

Optimizing Word Embeddings for Patient Portal Message Datasets with a Small Number of Samples

Qingyuan Song

Vanderbilt University

Congning Ni

Vanderbilt University

Jeremy L. Warner

Brown University

Qingxia Chen

Vanderbilt University Medical Center

Lijun Song

Vanderbilt University

S. Trent Rosenbloom

Vanderbilt University

Bradley A. Malin

Vanderbilt University

Zhijun Yin

`zhijun.yin.1@vumc.org`

Vanderbilt University

Article

Keywords: breast cancer, hormonal therapy, natural language processing, patient portal messages, word embedding models, word2vec

Posted Date: May 15th, 2024

DOI: <https://doi.org/10.21203/rs.3.rs-4350387/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background Patient portal messages often relate to specific clinical phenomena (e.g., patients undergoing treatment for breast cancer) and, as a result, have received increasing attention in biomedical research. These messages require natural language processing and, while word embedding models, such as word2vec, have the potential to extract meaningful signals from text, they are not readily applicable to patient portal messages. This is because embedding models typically require millions of training samples to sufficiently represent semantics, while the volume of patient portal messages associated with a particular clinical phenomenon is often relatively small.

Objective We introduce a novel adaptation of the word2vec model, PK-word2vec, for small-scale messages.

Methods PK-word2vec incorporates the most similar terms for medical words (including problems, treatments, and tests) and non-medical words from two pre-trained embedding models as prior knowledge to improve the training process. We applied PK-word2vec on patient portal messages in the Vanderbilt University Medical Center electronic health record system sent by patients diagnosed with breast cancer from December 2004 to November 2017. We evaluated the model through a set of 1000 tasks, each of which compared the relevance of a given word to a group of the five most similar words generated by PK-word2vec and a group of the five most similar words generated by the standard word2vec model. We recruited 200 Amazon Mechanical Turk (AMT) workers and 7 medical students to perform the tasks.

Results The dataset was composed of 1,389 patient records and included 137,554 messages with 10,683 unique words. Prior knowledge was available for 7,981 non-medical and 1,116 medical words. In over 90% of the tasks, both reviewers indicated PK-word2vec generated more similar words than standard word2vec ($p=0.01$). The difference in the evaluation by AMT workers versus medical students was negligible for all comparisons of tasks' choices between the two groups of reviewers ($p=0.774$ under a paired t-test).

Conclusions. PK-word2vec can effectively learn word representations from a small message corpus, marking a significant advancement in processing patient portal messages.

1. Introduction

Patient portals have become standard components of modern electronic health record (EHR) systems.¹ They enable patients to access health information, manage clinical appointments, and communicate securely with care providers in an asynchronous manner.^{2,3} The messaging functionality of patient portals has gained in popularity over time^{4,5} due to various benefits, such as the ability to increase patient involvement in medical decision-making⁶. At the same time, the information conveyed in patient portal messages can serve as a foundation for clinical research, such as studies of initiation and

discontinuation of cancer therapies^{7,8}, assessing readmission risk^{9,10}, and creating taxonomies of patient engagement¹¹.

However, maximizing the utility of patient portal messages requires a means to extract meaningful signals from unstructured text.¹² Rule-based methods, which rely on manually pre-defined rules to perform a task, are ineffective as the volume of data or complexity of the task increases¹². While specific computational methods, such as topic modeling, can be applied in this regard, these approaches typically rely on word frequency to determine the most representative words for each topic. By contrast, word embedding models, such as word2vec¹³, have been widely adopted by natural language processing (NLP) applications. These models can learn vector representations, typically from millions of documents, to capture each word's semantic and linguistic meaning. However, it is challenging to procure a cohort that is sufficiently large to train a specific word embedding model. As an illustration, in this paper, we focus on breast cancer as it serves as an exemplar of a complex care coordination setting between various specialists who often communicate with patients via the patient portal¹⁴.

The current approaches for generating word embeddings on relatively small corpora may learn semantic meaning inaccurately due to lack of sufficient occurrences of word usage. A natural solution to this problem is to fine-tune a word embedding model that is pre-trained on a larger data corpus. This is because the fitted parameters of a pre-trained model can serve as a better starting point for training than random initialization. Yet this strategy presents certain limitations as well. In particular, pre-trained models are incapable of handling words that are absent from the pre-trained model's vocabulary. Moreover, inconsistencies arise due to the incompatible linguistic contexts between the data used to fit the pre-trained model and the data for a specific domain of interest.¹⁵

In this study, we introduce a novel approach, PK-word2vec, to train word2vec models on a patient portal message corpora of relatively small sample size. Our model leverages prior information, as defined by the most similar words sampled from pre-trained embedding models, to regularize the model training. We demonstrate the utility of this approach using a dataset of approximately 137,000 patient portal messages from 1,389 patients diagnosed with breast cancer at a large academic medical center.¹⁴¹⁶ Our experiments show that the regularization process of PK-word2vec can guide and constrain the model training to learn word semantic meanings more effectively than standard word2vec models.

2. Methods

Figure 1 depicts the workflow for this study: 1) data extraction and cleaning; 2) prior knowledge extraction from two pre-trained models, Snomed2vec¹⁶ and the Google pre-trained word2vec (Google-word2vec)¹³ for medical and non-medical words, respectively; 3) model training regularized by prior knowledge (PK-word2vec); and 4) similarity analysis and human evaluation of the model performance.

This study was deemed exempt from human subjects research by the Vanderbilt University Institutional Review Board.

2.1. Data Collection and Preprocessing

We collected the patient portal messages from the EHR system of Vanderbilt University Medical Center (VUMC), a large, private, non-profit, academic medical center in Nashville, Tennessee. VUMC launched its patient portal, *My Health at Vanderbilt (MHAV)*, in 2004, which was migrated to Epic's MyChart platform in 2017. MHAV supports secure messaging between healthcare providers and patients, appointment scheduling, billing management, and access to laboratory results or other EHR data.⁵

In this study, we focused on the patient portal messages for patients with breast cancer who were prescribed the most common hormonal therapy medications: *anastrozole*, *exemestane*, *letrozole*, *raloxifene*, and/or *tamoxifen*. We masked URLs, email addresses, phone numbers, timestamps, and numerals by using the *ekphrasis* (v0.5.1)¹⁷. Afterwards, all personal names and usernames were manually masked, and misspelled words were corrected with the guidance of two domain experts (STR, JLW) by referring to the contexts from randomly selected messages that contain these misspelled words. We applied the CLAMP software package (v1.6.4) to detect medical words pertaining to diseases, laboratory tests, and medications from messages¹⁸.

