# Distributional Semantic Representation for Text Classification and Information Retrieval

[NLP\_CEN\_AMRITA@MSIR-FIRE-2016]

Barathi Ganesh HB Artificial Intelligence Practice Tata Consultancy Services Kochi - 682 042 India barathiganesh.hb@tcs.com

# ABSTRACT

The objective of this experiment is to validate the performance of the distributional semantic representation of text in the classification (Question Classification) task and the Information Retrieval task. Followed by the distributional representation, first level classification of the questions is performed and relevant tweets with respect to the given queries are retrieved. The distributional representation of text is obtained by performing Non - Negative Matrix Factorization on top of the Document - Term Matrix in the training and test corpus. To improve the semantic representation of the text, phrases are also considered along with the words. This proposed approach achieved 80% as a F-1 measure and 0.0377 as a mean average precision against the its respective Mixed Script Information Retrieval task1 and task 2 test sets.

### Keywords

Distributional Semantics; Non-Negative Matrix Factorization; Term - Document Matrix; Text Classification

# 1. INTRODUCTION

Information Retrieval (IR) and Text classification are the classic applications in text analytics domain, that is utilized in the multiple domains and industries in various forms. Given a text content, the classifier must have the capability of classifying it into the predefined set of classes and given a query, the search engine must have the capability of retrieving relevant text content within the stored collection of text [1][12]. This task becomes more complex, when the text contents are represented in more than one language. This introduces the problem during the representation as well as while mining information out of it.

The fundamental component in classification and retrieval task is text representation, which tries to represent the given text into its equivalent form of numerical components. Later, these numerical components are utilized directly to perform the further actions or will be used to extract the features required for performing further action.

This text representation methods evolved over the time to improve the originality of representation, which paves way to move from the frequency based representation methods to Anand Kumar M and Soman KP Center for Computational Engineering and Networking (CEN) Amrita School of Engineering, Coimbatore Amrita Vishwa Vidyapeetham Amrita University, India m anandkumar@cb.amrita.edu, kp soman@.amrita.edu

the semantic representation methods. Though other methods like set-theoretic Boolean systems are also available, this paper focuses only on Vector Space Model (VSM) and Vector Space Model of Semantics (VSMs) [13].

In VSM, the text is represented as a vector, based on the occurrence of terms (binary matrix) or frequency of the occurrence of terms (Term - Document Matrix) present in the given text. The given text is represented as a vector, based on frequency of terms that occur in the text. Here, 'terms' represents words or phrases [9]. Considering only the term frequency is not sufficient, since it ignores syntactic and semantic information that lies within the text.

The term documents matrix is inefficient due to the biasing problem (i.e. few terms gets higher weight because of unbalanced and uninformative data). To overcome this, Term Frequency - Inverse Document Frequency (TF-IDF) representation method is introduced, which re-weighs the terms based upon its presence across the documents [7]. It has a tendency to give higher weights to the rarely occurring words, wherein these words may be misspelled which is obvious with social media texts.

The Vector Space Model of Semantics (VSMs) overcomes the above mentioned shortcomings by weighing terms based on the context. This is achieved by applying TDM on matrix factorization methods like Singular Value Decomposition (SVD) and Non - Negative Matrix Factorization (NMF) [10, 15, 11]. This has the ability of weighing terms though it is not present in a given query. This is because, matrix factorization leads to represent the TDM matrix with its basis vectors [5]. This representation does not include the syntactic information which requires large data and is computationally high because of its high dimension.

Word Embeddings along with the structure of the sentence are utilized to represent the short texts. This requires very less data and the dimension of the vector can be controlled. But to develop the Word to Vector (Word2Vec) model it requires a very large corpus [14][4]. Here we are not considering it since we do not have large size mixed script text data. Followed by representation, similarity measures is carried on in-order to perform the question classification task. Here similarity measures are distance measure (Cosine distance, Euclidean distance, Jaccard distance, etc.) and correlation measure (Pearson correlation coefficient) [6].



