ISSN 2347 - 3983
Volume 8. No. 7, July 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter79872020.pdf
https://doi.org/10.30534/ijeter/2020/79872020

Arabic Sentiment Analysis using Different Representation Models

Mohammed Bekkali[1], Abdelmonaime Lachkar[2]
[1] ENSA, USMBA, Fez, Morocco, bekkalimohammed@gmail.com
[2] ENSA, AEU, Tangier, Morocco, abdelmonaime_lachkar@yahoo.fr

ABSTRACT

Social network users generate a large number of reviews and comments, these reviews and comments express their opinions about on different topics. As a result, there is a great need to understand and classify these opinions. Sentiment analysis is a good way to overcome this problem. A Sentiment Analysis System (SAS) receives a text and returns the reflected polarity, that is, positive or negative; and sometimes the neutral polarity may be added. Several works have been developed on languages such as English, Spanish and even Chinese. Unfortunately, little attempt has been made for the Arabic language. In this paper, we propose to carry out a comparative study on four representation models in SAS, named Bag of Word (BoW), Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA). The four representation models have been tested on an Arabic sentiment analysis collected from Twitter and the obtained results show that the BoW representation outperforms the other models.

Key words : Sentiment Analysis, Arabic Language, BoW, LSA, PLSA, LDA.

1. INTRODUCTION

Sentiment analysis has been extensively studied in the past decade, due to the volume of publications on many social networks, providing a rich source of ideas and opinions about different topics. Given the huge amount of information shared in social network services to express opinions, analyzing the content of these social networks for Arabic and many other languages has become a necessity.

Among the major problems encountered in the sentiment analysis is shortness and sparseness. Because, these opinions are usually expressed as a short text, and they are different from traditional documents. As a result, they tend to be ambiguous without enough contextual information [1]. Thus, the text representation techniques play a vital role and may affect positively or negatively the accuracy of any Text Mining task, Sentiment Analysis in our case. The Arabic language makes these problems even more serious, because it is among the richest languages morphologically.

Sentiment analysis system can be considered as a Text Categorization (TC) system. TC is the task of assigning a text to a predefined category based on its content. Machine Learning is the tool that allows deciding whether a text belongs to a set of predefined categories [14] [15]. In the process of TC, the document must pass through a series of steps (Figure 1): transforming documents into brut text; removing the stop words, which are considered irrelevant word; and finally, all words must be stemmed. To represent the internal of each document, the document must have passed by the process consists of three phases [8]:
a) Defining the term set containing all the terms existing in the dataset;
b) Term selection, which is the process of selecting a subset of relevant terms without using any information;
c) Term weighting, which is the process of calculating a weight for each term selected in phase (b). The weight may be calculated using a weighting scheme such as TF-IDF, which combine the definition of term frequency and inverse document frequency.

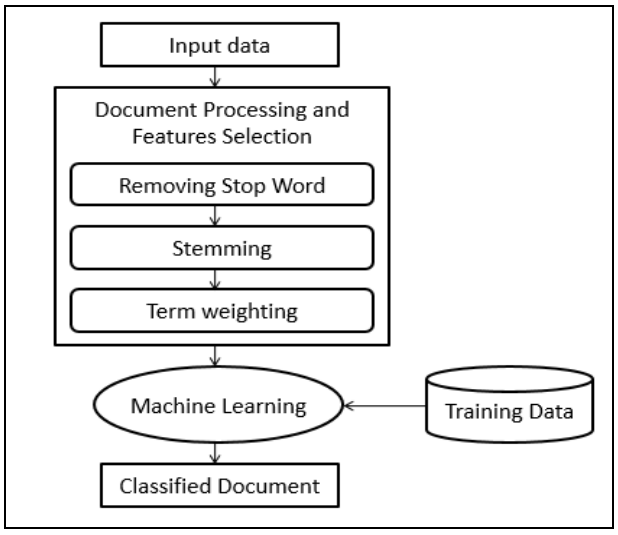

Figure. 1: Architecture of a Text Categorization System

Finally, the classifier is built by learning the characteristics of each category from a training set of documents. After building

of classifier, its effectiveness is tested by applying it to the test set and verifies the degree of correspondence between the obtained results and those encoded in the corpus.

Note that, one of the major problems in TC is the document's representation where we still limited only by the terms or words that occur in the document. In this paper, we propose to conduct a comparative study on representation models, to know the most adapted model for sentiment analysis in Arabic language. To this end, we present four sentiment analysis systems based on four representation models namely: BoW, LSA [2], PLSA [3], and LDA [4]. The obtained results show that, the most efficient representation model is BoW, given the best results and outperforms the other models.

The rest of this paper is organized as follows: we begin with a brief survey on related work about Arabic sentiment analysis in the next section. Section 3 gives the basic concepts of Bow, LSA, PLSA and LDA models. Section 4 describes our proposed system and leads the results of the experiments. Finally, section 5 concludes this paper and presents some perspectives.

