A NMF-based Approach to Topic and Sentiment Analysis of Short Texts with Prior Knowledge

Suman Roy, Siddhartha Asthana, Anurag Miglani, and Madhvi Gupta

Optum Global Services Pvt. Ltd., # 44 Electronics City, Hosur Road, Bangalore 560 100, India

Abstract. Huge amount of short texts are generated on the social media across the internet. Retrieving information from them has attracted lot of attention in the recent past as mining these texts can extract customers' sentiment and opinion towards various objects. In this work we focus on sentiment prediction problem for short texts through a joint sentiment/topic model in a semi-supervised setting which is based on NMF-based techniques. Our work offers a couple of advantages over the competing techniques meant for finding out joint sentiment-topic models. One novelty is that we indeed take into account the shortness of the data and learn topics (term-topic matrix) by factoring term context correlation data using non-negative matrix factorization (NMF), rather than the high-dimensional and sparse term occurrence information in texts. Then we learn topic-document matrix from term-topic matrix. In the last step we use tri-matrix factorization to learn sentiment values for unlabeled texts. We consider different short texts from different sources, some of which are labeled with real values using suitable prediction engines and rest are unlabeled, and employ our technique to extract a topic-sentiment mixture model. Our experimental results demonstrates that our method performs better than other topic learning/sentiment classification methods.

Keywords: Short texts, Sentiment Classification, Sentiment prediction, NMF, tri-NMF, Topic.

1 Introduction

Short texts are very popular medium of communication in the social networks that are spread across the internet and appear abundant in different applications. Examples of them are twitter, movie reviews, weblogs, feedback etc. Mining these short texts makes sense as it helps finding customers' sentiment and opinion towards the quality of service they have received from the relevant vendors.

Sentiment classification is a prime opinion mining task which determines if the sentiment orientation in a text is positive, negative or neutral. Most existing approaches to sentiment classification use supervised learning in which labeled documents are used to mark new texts [13] with sentiment classes. While such sentiment classification for texts provides useful information the general association of texts with some topics reveals more intuitive information in certain applications. For example in healthcare application, a customer feedback saying "coordinate the request with the doctor's office directly instead of me being the middle man" reveals a negative orientation and it is

probably about topic related to visit to doctor's clinic. Whereas a feedback with a partial content, "the report is with the post office and mail man handled it beautifully", has a positive orientation and can be related to report mailing topic. In these scenarios, discovering simultaneously both topic and sentiment would aid in opinion mining and summarization of short texts. In lot of sentiment analysis problems, topic detection and sentiment classification are done in a two-step approach in which topics are discovered first followed by assigning sentiment polarities to texts and thus topics [9, 8]. We propose a similar approach for a combined sentiment topic analysis.

In particular in this paper, we address sentiment prediction problem of short texts some of which are labeled with real numbers lying in a particular range, using a semi-supervised approach of extracting joint sentiment/topic model. We base our models on a constrained non-negative tri-factorization of the term-document matrix which can be implemented using novel yet simple update rules. Although our method is similar in spirit to [7] it differs on two aspects. We consider the shortness of data and learn topics by utilizing term context correlation data using non-negative matrix factorization (NMF), rather than the high-dimensional and sparse term occurrence information in short texts. The authors in [7] do mention about the sparsity of the term-document matrix and recommend using sparse matrix multiplications for circumventing this problem. As we determine term-topic matrix using term context correlation matrix our non-negative tri-factorization is more efficiently employed with one less update rule. Subsequently, we learn topic-document matrix from term-document matrix which helps assigning short texts to topics. Finally, we use tri-matrix factorization to learn sentiment value for unlabeled texts. For experimental purposes, we consider a collection of short texts from different applications some of which are labeled and the rest unlabeled, and employ our technique to extract a topic-sentiment model. We also compare our approach with other competing approaches and report on some promising results obtained through our experiments.

Organization of the paper: The paper is organized as follows. In the next Section 1.1 we quickly dispense with the available prior work. In Section 2 we discuss our topic derivation approach for short texts using NMF. We introduce our learning technique to learn sentiment values for unlabeled corpora of short texts in Section 3. We describe our experiment efforts on three different data sets in Section 4. Finally we conclude in Section 5.

1.1 Related Work

There is a plethora of work available on sentiment classification using different machine learning techniques. Most of the classical techniques of sentiment classification are described in a book by Pang and Lee [12]. An unsupervised learning algorithm is proposed for determining opinion polarity of reviews by discovering the average semantic orientation of the phrases containing adjectives or adverbs [18]. In another work Pang *et al.* have used classical techniques for semantically classifying movie reviews, however, these algorithms do not perform well on the sentiment classification problem. Kim and Hovy have designed a system by which they could determine the sentiment of a given sentence by combining the individual world-level sentiment [4]. In another

work [20] the authors handle a similar problem of sentiment classification using Conditional Random Field (CRF) which utilize both contextual dependency and label redundancy. In [7] the authors learn high-quality sentiment models using the model of constrained non-negative tri-factorization.