**Distributional Representation** 

#### Figure 1: Model Diagram for Distributional Representation of Text

Considering above said pros and cons, here the proposed approach is experimented to observe the performance of distributional semantic representation of text in the classification and retrieval task. The given questions are represented as a TDM matrix after the necessary preprocessing steps and NMF is applied on it to get the distributional representation. Thereafter, distance measure and correlation measures between entropy vector of each class and vector representation of the question are computed in order to perform the question classification task and in order to retrieve the relevant tweets with respect to the given query, cosine distance between query and tweets are measured.

### 2. DISTRIBUTIONAL REPRESENTATION

This section describes about the distributional representation of the text, which is used further for the question classification and retrieval tasks. The systematic approach for the distributional representation is given in Figure 1.

### 2.1 **Problem Definition**

Let,  $Q = q_1, q_2, q_3, ..., q_n$  are the questions  $(q_i \text{ represents})$ the  $i^{th}$  question),  $C = c_1, c_2, ..., c_n$  are the classes which the questions falls under and n is size of corpus.  $T = t_1, t_2, ..., t_n$ are the tweets which the questions are related and n is size of corpus. The objective of the experimentation is to classify each query into its respective predefined classes in task 1 and retrieving the relevant tweets with respect to the input query in task 2.

### 2.2 Preprocessing

Few of the terms that appears across multiple classes will shows conflict towards the classification, where the terms generally gets low weighs in TF-IDF representation. Hence these terms are eliminated if it occurs more than 3/4 times across the classes and in order to avoid the sparsity of the representation, terms with the document frequency of one are eliminated. Here TF-IDF representation not considered. Because, it has a tendency to provide weighs for the rare words which is more common in mixed script texts. Here, advantage of the TF-IDF representation is indirectly obtained by handling document frequency of the terms.

# 2.3 Vector Space Model : Term - Document Matrix

In TDM, vocabulary has been computed by finding unique words present in the given corpus. Then the number of times term presents (term frequency) in each question is computed against the vocabulary formed. The terms present in this vocabulary acts as a first level features.

$$\left[A\right]_{i,i} = TDM(Corpus) \tag{1}$$

$$[A]_{i} = termfrequency(q_{i}) \tag{2}$$

Where, i represents the  $i^{th}$  question and j represents the  $j^{th}$  term in the vocabulary. In-order to improve the representation, along with the unigram words, the bi-gram and tri-gram phrases also considered after following above mentioned preprocessing steps.

### 2.4 Vector Space Model of Semantics : Distributional Representation

The above computed TDM is applied on NMF to get the distributional representation of the given corpus.

$$\left[W\right]_{i,r} = nmf(\left[A\right]_{i,i}) \tag{3}$$

In general matrix factorization is done to get the product of matrices, subject to their reconstruction that the error needs to be low. The product components from the factorization gives the characteristics of the original matrix [10, 15]. Here NMF is incorporated along with the proposed model to get the principal characteristic of the matrix, known as basis vector. Sentences may vary in its length but their representation needs to be of fixed size for its use in various applications. TDM representation followed by the Non - Negative Matrix Factorization (NMF) will achieve this [16]. Mathematically it can be represented as,

$$A \approx W H^T \tag{4}$$

If A is  $m \times n$  original TDM matrix, then W is  $i \times r$  basis matrix and H is  $j \times r$  mixture matrix. Linear combination of basis vectors (column vectors) of W along with the weights of H gives the approximated original matrix A. While factorizing, initially random values are assigned to W and H then the optimization function is applied on it to compute appropriate W and H.

$$minf_r(W,H) \equiv \left\| V - WH^T \right\|_F^2$$
(5)  
s.t.  $W, H \ge 0$ 

Here F is Forbenius norm and r is parameter for dimension reduction, which is set to be 10 to have  $i \times 10$  fixed size vector for each question.



Question Class

Figure 2: Model Diagram of Proposed Approach

Here NMF is used for finding out the basis vector for the following reasons: the non-negativity constraints makes interpretability straight forward than the other factorization methods; selection of r is straight forward; and the basis vector in semantic space is not constrained to be orthogonal, which is not affordable by finding singular vectors or eigen vectors [8].

