UPM
UNIVERSITI PUTRA MALAYSIA
BERILMU BERBAKTI

UNIVERSITI PUTRA MALAYSIA
AGRICULTURE • INNOVATION • LIFE

Improvement of Sentiment Analysis based on Clustering of Word2Vec Features

Eissa Alshari, Azreen Azman, Norwati Mustapha, Shyamala Doraisamy and Mustafa Alkeshr

29 August 2017

TIR 2017

Outline

- Introduction
- Feature Extraction Method based on Clustering for Word2Vec
- Results
- Conclusion

Introduction

- More users rely on **online reviews** or **comments** to make everyday decision on products and services.
- **Summarizing** the **overall sentiment** on a product or service is still a challenge to researchers.

Introduction

- The **features** used in the classification of text for sentiment analysis plays an important role in its success.
- Several type of features have been investigated:
  - **Discrete distribution** such as LDA, LSA and bag-of-words (BoW).
  - **Continuous distribution** such as Word2Vec, Doc2Vec and other NN based approaches.

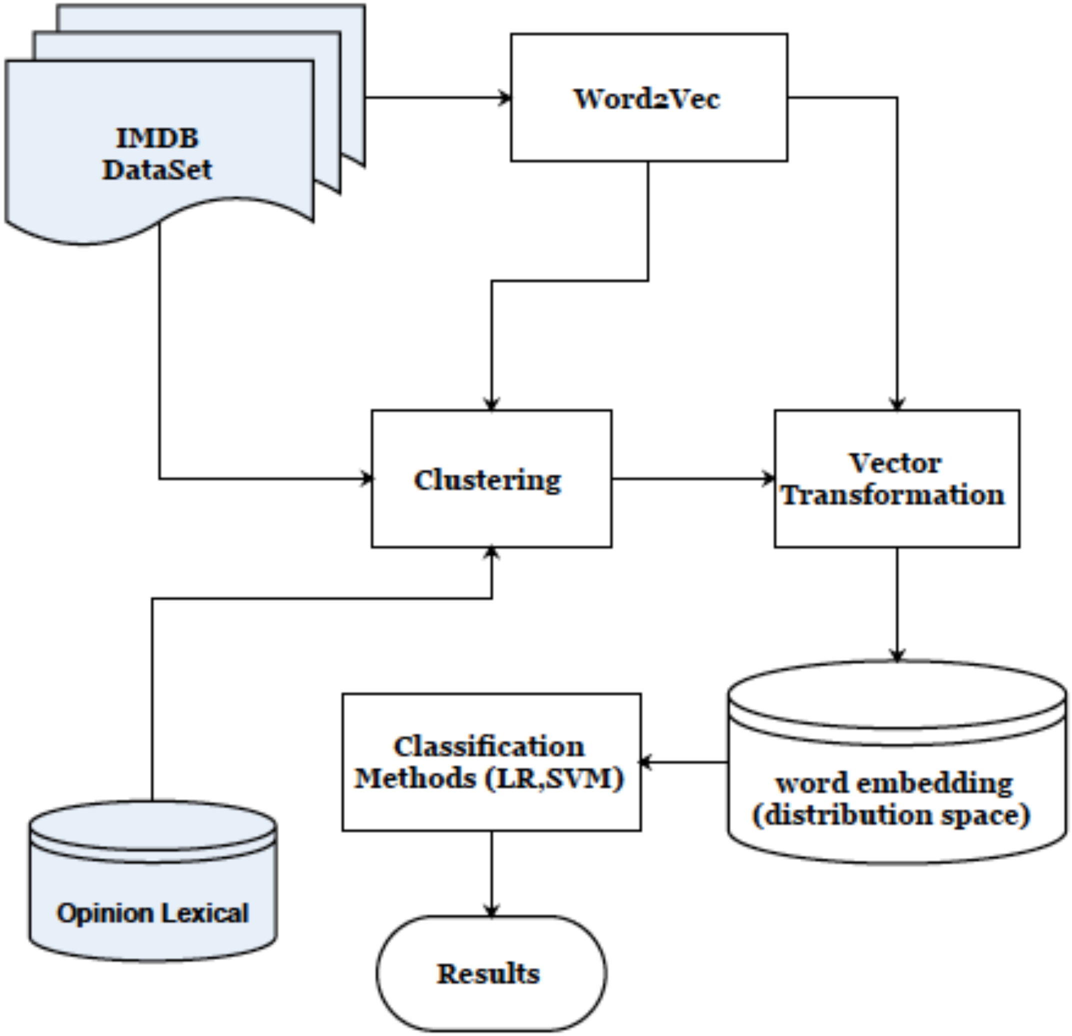
Introduction

- The **features** used in the classification of text for sentiment analysis plays an important role in its success.
- Several type of features have been investigated:
  - **Discrete distribution** such as LDA, LSA and bag-of-words (BoW).
  - **Continuous distribution** such as Word2Vec, Doc2Vec and other NN based approaches.

Problem Statement

- **Word2Vec** has been **effective** as features for sentiment analysis. However, its **high dimensionality** increases the **complexity** of the classifier.
- **Doc2Vec** reduces the complexity of the features BUT is not effective to deal with short sentences.

The Proposed Method



The Proposed Method (cont.)

1. Learning Word Representation based on Word2Vec
   - Using **Skip-gram** technique of the Word2Vec
   - The resulting vectors are **highly dimensional**

The Proposed Method (cont.)

