Searching for the X-Factor: Exploring Corpus Subjectivity for Word Embeddings

Maksim Tkachenko and **Chong Cher Chia** and **Hady W. Lauw**
School of Information Systems
Singapore Management University
maksim.tkatchenko@gmail.com
{ccchia.2014,hadywlauw}@smu.edu.sg

Abstract

We explore the notion of subjectivity, and hypothesize that word embeddings learnt from input corpora of varying levels of subjectivity behave differently on natural language processing tasks such as classifying a sentence by sentiment, subjectivity, or topic. Through systematic comparative analyses, we establish this to be the case indeed. Moreover, based on the discovery of the outsized role that sentiment words play on subjectivity-sensitive tasks such as sentiment classification, we develop a novel word embedding *SentiVec* which is infused with sentiment information from a lexical resource, and is shown to outperform baselines on such tasks.

1 Introduction

Distributional analysis methods such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) have been critical for the success of many large-scale natural language processing (NLP) applications (Collobert et al., 2011; Socher et al., 2013; Goldberg, 2016). These methods employ distributional hypothesis (i.e., words used in the same contexts tend to have similar meaning) to derive distributional meaning via context prediction tasks and produce dense word embeddings.

While there have been active and ongoing research on improving word embedding methods (see Section 5), there is a relative dearth of study on the impact that an input corpus may have on the quality of the word embeddings. The previous preoccupation centers around corpus size, i.e., a larger corpus is perceived to be richer in statistical information. For instance, popular corpora include Wikipedia, Common Crawl, and Google News.

We postulate that there may be variations across corpora owing to factors that affect language use. Intuitively, the many things we write (a work email, a product review, an academic publication, etc.) may each involve certain stylistic, syntactic, and lexical choices, resulting in meaningfully different distributions of word cooccurrences. Consequently, such factors may be encoded in the word embeddings, and input corpora may be differentially informative towards various NLP tasks.

In this work, we are interested in the notion of *subjectivity*. Some NLP tasks, such as sentiment classification, revolve around subjective expressions of likes or dislikes. Others, such as topic classification, revolve around more objective elements of whether a document belongs to a topic (e.g., science, politics). Our central hypothesis is that word embeddings learnt from input corpora of contrasting levels of subjectivity perform differently when classifying sentences by sentiment, subjectivity, or topic. As the *first contribution*, we outline an experimental scheme to explore this hypothesis in Section 2, and conduct a series of controlled experiments in Section 3 establishing that there exists a meaningful difference between word embeddings derived from objective vs. subjective corpora. We further systematically investigate factors that could potentially explain the differences.

Upon discovering from the investigation that sentiment words play a particularly important role in subjectivity-sensitive NLP tasks, such as sentiment classification, as the *second contribution*, in Section 4 we develop *SentiVec*, a novel word embedding method infused with information from lexical resources such as a sentiment lexicon. We further identify two alternative lexical objectives: *Logistic SentiVec* based on discriminative logistic regression, and *Spherical SentiVec* based on soft clustering effect of von Mises-Fisher distributions. In Section 6, the proposed word embeddings show

evident improvements on sentiment classification, as compared to the base model Word2Vec and other baselines using the same lexical resource.

2 Data and Methodology

We lay out the methodology for generating word embeddings of contrasting subjectivity, whose effects are tested on several text classification tasks.

2.1 Generating Word Embeddings

As it is difficult to precisely quantify the degree of subjectivity of a corpus, we resort to generating word embeddings from two corpora that contrast sharply in subjectivity, referring to them as the *Objective Corpus* and the *Subjective Corpus*.

Objective Corpus As virtually all contents are written by humans, an absolutely objective corpus (in the philosophical sense) may prove elusive. There are however exemplars where, by construction, a corpus aspires to be as objective as possible, and probably achieves that in practical terms. We postulate that one such corpus is Wikipedia. Its list of policies and guidelines[1], assiduously enforced by an editorial team, specify that an article must be written from a *neutral point of view*, which among other things means *"representing fairly, proportionately, and, as far as possible, without editorial bias, all of the significant views that have been published by reliable sources on a topic."*. Moreover, it is a common resource for training distributional word embeddings and adopted widely by the research community to solve various NLP problems. Hence, in this study, we use Wikipedia as the *Objective Corpus*.

Subjective Corpus By extension, one may then deem a corpus subjective if its content does not at least meet Wikipedia's *neutral point of view* requirement. In other words, if the content is replete with personal feelings and opinions. We posit that product reviews would be one such corpus. For instance, Amazon's Community Guideline[2] states that *"Amazon values diverse opinions"*, and that *"Content you submit should be relevant and based on your own honest opinions and experience."*. Reviews consist of expressive content written by customers, and may not strive for the neutrality of an encyclopedia. We rely on a large corpus of Amazon reviews from various categories (e.g., electronics, jewelry, books, and etc.) (McAuley et al., 2015) as the *Subjective Corpus*.

Word Embeddings For the comparative analysis in Section 3, we employ *Word2Vec* (reviewed below) to generate word embeddings from each corpus. Later on in Section 4, we will propose a new word embedding method called *SentiVec*.

