Contextual Transformation of Short Text for Improved Classifiability

Anirban Chatterjee* Smaranya Dey† Uddipto Dutta ‡

March 23, 2019

Abstract

Text classification is the task of automatically sorting a set of documents into predefined set of categories. This task has several applications including separating positive and negative product reviews by customers, automated indexing of scientific articles, spam filtering and many more. What lies at the core of this problem is to extract features from text data which can be used for classification. One of the common techniques to address this problem is to represent text data as low dimensional continuous vectors such that the semantically unrelated data are well separated from each other. However, sometimes the variability along various dimensions of these vectors is irrelevant as they are dominated by various global factors which are not specific to the classes we are interested in. This irrelevant variability often causes difficulty in classification. In this paper, we propose a technique which takes the initial vectorized representation of the text data through a process of transformation which amplifies relevant variability and suppresses irrelevant variability and then employs a classifier on the transformed data for the classification task. The results show that the same classifier exhibits better accuracy on the transformed data than the initial vectorized representation of text data.

1 Introduction

Text classification is the task of automatically sorting a set of documents into predefined set of categories. Despite considerable work in the area of numerical and categorical data analysis, the processing of textual data is still left with a number of challenges [1]. Extensive theoretical and empirical study has been carried out in last decade in the field of machine learning and text analytics. The problems which lie at the core of text analytics are essentially text clustering and text classification. Text clustering is the application of statistical cluster analysis as well as data mining techniques to unstructured digital text documents [2] while text classification is defined as the task of classifying a group of documents under multiple headings. Currently automatic text classification is widely used in spam filtering, topic modelling, automatic title generation, sentiment analysis and many more. The importance of applying text mining methods in the information systems discipline has been highlighted in Debortoli *et al.* [3].

Recent popularization of e-commerce and online communication has triggered interest in classification of short texts. Several applications such as chat messages, twitter feeds, online product reviews, feedback mechanism, news comments, online question answering and many more, involve massive amount of short texts. There has been work to categorize text and include the categories in a set of subsequent statistical analysis to explore the relationships among the categories. However this subsequent statistical analysis is sensitive to mis-classification; thus accurate categorization is imperative for effective modelling that relies

*WalmartLabs, Anirban.Chatterjee@walmartlabs.com
†WalmartLabs, Smaranya.Dey@walmart.com
‡WalmartLabs, Uddipto.Dutta@walmartlabs.com

on categorization results.

There are several reasons the problem of short text classification is deemed to be challenging. Firstly, enough numbers of words are not present to indicate the context of short text. It is problematic to select powerful language features since shared context and word co-occurrences are insufficient for using valid distance measures. Secondly, short text always appears in large quantity, resulting in conventional document classification running into problems such as labelling bottlenecks. It is too cumbersome to assign labels manually in a sizable training set. Thirdly, the common existence of non-standard terms and noise such as misspellings, grammatical errors, abbreviations, slang words or even foul language makes it even more difficult to make the model tolerant of a certain degree of such "anomalies". Consequently, how to reasonably represent and select discriminatory features (Forsyth *et al.* [4]), effectively reduce spatial dimensionality and noise, and make the best use of those limited hand-labelled instances are stimulating questions for short text classification.

In this paper, our focus is on learning a transformation function which can project the original text data into a new space to provide better classifiability without introducing any further complexity in the classification model. We set this problem in a semi-supervised learning setting where we know the label information of a small amount of short text and use this information to learn the transformation function. We demonstrate how the text data after transformation exhibits better classification accuracy with same model.

The rest of the paper is organized as follows: the next section presents a review of related literature on text clustering, classification and guided data transformation. This is followed by introduction on proposed data transformation method and how it is applied to text data. The subsequent section gives an overview of data used for training and testing along with configurations and experimental results.