# 3. QUESTION CLASSIFICATION

Question Answering (QA), systems becoming necessary units in all the industry as well as the non - industrial domains. Especially, they try to automate the manual efforts required in the personal assistance systems and virtual agents. With this information the remaining part of the section describes about the proposed approach in question classification task.

For this experiment the data set has been provided by Mixed Script Information Retrieval (MSIR) task committee [3, 2]. The detailed statistics about the training and the testing set are given in Table 1.

The objective of task is to classify the given question into its corresponding class. The distributional representation of the given training and testing corpus are computed as described in the previous section. The systematic diagram for the remaining approach is given in Figure 2.

After the representation, the reference vector for the each class is computed by summing up the question vectors in that class. This reference vector acts as a entropy vector for its corresponding class. This is mathematically represented as,

$$r_c = \sum^c q_i \tag{6}$$

s.t. 
$$q_i \in c$$

Description	Train file	Test file
# Questions	330.0	180.0
Total # Unique Words	407.0	282.0
# Words after Filtering	220.0	133.0
Average # Words per Question	0.67	0.74
# Bi-Grams after Filtering	207.0	118.0
Average # Bi-Grams per Question	0.63	0.66
# Tri-Grams after Filtering	92.0	55.0
Average # Tri-Grams per Question	0.28	0.31
Total # Features	519.0	306.0

 Table 1: Data-set Statistics



 Table 2: Measured Similarity Features

$$R_c = r_{c_1}, r_{c_2}, \dots, r_{c_n} \tag{7}$$

Then the similarity measures between question vector  $q_i$ and reference vectors in R are computed. Similarity measures computed are given in table 2. These similarity measures that is computed are taken as the attributes for the supervised classification algorithm.

The Random Forest Tree (RFT) with  $nC_{\sqrt{n}}$  number of trees are utilized to perform the supervised classification. In order to ensure the performance, 10-fold 10-cross validation performed during the training and this yields 82% as precision. Proposed approach yields 79.44% as accuracy measure against the test set and statistics about the results are tabulated in Table 3. There are totally 3 runs were submitted to the task committee, which has changes in final classification algorithm. In this paper we described about the approach that yields best performance.

### 4. INFORMATION RETRIEVAL

The information shared through the social media is huge and it has various challenges in its representation. This induces to carryout research in order to obtain useful in-

Team	Accuracy	PER	LOC	ORG	NUM	TEMP	MONEY	DIST	OBJ	MISC
AmritaCEN	80.55	0.82	0.81	0.56	0.92	1.00	0.77	0.97	0.57	NA
AMRITA-CEN-NLP	79.44	0.80	0.89	0.60	0.85	0.97	0.72	0.95	0.58	NA
Anuj	81.11	0.81	0.81	0.55	0.98	0.97	0.96	0.95	0.50	NA
BITS_PILANI	81.11	0.72	0.74	0.74	0.96	0.92	0.94	0.95	0.50	0.20
IINTU	83.33	0.80	0.89	0.65	0.94	1.00	0.81	0.97	0.53	NA
IIT(ISM)D*	80.00	0.77	0.89	0.61	0.89	0.94	0.76	0.95	0.58	NA
NLP-NITMZ	78.88	0.85	0.81	0.70	0.89	0.95	0.74	0.92	0.33	0.20

### Table 3: Results

formation out of it. IR is being part of such a research, which is basic component in text analytics and serves as a input to the other applications. One of the major problem is handling the transliterated texts in IR. These transliterated texts introduces more complex problem especially with representation.

For this experiment the data set has been provided by Mixed Script Information Retrieval (MSIR) task committee [3]. The detailed statistics about the training and the testing set are given in Table 4.