For Word2Vec, we use the Skip-gram model to train distributional word embeddings on the *Objective Corpus* and the *Subjective Corpus* respectively. Skip-gram aims to find word embeddings that are useful for predicting nearby words. The objective is to maximize the context probability:

$$\log \mathcal{L}(W;C) = \sum_{w \in W} \sum_{w' \in C(w)} \log \mathrm{P}(w'|w), \qquad (1)$$

where W is an input corpus and $C(w)$ is the context of token w. The probability of context word w', given observed word w is defined via softmax:

$$\mathrm{P}(w'|w) = \frac{\exp\left(v_{w'} \cdot v_w\right)}{\sum_{\hat{w} \in V} \exp\left(v_{\hat{w}} \cdot v_w\right)}, \qquad (2)$$

where v_w and $v_{w'}$ are corresponding embeddings and V is the corpus vocabulary. Though theoretically sound, the formulation is computationally impractical and requires tractable approximation.

Mikolov et al. (2013) propose two efficient procedures to optimize (1): Hierarchical Softmax and Negative Sampling (NS). In this work we focus on the widely adopted NS. The intuition is that a "good" model should be able to differentiate observed data from noise. The differentiation task is defined using logistic regression; the goal is to tell apart real context-word pair (w', w) from randomly generated noise pair $(\hat{w}, w)$. Formally,

$$\log \mathcal{L}_{[w',w]} = \log \sigma\left(v_{w'} \cdot v_w\right) + \sum_{i=1}^{k} \log \sigma\left(-v_{\hat{w}_i} \cdot v_w\right), \qquad (3)$$

where $\sigma(\cdot)$ is a sigmoid function, and $\{\hat{w}_i\}_{i=1}^{k}$ are negative samples. Summing up all the context-word pairs, we derive the NS Skip-gram objective:

$$\log \mathcal{L}_{word2vec}(W;C) = \sum_{w \in W} \sum_{w' \in C(w)} \log \mathcal{L}_{[w',w]}. \qquad (4)$$

Training word embeddings with Skip-gram, we keep the same hyperparameters across all the runs: 300 dimensions for embeddings, $k = 5$ negative samples, and window of 5 tokens. The *Objective*

[1] https://en.wikipedia.org/wiki/Wikipedia:List_of_policies_and_guidelines

[2] https://www.amazon.com/gp/help/customer/display.html?nodeId=201929730

and *Subjective* corpora undergo the same preprocessing, i.e., discarding short sentences (< 5 tokens) and rare words (< 10 occurrences), removing punctuation, normalizing Unicode symbols.

2.2 Evaluation Tasks

To compare word embeddings, we need a common yardstick. It is difficult to define an inherent quality to word embeddings. Instead, we put them through several evaluation tasks that can leverage word embeddings and standardize their formulations as binary classification tasks. To boil the comparisons down to the essences of word embeddings (which is our central focus), we rely on standardized techniques so as to attribute as much of the differences as possible to the word embeddings. We use logistic regression for classification, and represent a text snippet (e.g., a sentence) in the feature space as the average of the word embeddings of tokens in the snippet (ignoring out-of-vocabulary tokens). The evaluation metric is the average accuracy from 10-fold cross validation.

There are three evaluation tasks of varying degrees of hypothetical subjectivity, as outlined below. Each may involve multiple datasets.

Sentiment Classification Task This task classifies a sentence into either *positive* or *negative*. We use two groups of datasets as follows.

The first group consists of 24 datasets from *UCSD Amazon product data*[3] corresponding to various product categories. Each review has a rating from 1 to 5, which is transformed into *positive* (ratings 4 or 5) or *negative* (ratings 1 or 2) class. For each dataset respectively, we sample 5000 sentences each from the positive and negative reviews. Note that these sentences used for this evaluation task have not participated in the generation of word embeddings. Due to space constraint, in most cases we present the average accuracy across the datasets, but where appropriate we enumerate the results for each dataset.

The second is *Cornell's sentence polarity dataset v1.0*[4] (Pang and Lee, 2005), made up of 5331 each of positive and negative sentences from Rotten Tomatoes movie reviews. The inclusion of this out-of-domain evaluation dataset is useful for examining whether the performance of word embeddings from the *Subjective Corpus* on the first group above may inadvertently be affected by in-domain advantage arising from its Amazon origin.

Subjectivity Classification Task This task classifies a sentence into *subjective* or *objective*. The dataset is *Cornell's subjectivity dataset v1.0*[5], consisting of 5000 subjective sentences derived from Rotten Tomatoes (RT) reviews and 5000 objective sentences derived from IMDB plot summaries (Pang and Lee, 2004). This task is probably less sensitive to the subjectivity within word embeddings than sentiment classification, as determining whether a sentence is subjective or objective should ideally be an objective undertaking.

Topic Classification Task We use the 20 Newsgroups dataset[6] ("bydate" version), whereby the newsgroups are organized into six subject matter groupings. We extract the message body and split them into sentences. Each group's sentences then form the *in-topic* class, and we randomly sample an equivalent number of sentences from the remaining newsgroups to form the *out-of-topic* class. This results in six datasets, each corresponding to a binary classification task. In most cases, we present the average results, and where appropriate we enumerate the results for each dataset. Hypothetically, this task is the least affected by the subjectivity within word embeddings.

3 Comparative Analyses of Subjective vs. Objective Corpora

We conduct a series of comparative analyses under various setups. For each, we compare the performance in the evaluation tasks when using the *Objective Corpus* and the *Subjective Corpus*. Table 1 shows the results for this series of analyses.