2.2. Skip-Gram Model with Negative Sampling

PK-word2vec is built on a skip-gram model with negative sampling (SGNS)¹³, a standard word2vec. In SGNS, each word has two vector representations: 1) a center vector when it is processed as a center word, and 2) a context vector when it serves as a neighbor of a center word. Figure 2 illustrates the center word and neighbor word within a context window size of 1 in the sentence "*Please reschedule my appointment to Friday*". In this example, *reschedule* is one of the neighbors of the center word *appointment*. Formally, SGNS structures the model training as a binary classification problem, where the objective is to predict whether a certain pair of words are neighbors. Formally, given a sequence of words $w_1, w_2, \dots, w_T$ and a context window of size m , the SGNS objective function is defined as:

$$Loss_{SGNS} = \sum_{t=1}^T \left[\sum_{-m \leq j \leq m} \left[-\log \sigma \left(\mathbf{u}_{t+j}^T \mathbf{v}_t \right) - \sum_{k=1, w_k \sim P(w_t)}^K \log \sigma \left(-\mathbf{u}_k^T \mathbf{v}_t \right) \right] \right]$$

1

where $\mathbf{u}_t$ and $\mathbf{v}_t$ represent the context and center vector of w_t , which are the parameters optimized in the loss function; w'_k represents a random word that does not appear in the neighboring context of w_t . w'_k is sampled from $P(w_t)$, a unigram distribution defined by the frequency of occurrence of each

word¹⁹; K is a hyperparameter that specifies the number of negative words sampled for w_t ; and σ is the sigmoid function. The first term in Eq. (1) aims to maximize the probability of occurrence for words that are in the context window. By contrast, the second term in Eq. (1) aims to minimize the probability of occurrence for words that are not in the context window.

2.3. Prior Knowledge Word2vec Model

The challenge of training a word2vec model on a small dataset arises from the insufficient context that such datasets provide. To address this issue, we define the prior knowledge of a word as its most similar words identified from other pre-trained models. Subsequently, the prior knowledge is applied to regularize the PK-word2vec model training. More formally, if a word in the patient portal messages w_t also appears in a pre-trained embedding vocabulary, the prior knowledge of w_t is defined by another word w_j that is sampled with probability p_{tj} from the patient portal message vocabulary. In this respect, the more similar w_j is to w_t in the pre-trained embedding model, the greater the likelihood that w_j will be sampled.

We introduce a prior knowledge loss function for the entire dataset, which is defined as follows:

$$\text{Loss}_{\text{PK}} = \begin{cases} \sum_{t=1}^T [\gamma(w_t) \bullet \Psi(\mathbf{u}_t, \mathbf{u}_j, \mathbf{v}_t, \mathbf{v}_j)] & w_t \text{ with prior knowledge} \\ 0 & w_t \text{ without prior knowledge} \end{cases}$$

2

where $\gamma(w_t) = \tau / \text{freq}(w_t)$ discounts the words proportionally to their frequency, τ is a hyperparameter, and $\Psi(\cdot)$ is a regularization factor to be defined in Eq. (4) below. Note that if a word does not exist in the pre-trained model, then it has no prior knowledge and its corresponding term in Loss_{PK} is set to zero.

In this work, we calculate word similarity as the cosine similarity of two vectors $\mathbf{u}$ and $\mathbf{v}$, $\cos(u, v) = \mathbf{u} \bullet \mathbf{v} / (|\mathbf{u}| \bullet |\mathbf{v}|)$. We require that each sampled w_j have a similarity score with w_t above a pre-defined threshold θ . As such, p_{tj} is defined as:

$$p_{tj} = \begin{cases} \text{softmax}[\cos(\mathbf{u}_t, \mathbf{u}_j)] & \cos(\mathbf{u}_t, \mathbf{u}_j) \geq \theta \\ 0 & \cos(\mathbf{u}_t, \mathbf{u}_j) < \theta \end{cases}$$

3

In our experiments, for each pre-trained model, we set θ to the average of the top 10 similarity scores for words with prior knowledge in patient portal messages. In practice, each time w_t is processed in model training using gradient descent optimization, a word w_j is sampled and its cosine similarity with w_t in both context and center vectors is calculated as its prior knowledge. Then, the corresponding item for w_t that needs to be minimized in the PK-word2vec loss function in Eq. (2) is:

$$\Psi(\mathbf{u}_t, \mathbf{u}_j, \mathbf{v}_t, \mathbf{v}_j) = [1 - \cos(\mathbf{u}_t, \mathbf{u}_j)] + [1 - \cos(\mathbf{v}_t, \mathbf{v}_j)]$$

There are several reasons for incorporating context vector similarity into Eq. (4). First, a key characteristic of high-quality word embeddings is that the distance between the context vectors for similar words is relatively small²⁰. Second, since the context vectors are applied to estimate the center vectors of neighboring words, the regularization of the context vectors can propagate to the center vectors of neighboring words without prior knowledge. Finally, the overall objective function for PK-word2vec can be defined as:

$$\text{Loss} = \text{Loss}_{\text{SGNS}} + \alpha \times \text{Loss}_{\text{PK}}$$

2

where α is a hyperparameter that determines the degree to which the training of PK-word2vec relies on prior knowledge. A higher α indicates a stronger dependency on prior knowledge. To evaluate the impact of α , we compared our approach to a baseline model where $\alpha = 0$, a condition corresponding to SGNS. We generate prior knowledge for non-medical and medical words using two different models. Specifically, we applied Google-word2vec, which was trained on the Google News dataset (about 100 billion words), to build prior knowledge for non-medical words. We turned to a *Snomed2vec* model, which was trained on the SNOMED-CT knowledge graph¹⁶ to obtain prior knowledge for medical words.

2.4. Hyperparameter Selection

Hyperparameter tuning for word2vec models is challenging because it is difficult to define similarity. Since Google-word2vec was trained on a vast corpus, we selected hyperparameters, the dimension of word embedding d and the prior knowledge weight α , such that PK-word2vec is similar to Google-word2vec in terms of the distribution of the similarity scores between a given word and its 10 most similar words. In this paper we name this distribution as similarity distribution for simplicity. We did not use Snomed2vec as a reference because it is inherently a node2vec²¹ trained on a semantic network in the Unified Medical Language System²² rather than text data.

We set the embedding size d with a value that corresponds to the smallest Wasserstein distance²³ in the similarity distributions between the two models. To select α , we first categorized words into high-, mid-, and low-frequency ranges based on their tertiles. Then, we set α with a value that corresponds to the smallest Wasserstein distance in the similarity distribution between three distinct word frequency ranges. This resulted in a comparison in an unbiased manner, since high-frequency words tended to receive better representations. When training PK-word2vec, the θ threshold was set to the average word similarity between all the words and their 10 most similar words in the pre-trained models. We set the threshold τ in Eq. (2) to 0.00001, batch size to 10000, and trained the model for 10 epochs.