However, all these above-mentioned methods suffer from a few limitations like they only address the problem of sentiment classification without considering the mixture of topics in the text and most of the approaches advocate the use of supervised techniques requiring labeled corpus for training. An earlier work related to joint modeling of topic with sentiment is the Topic-Sentiment Model (TSM) proposed in [10], in which the authors combine the topic extraction and sentiment prediction for the entire document. As TSM is primarily based on pLSI [3] it suffers from the common drawbacks of the latter such as inferencing problem on new document and over-fitting. To overcome these limitations Lin *et al.* propose an unsupervised hierarchical Bayesian model (JST) which is capable of classifying document sentiment and extracting topic models simultaneously [8]. For the same purpose, researchers have proposed hidden topic-sentiment models (HTSM), a model that explicitly captures topic coherence and sentiment consistency in an opinionated text document [16], using a framework (SURF) that identifies opinions expressed in a review, and then finds similar opinions from other reviews [14] etc. Very recently in [17], Shi *et al.* have introduced semantics-assisted non-negative matrix factorization (SeaNMF) model to discover topics for short texts that can address the inherent weaknesses of short texts like sparseness, ambiguity, nosiness, limited contextual information etc. Our work is mainly motivated by short text corpus (consisting of both unlabeled and unlabeled documents) for which we adopt semi-supervised framework of learning sentiments and discovering topics using tri-matrix factorization.

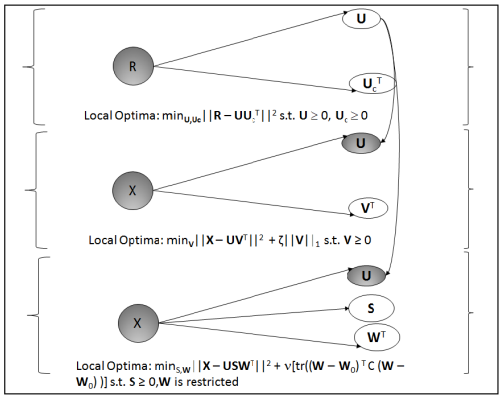


Fig. 1. Different NMF steps

2 The NMF-based Approach for topic discovery

In this work we discover topics from term context correlation matrix [17] instead of term-document matrix [1] which is very sparse as most of the short texts consist of only 2-3 sentences. For this purpose a semantic term correlation matrix is constructed by capturing the relationship between each term and its associated context. Below we denote the elements of a matrix $\mathbf{M} \in \mathbb{R}_+^{p \times q}$ as $[m_{ij}]_{\{1 \leq i \leq p, 1 \leq j \leq q\}}$.

2.1 Correlation between Terms and Context

We know terms are correlated through topics. It has been observed in [1] that the size of distinct terms appearing in short text corpus is relatively small and stable as the size of the corpus becomes bigger, by which the authors conclude that it is more realistic to directly derive the topics from term correlation data rather than the sparse term-document matrix. However, we have observed that the term correlation matrix is even sparse for short texts. For that we adopt the term-context correlation matrix $\mathbf{R}$ (word-context correlation matrix) from [17] which is defined using Skip-gram view of the corpus as,

$$r_{ij} = \max\left[\log\left(\frac{\#(t_i, c_j)}{\#(t_i) \cdot p(c_j)}\right) - \log\kappa, 0\right], 1 \leq i, j \leq m$$

The notation $\mathbb{V}$ denotes the the overall vocabulary of terms and contexts, $\#(t_i, c_i)$ denotes the number of times t_i appears with context c_i in text corpora. Further $\#(t_i) = \sum_{c_j \in \mathbb{V}} \#(t_i, c_j)$ and $\#(c_j) = \sum_{t_i \in \mathbb{V}} \#(t_i, c_j)$ represent the number of times t_i and c_j occur in all possible term-context pairs respectively, and κ is the number of negative samples. Finally, $p(c_j)$ is a unigram distribution for sampling a context c_j defined as $p(c_j) = \frac{\#(c_j)}{\sum_{c_j \in \mathbb{V}} \#(c_j)}$.

We choose a context as a long pseudo-text by aggregating short texts belonging to a cluster. Such a collection of clusters is formed by assuming word embedding representation of short texts and using Word Mover's Distance (WMD) between short texts [15].