Initial Condition Setup I seeks to answer whether there is any difference between word embeddings derived from the *Objective Corpus* and the *Subjective Corpus*. The word embeddings were trained on the whole data respectively. Table 1 shows the corpus statistics and classification accuracies. Evidently, the *Subjective* word embeddings outperform the *Objective* word embeddings on all the evaluation tasks. The margins are largest for sentiment classification (86.5% vs. 81.5% or +5% Amazon, and 78.2% vs. 75.4% or +2.8% on Rotten Tomatoes or RT). For subjectivity and topic classifications, the differences are smaller.

[3] http://jmcauley.ucsd.edu/data/amazon/

[4] http://www.cs.cornell.edu/people/pabo/movie-review-data/rt-polaritydata.README.1.0.txt

[5] http://www.cs.cornell.edu/people/pabo/movie-review-data/subjdata.README.1.0.txt

[6] http://qwone.com/~jason/20Newsgroups/

Setup	Corpus	Corpus Statistics			Classification (Accuracy)			
		# types	# tokens	# sentences	Sentiment Amazon	Sentiment RT	Subjectivity	Topic
I	Objective	1.34M	1.81B	89M	81.5	75.4	90.5	83.2
	Subjective	1.47M	5.49B	313M	86.5	78.2	91.1	83.4
II	Objective	1.34M	1.81B	89M	81.5	75.4	90.5	83.2
	Subjective	0.59M	1.56B		85.5	77.9	90.7	82.8
III	Objective	0.29M	1.75B	89M	81.6	75.6	90.6	83.4
	Subjective		1.54B		85.4	77.9	90.6	82.8

Table 1: Controlled comparison of Objective and Subjective corpora

As earlier hypothesized, the sentiment classification task is more sensitive to subjectivity within word embeddings than the other tasks. Therefore, training word embeddings on a subjective corpus may confer an advantage for such tasks. On the other hand, the corpus statistics show a substantial difference in corpus size, which could be an alternative explanation for the outperformance by the *Subjective Corpus* if the larger corpus contains more informative distributional statistics.

Controlling for Corpus Size In Setup II, we keep the number of sentences in both corpora the same, by randomly downsampling sentences in the *Subjective Corpus*. This procedure consequently reduces the number of types and tokens (see Table 1, Setup II, Corpus Statistics). Note that the number of tokens in the *Subjective* corpus is now fewer than in the *Objective*, the latter suffers no change. Yet, even after a dramatic reduction in size, the *Subjective* embeddings still outperform the *Objective* significantly on both datasets of the sentiment classification task (+4% on Amazon and +2.5% on RT), while showing similar performance on subjectivity and topic classifications.

This bolsters the earlier observation that sentiment classification is more sensitive to subjectivity. While there is a small effect due to corpus size difference, the gap in performance between *Subjective* and *Objective* embeddings on sentiment classification is still significant and cannot be explained away by the corpus size alone.

Controlling for Vocabulary While the *Subjective Corpus* has a much smaller vocabulary (i.e., # types), we turn a critical eye on whether its apparent advantage lies in having access to special word types that do not exist in the *Objective Corpus*. In Setup III, we keep the training vocabulary the same for both, removing the types that are present in one corpus but not in the other, so that out-of-vocabulary words are ignored in the training phase. Table 1, Setup III, shows significant reduction in types for both corpora. Yet, the outperformance by the *Subjective* embeddings on the sentiment classification task still stands (+3.8% on Amazon and +2.3% on RT). Moreover, it is so for both Amazon and Rotten Tomatoes datasets, implying that it is not due to close in-domain similarity between the corpora used for training the word embeddings and the classification tasks.

Objective Corpus	*Subjective Corpus*
waste, money, return, love, great, and, loves, refund, Great, This, product, recommend, this, even, Very, returned, easy, not, send, sent, customer, item, broke, defective, her	money, waste, return, and, Great, love, refund, recommend, great, this, loves, even, product, This, Very, easy, item, junk, anyone, Don't, horrible, gift, poor, Do, returned

Table 2: Top words of misclassified sentences

Significant Words To get more insights on the difference between the *Subjective* and *Objective* corpora, we analyze the mistakes word embeddings make on the development folds. At this point we focus on the sentiment classification task and specifically on the Amazon data, which indicates the largest performance differences in the controlled experiments (see Table 1, Setup III).

As words are still the main unit of information in distributional word embeddings, we extract words strongly associated with misclassified sentences. We employed log-odds ratio with informative Dirichlet prior method (Monroe et al., 2008) to quantify this association. It is used to contrast the words in misclassified vs. correctly classified sentences, and accounts for the variance of words and their prior counts taken from a large corpus.

Table 2 shows the top 25 words most associated with the misclassified sentences, sorted by their association scores. On average 50% of the mistakes overlap for both word embeddings, therefore, some of the words are included in both lists. $40 - 44\%$ of these words carry positive or negative sentiment connotations in general (see the underlined words in Table 2), while other words like *return* or *send* may carry sentiment connotation in e-commerce context. We check if a word carries sentiment connotation using sentiment lexicon compiled by Hu and Liu (2004), including 6789 words along with positive or negative labels.