2.5. Human Evaluation

To compare PK-word2vec with SGNS, we randomly sampled 1,000 words from the patient portal message vocabulary, with the same proportion of medical to non-medical words. For each term of interest, we created one group of the five most similar words generated from PK-word2vec and SGNS,

respectively. To evaluate the system, we asked each reviewer to indicate which word group was more related to the word of interest. We randomized the order of the two groups in each task to avoid framing biases.

We recruited two groups of reviewers for the model evaluation. The first group consisted of 330 Amazon Mechanical Turk (MTurk) workers²⁴, recognized as “masters” due to their consistently submission of high-quality results in their history annotations. Each MTurk worker completed 100 out of the 1000 tasks and each task was answered by 33 different MTurk workers. The second group consisted of 7 medical students recruited through the Vanderbilt Institute for Clinical and Translational Research. Each medical student completed all 1,000 tasks. We set a 30-second time limit for each task; task not completed within this time limit were reassigned to a new worker.

To analyze the human evaluation data, we first calculated the support rate, defined as the proportion of the 1,000 tasks for which MTurk workers or medical students preferred PK-word2vec by a majority vote. Next, we employed sample skewness²⁵ to examine the distribution of the proportion of reviewers who favored PK-word2vec in each task. This provided insights into the extent to which PK-word2vec was favored over SGNS. A larger value for left sample skew indicates a preference of PK-word2vec over SGNS. Finally, to account for the heterogeneity, we fitted a mixed-effects logistic regression model to assess which word embedding model was preferred by MTurk workers. The type of model (PK-word2vec or SGNS) and the category of the term of interest (medical or non-medical words) were variables with fixed effects. We assigned a value of 1 to the PK-word2vec model and medical words and a value of 0 to the SGNS and non-medical words in the corresponding variables.

3. Results

3.1. Data Summary

The dataset was composed of 137,554 messages sent by 1,389 patients with breast cancer. After preprocessing, there was a total of 10,683 unique words in the vocabulary, of which 8,878 (80%) and 1,895 (20%) were non-medical and medical words, respectively. Prior knowledge was available for 7,981 (90%) non-medical and 1,116 (59%) medical words.

3.2. Hyperparameter Selection

Embedding size (d). The word embedding size was selected from a range of 10 to 300. Figure 3 depicts the Wasserstein distance between SGNS and the Google-word2vec as a function of the word embedding size. In the following experiments, we present results with an embedding size of 45 for SGNS and PK-word2vec because it corresponds to the smallest distance among all the candidate vector sizes.

Prior knowledge weight (α). Figure 4a shows the similarity distribution for PK-word2vec as a function of α . When $\alpha = 0$, this model corresponds to SGNS and most of the density concentrates around 0.6. As α increases, the prior knowledge becomes dominant, and the density gradually shifts toward two

directions, forming two peaks. Figure 4b shows the Wasserstein distance in the similarity distribution between Google-Word2vec (three dashed lines at the bottom of the figure) and PK-word2vec (three solid lines at the top of the figure) in different word frequency ranges. The distributions for Google-word2vec were nearly flat and maintained relatively small values. Based on this observation, we set α to 5×10^{-6} for the following experiments because it corresponds to the smallest Wasserstein distances as well as smallest variance across three pairs of comparisons.

3.3. Word Similarity

Context vector regularization. Figure 5 depicts the similarity distribution for the models with (panels a-d) and without (panels e-g) context vector regularization. The blue and orange lines in each graph correspond to words with and without prior knowledge, respectively. The comparisons between two similarity distributions are illustrated in settings with different pre-selected α values. As extreme examples, Figs. 5d and 5g show that when $\alpha = 10^{-4}$, context vector regularization mitigates the skew of the similarity distribution for the words with prior knowledge. While a large α value shifts the mode of the similarity distribution for words without prior knowledge in the model without context vector regularization to the right, the variance of similarity distribution is not significantly altered. Figures 5c and 5f depict the scenario where α is set to its optimal value of 5×10^{-6} , where the difference between the two distributions is in an intermediate state compared with other α values. These results show that context vector regularization mitigates the impact of prior knowledge on the similarity distribution between words with and without prior knowledge.

3.4. Human Evaluation

Figure 6 displays the human evaluation results based on a majority vote from the MTurk workers and medical students. In addition to an overall comparison, the results were stratified by the word type, and the existence of prior knowledge. The difference was negligible for all comparisons between the two groups of reviewers ($p = 0.774$ under a paired t-test). In over 90% of the tasks, both groups of reviewers indicated that PK-word2vec generates more similar words than SGNS. For medical words, the support rate for PK-word2vec was 78.8% for MTurk workers and 71.8% for medical students. For non-medical words, the support rate for PK-word2vec was 93.6% for MTurk workers and 93.4% for medical students. For words lacking prior knowledge, the support rate for PK-word2vec was 66.7% from MTurk workers and 61.4% from medical students.

Figure 7 shows the distribution of the proportion of reviewers that favor PK-word2vec in each task. The three curves represent the proportion distribution for the low-, mid-, and high-frequency words, respectively. All distributions are strongly left-skewed and most of the density falls between 0.8 and 1.0, indicating reviewers consistently favored PK-word2vec. Notably, in the responses from MTurk workers, the sample skew was 1.19, 1.65, and 1.92 for the high-, mid-, and low-frequency words, respectively. In the responses from the medical students, the sample skew was 0.88, 1.46, and 1.84 for the high, mid-, and low-frequency words, respectively. In other words, the degree to which the distribution was left-skewed was inversely proportional to the word frequency.

The results of the mixed-effect model showed that PK-word2vec was significantly more favored than SGNS ($\beta = 2.702$, p-value = 0.01). By contrast, no difference was found between medical word and non-medical word ($\beta = 0$, p-value = 1).

4. Discussion

This study has several findings that are worth highlighting. First, relying on external word pairs to tune hyperparameters, as has been done in many prior studies^{13,26,27}, is suboptimal for smaller datasets. This limitation arises from the fact that many word pairs in external datasets may not be included in the target datasets. To address this issue, we selected hyperparameters by minimizing the distance in the similarity distribution between Google-word2vec and PK-word2vec. Our experiments showed that the prior knowledge weight α changed the word similarity distribution from a unimodal to a bimodal form, suggesting that introducing too much prior knowledge might degrade the quality of word embedding models. However, when comparing the similarity distribution between words across different frequency ranges, we observed that when the prior knowledge weight was very small or very large, it caused the similarity distribution to become dissimilar amongst these word frequency ranges. This finding highlights the necessity of carefully balancing the weight of prior knowledge within the model training.