2. Clustering of Term Vectors based on Sentiment Lexical Dictionary
   - Centroids:
     - a list of opinion words from a **sentiment lexical dictionary** (+ve :2005 & −ve:4783)
     - BUT, only those opinion words that are also exist in the Word2Vec vocabulary (almost 600 opinion words are ignored)

The Proposed Method (cont.)

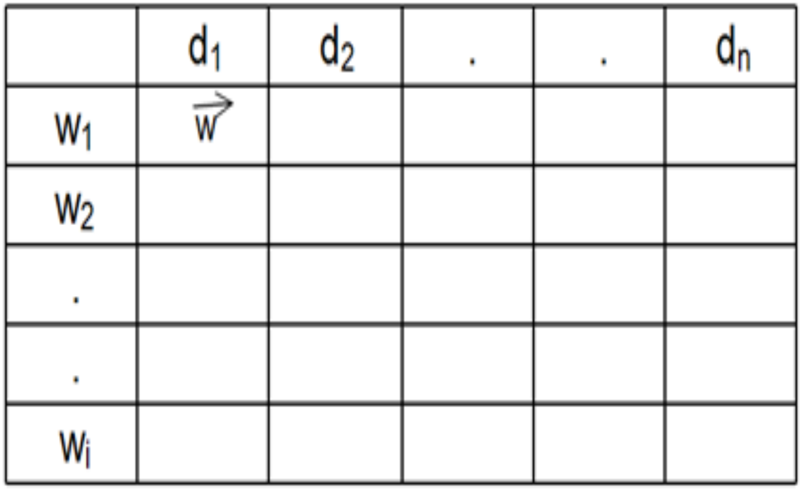
- Clustering:
  - ***cosine* similarity** between non opinion words with all centroids are measured.
  - Non opinion words are added to the cluster with the **most similar centroid**.
- Transformation:
  - For those terms belonging to the **negative clusters**, a **simple transformation** is applied to those vectors in order to separate the distribution in the space.
  - **NOTE**: it is later found that this step is **less significant** and can be **omitted**.

The Proposed Method (cont.)

3. Feature Extraction based on Polarity Clusters
   - The **dimension** of the vectors is based on the number of **clusters**.
   - For a given text, the terms appear in each cluster is observed. If the cluster contains the terms from the text, the **mean** of *cosine* similarity of those terms is used as the value of the vector. If not, the value is set to zero.

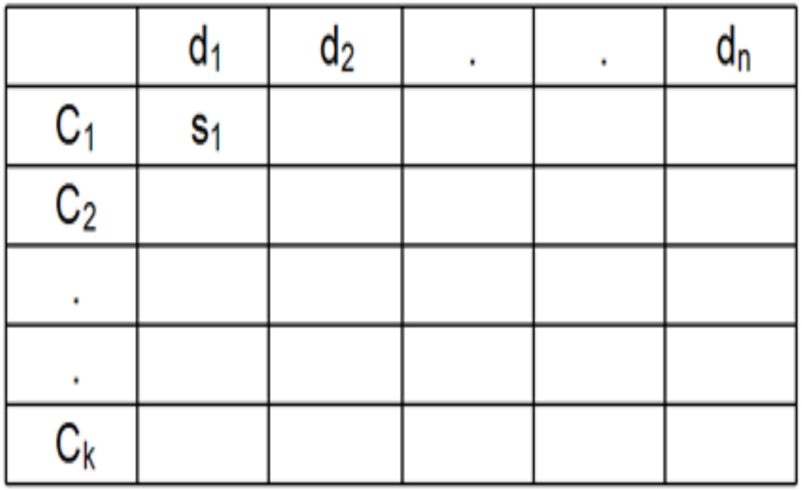
The Proposed Method (cont.)

- Matrix comparison



	d_1	d_2	.	.	d_n
w_1	$\vec{w}$				
w_2					
.					
.					
w_i					

i is number of vocabulary size



	d_1	d_2	.	.	d_n
C_1	s_1				
C_2					
.					
.					
C_k					

k is number of Cluster size

Results (Accuracy)

	Word2Vec	Doc2Vec	Bag-of-Words	Proposed Method
Logistic Regression (LR)	83.10	86.80 (+4.5%)	89.15 (+7.3%)	**93.80** **(+12.9%)**
Support Vector Machine (SVM)	70.25	86.20 (+22.7%)	83.60 (+19%)	**86.60** **(+23.3%)**

Conclusion

- The proposed method for feature extraction in sentiment analysis is more **effective** and **efficient** than the used of Word2Vec features.
- It is also more **effective** than other similar approaches.
- In future:
  - Optimization of parameters for classification

Selected References

- T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in Advances in neural information processing systems, 2013, pp. 3111–3119.
- T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” Nips, pp. 1–9, 2013.
- M. P. Villegas, M. Jose´, G. Ucelay, J. P. Ferna´ndez, M. A. A´ lvarez Carmona, M. L. Errecalde, and L. C. Cagnina, “Vector-based word representations for sentiment analysis: a comparative study,” pp. 785-793.
- A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, “Learning word vectors for sentiment analysis,” in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA: Association for Computational Linguistics, June 2011, pp. 142–150.

TERIMA KASIH / *THANK YOU*

www.upm.edu.my