2.2 Topic Learning for texts

We assume a semantic representation of terms by considering a latent matrix for terms $\mathbf{U} \in {\mathbb{R}_+}^{m \times k}$. To capture the semantic relationship between terms and their contexts we consider a latent matrix for contexts as $\mathbf{U}_c \in {\mathbb{R}_+}^{n \times k}$. In the above, m and n are the number of terms and documents respectively and k is the number of topics (or classes). The topic learning problem is formulated as finding the term-topic matrix $\mathbf{U}$ by minimizing the following objective function:

$$J_1(\mathbf{U}) = \frac{1}{2}\left\|\mathbf{R} - \mathbf{U}\mathbf{U}_c^T\right\|^2, \text{ s.t. } \mathbf{U} \geq 0, \mathbf{U}_c \geq 0 \tag{1}$$

The minimization problem on the objective function in Eqn. 1 is a NNLS[1] problem which can be solved using multiplicative update algorithm (MUA) [19, 5].

2.3 Deriving Representative terms for a topic

In the second step we generate a representation of a topic in terms of texts. One obtains this representation by factorizing the term-document matrix $\mathbf{X}$ into term-topic matrix $\mathbf{U}$ and document-topic matrix $\mathbf{V}$. Our method utilizes term-topic matrix $\mathbf{U}$ (obtained using Eqn 1) to derive the representative terms. It is desired that a short text should be associated with only a few topics, and hence $\mathbf{V}$ should be sparse. This problem is

[1] NMF is a kind of Non-negative Least Square (NNLS) problem

formulated as finding a non-negative matrix $\mathbf{V}$ (of dimension $m \times k$) by minimizing the loss function as the difference between the product $\mathbf{UV}$ and $\mathbf{X}$ with an additional sparsity constraint on $\mathbf{V}$:

$$J_2(\mathbf{V}) = \frac{1}{2} \left\| \mathbf{X} - \mathbf{U}\mathbf{V}^T \right\|^2 + \zeta \left\| \mathbf{V} \right\|_1, \text{ s.t. } \mathbf{V} \geq 0, \mathbf{U} \text{ is given} \tag{2}$$

The last term in Equation 2 is a regularization term in which ζ is a regularization parameter. The sparse NMF formulation can be easily solved using the MUA by breaking the original problem into two sub-problems, see [6] for details.

3 Sentiment learning by incorporating lexical knowledge

It may be noted that a subset of short texts in the repository considered is labeled with sentiment values which are real and may vary between $-\kappa$ and κ, $\kappa \in \mathbb{N}$, or may vary between 0 and κ' (we consider the latter which can be always obtained by changing scale). These partial labels (in terms of sentiment values) of texts can be captured using a matrix $\mathbf{W}_0 \in {\mathbb{R}_+}^{n \times l}$, where l is the number of sentiment classes. As pre-labeled texts take only real values we assume there is only one sentiment class, that is, $l = 1$. The overall document sentiment class matrix will be given by $\mathbf{W} \in {\mathbb{R}_+}^{n \times l}$. Let us write the structure of $\mathbf{W}$ and $\mathbf{W}_0$ as follows.

$$\mathbf{W} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_h \\ w_{h+1} \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} \mathbf{W}_h \\ \hline \mathbf{W}_{nh} \end{pmatrix} \text{ and } \mathbf{W}_0 = \begin{pmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_h \\ * \\ \vdots \\ * \end{pmatrix} = \begin{pmatrix} \mathbf{\Omega}_h \\ \hline \mathbf{\Omega}_{nh} \end{pmatrix}$$

Above, $\mathbf{W}$ is also partitioned into two column vectors $\mathbf{W}_h$ and $\mathbf{W}_{nh}$, with the former denoting the first h row entries and the latter denoting the second $n - h$ row entries. Similarly, $\mathbf{W}_0$ is partitioned into into two column vectors $\mathbf{\Omega}_h$ and $\mathbf{\Omega}_{nh}$ ($\mathbf{\Omega}_{nh}$ contains only null values) as above. Now, we formulate the semi-supervised learning problem with pre-labeled texts as:

$$\begin{aligned} \min \ \frac{1}{2} \left[\left\| \mathbf{X} - \mathbf{U}\mathbf{S}\mathbf{W}^T \right\|^2 + \nu \left[tr((\mathbf{W} - \mathbf{W_0})^T \mathbf{C} (\mathbf{W} - \mathbf{W_0})) \right] \right] \\ \text{subject to, } \ u_{ij} \geq 0, s_{ij} \geq 0, \\ w_{ij} \geq 0, \ w_{ij} \leq \kappa, \kappa \in \mathbb{N} \end{aligned} \tag{3}$$

Above, we assume $\nu \geq 0$ to be a user-defined constant which determines to what extent we impose $\mathbf{W} = \mathbf{W}_0$. Also we use a diagonal matrix $\mathbf{C} (= [c_{ij}]) \in {\mathbb{R}_+}^{n \times n}$, which is defined by $c_{ii} = 1$ if the sentiment value of i-th (text) row is known, and $c_{ij} = 0$ otherwise.

