NLP_CEN_AMRITA @ SMM4H: Health Care Text Classification through Class Embeddings

Barathi Ganesh Hullathy Balakrishnan, Vinayakumar, Anand Kumar Madasamy, Soman Kotti Padannayil
Center for Computational Engineering and Networking (CEN), Amrita School of Engineering Coimbatore, Amrita Vishwa Vidyapeetham, Amrita University, India,
barathiganesh.hb@gmail.com, *vinayakumarr77@gmail.com*,
m_anandkumar@cb.amrita.edu, *kp_soman@amrita.edu*

Abstract

Artificial Intelligence has been a major breakthrough in many domains. Now, it has started automating health care domain through Natural Language Processing and Computer Vision applications. As a part of it, researchers are now focusing more on mining health related information from the text shared through social media and clinical trials. This paper explains about our system for health care text classification tasks conducted by Health Language Processing (HLP) Lab. We experimented with representing the target classes available in task 1 and task 2 as vectors. The classification has been performed using Support Vector Machine. To compute the representation for target classes, we used traditional methods available in Vector Space Models and Vector Space Models of Semantics. In this shared task, the task 1 is about distinguishing the tweets mentioning "adverse drug reaction" from the ones which do not. The task 2 is about distinguishing the tweets that includes personal medication intake, possible medication intake and non-intake. The preliminary results are satisfying in-order to continue the research in developing a representation method for target classes.

Introduction

Making sense out of information shared (as a text) through social media is becoming a part of many applications. This shared information is directly from the user and are considered to be highly reliable. The users profile information as well as the health related information from these shared texts will help us in developing an automated solution for personalized medicine[1].

Unlike the text from other domains, the health care domain text includes more content words than the functional words. The available patterns between the verb and the proper-noun in health texts are complex and vast in nature. Due to this, the matrix built out of the traditional representation methods becomes sparse. These are the core reasons for the requirement of an effective representation in-order to carry out further steps in developing any application. It becomes even more complex when these health texts are taken from social media for mining information out of it as the users tends to use short words, abbreviations and symbols etc.

Representation is an essential part for any Natural Language Processing (NLP) application[2]. Most of the available methods focuses on representing the direct text data. Similar to this text data, the target classes also includes the latent information and can be represented as a vector. Using this class vector, we can build further applications.

In this paper, we have experimented to compute the representation scheme for target classes from the traditional representation methods available in Vector Space Models (VSM) and Vector Space Models of Semantics (VSMs). In this shared task, the task 1 is about distinguishing Adverse Drug Reaction (ADR) mentioned tweets from those that do not and task 2 is about distinguishing the tweets that includes personal medication intake, possible medication intake and non-intake[8,9].

Representation

The objective here is to represent the given tweets into its equivalent numerical representation in-order to carry out classification.

Representation : Vector Space Models

Document - Term Matrix (DTM) and Term Frequency - Inverse Document Frequency (TF-IDF) representation methods are used in which the given tweets $T = t_1, t_2, t_3, ..., t_n$ are presented as a matrix D with the dimension $m \times n$. Here m represents the number of tweets and n represents the number of unique words present in the tweet collection T.

$$D = dtm(T) \tag{1}$$

$$D = tfidf(T) \tag{2}$$

In DTM the frequency count of the words alone are considered to form the representation for tweets[3]. In TF-IDF, along with the frequency count of the words, frequency count of the words appearing across the tweets (inverse document frequency) are also taken into the consideration[4]. This re-weighting scheme in TF-IDF gives higher weights to the rarely occurring word and lower weights to the frequently occurring word.

Representation : Vector Space Models of Semantics

The matrix computed from the previous section undergoes matrix factorization to get the distributional representation of tweets. These vectors can be seen as a semantic representation of tweets, as the vector produced out of matrix factorization becomes the basis vector representation of matrix D. Here Singular Value Decomposition (SVD) is used to perform the matrix factorization[5],[6].

$$U\Sigma V^T = svd(D) \tag{3}$$

In the above equation, U represents the distributional representation of tweets with the dimension of $m \times m$, V^T represents the distributional representation of the words with the dimension of $n \times n$ and Σ represents the significance of the basis vectors present in U and V^T. In detail, column vectors in U are the Eigen vector of DD^T which represents the column space, column vector in V^T are the Eigen vector of $D^T D$ which in turns represents the row space and the diagonal element of Σ are the squared Eigen values of DD^T and $D^T D$. The computation of DD^T finds the cross co-occurrence of the words in the Matrix D. Finally, the resultant column vector in U is taken as D to for further steps.

Representation : Class Embedding

We have experimented to represent the target classes as an entropy vector by summing up the tweets vectors available in the matrix D with respect to the target class. This can be mathematically represented as,

$$C_e = \sum_{i=1}^{m} D[i,:] \; if \; t_c = C \tag{4}$$

In above equation, C_e represents the class embedding (entropy vector of the class) , t_c represents the target classes per tweet and C represents the available unique target classes. The dimensions of the class embedding in Vector Space Model representation is $1 \times n$ and $1 \times m$ in Vector Space Models of Semantics representation.

Representation : Feature Learning

The distance, similarity and correlation between the class embedding and tweet vectors are measured to get the feature matrix in-order to perform the final prediction.

$$F = features(D, C_e) \tag{5}$$

Here F is the feature matrix with the dimension $m \times (5 \times numberofuniquetargetclasses)$. The measured features are Dot Product, Euclidean Distance, Chebyshve Distance, Bray Curtis Dissimilarity and Correlation[7].