2 Related Work

While tremendous achievements have been obtained in numerical or categorical data analysis, the processing of textual data still remains to be perfected. A lot of work has taken place on text embeddings which are distributed representations of text data in the form of low dimensional continuous vectors. Various embedding techniques are commonly used in text classification task. Some of the commonly used methods are embedding layer, word2vec [10], GloVE [5] etc. For instance, word2vec is a statistical method for learning the embeddings of a word given a text corpus. The principal advantage of using word2vec is that, high-quality word embeddings can be learned efficiently (low space and time complexity). It allows learning of higher dimensional embeddings from large corpora of text (billions of words). Continuous Bag-of-Words Model [10] and Continuous Skip-Gram Model [10] are two different learning models that were introduced as a part of the word2vec approach. The CBOW model learns the embedding by predicting the current word based on its context. The continuous skip-gram model learns by predicting the surrounding words given a current word. However, embeddings done using these methods often bring about irrelevant variability in various dimensions and hence often perform sub-optimally when it comes to classification of text. Much of the literature has focused on long text contexts. These successful results now need to be investigated in short text contexts for their applicability and potential improvement. One example of this work is presented by Sun *et al.* [6] using a straightforward but scalable method which achieved favourable categorization accuracy. The approach started with a manual extraction of representative word combinations. A Term Frequency and Inverse Document Frequency (TFIDF) weighting scheme were adopted to ensure topic-specificity of the query word sequences that can represent as much content as possible. Then it searched for a limited set of hand-coded texts that are most relevant to the query instances and determined the category heading according to the highest score and vote based on previous search results. In contrast, Li *et al.* [7] subscribed to the view that the classical TFIDF weighting factor is not effective for short text classification. They argued that even the refined TFITF algorithm (ITC) which substitutes term frequency with its logarithmic form has conspicuous imperfections on account of the high dependence on the quality of training collection [7] . Hence, the authors

demonstrated a solution which overcomes the deficiencies to a large extent. The new hybrid functions amalgamated the Document Distribution Entropy algorithm as well as the Position Distribution Weight algorithm together, and it outperformed the conventional methods.

3 Methodology

Representation of text is one of the fundamental problems in natural language processing. The concept of word vectors offers a feasible approach to compute text vector. The word vectors are computed such a way that the semantically related words are located close to each other in the vector space. One of the ways to compute text vector is to aggregate the word vectors constituting the text. Below are the overview of how contextual transformation of naively computed text vectors can potentially improve the text classification accuracy without introducing any further complexity in the classification model.

Word-to-Vector Based Text Classification Consider a text classification: map an input sequence of tokens X to one of k labels. We first apply a composition function g to the sequence of word embeddings $\boldsymbol{v}_w$ for $w \in X$. The output of this composition function is a vector $\boldsymbol{z}$ that serves as input to a logistic regression function.
In our instantiation of g, it averages word embeddings

$$\boldsymbol{z} = g(w \in X) = \frac{1}{|X|} \sum_{w \in X} \boldsymbol{v}_w.$$

Feeding $\boldsymbol{z}$ to softmax layer induces estimated probabilities for each output label

$$\hat{y} = \text{softmax } (\boldsymbol{W}_s.\boldsymbol{z} + \boldsymbol{b}),$$

where the softmax function is

$$\text{softmax}(q) = \frac{\exp \boldsymbol{q}}{\sum_{j=1}^{k} \exp \boldsymbol{q}_j},$$

$\boldsymbol{W}_s$ is a $k \times d$ matrix for a dataset with k output labels and $\boldsymbol{b}$ is a bias term.

Transformation of Text Vectors In transformation operation, some of the dimensions of the text vectors are upscaled and some are downscaled to minimize the irrelevant variation in the data for the intended classification task because irrelevant variation often adds noise to the data blurring the separation between dissimilar data-points. This operation essentially means projecting the original data into a new space which exhibits better separation among data-points belonging to different classes and more proximity among data-points which belong to same class.
More formally, when an input X is transformed into a new representation Y, it seeks to maximize the mutual information $I(X, Y)$ between X and Y such that for every class the total sum-of-square distance among the data-points belonging to that class has an upper-bound K. This leads us to solving an optimization problem with objective function and constraints are as follows