We also observe linguistic negations (i.e., *not*, *Don't*). For instance, the word most associated with the *Objective*-specific mistakes (excluding the *Subjective* misclassified sentences) is *not*, which suggests that perhaps *Subjective* word embedding accommodates better understanding of linguistic negations, which may partially explain the difference. However, our methodology as outlined in Section 2.2 permits exchangeable word order and is not intended to analyze structural interaction between words. We focus on further analysis of sentiment words, leaving linguistic negations in word embeddings for future investigation.

Controlling for Sentiment Words To control for the "amount" of sentiment in the *Subjective* and *Objective* corpora, we use sentiment lexicon compiled by Hu and Liu (2004). For each corpus, we create two subcorpora: *With Sentiment* contains only the sentences with at least one word from the sentiment lexicon, while *Without Sentiment* is the complement. We match the corpora on the number of sentences, downsampling the larger corpus, train word embeddings on each subcorpus, and proceed with the classification experiments. Table 3 shows the results, including that of random word embeddings for reference. Sentiment lexicon has a significant impact on the performance of sentiment and subjectivity classifications, and a smaller impact on topic classification. Without sentiment, the *Subjective* embeddings prove more robust, still outperforming the *Objective* on sentiment classification, while the *Objective* performs close to random word embeddings on Amazon .

In summary, evidences from the series of controlled experiments support the existence of some X-factor to the *Subjective* embeddings, which confers superior performance in subjectivity-sensitive tasks such as sentiment classification.

Corpus	Subcorpus Sentiment?	Sentiment		Subject-ivity	Topic
		Amazon	RT		
Objective	With	81.8	75.2	90.7	83.1
	Without	76.1	67.2	87.8	82.6
Subjective	With	85.5	78.0	90.3	82.5
	Without	79.8	71.0	89.1	82.2
Random Embeddings		76.1	62.2	80.1	71.5

Table 3: With and without sentiment

4 Sentiment-Infused Word Embeddings

To leverage the consequential sentiment information, we propose a family of methods, called *SentiVec*, for training distributional word embeddings that are infused with information on the sentiment polarity of words. The methods are built upon *Word2Vec* optimization algorithm and make use of available lexical sentiment resources such as SentiWordNet (Baccianella et al., 2010), sentiment lexicon by Hu and Liu (2004), and etc.

SentiVec seeks to satisfy two objectives, namely context prediction and lexical category prediction:

$$\log \mathcal{L} = \log \mathcal{L}_{word2vec}(W; C) + \lambda \log \mathcal{L}_{lex}(W, L), \quad (5)$$

where $\mathcal{L}_{word2vec}(W; C)$ is the Skip-gram objective as in (4); $\mathcal{L}_{lex}(W, L)$ is a lexical objective for corpus W and lexical resource L; and λ is a trade-off parameter. Lexical resource $L = \{X_i\}_{i=1}^{n}$ comprises of n word sets, each X_i contains words of the same category. For sentiment classification, we consider *positive* and *negative* word categories.

4.1 Logistic SentiVec

Logistic SentiVec admits lexical resource in the form of two disjoint word sets, $L = \{X_1, X_2\}$, $X_1 \cap X_2 = \emptyset$. The objective is to tell apart which word set of L word w belongs to:

$$\log \mathcal{L}_{lex}(W, L) \quad (6)$$
$$= \sum_{w \in X_1} \log \mathrm{P}(w \in X_1) + \sum_{w \in X_2} \log \mathrm{P}(w \in X_2).$$

We further tie these probabilities together, and cast the objective as a logistic regression problem:

$$\mathrm{P}(w \in X_1) = 1 - \mathrm{P}(w \in X_2) = \sigma(v_w \cdot \tau), \quad (7)$$

where v_w is a word embedding and τ is a direction vector. Since word embeddings are generally invariant to scaling and rotation when used as downstream feature representations, τ can be chosen randomly and fixed during training. We

experiment with randomly sampled unit length directions. For simplicity, we also scale embedding v_w to its unit length when computing $v_w \cdot \tau$, which now equals to cosine similarity between v_w and τ.

When v_w is completely aligned with τ, the cosine similarity between them is 1, which maximizes $\mathrm{P}(w \in X_1)$ and favors words in X_1. When v_w is opposite to τ, the cosine similarity equals to -1, which maximizes $\mathrm{P}(w \in X_2)$ and predicts vectors from X_2. Orthogonal vectors have cosine similarity of 0, which makes both $w \in X_1$ and $w \in X_2$ equally probable. Optimizing (6) makes the corresponding word embeddings of X_1 and X_2 gravitate to the opposite semispaces and simulates clustering effect for the words of the same category, while the *Word2Vec* objective prevents words from collapsing to the same directions.

Optimization The objective in (6) permits simple stochastic gradient ascent optimization and can be combined with negative sampling procedure for Skip-gram in (5). The gradient for unnormalized embedding v_w is solved as follows:

$$\begin{aligned}&\left(\log \mathcal{L}_{[w\in X_1]}(D,L)\right)'_{v_{wi}} = \left(\log \mathrm{P}\left(x \in X_1\right)\right)'_{v_{wi}} \\ &= \frac{1}{\|v_w\|^2}\,\sigma\left(-\frac{v_w\cdot\tau}{\|v_w\|}\right)\left(\tau_i\,\|v_w\| - v_{wi}\frac{v_w\cdot\tau}{\|v_w\|}\right)\end{aligned} \tag{8}$$

The optimization equation for v_w, when $w \in X_2$, can be derived analogously.