2. RELATED WORKS

Users of the different social media platforms search for timely and social information. As in the rest of the world, users in Arab countries engage in social media applications for interacting and posting information, opinions, and ideas. As results, there is a great interest Arabic sentiment analysis and opinion mining currently.

Recently, Alahmary et al. proposed an efficient method using deep leaning in sentiment analysis of Saudi dialect. They have collected a corpus of 32063 tweets and applied two deep learning techniques: Long short-term memory (LSTM) and bidirectional long short-term memory (Bi-LSTM). The obtained results was very encouraging [5]. Alharbi and Khan introduced an investigation about the identification of comparative sentence from non-comparative ones in Arabic texts. They have used a dataset of Youtube comments and the obtained results were promising [6]. Gamal et al. presented an Arabic benchmark dataset for sentiment analysis showing the gathering methodology of the most recent tweets in different Arabic dialects. This dataset includes more than 151,000 different opinions in variant Arabic dialects, which labeled into two balanced classes, namely, positive and negative [7]. Bekkali and Lachkar proposed a sentiment analysis method based on conceptual representation, where they integrated LDA to overcome the coverage problem in Arabic semantic knowledge resources [10]. Alayba et al have used different machine learning algorithms, neural networks and featurs selection in order to construct a word2vec representation from a large Arabic dataset derived from different newspapers. They have tested their method on publicly available Arabic health sentiment dataset and the obtained results present a significant improvement [11]. Al-Azani and El-Alfy have presented a study where they compare different classifiers for opinion mining in highly imbalanced short text datasets; they have used features learned by word embedding. On the other side, they have tested the impact of using SMOTE (Synthetic Minority Over-Sampling Technique) on the training data. Their work has been tested on Arabic dialects dataset, and the obtained results rich a significant improvement [12]. Abdul-Mageed et al. present a Subjectivity and Sentiment Analysis (SSA) system for Arabic social media genres as one of the most morphologically complex languages. They present SAMAR, a sentence-level SSA system for Arabic social media texts. The advantage of this study is that the authors have created a multi-genre corpus (from four types of social media) of a text written in Standard Arabic (MSA) and Dialectal Arabic (DA) [13].

3. REPRESENTATION MODELS

In this section, we describe the different representation models, which will be used by our Arabic sentiment system later.

3.1 Bag of Word "BoW"

In the BoW representation, a vector containing the words of the initial document will represent each document. Thus, collection of documents can be represented by a matrix whose columns are the terms that appear at least once and the lines are the documents of this collection. A large number of researchers have chosen to use a vector representation in which each document is represented by a vector of n weighted terms. The n terms are simply the n different words that appear in the training set documents.

In the categorization of documents, we transform the document D_j into a vector $D_j = (w_{1j}, w_{2j}, ..., w_{1n})$, where n is the set of terms (descriptors) that appear at least once in the corpus. The weight w_{kj} corresponds to the contribution of the terms t_k to the semantics of the tweet D_j.

3.2 Latent Semantic Analysis

LSA is a method of Natural Language Processing (NLP), in the context of vector semantics. The LSA was patented in 1988 and published in 1990 [2].

The central point of LSA is a lexical table that contains the number of occurrences of each word in each document. To derive from a lexical table, the semantic relations between words, the simple analysis of raw co-occurrences comes up with a major problem. Even in a large body of opinions (which are considered as short texts), most of the words are relatively rare; and it seems that co-occurrences are even more rare. Their sparsity makes them particularly sensitive to random variations. LSA overcomes this problem by replacing the original frequency table with an approximation that

produces a kind of smoothing of the associations. To this end, the frequency table is decomposed into singular values before being recomposed from only a fraction of the information it contains. Linear combinations or 'semantic dimensions' on which the original words can be located thus replace the thousands of words characterizing the texts. Unlike the classical factor analysis, the extracted dimensions are very numerous (several hundreds) and not interpretable. They can, however, be seen as analogous to the frequently postulated semantic traits in describing the meaning of words.

LSA uses a matrix X whose lines correspond to the documents, the columns correspond to the terms and the components give the presence of the term (or rather, its importance) in each document. A single value decomposition is then performed on X, which gives two orthonormal matrices U and V and a diagonal matrix Σ (Figure 2). The values of Σ are the singular values of X.

