VYTAUTAS
MAGNUS
UNIVERSITY

Faculty of
Informatics

Extrinsic evaluation of word embedding models using semisupervised NLP tasks: the case of sentiment analysis

M. Petkevičius[1], D. Vitkutė-Adžgauskienė[2]

Vytautas Magnus University [1]mindaugas.petkevicius@vdu.lt [2]daiva.vitkute-adzgauskiene@vdu.lt

ABSTRACT

Two main approaches are used for evaluating the performance of a word embedding model - intrinsic and extrinsic evaluation.

Model can be evaluated by applying it to a Natural Language Processing (NLP) task, for example, a sentiment analysis task.

Training a classifier in a supervised learning approach requires large amounts of labeled data in order to achieve acceptable quality measures.

Meanwhile, a semi-supervised model can achieve comparable results with a small amount of labeled data and a large amount of unlabeled data.

In this research, pre-trained models are tested and compared by applying them to a semi-supervised social text sentiment analysis task for Lithuanian language social texts.

A small dictionary, initially derived from a relatively small training dataset of labeled 500 reviews, is further expanded using several word embedding models.

METHODS

A semi-supervised learning approach is used to build a text polarity classifier. This is a multi-stage approach:

1. A small sample of reviews is selected as a training set.
2. Reviews are pre-processed (lowercased, stop-word removal), and sentiment (polarity) labels are assigned.
3. A document-term matrix is created for each review.
4. A linear regression model is used for assigning keyword weights, i.e. sentiment scores for words for building a base sentiment dictionary.
5. Base sentiment dictionary is expanded, using an embedding model.
6. Text polarity classifier is built using the expanded sentiment dictionaries.

Lasso regression model is used for keyword weight assignment: independent variables (word vectors in a document-word matrix) are linearly integrated to predict dependent variables – sentiment scores.

$$\min_{\beta_0,\beta}\left\{\sum_{i=1}^{N}(y_i-\beta_0-x_i^T\beta)^2\right\}\text{ subject to }\sum_{j=1}^{p}|\beta_j|\leq t.$$

where N cases, p covariates and a single outcome, y_i be the outcome and $x_i := (x_1, x_2, \ldots, x_p)_i^T$ be the covariate vector for the i th case
$\beta := (\beta_1, \beta_2, \ldots, \beta_p)$ is the coefficient vector

Different sets of reviews are selected for training, repeating the process several times and averaging the word sentiment scores.

Table 1. Positive keywords

Word	Sentiment value
skanus	0.0500
skanu	0.0492
puikus	0.0399
geras	0.0299
malonus	0.0245
gerai	0.0229
super	0.0213
patiko	0.0209
puiku	0.0179

Table 2. Negative keywords

Word	Sentiment value
prastas	-0.0301
lėtas	-0.0164
blogas	-0.0151
laukti	-0.0143
buvo	-0.0141
nemalonus	-0.0141
letas	-0.0141
mėsa	-0.0139
stalus	-0.0130

EXPANDING DICTIONARY

Experiments with different embedding models were carried out to expand the dictionary. For each of the words in the base dictionary, we found the most similar word by using word vector cosine similarity.

$$S_C(A,B) := \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2}\sqrt{\sum\limits_{i=1}^{n} B_i^2}},$$

where *A* and *B* are word vectors

When adding words to the dictionary, sentiment scores are multiplied by their similarity.