4.2 Spherical SentiVec

Spherical SentiVec extends *Logistic SentiVec* by dealing with any number of lexical categories, $L = \{X_i\}_{i=1}^n$. As such, the lexical objective takes on generic form:

$$\log \mathcal{L}_{lex}(W,L) = \sum_{i=1}^{n}\sum_{w\in X_i} \log \mathrm{P}\left(w \in X_i\right), \tag{9}$$

Each $\mathrm{P}\left(w \in X_i\right)$ defines embedding generating process. We assume each length-normalized v_w for w of L is generated w.r.t. a mixture model of von Mises-Fisher (vMF) distributions. vMF is a probability distribution on a multidimensional sphere, characterized by parameters μ (mean direction) and κ (concentration parameter). Sampled points are concentrated around μ; the greater the κ, the closer the sampled points are to μ. We consider only unimodal vMF distributions, restricting concentration parameters to be strictly positive. Hereby, each $X_i \in L$ is assigned to vMF distribution parameters (μ_i, κ_i) and the membership probabilities are defined as follows:

$$\mathrm{P}(w \in X_i) = P(v_w;\mu_i,\kappa_i) = \frac{1}{Z_{\kappa_i}} e^{\kappa_i \mu_i \cdot v_w}, \tag{10}$$

where Z_κ is the normalization factor.

The *Spherical SentiVec* lexical objective forces words of every $X_i \in L$ to gravitate towards and concentrate around their direction mean μ_i. As in *Logistic SentiVec*, it simulates clustering effect for the words of the same set. In comparison to the direction vector of *Logistic SentiVec*, mean directions of *Spherical SentiVec* when fixed can substantially influence word embeddings training and must be carefully selected. We optimize the mean directions along with the word embeddings using alternating procedure resembling K-means clustering algorithm. For simplicity, we keep concentration parameters tied, $\kappa_1 = \kappa_2 = \ldots = \kappa_n = \kappa$, and treat κ as a hyperparameter of this algorithm.

Optimization We derive optimization procedure for updating word embeddings assuming fixed direction means. Like *Logistic SentiVec*, *Spherical SentiVec* can be combined with the negative sampling procedure of Skip-gram. The gradient for unnormalized word embedding v_w is solved by the following equation:

$$\left(\log \mathcal{L}_{[w\in X_i]}\left(W,L\right)\right)'_{v_{wj}} = \kappa_i \frac{\left(\mu_{ij}\,\|v_w\| - v_{wj}\frac{v_w\cdot\mu_i}{\|v_w\|}\right)}{\|v_w\|^2} \tag{11}$$

Once word embedding v_w $(w \in X_i)$ is updated, we revise direction mean μ_i w.r.t. maximum likelihood estimator:

$$\mu_i = \frac{\sum_{w\in X_i} v_w}{\left\|\sum_{w\in X_i} v_w\right\|}. \tag{12}$$

Updating the direction means in such a way ensures that the lexical objective is non-decreasing. Assuming the stochastic optimization procedure for $\mathcal{L}_{word2vec}$ complies with the same non-decreasing property, the proposed alternating procedure converges.

5 Related Work

There have been considerable research on improving the quality of distributional word embeddings. Bolukbasi et al. (2016) seek to debias word embeddings from gender stereotypes. Rothe and Schütze (2017) incorporate WordNet

lexeme and synset information. Mrkšic et al. (2016) encode antonym-synonym relations. Liu et al. (2015) encode ordinal relations such as hypernym and hyponym. Kiela et al. (2015) augment Skip-gram to enforce lexical similarity or relatedness constraints, Bollegala et al. (2016) modify GloVe optimization procedure for the same purpose. Faruqui et al. (2015) employ semantic relations of PPDB, WordNet, FrameNet to *retrofit* word embeddings for various prediction tasks. We use this *Retrofitting* method[7] as a baseline.

Socher et al. (2011) derive multi-word embeddings for sentiment distribution prediction, while we focus on lexical distributional analysis. Maas et al. (2011) and Tang et al. (2016) use document-level sentiment annotations to fit word embeddings, but document annotation might not always be available for distributional analysis on neutral corpora such as Wikipedia. *SentiVec* relies on simple sentiment lexicon instead. *Refining* (Yu et al., 2018) aligns the sentiment scores taken from lexical resource and the cosine similarity scores of corresponding word embeddings. The method generally requires fine-grained sentiment scores for the words, which may not be available in some settings. We use *Refining* as a baseline and adopt coarse-grained sentiment lexicon for this method.

Villegas et al. (2016) compare various distributional word embeddings arising from the same corpus for sentiment classification, whereas we focus on the differentiation in input corpora and propose novel sentiment-infused word embeddings.

6 Experiments

The objective of experiments is to study the efficacy of *Logistic SentiVec* and *Spherical SentiVec* word embeddings on the aforementioned text classification tasks. One natural baseline is *Word2Vec*, as *SentiVec* subsumes its context prediction objective, while further incorporating lexical category prediction. We include two other baselines that can leverage the same lexical resource but in manners different from *SentiVec*, namely: *Retrofitting* (Faruqui et al., 2015) and *Refining* (Yu et al., 2018). For these methods, we generate their word embeddings based on Setup III (see Section 3). All the methods were run multiple times with various hyperparameters, optimized via grid-search; for each we present the best performing setting.