Second, including context vector regularization into the PK-word2vec training helped reduce the difference in the similarity distribution between words with and without prior knowledge. Specifically, in the model with context vector regularization, the distribution of words with prior knowledge exhibited a longer tail as the value of α increased. By contrast, without context vector regularization, the similarity distribution for words without prior knowledge was right shifted without significant changes to its shape. To investigate the root cause, we inspected the most similar words for those without prior knowledge and located in the tail of the similarity distribution around cosine similarity of 0.9. For example, consider the word *miscarriage*. In the model with context vector regularization, its most similar words were *trimester*, *gestational*, *gestation*, *fetal*, and *fetus*, all of which are directly related to pregnancy. Conversely, without context vector regularization, the most similar words were *tube*, *defects*, *defect*, *homocysteine*, and *suction*. While these terms are associated with miscarriage, they revealed a different pattern of similarity compared to the model with context vector regularization. This effect suggests the importance of context vector regularization in enhancing the model's compatibility with a small sample size of data.

Third, despite using a relatively small dataset, our human reviewers preferred the word groups produced by PK-word2vec over those produced by SGNS, and had a higher support rate for PK-word2vec for non-medical words than medical words. There are several possible explanations for this result. First, the medical words are inherently complex and more difficult for the reviewers to evaluate. Second, the number of medical words was relatively small, such that some medical words are associated with fewer than five other words with high similarity scores. In addition, the new PK-word2vec helped improve the vector representation, especially for low-frequency words (see Fig. 6). While medical students likely had more medical training than MTurk workers on average, our analysis suggested that the two types of

reviewers were similar regarding their preferences for non-medical and medical words. This implies that for simple medical surveys, engaging a large number of participants pool can obtain accurate results, mitigating the need for, and potential higher costs associated with, specialized professionals.

In the context of customizing word representation models, our PK-word2vec method offers a distinct approach compared to previous studies. Several earlier studies have adapted SGNS to enhance word representation, primary focusing on aligning the vector space of a general, pre-trained embedding model to a specific problem domain^{28,29}. Other refined strategies included improving the embedding quality by retrofitting the pre-trained word vectors using lexical relational resources³⁰, and augmenting SGNS with additional knowledge-based graph models and `anchor context models³¹. Different from these approaches, our PK-word2vec method does not require a large number of training samples to fine-tune the word embedding model. This significant advantage could substantially facilitate research within the field of biomedical informatics, mainly when working with small study cohorts.

Despite the encouraging findings, this study has several limitations that can serve as a basis for future research. First, PK-word2vec was examined solely on patients with breast cancer. It has yet to be determined whether the method generalizes to clinical communications or notes for other types of patients. Second, we generated prior knowledge from two pre-trained models only. It is unknown if incorporating additional pre-trained models can enhance the generalizability of PK-word2vec. Third, complex medical concepts are often represented by long phrases rather than words, which begs the question of how to include more complex concepts in human evaluations. Finally, the model should be tested in downstream NLP tasks by jointly tuning the prior knowledge weight and other hyperparameters.

5. Conclusions

This paper introduces PK-word2vec, an adapted word embedding model trained on a relatively small number of patient portal messages. This was achieved by incorporating prior knowledge from two pre-trained embedding models into the model training. Our evaluation with MTurk workers and medical students demonstrated that PK-word2vec outperformed the standard word2vec model in generating more similar words from small-size patient portal messages. While this study focused on extracting prior knowledge from Google-word2ec and Snomed2vec for patient portal messages, the proposed PK-word2vec can leverage prior knowledge from any pre-trained models, irrespective of the vector size or representation space, to build vector representations for other clinical text data.

Declarations

Author Contribution

Q.S. and Z.Y. wrote the main manuscript text and Q.S. prepared figures. All authors reviewed the manuscript.

Acknowledgments

This work was supported by the National Cancer Institute under award number R37CA237452.

Data Availability

The data that support the findings of this study are available on the Vanderbilt University Medical Center Research Derivative, but restrictions apply to the availability of these data due to privacy issue, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Vanderbilt University Office of Research Informatics.

References

1. Dendere R, Slade C, Burton-Jones A, Sullivan C, Staib A, Janda M. Patient Portals Facilitating Engagement With Inpatient Electronic Medical Records: A Systematic Review. *J Med Internet Res*. 2019;21(4):e12779. doi:10.2196/12779
2. Goel MS, Brown TL, Williams A, Cooper AJ, Hasnain-Wynia R, Baker DW. Patient reported barriers to enrolling in a patient portal. *Journal of the American Medical Informatics Association*. 2011;18(Supplement_1):i8–i12.
3. Kruse CS, Bolton K, Freriks G. The effect of patient portals on quality outcomes and its implications to meaningful use: a systematic review. *J Med Internet Res*. 2015;17(2):e44.
4. Ralston JD, Carrell D, Reid R, Anderson M, Moran M, Hereford J. Patient web services integrated with a shared medical record: patient use and satisfaction. *Journal of the American Medical Informatics Association*. 2007;14(6):798–806.
5. Osborn CY, Rosenbloom ST, Stenner SP, et al. MyHealthAtVanderbilt: policies and procedures governing patient portal functionality. *Journal of the American Medical Informatics Association*. 2011;18(Supplement_1):i18–i23.
6. Griffin A, Skinner A, Thornhill J, Weinberger M. Patient portals. *Appl Clin Inform*. 2016;7(02):489–501.
7. Yin Z, Warner JL, Chen Q, Malin BA. Patient Messaging Content Associated with Initiating Hormonal Therapy after a Breast Cancer Diagnosis. In: *AMIA Annual Symposium Proceedings*. Vol 2019.; 2019:962.
8. Yin Z, Harrell M, Warner JL, Chen Q, Fabbri D, Malin BA. The therapy is making me sick: how online portal communications between breast cancer patients and physicians indicate medication discontinuation. *Journal of the American Medical Informatics Association*. 2018;25(11):1444–1451. doi:10.1093/jamia/ocy118
9. Sulieman L, Yin Z, Malin BA. Why patient portal messages indicate risk of readmission for patients with ischemic heart disease. In: *AMIA Annual Symposium Proceedings*. Vol 2019.; 2019:828.

