34**	**0.48**	**0.48**	**0.48**	**0.47**

Table B.7: Results (F1 score) of distance-based text classification with class concepts (with size of 5 adjacent words) and class label instances (with size of 5 adjacent words) for Dutch corpora using word2vec Wiki and Doc2Vec without and with filtering out of vocabulary words.

Appendix C. Miscellaneous

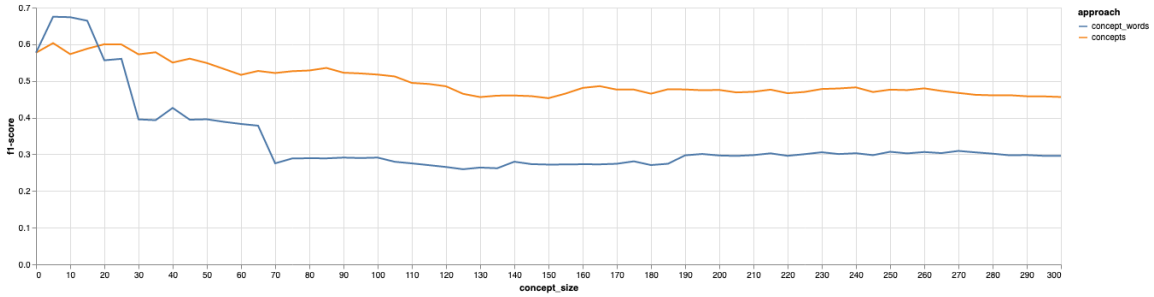

Figure C.1: Results (F1 score) of distance-based text classification with class concept and class label instances with word2vec model with size 0-300 for AG News corpus for texts.

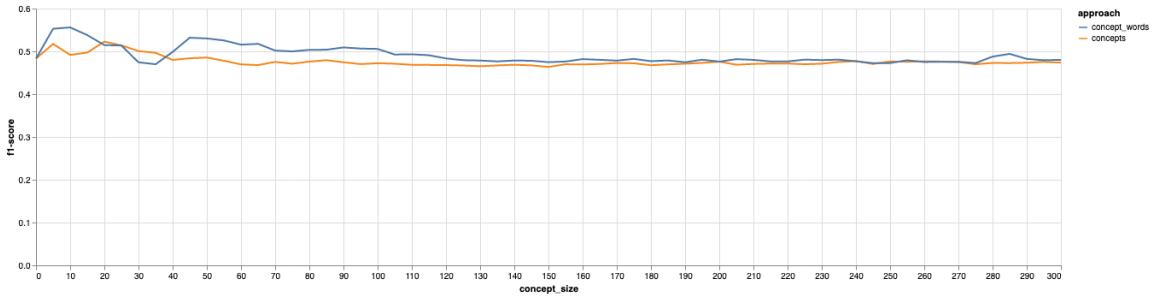

Figure C.2: Results (F1 score) of distance-based text classification with class concept and class label instances with word2vec model with size 0-300 for AG News corpus for headlines.

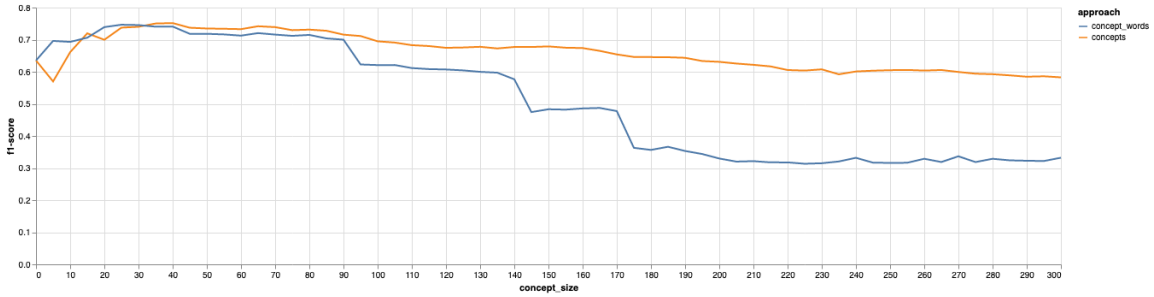

Figure C.3: Results (F1 score) of distance-based text classification with class concept and class label instances with word2vec model with size 0-300 for BBC corpus for texts.

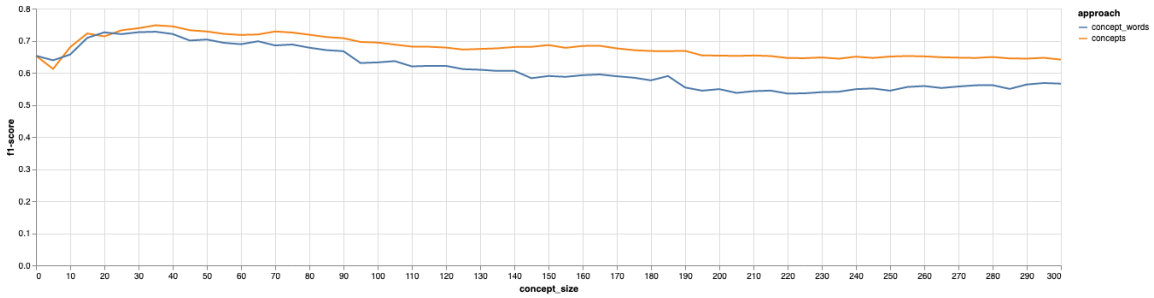

Figure C.4: Results (F1 score) of distance-based text classification with class concept and class label instances with word2vec model with size 0-300 for BBC corpus for abstracts.

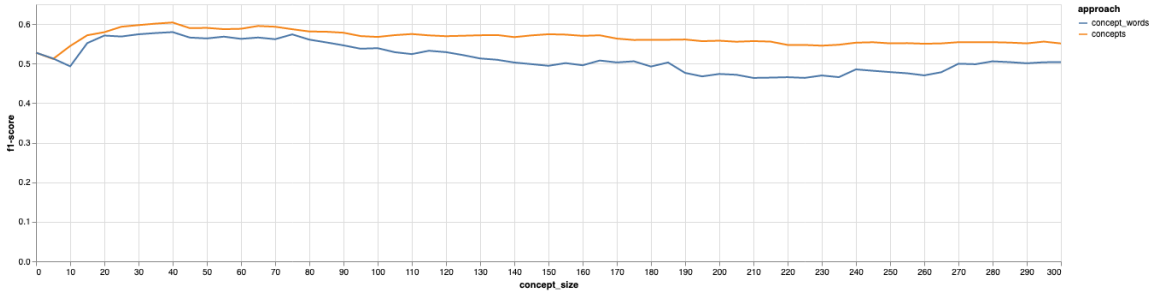

Figure C.5: Results (F1 score) of distance-based text classification with class concept and class label instances with word2vec model with size 0-300 for BBC corpus for headlines.