The Lagrange function is now given by,

$$L(\mathbf{U},\mathbf{S},\mathbf{W}) = \frac{1}{2}\left[\left\|\mathbf{X}-\mathbf{USW}^T\right\|^2 + \nu\left[\mathrm{tr}((\mathbf{W}-\mathbf{W_0})^T\mathbf{C}(\mathbf{W}-\mathbf{W_0}))\right]\right] + \mathrm{tr}(\alpha\mathbf{U}^T) + \mathrm{tr}(\beta\mathbf{S}^T) + \mathrm{tr}(\gamma\mathbf{W}^T) + \mathrm{tr}(\delta(\kappa\mathbf{E}-\mathbf{W})^T) \quad (4)$$

where $\mathbf{E}$ is a matrix of dimension $n \times l$ having all entries equal to 1. Further $\alpha = [\alpha_{ij}], \beta = [\beta_{ij}]$ are Lagrange multipliers for the constraints $u_{ij} \geq 0, s_{ij} \geq 0$ respectively, and $\gamma = [\gamma_{ij}], \delta = [\delta_{ij}]$ are Lagrange multipliers for the constraints $w_{ij} \geq -\kappa,\ w_{ij} \leq \kappa$ respectively. Differentiating $L(\mathbf{U},\mathbf{S},\mathbf{W})$ *wrt* $\mathbf{U},\mathbf{S}$ and $\mathbf{W}$ we get:

$$\frac{\partial L}{\partial \mathbf{U}} = -\mathbf{XWS}^T + \mathbf{USW}^T\mathbf{WS}^T + \alpha; \quad (5)$$

$$\frac{\partial L}{\partial \mathbf{S}} = -\mathbf{U}^T\mathbf{XW} + \mathbf{U}^T\mathbf{USW}^T\mathbf{W} + \beta \quad (6)$$

$$\frac{\partial L}{\partial \mathbf{W}} = -\mathbf{X}^T\mathbf{US} + \mathbf{WS}^T\mathbf{U}^T\mathbf{US} + \nu\mathbf{C}(\mathbf{W}-\mathbf{W}_0) + \gamma - \delta \quad (7)$$

At some optimal solution $\mathbf{U}^*, \mathbf{S}^*$ and $\mathbf{W}^*$ all these partial derivatives will vanish, that is, $\frac{\partial L}{\partial \mathbf{U}}\ |_{\mathbf{U}^*} = 0$, $\frac{\partial L}{\partial \mathbf{U}}\ |_{\mathbf{S}^*} = 0$ and $\frac{\partial L}{\partial \mathbf{U}}\ |_{\mathbf{W}^*} = 0$.

As there are four Lagrange multipliers there would be 4 complementary slackness conditions. Correspondingly there will be $2^4 = 16$ cases to be examined to find the optimal solution for the minimization problem. However, we rule out a couple of cases corresponding to multipliers α and β. Suppose $\alpha^*_{ij} > 0, \beta^*_{ij} > 0$. It does not make sense to consider the optimal values of $\mathbf{U}$ and $\mathbf{S}$ as $u^*_{ij} = 0$ and/or $s^*_{ij} = 0$ since this will lead to a meaningless minimization problem. Hence we shall consider the cases when the corresponding inequality constraints are inactive, that is, $\alpha^*_{ij} = 0$ and $\beta^*_{ij} = 0$. [2]

As we learn $\mathbf{U}$ from Eqn 1 we would not need any update algorithm for $\mathbf{U}$ and treat $\mathbf{U}$ as constant in this learning task.

We now directly solve for $\mathbf{S}$. From Eqn 6 along with the zero gradient condition for $\mathbf{S}$ we get, $\mathbf{U}^T\mathbf{XW} = \mathbf{U}^T\mathbf{USW}^T\mathbf{W}$, solving which we get: $\mathbf{S} = (\mathbf{U}^T\mathbf{U})^{-1}(\mathbf{U}^T\mathbf{XW})(\mathbf{W}^T\mathbf{W})^{-1}$. As $l = 1$ we can drop l and write:

$$s_i \leftarrow \left((\mathbf{U}^T\mathbf{U})^{-1}(\mathbf{U}^T\mathbf{XW})(\mathbf{W}^T\mathbf{W})^{-1}\right)_i \quad (8)$$

The complimentary conditions involving $\mathbf{W}$ are

$$\gamma^*_{ij}(w^*_{ij} + \kappa) = 0 \quad (9)$$

$$\delta^*_{ij}(\kappa - w^*_{ij}) = 0 \quad (10)$$

Correspondingly, we consider 4 cases.

