Distributional Semantic Representation in Health Care Text Classification

NLP_CEN_AMRITA@CHIS-FIRE-2016

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

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

ABSTRACT

This paper describes about the our proposed system in the Consumer Health Information Search (CHIS) task. The objective of the task 1 is to classify the sentences in the document into relevant or irrelevant with respect to the query and task 2 is analysing the sentiment of the sentences in the documents with respect to the given query. In this proposed approach distributional representation of text along with its statistical and distance measures are carried over to perform the given tasks as a text classification problem. In our experiment, Non - Negative Matrix Factorization utilized to get the distributed representation of the document as well as queries, distance and correlation measures taken as the features and Random Forest Tree utilized to perform the classification. The proposed approach yields 70.19% in task 1 and 34.64% in task 2 as an average accuracy.

Keywords

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

1. INTRODUCTION

Over the past few years, tremendous amount of investment and research carried on to enhance the predictive analytics through text analytics in health care domain [11, 10]. Health care information are available as a text (Clinical Trails) in the form of admission notes, literature, reports and summaries [1][2][3]. Unlike traditional structure of text resources, the unstructured nature of clinical trial's text sources are introduces more challenges while mining information out of it. These available challenges induces researchers to carry out the text analytics research to enhance the developed model and to create the new models.

The informations explicitly available in Electronics Health Records (EHR) but implicitly available in clinical trails as a form of text. Now, our primary problem is becomes, representing text that can be easily and effectively used for further application. The application may be a sequential modeling tasks (Information Extraction) or text classification tasks (Document Retrieval, sentiment analysis on retrieved documents and Validation of retrieved documents).

Document retrieval is primary task in text analytics application in which the Consumer Health Information Search (CHIS) is focused on validating the retrieved results (Relevant or Irrelevant) and performing sentiment analysis on retrieved results (Support, Oppose and Neutral). The given problem can be viewed as a text classification problem with the target classes as mentioned in above two tasks.

Text classification is a classic application 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 [1]. This task becomes more complex, when the text contents includes medical descriptions (Drug names, Measurements and Dosages). This introduces the problem during the representation as well as while mining information out of it.

The fundamental component in classification 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 for the classification or will be used to extract the features required to perform the classification task. 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 the semantic representation methods. Though other methods 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 within the text by having vocabulary built across the entire corpus. Here, 'terms' represents the words or the phrases [8]. Considering only the term frequency is not sufficient, since it ignores the 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 un-

[1]https://medlineplus.gov/
[2]https://clinicaltrials.gov/
[3]https://clinicaltrials.gov/

balanced and uninformative data). To overcome this, Term Frequency - Inverse Document Frequency (TF-IDF) representation method is introduced, which re-weighs the term frequency based upon its presence across the documents [5]. It has a tendency to give higher weights to the rarely occurring words, wherein these words may be misspelled or uninformative words with respect to the classification task which is obvious with clinical trail 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) [9, 15, 12]. 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 [3]. 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. To develop the Word to Vector (Word2Vec) model it requires a very large corpus [14][2]. Here we are not considering it since we do not have large size clinical trails text data. Followed by the representation, similarity measures is carried on between the query and text documents to achieve the objective. Here similarity measures are distance measure (Cosine distance, Euclidean distance, Jaccard distance, etc.) and correlation measure (Pearson correlation coefficient) [4].

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 task. The given query and documents 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 query vector of each document and vector representation of the sentences in the documents are computed in order to perform the classification task.

2. DISTRIBUTIONAL REPRESENTATION

This section describes about the distributional representation of the text, which is used further for the classification task. The distributional representation aims to compute the basis vector from the term frequency vector by applying NMF on the TDM. The systematic approach for the distributional representation is given in Figure 1.

2.1 Problem Definition

Let, $d_k = s_1, s_2, s_3, ..., s_n$ are the sentences in the k th document in the document set $D = d_1, d_2, d_3, ...d_n$, q_i represents the i^{th} query and $C = c_1, c_2, ..., c_n$ are the classes in which s falls under with respect to the q and n is the size of corpus. The objective of the experimentation is to classify each sentence in the document into its respective predefined classes.