Description	Train file	Test file
# Questions	11238.0	5090.0
Total # Words	19654.0	11994.0
# Words after Filtering	13616.0	6756.0
Average # Words per Question	1.21	1.32
# Bi-Grams after Filtering	46673.0	16494.0
Average # Bi-Grams per Question	4.15	3.24
# Tri-Grams after Filtering	56513.0	11961.0
Average # Tri-Grams per Question	5.03	2.35
Total # Features	116802.0	35211.0

Table 4: Data-set Statistics

The objective of this task is to retrieve the top 20 relevant tweets from the corpus with respect to the input query. Primarily queries and corpus are distributionally represented as described in the section 3.

$$Q = q_1, q_2, q_3, \dots, q_n \tag{8}$$

$$T = t_1, t_2, \dots, t_n \tag{9}$$

Then the cosine distance between the query and the corpus vectors are calculated to retrieve the top 20 tweets with minimum distance. Mathematically it is expressed as,

$$similarity = \frac{\sum_{i=1}^{n} q_i t_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$
(10)

$$distance = \frac{\cos^{-1}(similarity)}{\pi} \tag{11}$$

The proposed distributional representation based approach yields 0.0377 mean average precision against the test queries, which is best amongst the other approaches proposed in this task [2]. The statistics about the obtained results are given in Table 5.

Team	Mean Average Precision
UB	0.0217
Anuj	0.0209
Amrita_CEN	0.0377
NLP-NITMZ	0.0203
NITA_NITMZ	0.0047
CEN@Amrita	0.0315
IIT(ISM)D	0.0083

Table 5: Results

# 5. CONCLUSION

The classification task and retrieval task are developed based on the distributional representation of the text by utilizing term - document matrix and non-negative matrix factorization. The proposed approach outperformed well in both the task, but there is still room for the improvement. Though the distributional representation methods performed well, it suffers from the well known problem 'Curse of Dimensionality'. It requires a much research in feature engineering, which directly reduces the dimension of the term - document matrix. Hence the future work will be focused on improving performance of the retrieval and reducing the dimensionality of the representation basis vectors.

### 6. **REFERENCES**

- C. C. Aggarwal and C. Zhai. A survey of text classification algorithms. *InMining text data*, pages 163–222, 2012.
- [2] S. Banerjee, S. Naskar, P. Rosso, S. Bandyopadhyay, K. Chakma, A. Das, and M. Choudhury. Msir@fire: Overview of the mixed script information retrieval. In Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation, Kolkata, India, December 7-10, 2016, CEUR Workshop Proceedings. CEUR-WS.org, 2016.
- [3] S. Banerjee, N. Sudip Kumar, P. Rosso, and S. Bandyopadhyay. The first cross-script code-mixed question answering corpus. *In Modelling, Learning and mining for Cross/Multilinguality Workshop*, pages 56–65, 2016.
- [4] H. B. Barathi Ganesh, M. Anand Kumar, and K. P. Soman. Amrita cen at semeval-2016 task 1: Semantic

relation from word embeddings in higher dimension. Proceedings of SemEval-2016, pages 706–711, 2016.

- [5] W. Blacoe and M. Lapata. A comparison of vector-based representations for semantic composition. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 546–556, 2012.
- [6] S.-H. Cha. Comprehensive survey on distance/similarity measures between probability density functions. *City*, 1, 2007.
- [7] R. Juan. Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning, 2003.
- [8] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. 1999.
- [9] A. Manwar, H. Mahalle, K. Chinchkhede, and V. Chavan. A vector space model for information retrieval: A matlab approach. *Indian Journal of Computer Science and Engineering*, 3:222–229, 2012.
- [10] R. Pat. An introduction to latent semantic analysis. Indian Journal of Computer Science and Engineering.
- [11] U. Reshma, H. B. Barathi Ganesh, and M. Anand Kumar. Author identification based on word distribution in word space. In Advances in Computing, Communications and Informatics (ICACCI), pages 1519–1523, 2015.
- [12] A. Roshdi and A. Roohparvar. Review: Information retrieval techniques and applications.
- [13] G. Salton, W. Anita, and Y. Chung-Shu. A vector space model for automatic indexing. *Communications* of the ACM, 18:613–620, 1975.
- [14] R. Socher, E. Huang, J. Pennin, C. Manning, and A. Ng. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. Advances in Neural Information Processing Systems, pages 801–809, 2011.
- [15] W. Xu, X. Liu, and Y. Gong. Xu w, liu x, gong y. document clustering based on non-negative matrix factorization. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, pages 267–273, 2003.
- [16] Y. Ye. Comparing matrix methods in text-based information retrieval. Tech. Rep., School of Mathematical Sciences, Peking University, 2000.