Case 1: Both the constraints in Eqns 9 and 10 are active and $\gamma_{ij} > 0, \delta_{ij} > 0$.

This means $w^*_{ij} + \kappa = 0$ and $\kappa - w^*_{ij} = 0$. This cannot happen simultaneously.

[2] Note that using the zero gradient conditions at optimal values of u^*_{ij} we can get an update rule for $\mathbf{U}$ when it is not fixed, as: $u_{ij} \leftarrow u_{ij}\frac{(\mathbf{XWS}^T)_{ij}}{(\mathbf{USW}^T\mathbf{WS}^T)_{ij}}$

Case 2: Both the constraints in Eqns 9 and 10 are inactive. Consequently, $\gamma_{ij} = 0, \delta_{ij} = 0$.

From the zero gradient condition in Eqn. 7 we get

$$\mathbf{X}^T\mathbf{US} + \nu\mathbf{CW}_0 = \mathbf{W}\underbrace{\mathbf{S}^T\mathbf{U}^T\mathbf{US}}_{\text{scalar}} + \nu\mathbf{CW} \tag{11}$$

Assume $\mathbf{S}^T\mathbf{U}^T\mathbf{US} = \|\mathbf{US}\|^2 = \eta$, where η is a scalar. Also $\mathbf{X}^T\mathbf{US}$ is a column vector. Suppose $\mathbf{X}^T\mathbf{US} = [\alpha_1 \cdots \alpha_n]^T$. Now we can simplify Eqn. 11 and write

$$\alpha_i + \nu * \omega_i = (\eta + \nu) * w_i, \text{ or, } w_i \leftarrow \frac{\alpha_i + \nu * \omega_i}{\eta + \nu}, 1 \leq i \leq h \tag{12}$$

$$\alpha_i = \eta * w_i, \text{ or } w_i \leftarrow \frac{\alpha_i}{\eta}, \overline{h+1} \leq i \leq n \tag{13}$$

While Eqn. 12 provides the updated value of sentiment for the first h rows of $\mathbf{W}$, Eqn. 13 supplies the values of unlabeled texts corresponding to the remaining rows in $\mathbf{W}$.

Case 3: The constraint in Eqn. 9 is inactive and $\gamma_{ij} = 0$, while the other constraint is active and $\delta_{ij} > 0$.

That is from Eqn. 10, $\kappa - w_{ij} = 0$ which gives, $w_{ij} = \kappa$. Replacing $\mathbf{W}$ in its zero gradient condition for in Eqn 7

$$\delta = -\mathbf{X}^T\mathbf{US} + \kappa\mathbf{ES}^T\mathbf{U}^T\mathbf{US} + \nu\mathbf{C}(\mathbf{W} - \mathbf{W}_0) \tag{14}$$

Now check if $\delta > 0$.

Case 4: This is similar to Case 3, in which the constraint in Eqn. 10 is inactive while the other constraint in Eqn. 9 is active. This can be dealt similarly as above.

For practical reasons we would not investigate Cases 3 and 4 further. A uniform assignment of a constant sentiment value to all the unlabeled texts in the collection is unrealistic.

The optimization problem in Eqn. 3 under the assumption of fixed $\mathbf{U}$ is tackled using Eqn. 8 for solving $\mathbf{S}$ and Eqns. 12, and 13 for $\mathbf{W}$ until convergence. We call our approach Tri-Matrix Factorization with Prior Knowledge (TMFPK). The different NMF steps that are performed are shown in Figure 1.

Algorithm 1: Tri-Matrix Factorization with Prior Knowledge (TMFPK)

Begin	:
Input	: Number of topics k, Term-document matrix $\mathbf{X} \in {\mathbb{R}_+}^{m \times n}$ and; term-topic matrix $\mathbf{U} \in {\mathbb{R}_+}^{m \times k}$
Initialization	: Initialize $\mathbf{W} = \mathbf{W}'_0$; $\mathbf{S} = (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{XW}(\mathbf{W}^T\mathbf{W})^{-1}$;
Iteration	: Update $\mathbf{S}$: fixing $\mathbf{W}$, (note we update $\mathbf{S}$ using Eqn. 8) ; Update $\mathbf{W}$: fixing $\mathbf{S}$, (note we update $\mathbf{W}$ using Eqns 12 and 13) ; Until convergence
End	:

We update $\mathbf{W}$ and $\mathbf{S}$ using the rules listed in Algorithm 1 which guarantees an asymptotic convergence to a local minimum, the proof can be carried out using ideas

from $[2, 7]$. The vector $\mathbf{W}_0'$ has the same top h rows as of $\mathbf{W}_0$ and the values for the rest $n-h$ rows are randomly chosen between $-\kappa$ and κ other than 0.
Remark 1: The role of $\mathbf{S}$ should be clarified. It can be used to guess gross sentiment value of a topic.
Remark 2: It is interesting to see how the sentiment values of the labeled texts are modified in this formulation. They change only by the error in the approximation function for $\mathbf{X}$, while the unlabeled ones assume their values by driving the same error towards zero.