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

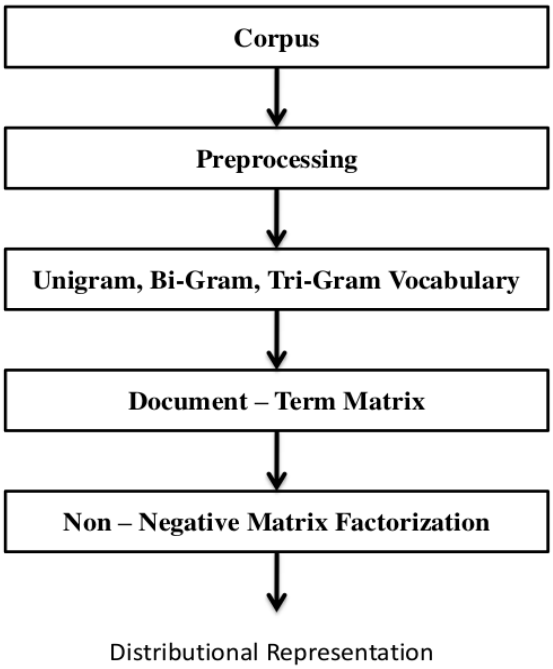

Figure 1: Model Diagram for Distributional Representation of Text

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 clinical texts (Drug names, Measurements and Dosage levels). 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.

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

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

Where, i represents the i^{th} sentence 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.

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

In general matrix factorization is done to get the product

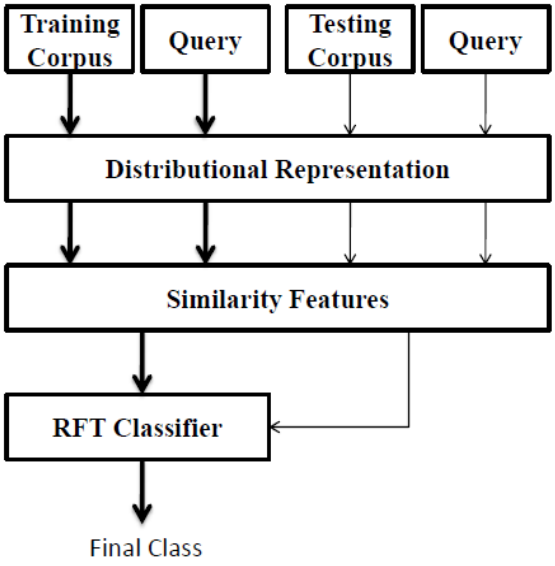

Figure 2: Model Diagram of Proposed Approach

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 [9, 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,

$$A \approx WH^T \tag{4}$$

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.

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

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. 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 [6].

3. TEXT CLASSIFICATION

For this experiment the data set has been provided by Consumer Health Information Search (CHIS) task committee [7]. The detailed statistics about the training and the testing set are given in Table 1.

Task 1 : This task is becoming necessary unit in-order to filter the retrieved results from Information Retrieval (IR) application. This ensures the recall of the Search Engine which is mandatory in health care domain text analytics applications. With this information the remaining part of the section describes about the proposed approach in text classification in task 1.

Let, $d_k = s_1, s_2, s_3, ..., s_n$ are the sentences in the k th document in the document set as mentioned in the Table 1 ($D = skincare, MMr, HRT, Ecig, Vitc$), q_i represents the i^{th} query and $C = Relevant, Irrelevant$ are the classes which the s falls under with respect to the q. n is size of corpus and this is also mentioned in Table 1.

The objective of task is to classify the given question into its corresponding classes (Relevant, Irrelevant). 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 similarity measures between query vector q_i and sentence vectors in D are computed. The computed similarity measures are given in table 3. These similarity measures that is computed are taken as the attributes for the supervised classification algorithm which is Random Forest Tree (RFT).

By having typical $fC_{\sqrt{f}}$ number of trees, output labels $Y = y1, y2, y3, ..., yn$ (Relevant, Irrelevant) and feature set $F = f1, f2, f3, ..., fn$ the bagging repeatedly (B times - Number of trees) done by selecting random samples and attributes from the training set and builds the decision tree for each set. Then the predictions for test set can be find by averaging the predictions from all the individual decision trees built through the train set. It can be interpreted as following:

$$f_b = f(W_b, Y_b, F_b) \tag{6}$$

$$Y = \frac{1}{B}\sum_{b=1}^{B} f_b(\hat{W}\hat{F}) \tag{7}$$

In order to ensure the performance, 10-fold 10-cross validation performed during the training and this yields near 72% as a precision and it yields 68.12% against the test set.