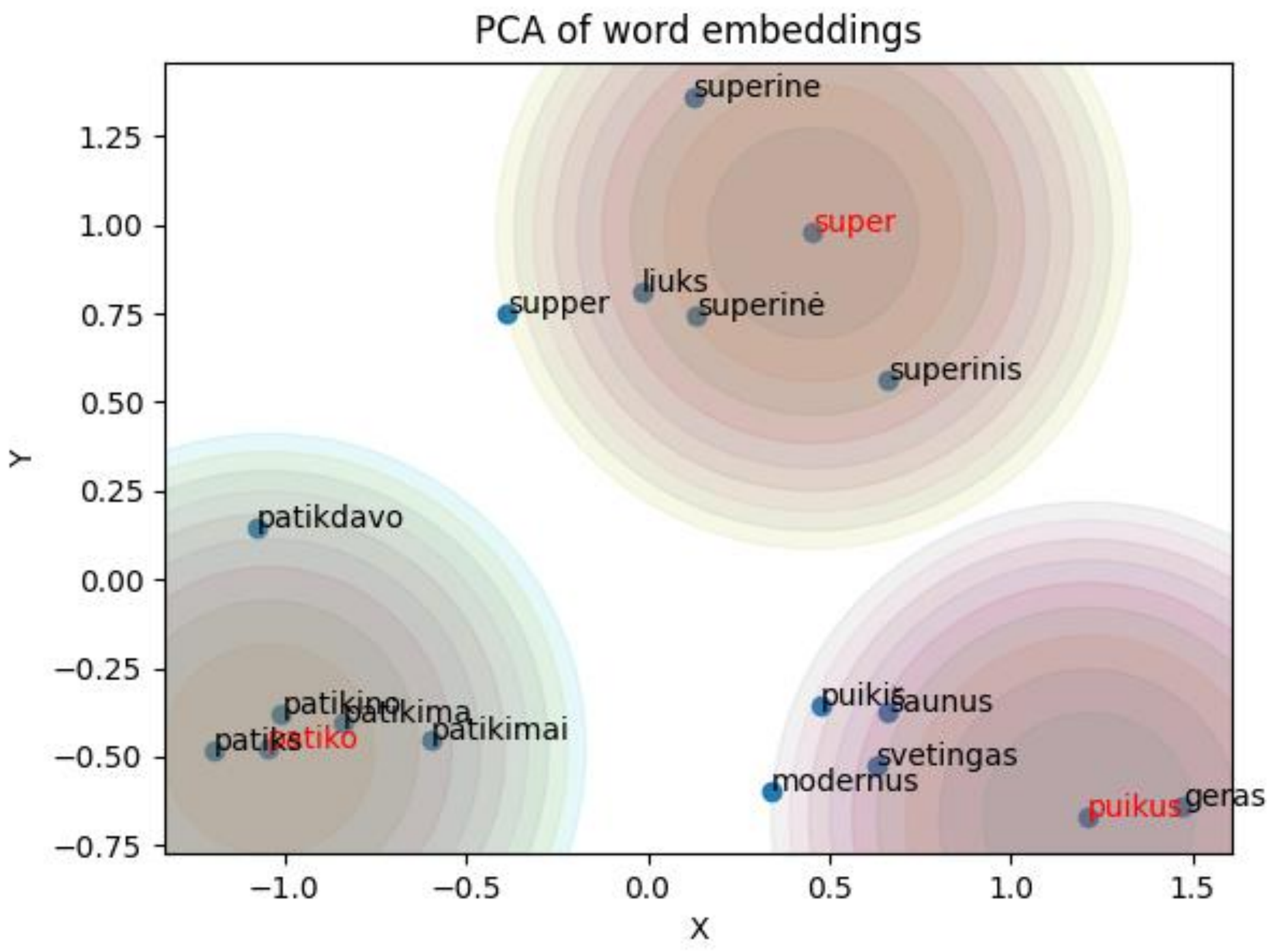

Figure 1. Red words are in the base dictionary; blue words are not.

The higher score assigned to the keywords, the closer they are to the base dictionary word. If more than one base dictionary word is nearby, the weighted scores of those words are summed up.

For example, if the word "*super*" has a score of **0.3** and the cosine similarity with "*superinis*" is **0.8**, the estimated sentiment value is **0.3** * **0.8** = **0.24**.

DATA RESOURCES

For this study, customer reviews from two different e-commerce websites were selected, namely, electronic store and restaurant reviews. The reviews were assigned values ranked from worst to best on a scale of 1 to 5.

The compiled review dataset contains 10 000 reviews, and each category contains 2000 reviews.

The reviews were divided into two categories: negative and positive.

Reviews contain typos and irregular words. Texts lack comprehensive scoring. Two people can write identical reviews, yet give them different ratings.