Using appropriate partitioned matrices (like $\mathbf{X} = \begin{pmatrix} \mathbf{X}_h \\ \hline \mathbf{X}_{nh} \end{pmatrix}$ etc) Eqn. 11 can be rewritten as:

$$\begin{pmatrix} {\mathbf{X}_h}^T - \mathbf{W}_h \mathbf{S}^T \mathbf{U}^T \\ \hline {\mathbf{X}_{nh}}^T - \mathbf{W}_{nh} \mathbf{S}^T \mathbf{U}^T \end{pmatrix} \mathbf{US} = \begin{pmatrix} \nu(\mathbf{W}_h - \mathbf{\Omega}_h) \\ \hline \mathbf{0} \end{pmatrix} \tag{15}$$

From Eqn. 15 we can see that unlabeled texts are provided with sentiment values that drive the approximation error of $\mathbf{X}_{nh}$ (involving those unlabeled ones) to zero. The set of labeled texts get a modified value depending on the value of $({\mathbf{X}_h}^T - \mathbf{W}_h \mathbf{S}^T \mathbf{U}^T)$ that is achieved. An efficient optimization will result in very small change in the sentiment values of labeled corpora.

4 Experiments

In this section we demonstrate the efficacy of our semi-supervised approach by conducting experiments on practical data sets. The processed data sets and the python code can be found in "https://github.com/SumanRoy68/SentimentPrediction".

4.1 Data set description

We choose three real-world short text data sets corresponding to different types of applications, - feedback, movie reviews, as discussed below.

Data set	#doc	#terms	density	doc-length
OHF	9999	5428	0.00384	26.54
ST	238888	32516	0.000198	6.7
IMDb	50000	88932	0.0015	226.19

Table 1. Basic statistics of the data set used

Optum Healthcare feedback data set (OHF): These feedback texts are provided by customers in certain healthcare domains. The labeled feedback data set is evaluated employing certain criteria and labeled with sentiment values ranging between -5 and 5 using Clarabridge sentiment scoring tool. Clarabridge indexes the sentiment score on an 11-point scale, however, it gives equips with the ability to adjust the sentiment scores to be even more business-specific, for example, user can choose to assign real values as scores.
Stanford treebank data set (ST): The data set in Stanford treebank is from the corpus of movie review excerpts appearing in the 'rottentomatoes.com' website which have been originally collected and published by Pang and Lee in [11]. The data set contains phrases. A raw score in the the form of an integer between 1 and 25 is assigned to each of the phrases, with 1 being the most negative and 25 the most positive. Each phrase can

have 3-6 raw scores from different sources. Finally one computes the sentiment value of a text based on how words compose the meaning of longer phrases, which is then mapped to a real value between 0 and 1.

IMDb data set (IMDb): This data set contains movie reviews from IMDB website. A star rating on a scale of 1 to 10 is provided against each movie review. Although the ratings are integer-valued, we relax the sentiment prediction problem assuming real sentiment values and finally approximate the predicted value as an integer value.

We consider the collection of short texts for each data set for extracting their feature space. We delete stop words from the collection of texts. Then we use Stanford NLP tool to tokenize the texts and perform POS tagging on the tokens. Once the POS tagging is performed we lemmatize the tokens. All the nouns are added as unigrams (also called keywords). Certain combinations of adjectives, verbs and nouns are used to form the bigrams using certain heuristics. We produce some basic statistics about the data set in Table 1. We adopt following notations, '#docs' is the number of documents/texts in each data set, '#terms' is the number of unigrams selected as features in the vocabulary, 'doc-length' is the average length of a document (number of words) in the data set etc. Further, 'density' is defined as $\frac{\#\text{non}-\text{zero}}{\#\text{docs}*\#\text{terms}}$, where '#non-zero' is the number of non-zero elements in the term-document matrix.

4.2 Topic Model Extraction

Most of the existing work in the literature evaluate the quality of derived topics by performing comparison with labeled data sets. Our industrial data are not partitioned a-priori and hence not labeled. Hence we shall use the metric Topic Coherence (T.C.) [17] to determine the quality of topic derivation which do not require the data to be labeled. T.C. is computed using the average PMI score over all the topics which can help evaluating the quality of topic models.