[7]Original code is available at: https://github.com/mfaruqui/retrofitting

First, we discuss the sentiment classification task. Table 4 shows the unfolded results for the 24 classification datasets of Amazon, as well as for Rotten Tomatoes. For each classification dataset (row), and for the *Objective* and *Subjective* embedding corpora respectively, the best word embedding methods are shown in bold. An asterisk indicates statistically significant[8] results at 5% in comparison to *Word2Vec*. Both *SentiVec* variants outperform *Word2Vec* in the vast majority of the cases. The degree of outperformance is higher for the *Objective* than the *Subjective* word embeddings. This is a reasonable trend given our previous findings in Section 3. As the *Objective Corpus* encodes less information than the *Subjective Corpus* for sentiment classification, the former is more likely to benefit from the infusion of sentiment information from additional lexical resources. Note that the sentiment infusion into the word embeddings comes from separate lexical resources, and does not involve any sentiment classification label.

SentiVec also outperforms the two baselines that benefit from the same lexical resources. *Retrofitting* does not improve upon *Word2Vec*, with the two embeddings essentially indistinguishable (the difference is only noticeable at the second decimal point). *Refining* makes the word embeddings perform worse on the sentiment classification task. One possible explanation is that *Refining* normally requires fine-grained labeled lexicon, where the words are scored w.r.t. the sentiment scale, whereas we use sentiment lexicon of two labels (i.e., positive or negative). *SentiVec* accepts coarse-grained sentiment lexicons, and potentially could be extended to deal with fine-grained labels.

As previously alluded to, topic and subjectivity classifications are less sensitive to the subjectivity within word embeddings than sentiment classification. One therefore would not expect much, if any, performance gain from infusion of sentiment information. However, such infusion should not subtract or harm the quality of word embeddings either. Table 5 shows that the unfolded results for topic classification on the six datasets, and the result for subjectivity classification are similar across methods. Neither the *SentiVec* variants, nor *Retrofitting* and *Refining*, change the subjectivity and topic classification capabilities much, which means that the used sentiment lexicon is targeted only at the sentiment subspace of embeddings.

[8]We use paired t-test to compute p-value.

Corpus/Category	Objective Embeddings					Subjective Embeddings				
	Word2Vec	Retrofitting	Refining	SentiVec		Word2Vec	Retrofitting	Refining	SentiVec	
				Spherical	Logistic				Spherical	Logistic
Amazon										
Instant Video	84.1	84.1	81.9	**84.9***	**84.9***	87.8	87.8	86.9	88.1	**88.2**
Android Apps	83.0	83.0	80.9	**84.0***	**84.0***	86.3	86.3	85.0	**86.6**	86.5
Automotive	80.7	80.7	78.8	81.0	**81.3**	**85.1**	**85.1**	83.8	84.9	85.0
Baby	80.9	80.9	78.6	82.1	**82.2***	84.2	84.2	82.8	84.4	**84.6**
Beauty	81.8	81.8	79.8	82.4	**82.7***	85.2	85.2	83.5	85.2	**85.4**
Books	80.9	80.9	78.9	81.0	**81.3**	85.3	85.3	83.6	85.3	**85.5**
CD & Vinyl	79.4	79.4	77.6	79.4	**79.9**	83.5	83.5	81.9	**83.7**	83.6
Cell Phones	82.2	82.2	80.0	82.9	**83.0***	86.8	86.8	85.3	86.8	**87.0**
Clothing	82.6	82.6	80.7	83.8	**84.0***	86.3	86.3	84.7	86.4	**86.8**
Digital Music	82.3	82.3	80.5	82.8	**83.0***	**86.3**	**86.3**	84.6	86.1	**86.3**
Electronics	81.0	81.0	78.8	80.9	**81.3**	85.2	85.2	83.6	**85.3**	**85.3**
Grocery & Food	81.7	81.7	79.4	**83.1***	**83.1***	85.0	85.0	83.7	85.1	**85.6***
Health	79.7	79.7	77.9	**80.4***	**80.4**	84.0	84.0	82.3	84.0	**84.3**
Home & Kitchen	81.6	81.6	79.5	**82.1**	**82.1**	**85.4**	**85.4**	83.9	85.3	**85.4**
Kindle Store	84.7	84.7	83.2	85.2	**85.4***	88.3	88.3	87.2	88.3	**88.6**
Movies & TV	81.4	81.4	78.5	81.9	**81.9**	85.2	85.2	83.5	85.4	**85.5**
Musical Instruments	81.7	81.6	79.7	**82.4**	**82.4**	85.8	85.8	84.1	**85.9**	85.7
Office	82.0	82.0	80.0	**83.0***	82.9	86.1	86.1	84.5	86.4	**86.5***
Garden	80.4	80.4	77.9	81.0	**81.5**	84.1	84.1	82.5	84.3	**84.6***
Pet Supplies	79.7	79.7	77.5	**80.4**	80.2	83.2	83.2	81.5	83.4	**83.8**
Sports & Outdoors	80.8	80.8	79.1	**81.3***	81.2	84.6	84.6	83.1	84.3	**84.7**
Tools	81.0	81.0	79.3	81.0	**81.3**	84.7	84.7	83.2	84.8	**84.9**
Toys & Games	83.8	83.8	82.0	84.7	**84.9***	87.2	87.2	85.7	87.1	**87.5**
Video Games	80.3	80.3	77.4	81.5	**81.7***	84.9	84.9	83.2	**85.0**	84.9
Average	81.6	81.6	79.5	82.2	**82.4**	85.4	85.4	83.9	85.5	**85.7**
Rotten Tomatoes	75.6	75.6	73.4	**75.8***	75.4	**77.9**	**77.9**	76.7	77.7	**77.9**