10. Sulieman L, Gilmore D, French C, et al. Classifying patient portal messages using Convolutional Neural Networks. *J Biomed Inform.* 2017;74:59–70. doi:<https://doi.org/10.1016/j.jbi.2017.08.014>
11. Glöggler M, Ammenwerth E. Development and Validation of a Useful Taxonomy of Patient Portals Based on Characteristics of Patient Engagement. *Methods Inf Med.* 2021;60(S 01):e44–e55.
12. Wang Z, Zhang W, Liu N, Wang J. Scalable rule-based representation learning for interpretable classification. *Adv Neural Inf Process Syst.* 2021;34:30479–30491.
13. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed Representations of Words and Phrases and their Compositionality. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, eds. *Advances in Neural Information Processing Systems*. Vol 26. Curran Associates, Inc.; 2013. <https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf>
14. Steitz BD, Levy MA. A Social Network Analysis of Cancer Provider Collaboration. *AMIA Annu Symp Proc.* 2016;2016:1987–1996.
15. Galea D, Laponogov I, Veselkov K. Sub-word information in pre-trained biomedical word representations: evaluation and hyper-parameter optimization. In: *Proceedings of the BioNLP 2018 Workshop.*; 2018:56–66.
16. Martinez Soriano I, Castro Peña JL, Fernandez Breis JT, San Román I, Alonso Barriuso A, Guevara Baraza D. Snomed2Vec: Representation of SNOMED CT Terms with Word2Vec. In: *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS).*; 2019:678–683. doi:10.1109/CBMS.2019.00138
17. Baziotis C, Pelekis N, Doukeridis C. DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis. In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017).* Association for Computational Linguistics; 2017:747–754. doi:10.18653/v1/S17-2126
18. Soysal E, Wang J, Jiang M, et al. CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association.* 2017;25(3):331–336. doi:10.1093/jamia/ocx132
19. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: *Advances in Neural Information Processing Systems.*; 2013:3111–3119.
20. Mitra B, Nalisnick E, Craswell N, Caruana R. A dual embedding space model for document ranking. *arXiv preprint arXiv:160201137.* Published online 2016.
21. Grover A, Leskovec J. node2vec: Scalable feature learning for networks. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*; 2016:855–864.
22. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 2004;32(suppl_1):D267–D270.
23. Villani C, others. *Optimal Transport: Old and New.* Vol 338. Springer; 2009.
24. Buhrmester M, Kwang T, Gosling SD. Amazon’s Mechanical Turk: A new source of inexpensive, yet high-quality data? Published online 2016.

25. Dean S, Illowsky B. Descriptive statistics: skewness and the mean, median, and mode. Connexions website. Published online 2018.
26. Jansen S. Word and phrase translation with word2vec. arXiv preprint arXiv:170503127. Published online 2017.
27. Chiu B, Crichton G, Korhonen A, Pyysalo S. How to train good word embeddings for biomedical NLP. In: Proceedings of the 15th Workshop on Biomedical Natural Language Processing.; 2016:166–174.
28. Sarma PK, Liang Y, Sethares WA. Domain adapted word embeddings for improved sentiment classification. arXiv preprint arXiv:180504576. Published online 2018.
29. Poerner N, Waltinger U, Schütze H. Inexpensive Domain Adaptation of Pretrained Language Models: Case Studies on Biomedical NER and Covid-19 QA. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics; 2020:1482–1490. doi:10.18653/v1/2020.findings-emnlp.134
30. Faruqui M, Dodge J, Jauhar SK, Dyer C, Hovy E, Smith NA. Retrofitting word vectors to semantic lexicons. arXiv preprint arXiv:14114166. Published online 2014.
31. Yamada I, Shindo H, Takeda H, Takefuji Y. Joint learning of the embedding of words and entities for named entity disambiguation. arXiv preprint arXiv:160101343. Published online 2016.

Figures

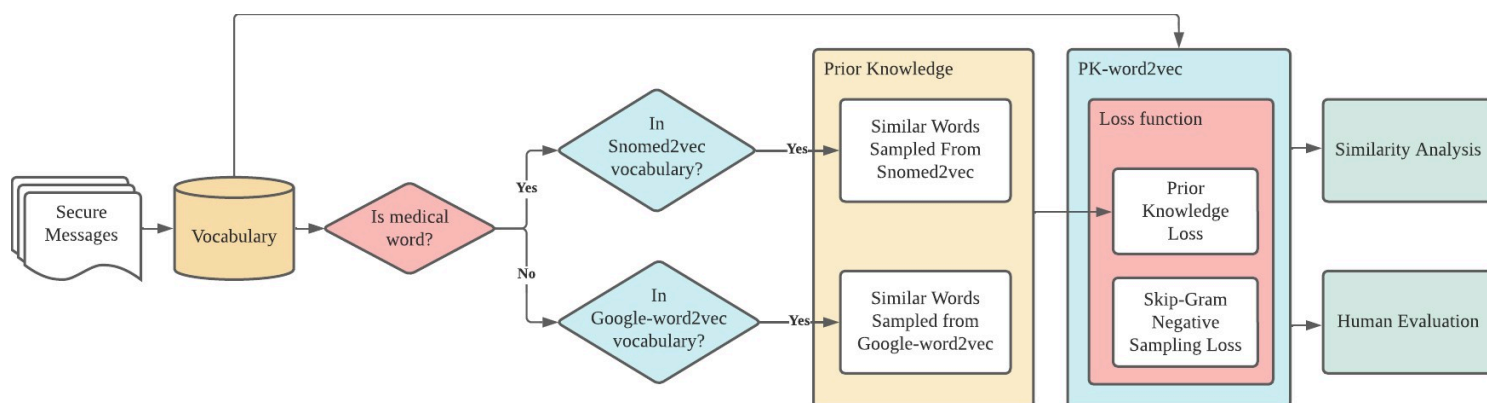


Figure 1

Workflow associated with the creation of the PK-word2vec model.