Experiments

This section details about how the proposed approach is applied on Task 1 and Task 2 data sets. Task 1 is a binary classification problem[8] and task 2 is a multi-class classification problem[9]. The dataset for both the tasks are provided by shared task organizers and its statistics are given in Table 1 and Table 2. Each task's data set includes training data, development data and test data.

Table 1: Task 1 Data Statistics

Data	**Total # Tweets**	**Total # Classes**	**# ADR Mentioned Tweets**	**# ADR not Mentioned Tweets**
Train	6725	2	721	6004
Dev	3535	2	240	3295
Test	9961	2	9190	771

Table 2: Task 2 Data Statistics

Data	**Total # Tweets**	**Total # Classes**	**Personal Medicine Intake**	**Possible Medicine Intake**	Non Intake
Train	1065	3	192	373	500
Dev	712	3	125	230	357
Test	7513	3	1731	2697	3085

The tweets in the given datasets are represented as a matrix using methods described in VSM and VSMs sections. The available target classes per class is mentioned in Table 1 and Table 2. The submitted runs varies only with representation but further classification remains same for all the runs. In task 1, the given data is represented as DTM in run1, TF-IDF in run2, DTM followed by a SVD in run3 and TF-IDF followed by a SVD in run4. While performing SVD we have taken column vectors from the U as a basis vector representation for tweets. The dimension of the vectors is equal to the number of instances. Similar to task 1, task 2 is also computed with the four types of representations.

The class embedding for target classes are computed by summing up the tweet vectors that belonged to the respective classes. In this way, for task 1 we have computed two class embeddings (ADR mentioned and ADR not mentioned). For task 2, we have computed three class embeddings (personal medication intake, possible medication intake and non-intake).

On successive computation of class embeddings, the features are computed between the tweet vectors and class embeddings as mentioned in Feature Learning Section. These measures are taken as the attributes and given to the classifier to make the final prediction. In task 1, one class SVM is used to handle the label biasing problem. In task 2, SVM with RBF kernel is used to make the final prediction.

In task 1, it has been observed that except TF-IDF, the other representation methods shows higher error in training the one class SVM with the ADR mention. Based on this, in submitted runs the training model is based on the tweets in which the ADR is not mentioned. The observed training error rate for task 1 is given in Table 3.

In task 2, applying SVD tends to appear as the over fitted model by giving constant accuracy for 10 - cross 10 - fold validation. Hence, we avoided to submit the multiple runs for task 2. We have submitted the model based on DTM and class embedding. The final submitted runs were evaluated by the shared task organizers and the obtained results are given in Table 4 and Table 5.

Conclusion

The preliminary approach to class representation method attains considerable accuracy in both the tasks. It has been observed that the imbalance in the target classes is the core reason for low score. Especially in the proposed class

Table 3: Task 1 Training Error Rate

Run	Training Module	Error Against Same Class	Error Against Opposite Class
1	ADR not mentioned	930	866
2	ADR mentioned	930	852
3	ADR not mentioned	195	2456
4	ADR not mentioned	856	792

Table 4: Task 1 Results

Run	ADR Precision	ADR Recall	ADR F-score
1	0.057	0.093	0.071
2	0.056	0.109	0.074
3	0.087	0.204	0.121
4	0.186	0.481	0.268

Table 5: Task 2 Results

Run	Micro-averaged precision for classes 1 and 2	Micro-averaged recall for classes 1 and 2	Micro-averaged F-score for classes 1 and 2
1	0.569	0.39	0.462

representation the entropy of the target class vector is directly dependent on the number of instances that belonged to the respective class. Hence the future work will be to focus on handling the label biasing problem, which is a common scenario with many practical applications.

References

1. Barathi Ganesh HB, Anand Kumar M, and Soman KP, Distributional Semantic Representation in Health Care Text Classification, Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation, Kolkata, India, December 7-10, 2016, 201–204. http://ceur-ws.org/Vol-1737/T5-3.pdf.
2. Barathi Ganesh HB, Anand Kumar M, and Soman KP, Vector Space Model as Cognitive Space for Text Classification, arXiv, 2017, http://arxiv.org/abs/1708.06068.
3. Antonellis Ioannis, and Efstratios Gallopoulos, Exploring term-document matrices from matrix models in text mining, arXiv preprint (2006).
4. Ramos Juan, Using tf-idf to determine word relevance in document queries, Proceedings of the first instructional conference on machine learning (2003).
5. Thomas K Landauer, Latent Semantic Analysis, Encyclopedia of Cognitive Science (2006).
6. Barathi Ganesh HB, Anand Kumar M, and Soman KP, Statistical Semantics in Context Space, Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016, 881–889, http://ceur-ws.org/Vol-1609/16090881.pdf.
7. Cha, Sung-Hyuk, Comprehensive survey on distance/similarity measures between probability density functions, City (2007).
8. Sarker, Abeed, and Graciela Gonzalez, Portable automatic text classification for adverse drug reaction detection via multi-corpus training, Journal of biomedical informatics 53 (2015): 196-207.
9. Klein, Ari, Abeed Sarker, Masoud Rouhizadeh, Karen O'Connor, and Graciela Gonzalez, Detecting Personal Medication Intake in Twitter: An Annotated Corpus and Baseline Classification System, BioNLP 2017 (2017): 136-142.