Table 4: Comparison of Sentiment-Infused Word Embeddings on Sentiment Classification Task

Corpus/Category	Objective Embeddings					Subjective Embeddings				
	Word2Vec	Retrofitting	Refining	SentiVec		Word2Vec	Retrofitting	Refining	SentiVec	
				Spherical	Logistic				Spherical	Logistic
Topic										
Computers	79.8	79.8	79.6	79.6	79.8	79.8	79.8	79.8	79.7	79.7
Misc	89.8	89.8	89.7	89.8	90.0	90.4	90.4	90.6	90.4	90.3
Politics	84.6	84.6	84.4	84.5	84.6	83.8	83.8	83.5	83.6	83.5
Recreation	83.4	83.4	83.1	83.1	83.2	82.6	82.6	82.5	82.7	82.8
Religion	84.6	84.6	84.5	84.5	84.6	84.2	84.2	84.2	84.1	84.2
Science	78.2	78.2	78.2	78.1	78.3	76.4	76.4	76.1	76.7	76.6
Average	83.4	83.4	83.2	83.3	83.4	82.8	82.8	82.8	82.9	82.8
Subjectivity	90.6	90.6	90.0	90.6	90.6	90.6	90.6	90.3	90.7	90.8

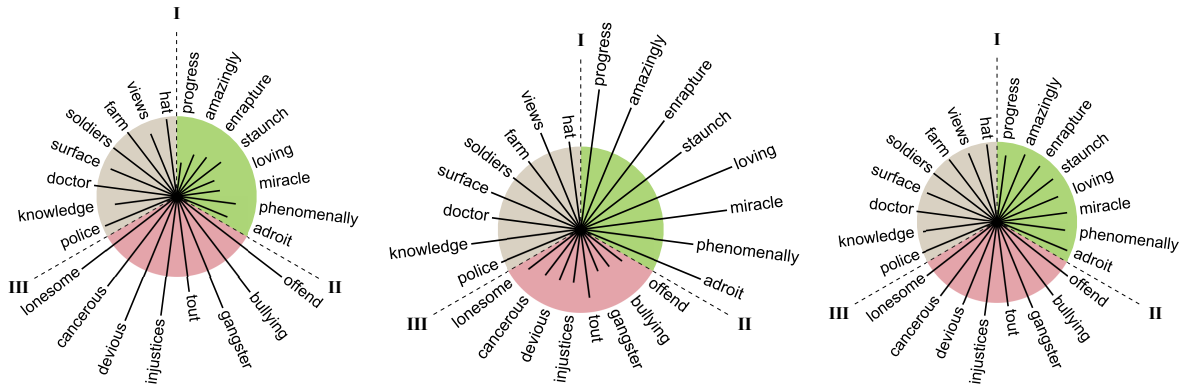
Table 5: Comparison of Word Embeddings on Subjectivity and Topic Classification Tasks

Illustrative Changes in Embeddings To give more insights on the difference between *SentiVec* and *Word2Vec*, we show "flower" diagrams in Figure 1 for *Logistic SentiVec* and Figure 2 for *Spherical SentiVec*. Each is associated with a reference word (e.g., *good* for Figure 1a), and indicates relative changes in cosine distances between the reference word and the testing words surrounding the "flower". Every testing word is associated with a "petal" or black axis extending from the center of the circle. The "petal" length is proportional to the relative distance change in two word embeddings: $\kappa = \frac{d_{SentiVec}(w_{ref}, w_{testing})}{d_{word2vec}(w_{ref}, w_{testing})}$, where $d_{SentiVec}$ and $d_{word2vec}$ are cosine distances between reference w_{ref} and testing $w_{testing}$ words in *SentiVec* and *Word2Vec* embeddings correspondingly. If the distance remains unchanged ($\kappa = 1$), then the "petal" points at the circumference; if the reference and testing words are closer in the *SentiVec* embedding than they are in *Word2Vec* ($\kappa < 1$), the "petal" lies inside the circle; when the distance increases ($\kappa > 1$), the "petal" goes beyond the circle.

The diagrams are presented for Objective Embeddings[9]. We use three reference words: *good* (positive), *bad* (negative), *time* (neutral); as well as three groups of testing words: green for words randomly sampled from positive lexicon (Sector I-II), red for words randomly sampled from negative lexicon (Sector II-III), and gray for frequent neutral common nouns (Sector III-I).