We present the topic coherence results of our model and LDA methos in Table 2. From the table we observe that compared with LDA TMFPK shows significant improvements, which implies that our models discover more coherent topics. Also our model performs better than the standard NMF which shows that TMFPK is effective for learning topics from short texts.

Method	OHF	ST	IMDb
LDA	1.35	1.8	1.53
NMF	1.29	0.52	2.18
TMFPK	1.59	2.16	2.81

Table 2. Topic coherence results

4.3 Comparison with other classification approaches

We compare our supervised method with those of three supervised classification methods: Naive Bayes (NB), Maximum Entropy (ME) and Support Vector Machines (SVM) [13]. In these approaches the sen-

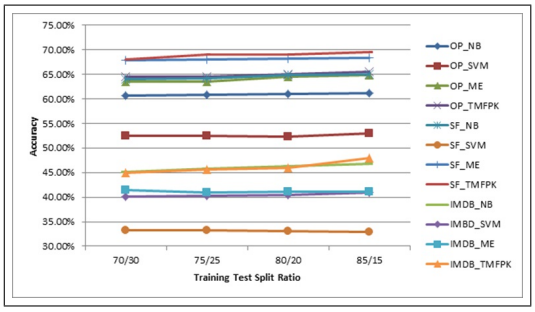


Fig. 2. Accuracy results on different training-test split ratio

timent values are clustered into a couple of classes in order to determine the accuracy of the methods. We can see that training test data split ratio does not impact much the accuracy for almost cases in each data set. The highest accuracy of 70% is achieved for Stanford treebank data set on 85-15 training train split ratio using TMFPK, while the same method achieves an average accuracy of 66% and 48% on Optum feedback and IMDb data sets respectively. The TMFPK performs better than ME approach by almost 2-3% for Stanford Treebank data set, better than ME approach by 2% for Optum feedback data and better than NB approach by about 1% for IMDb data set. Also the accuracy values show a pattern, Stanford treebank data set produces the maximum accuracy, followed by Optum feedback data set and then IMDb data set. We hope that some better tuning of optimization parameters will help achieve better accuracies which we leave for future work.

4.4 Comparison with prediction method

We also use a semi-supervised approach (SS) to predict sentiment value for a new feedback from labeled feedback using clustering and kNN-search. This method proceeds by partitioning feedback into clusters, placing the new feedback in the appropriate cluster to which it is closest to, choosing k nearest neighbors of the feedback in the cluster and finally computing the average of the sentiment scores of those k sentiments which is published as the sentiment value of the new feedback. Also we compute the root mean square error (RMSE) for each data set for a varying vocabulary size to compare the performance of TMFPK with the SS as captured in Figure 3. For IMDb data set SS produces a constant RMSE value of 3.50, while TMFPK realizes a decreasing error pattern with a high of 3.30 at 80% vocabulary size ratio and 3.15 at 95% vocabulary size ratio. For Optum feedback data SS produces a constant RMSE value of 1.80 and our method TMFPK produces a constant RMSE value of 1.70 almost across all vocabulary sizes. However, for Stanford treebank data set TMFPK shows a marginal improvement in RMSE value over SS method.

4.5 Sentiment analysis with prior knowledge: a comparison

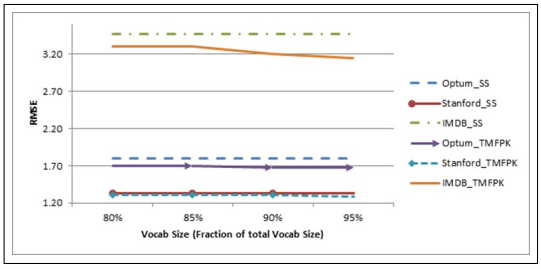
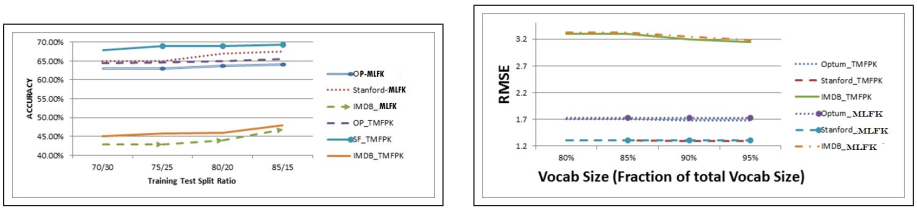


Fig. 3. RMSE for varying vocabulary size of the corpus

We compare our approach with a similar semi-supervised approach [7] called MLFK, which uses lexical prior knowledge in the labeled documents (with sentiment polarities) along with unlabeled documents. As non-negative tri-

matrix factorization is performed on the term-document matrix using simple update rules in this technique we replicate the same in our setting on our data sets, however assuming real values for sentiment scores. In this case we consider the loss function given in Eqn. 3 and use an appropriate update rule for computing $\mathbf{U}$ (like the one given in footnote in Pg. 6), use simple rule in Eqn. 8 for solving $\mathbf{S}$ and use Eqns. 12 and 13 for solving $\mathbf{W}$ to arrive at the optimal solution.