Task 2 : This task is also necessary unit, in-order interpret further information from the retrieved results. This is task is similar to the task 1 and carried on exactly similar to the task 1 with target class labels as $C = Oppose, Support, Neutral$. The classes in C are the final output label which the s falls under with respect to the q.

Here also 10-fold 10-cross validation performed during the training and this yields near 45% as a precision and it yields 38.53% against the test set. The detailed description about the results are given in Table 2.

4. CONCLUSION

The objective of the tasks (Consumer Health Information Search) are performed as a text classification problem based on the distributional representation of the text by utilizing

Document Types	# Training Sentences	# Task 1 Classes		# Task 2 Classes			# testing Sentences
		Relevant	Irrelevant	Oppose	Support	Neutral	
skincare	65	34	31	34	16	15	90
MMr	70	49	21	34	33	3	60
HRT	60	45	15	41	15	4	74
Ecig	82	71	11	33	27	22	66
Vitc	64	38	26	32	21	11	74

Table 1: Data-set Statistics

Document Types	Task 1 Results in %			Task 2 Results in %		
	Max	Min	Ours	Max	Min	Ours
skincare	79.55	48.86	48.86	73.8	23.86	23.86
MMr	89.66	56.89	88.89	68.97	32.75	34.72
HRT	93.06	38.89	75.86	54.16	22.2	43.10
Ecig	76.56	46.88	76.56	67.19	29.69	39.06
Vitc	78.38	55.41	60.81	50.00	31.08	32.43
Average	78.10	54.84	70.19	55.43	33.64	34.64

Table 2: Results Statistics

Measured Feature Functions
Similarity (Dot Product): $P^T * Q$
Euclidean Distance: $\sqrt{\sum_{i=1}^{d} \|P_i - Q_i\|^2}$
Bray Curtis Dissimilarity: $\frac{\sum_{i=0}^{d} \|P_i - Q_i\|}{\sum_{i=0}^{d} (P_i + Q_i)}$
Chebyshev Distance: $\min_i \|P_i - Q_i\|$
Correlation: $\sum_{i=1}^{d} \frac{(P_i - Q_i)^2}{Q_i}$

Table 3: Measured Similarity Features

term - document matrix and non-negative matrix factorization. Even though the proposed approach not yields the state of art performance in the tasks, the obtained results are good enough to continue the research. These results are described in the Table 2. Distributional semantic representation methods suffers from the well known problem 'Curse of Dimensionality'. Hence the future work will be focused on reducing the dimensionality of the representation basis vectors and including the dedicated feature engineering for health care domain.

5. REFERENCES

[1] C. C. Aggarwal and C. Zhai. A survey of text classification algorithms. *InMining text data*.

[2] 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.

[3] W. Blacoe and M. Lapata. A comparison of vector-based representations for semantic composition.

[4] S.-H. Cha. Comprehensive survey on distance/similarity measures between probability density functions. *City*, 1, 2007.

[5] R. Juan. Using tf-idf to determine word relevance in document queries. 2003.

[6] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. 1999.

[7] S. Manjira, M. Sandya, and R. Shourya. Chis@fire: Overview of the chis track on consumer health information search. In *Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation, Kolkata, India, December 7-10, 2016*, CEUR Workshop Proceedings. CEUR-WS.org, 2016.

[8] 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.

[9] R. Pat. An introduction to latent semantic analysis. *Indian Journal of Computer Science and Engineering*.

[10] F. Popowich. Using text mining and natural language processing for health care claims processing. 2005.

[11] W. Raghupathi and V. Raghupathi. Big data analytics in healthcare: promise and potential. volume 1, 2014.

[12] U. Reshma, H. B. Barathi Ganesh, and M. Anand Kumar. Author identification based on word distribution in word space. 2015.

[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. 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. pages 267–273, 2003.

[16] Y. Ye. Comparing matrix methods in text-based information retrieval. 2000.