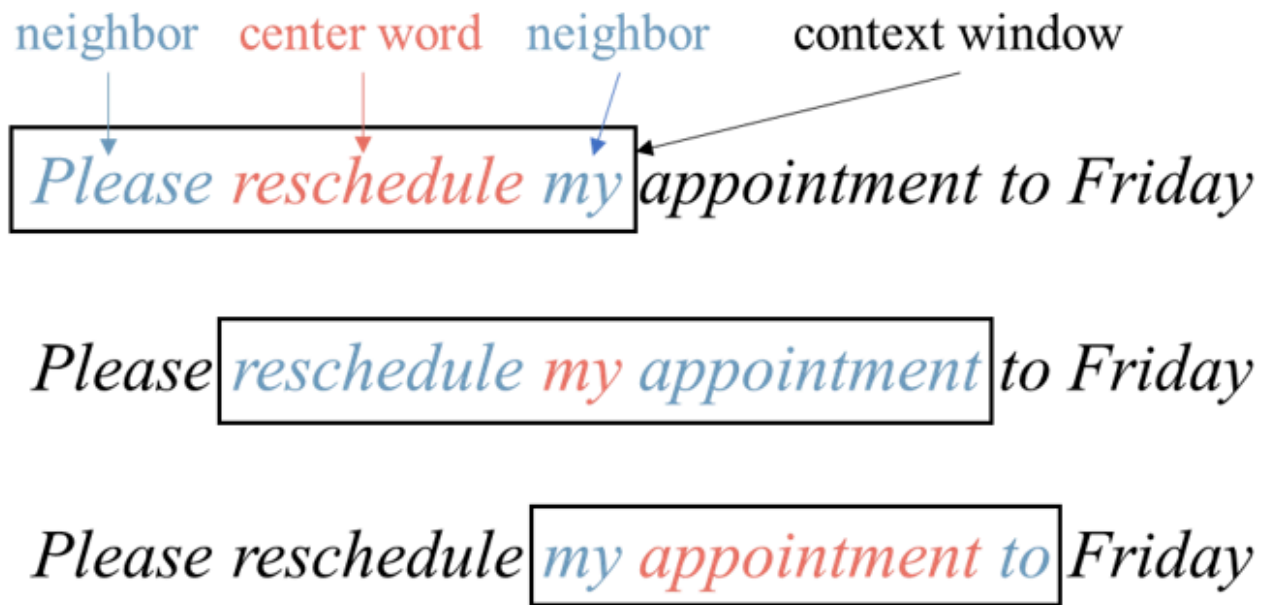


Figure 2

An illustration of a context window with a size of 1, a center word and its neighbors in the SGNS framework.

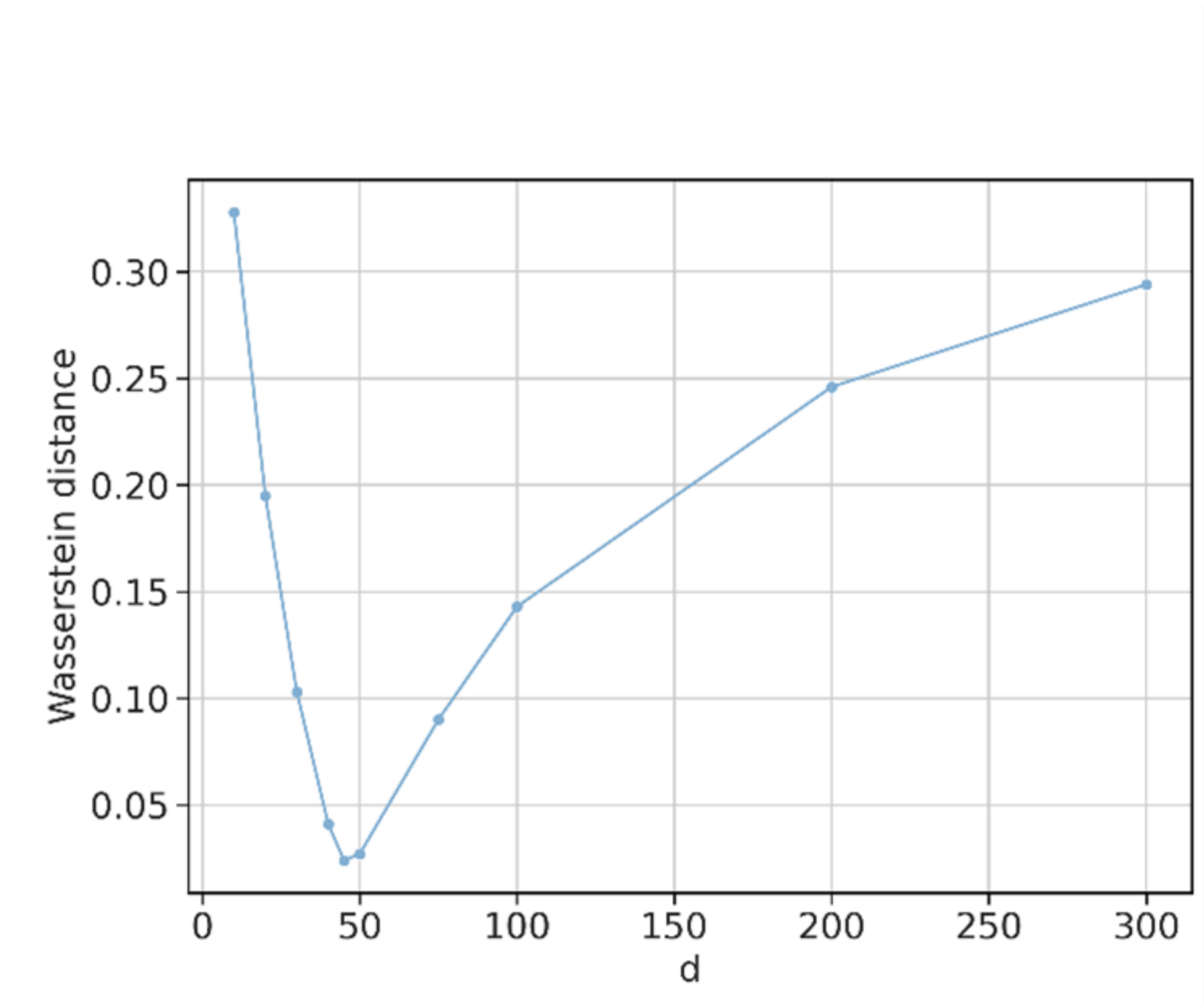


Figure 3

The Wasserstein distance in the similarity distribution between Google-word2vec and SGNS as a function of embedding size. A sensitivity analysis to optimize the word embedding size.