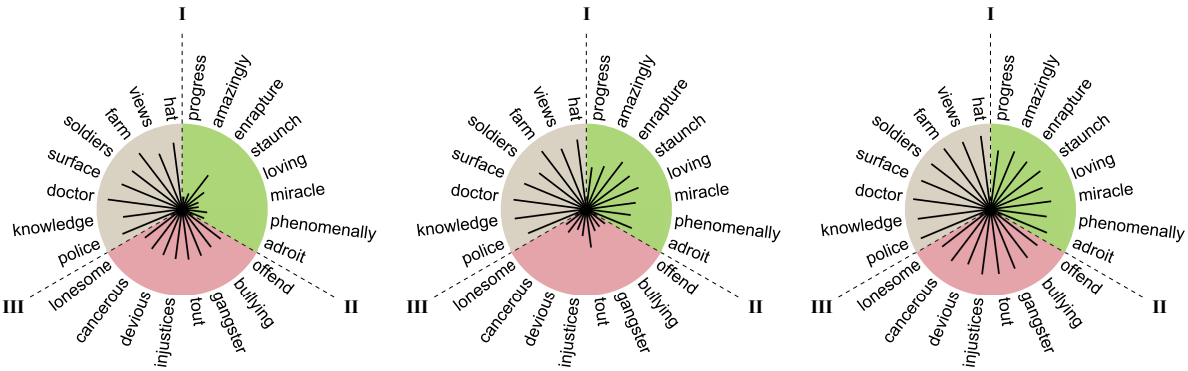
Figure 1 shows changes produced by *Logistic SentiVec*. For the positive reference word (Figure 1a), the average distance to the green words is shortened, whereas the distance to the red words increases. The reverse is observed for the negative reference word (Figure 1b). This observation

[9]The diagrams for Subjective Embeddings show the same trend, with the moderate changes.



(a) Reference word: **good** (positive) (b) Reference word: **bad** (negative) (c) Reference word: **time** (neutral)

Figure 1: Relative changes in cosine distances in *Logistic SentiVec* contrasted with *Word2Vec*



(a) Reference word: **good** (positive) (b) Reference word: **bad** (negative) (c) Reference word: **time** (neutral)

Figure 2: Relative changes in cosine distances in *Spherical SentiVec* contrasted with *Word2Vec*

complies with the lexical objective (7) of *Logistic SentiVec*, which aims to separate the words of two different classes. Note that the gray words suffer only moderate change with respect to positive and negative reference words. For the neutral reference word (Figure 1c), the distances are only moderately affected across all testing groups.

Figure 2 shows that *Spherical SentiVec* tends to make embeddings more compact than *Logistic SentiVec*. As the former's lexical objective (9) is designed for clustering, but not for separation, we look at the comparative strength of the clustering effect on the testing words. For the positive reference word (Figure 2a), the largest clustering effect is achieved for the green words. For the negative reference word (Figure 2b), as expected, the red words are affected the most. The gray words suffer the least change for all the reference words.

In summary, *SentiVec* effectively provides an advantage for subjectivity-sensitive task such as sentiment classification, while not harming the performance of other text classification tasks.

7 Conclusion

We explore the differences between objective and subjective corpora for generating word embeddings, and find that there is indeed a difference in the embeddings' classification task performances. Identifying the presence of sentiment words as one key factor for the difference, we propose a novel method *SentiVec* to train word embeddings that are infused with the sentiment polarity of words derived from a separate sentiment lexicon. We further identify two lexical objectives: *Logistic SentiVec* and *Spherical SentiVec*. The proposed word embeddings show improvements in sentiment classification, while maintaining their performance on subjectivity and topic classifications.

Acknowledgments

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its NRF Fellowship Programme (Award No. NRF-NRFF2016-07).

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*. volume 10.

Danushka Bollegala, Mohammed Alsuhaibani, Takanori Maehara, and Ken-ichi Kawarabayashi. 2016. Joint word representation learning using a corpus and a semantic lexicon. In *Proceedings of AAAI*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of NIPS*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *JMLR* 12(Aug).

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL-HLT*.

Yoav Goldberg. 2016. A primer on neural network models for natural language processing. *JAIR* 57.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM.

Douwe Kiela, Felix Hill, and Stephen Clark. 2015. Specializing word embeddings for similarity or relatedness. In *Proceedings of EMNLP*.

Quan Liu, Hui Jiang, Si Wei, Zhen-Hua Ling, and Yu Hu. 2015. Learning semantic word embeddings based on ordinal knowledge constraints. In *Proceedings of ACL-IJCNLP*. volume 1.

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of ACL-HLT*.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis* 16(4).

Nikola Mrkšic, Diarmuid OSéaghdha, Blaise Thomson, Milica Gašic, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of NAACL-HLT*.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL*.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of ACL*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*.

Sascha Rothe and Hinrich Schütze. 2017. Autoextend: Combining word embeddings with semantic resources. *Computational Linguistics* 43(3).

Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of EMNLP*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*.

Duyu Tang, Furu Wei, Bing Qin, Nan Yang, Ting Liu, and Ming Zhou. 2016. Sentiment embeddings with applications to sentiment analysis. *IEEE TKDE* 28(2).

María Paula Villegas, María José Garciarena Ucelay, Juan Pablo Fernández, Miguel A Álvarez Carmona, Marcelo Luis Errecalde, and Leticia Cagnina. 2016. Vector-based word representations for sentiment analysis: a comparative study. In *XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016)*..

L. C. Yu, J. Wang, K. R. Lai, and X. Zhang. 2018. Refining word embeddings using intensity scores for sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26(3).