$$\max_{f \in F} I(X, Y) \quad \text{s.t.} \quad \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} ||y_{ji} - m_j^y||^2 \leq K$$

where f is a transformation function, m_j^y is the centroid of the data-points belonging to class j in the transformed space, n is the total number of data points and K is a constant.
According to [8], the above optimization equations can be rewritten as follows

$$\max_{A} |A| \quad \text{s.t.} \quad \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} ||y_{ji} - m_j^y||^2 \leq K \quad \text{and} \quad A \succeq 0$$

where A is a linear transformation function with $Y = AX$. In the context of text data, X is a collection of all the text vectors and Y is the new representation of the all the text data.

Text Classification with Transformed Text Data In our new formulation, text embeddings are computed as follows

$$\boldsymbol{z}_{new} = g_{new}(w \in X) = \frac{1}{|X|} f\,(\sum_{w \in X} \boldsymbol{v}_w)$$

where f is the transformation function as explained above.
Once text vector $\boldsymbol{z}$ is computed as above, feeding $\boldsymbol{z}$ to softmax layer induces estimated probabilities for each output label

$$\hat{y} = \text{softmax}\ (\boldsymbol{W}_s.\boldsymbol{z}_{new} + \boldsymbol{b}),$$

4 Experiments

Data, Configurations and Results IMDB Movie Review Database [9] features 25,000 movie reviews for training and testing of binary sentiment classification. The reviews in this database are either positive or negative. 70% of the data were used for training the classifier and 30% for testing. Logistic regression and SVM classifier were used to measure the accuracy of classification against original and transformed data. Obtained results exhibit 12% improvement in accuracy against transformed data.

Classifier	**Accuracy with Original Data**	**Accuracy with Transformed Data**
Logistic Regression	72%	**80%**
SVM	68%	**74%**

5 Conclusion

In this paper we proposed an approach to perform contextual transformation of text data for better classifiability. Initial experimental results show improvement in classification accuracy after data transformation. This improvement suggests that contextual transformation is a promising research direction for text classification and related application. One of the possible future direction of research could be to address the scalability aspect of contextual transformation while dealing with ultra-high dimensional data.

References

[1] Losiewicz, P., Oard, D. W., & Kostoff, R. N. Textual data mining to support science and technology management. *Journal of Intelligent Information Systems*, 2000, 15(2), 99-119.

[2] Steinbach, M., Karypis, G., & Kumar, V. A comparison of document clustering techniques. Paper presented at the KDD workshop on text mining, 2000.

[3] Debortoli, S., Müller, O., Junglas, I., & vom Brocke, J. Text Mining For Information Systems Researchers: An Annotated Topic Modeling Tutorial. *Communications of the Association for Information Systems*, 2016, 39, pp-pp. https://doi.org/10.17705/1CAIS.03907.

[4] Forsyth, R. S., & Holmes, D. I. Feature-finding for text classification. *Literary and Linguistic Computing*, 1996, 11(4), 163-174.

[5] Pennington, Jeffrey & Socher, Richard & Manning, Christoper. Glove: Global Vectors for Word Representation. EMNLP. 2014, 14. 1532-1543. 10.3115/v1/D14-1162.

[6] Ma, J., Xu, W., Sun, Y.-h., Turban, E., Wang, S., & Liu, O. An ontology-based text-mining method to cluster proposals for research project selection. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions*, 2012, 42(3), 784-790.

[7] Li, L., & Qu, S. Short Text Classification Based on Improved ITC. *Journal of Computer and Communications*, 2013

[8] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning Distance Functions using Equivalence Relations. *Proc. International Conference on Machine Learning (ICML)*, 2003, pp. 11-18.

[9] Maas, A., Daly, R., Pham, P., Huang, D., Ng, A. and Potts, C. Learning Word Vectors for Sentiment Analysis. *Proc. Association for Computational Linguistics(ACL)*, 2011, pp.142–150.

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. *Proc. NIPS*, 2013.