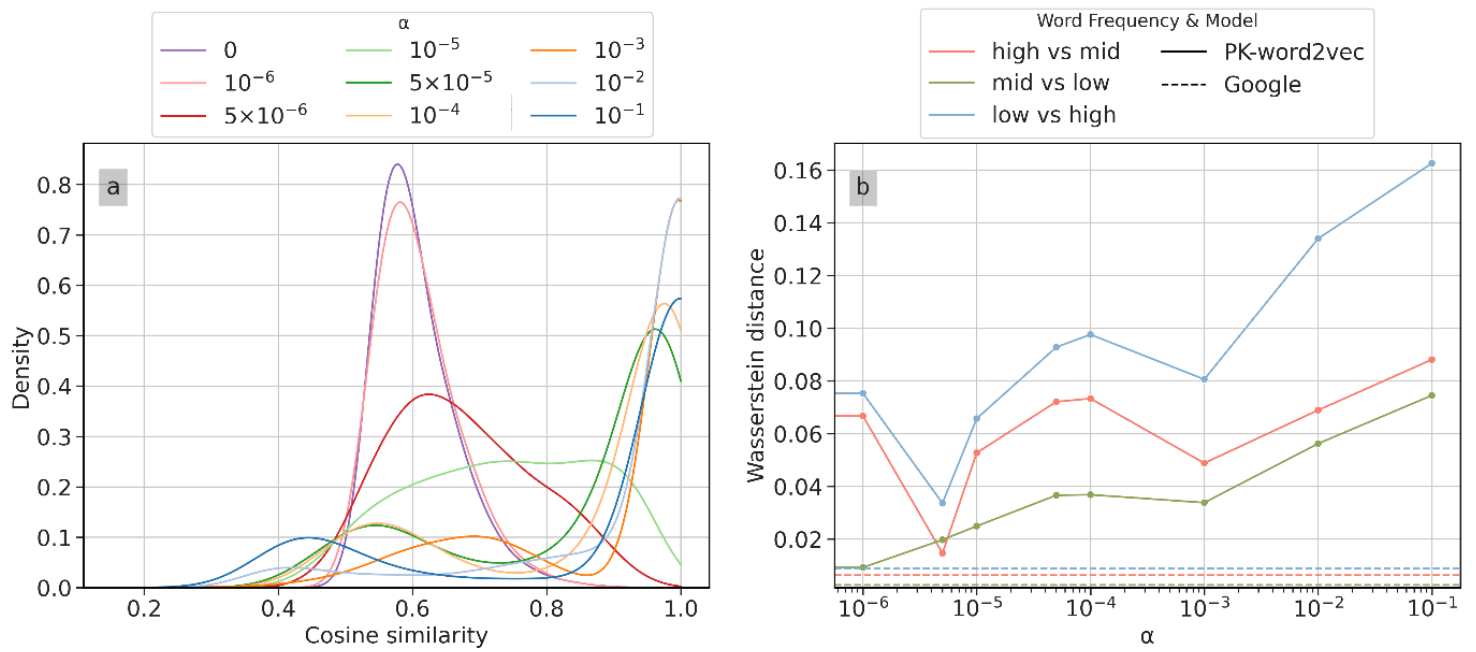


Figure 4

Sensitivity of similarity distribution (left) and Wasserstein distance between Google-word2vec and PK-word2vec on various word frequencies (right) across the range of α values.

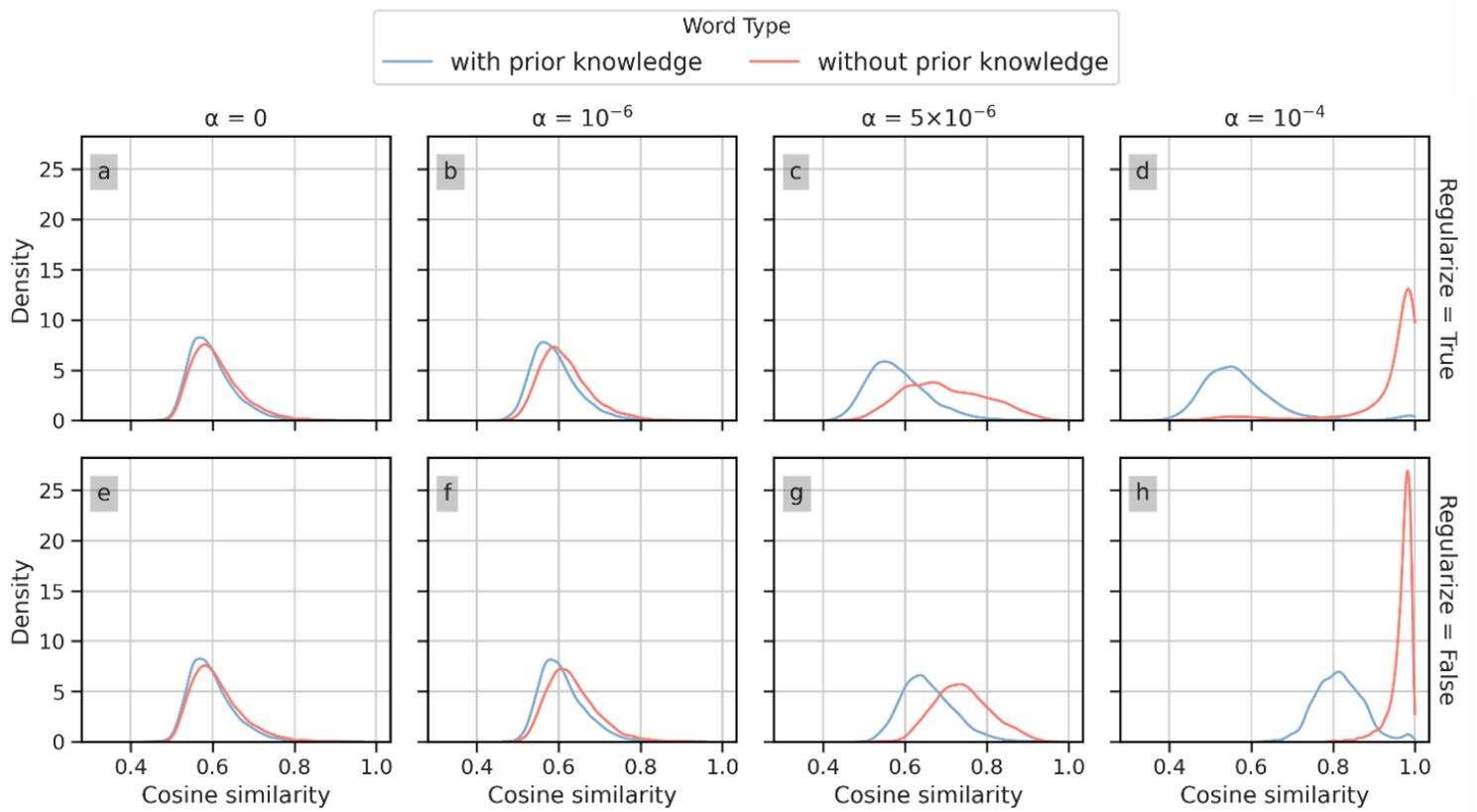


Figure 5

The distribution of the words with the 10 highest similarity scores of words with and without context vector regularization during model training.