As we study the accuracy values for different training test split ratios in Figure 4(a) the highest accuracy of 70% is observed for Stanford Treebank data set using TMFPK at 85-15 split ratio while MLFK produces 68% at the same split ratio. For Optum feedback data set we achieve an accuracy of 65% almost across all split ratios while MLFK produces a varying accuracy with a low of 62% at 70-30 split ratio and high of 63% at 85-15 split. We observe a pattern of increasing accuracy values for increasing training test split ratios for IMDb data set. At 70-30 split ratio TMFPK and MLFK show accuracy values of 45% and 43% respectively and at 85-15 they produce accuracy values of 48% and 47% respectively. From the plot of RMSE values in Figure 4(b), TMFPK shows marginal improvement over TMFPK for almost all data sets.



(a) Accuracy comparison of TMFPK with MFLK on different training-test split ratio

(b) RMSE values for TMFPK and MFLK on increasing vocabulary size

Fig. 4. Results for TMFPK in comparison with MFLK

5 Conclusion

In this work we have proposed a topic-sentiment mixture model based on NMF approaches. With this model we are able to extract topic and learn sentiment models. In most of the cases our method performs better than other competing methods. In some of the situations we could achieve marginal improvement, although we have to keep in mind that the techniques compared with are primarily designed for classification problems and not for predicting real values. It will be interesting to see whether our 3-step approach could lead to better results in sentiment learning in presence of labeled feedback having more than one polarity classes. An interesting direction would be to compare the bias of a topic with its sentiment value produced by topic-sentiment class matrix $\mathbf{S}$.

References

1. X. Cheng, J. Guo, S. Liu, Y. Wang, and X. Yan. Learning topics in short texts by non-negative matrix factorization on term correlation matrix. In *Proceedings of the 13th SIAM International Conference on Data Mining'13*, pages 749–757, 2013.

2. C. H. Q. Ding, T. Li, W. Peng, and H. Park. Orthogonal Non-negative Matrix Tri-factorizations for Clustering. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 126–135, 2006.
3. T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR '99*, pages 50–57. ACM, 1999.
4. S.-M. Kim and E. Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04. Association for Computational Linguistics, 2004.
5. D. Kuang, J. Choo, and H. Park. *Nonnegative Matrix Factorization for Interactive Topic Modeling and Document Clustering*, pages 215–243. Springer International Publishing, 2015.
6. D. Kuang, J. Choo, and H. Park. Nonnegative matrix factorization for interactive topic modeling and document clustering. In *Partitional Clustering Algorithms*, pages 215–243. Springer, 2015.
7. T. Li, Y. Zhang, and V. Sindhwani. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In *Proceedings of the 47th ACL'09 and the 4th AFNLP'09*, pages 244–252, 2009.
8. C. Lin and Y. He. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 375–384. ACM, 2009.
9. Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th International Conference on World Wide Web, WWW'07*, pages 171–180, 2007.
10. Q. Mei, X. Shen, and C. Zhai. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 490–499, 2007.
11. B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd ACL'05*, pages 115–124. Association for Computational Linguistics, 2005.
12. B. Pang and L. Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, Jan. 2008.
13. B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, EMNLP'02*, 2002.
14. L. Poddar, W. Hsu, and M. Lee. Author-aware aspect topic sentiment model to retrieve supporting opinions from reviews. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP'17*, pages 472–481, 2017.
15. J. Qiang, P. Chen, T. Wang, and X. Wu. Topic modeling over short texts by incorporating word embeddings. In *PAKDD'17, Proceedings, Part II*, pages 363–374, 2017.
16. M. M. Rahman and H. Wang. Hidden topic sentiment model. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 155–165, 2016.
17. T. Shi, K. Kang, J. Choo, and C. K. Reddy. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In *Proceedings of the World Wide Web Conference on World Wide Web, WWW'18*, pages 1105–1114, 2018.
18. P. D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)'04*, pages 417–424, 2002.
19. W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03. ACM, 2003.

20. J. Zhao, K. Liu, and G. Wang. Adding redundant features for crfs-based sentence sentiment classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 117–126. Association for Computational Linguistics, 2008.