Table 3. Sample of the dataset

Review	Score
nesilankau cia gyvenu sveikai	1/5
Piktai aptarnauja bet picos skanios	2/5
kainos trigubai pakilo euro ivedimo šašlikai niekad nekainavo 45litu dabar 13 50	4/5
šiaip neblogai labai garsi muzika sunku susikalbėti	4/5
labai užimta savaitgaliais gali tekti laukti staliuko	5/5

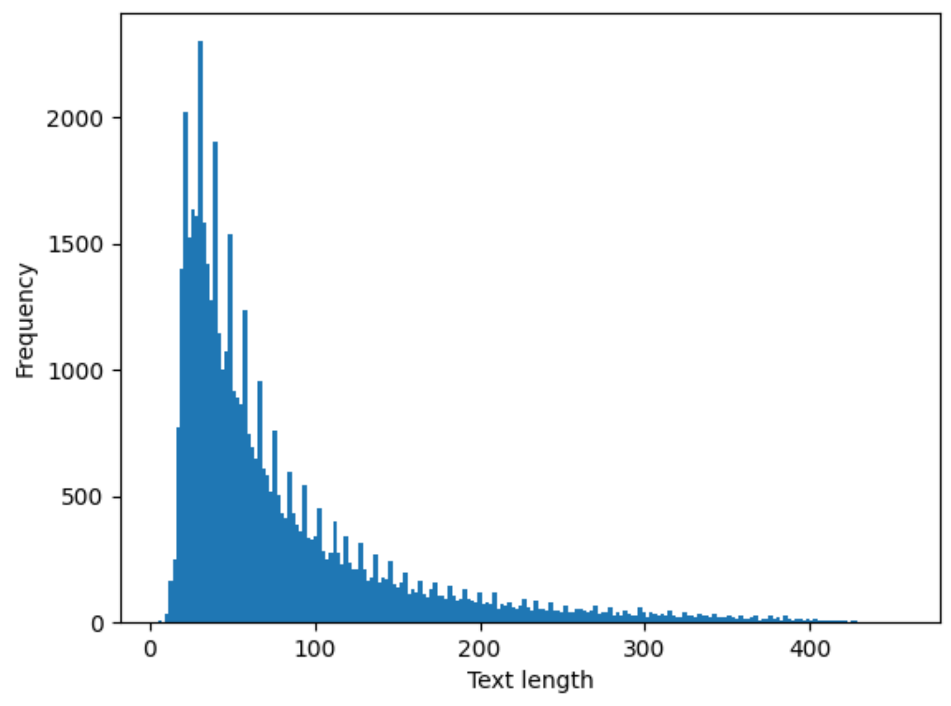

Figure 2. Distribution of dataset review lengths

EXPERIMENTS

Numerous experiments with varying degrees of complexity were conducted. For semi-supervised learning, different embedding models were used.

Embedding models:

- Word2vec
- FastText

Each model came in three variants:

- Pre-trained
- Pre-trained with additional social texts
- Trained on the dataset (domain trained)

Table 4. Comparison results

Model	F1 score
FastText (domain trained)	0.816
FastText (pre-trained with social)	0.807
Word2vec (domain trained)	0.805
FastText (pre-trained)	0.793
Word2vec (pre-trained with social)	0.783
Word2vec (pre-trained)	0.767

For comparison, supervised models were also trained on the whole dataset:

- Transformer based model based on Bert "bert-base-multilingual-cased" fine-tuned classifier.
- Convolutional Neural Network (CNN) with fastText embedding.

Table 5. Comparison to supervised model results

Model	F1 score
FastText (domain trained)	0.816
Transformer based model	0.864
CNN model	0.830
CNN model (1000 training dataset)	0.764

CONCLUSIONS

A semi-supervised learning approach with a fastText word embedding model, trained on a domain-specific dataset, was able to reach the F1 score of 81.6%, which is comparable to the F1 score of 83% for a CNN classifier, and the transformer-based model achieved the highest F1 score of 86% in our sentiment analysis experiment.

The results of the semi-supervised were comparable to those of supervised learning models, and the semi-supervised model required less time, fewer computational resources, and a smaller dataset to learn.

It is also shown that word embedding models for morphologically rich Lithuanian language, trained on a domain-specific dataset, outperform corresponding pre-trained word embedding models in semi-supervised learning tasks.

The accuracy of sentiment analysis using a semi-supervised learning approach with the expanded dictionary is shown to be significantly higher than the accuracy of a supervised learning approach with a small training dataset and close to the accuracy of a supervised learning approach with large training datasets.

FURTHER RESEARCH

The word context is particularly significant for sentiment analysis. Transformer-based models have demonstrated their ability to outperform dictionary-based models in terms of results. As part of our ongoing research, we will experiment with a variety of transformer models. Also, we will further expand our datasets to include more domains.

13th Conference **Data Analysis Methods for Software Systems**,
Druskininkai, Lithuania,