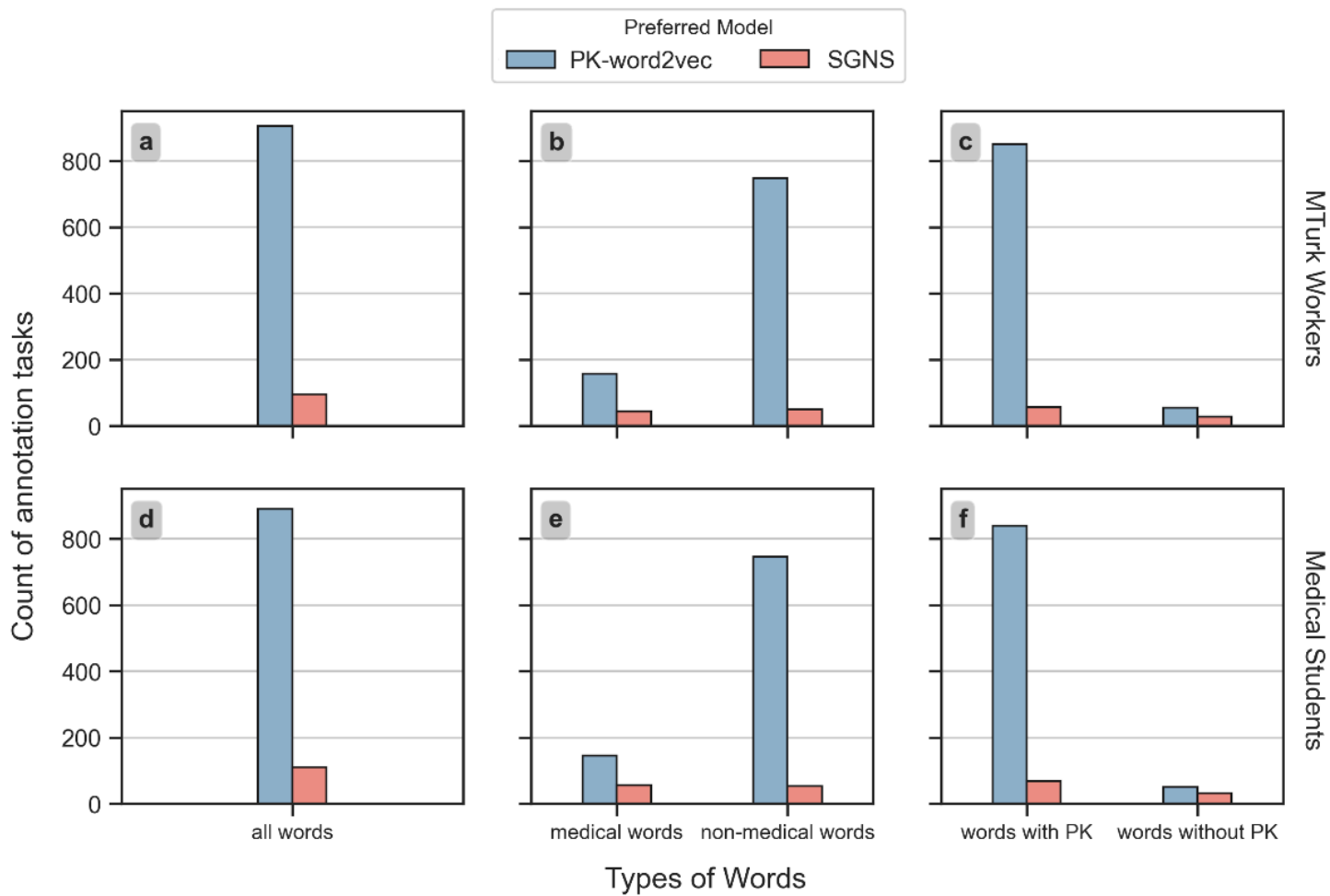


Figure 6

The number of tasks where one model was preferred over the other based on a majority vote by MTurk workers (upper row) and medical students (lower row).

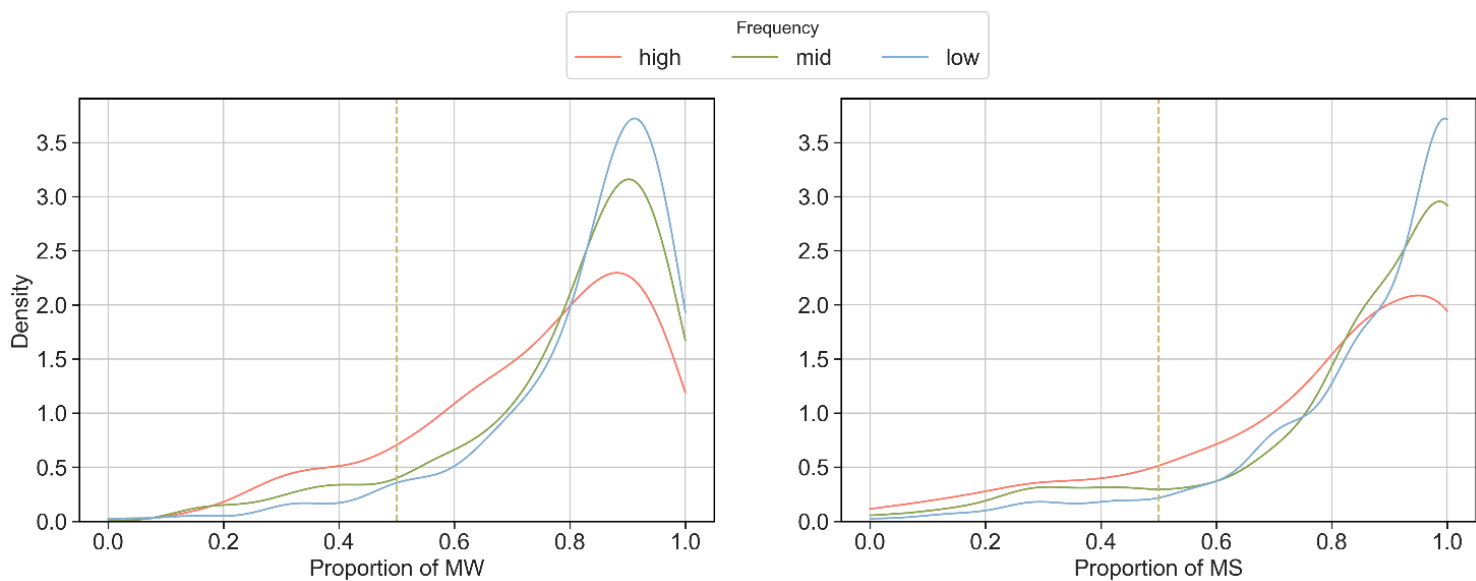


Figure 7

Distribution of the proportion of reviewers that choose PK-word2vec over SGNS per annotation task ($n = 1000$). Different colors indicate the word frequency ranges of the original word in each task. The vertical dashed line corresponds to 0.5, which indicate that there was no difference in the annotator preference.