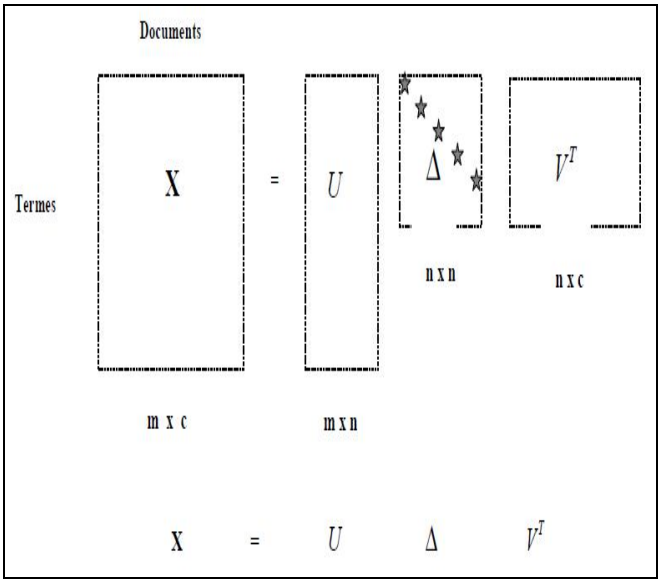

Figure. 2: Singular value decomposition of the term-document matrix

After having constructed the matrix of the occurrences, the LSA makes it possible to find a matrix of lower rank, which gives an approximation of this matrix of the occurrences.

3.3 Probabilistic Latent Semantic Analysis

PLSA is a method of automatic language processing inspired by latent semantic analysis [3]. It improves the latter by including a particular statistical model. It has been applied in filtering and information retrieval, natural language processing, machine learning and related fields.

Compared to the simple LSA, which derives from linear algebra to reduce the matrices of occurrences (by means of a singular value decomposition), the probabilistic approach uses a mixture of decompositions resulting from the analysis of the latent classes. This provides a more flexible approach, based on statistics.

Unlike BoW and LSA where a document is treated as a set of terms, PLSA aim to represent the document as a distribution of topics, and each topic as a distribution of terms.

3.4 Latent Dirichlet Allocation

LDA is a generative probabilistic model for a set of documents [4][16]. The central point of this model is that documents are represented as random mixtures over latent topics. Each topic is represented by a probability distribution over the terms. Each document is represented by a probability distribution over the topics.

LDA first samples a per-document multinomial distribution over topics from a Dirichlet distribution. Then it repeatedly samples a topic from this multinomial and samples a word from the topic. The topic discovered by LDA capture correlations among words. LDA defines a generative model for word by the following scheme (Figure 3) [4]:

- Pick a latent topic z with probability $p(z|\theta)$, where $p(z|\theta)$ denotes probability of the topic z from a multinomial distribution with parameter vector θ.
- Generate a word with probability $p(wt|z,\beta)$, where $p(wt|z,\beta)$ denotes the topic-conditional probability of a specific word wt conditioned on the unobserved topic variable z with a multinomial distribution parameter β
- Pick a multinomial distribution β for each topic z from a Dirichlet distribution $p(\beta|\eta)$ with parameter η.
- Pick multinomial distribution θd for document d from a Dirichlet distribution $P(\theta d|\alpha)$ with parameter α.

Thus, the likelihood of generating a corpus D, whose vocabulary size is V, is

$$p(D) = \prod d\in D \{ \int p(\theta d|\alpha)\, p(\beta|\eta) \prod t=1,V \sum k=1,K\, p(zt = k|\theta d)p(wt|zt = k,\beta)\, d\theta d\, d\eta \}$$

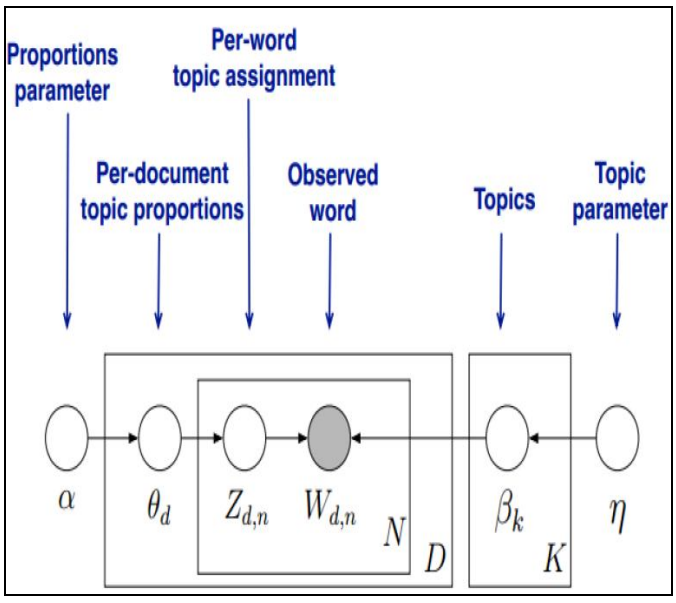

Figure 3: Graphic model of a typical generative model LDA

In this section, we have presented the basics of the four representation models: BoW, LSA, PLSA and LDA, which we are going to use in our Arabic sentiment analysis system. In the next section, we describe our experiments and the obtained results down to the smallest detail.

4. ARABIC SENTIMENTS ANALYSIS

In this section, we present the obtained results of our Arabic sentiment analysis system (Figure 4). The system includes two main steps: preparing the opinions and choosing the representation model to use.

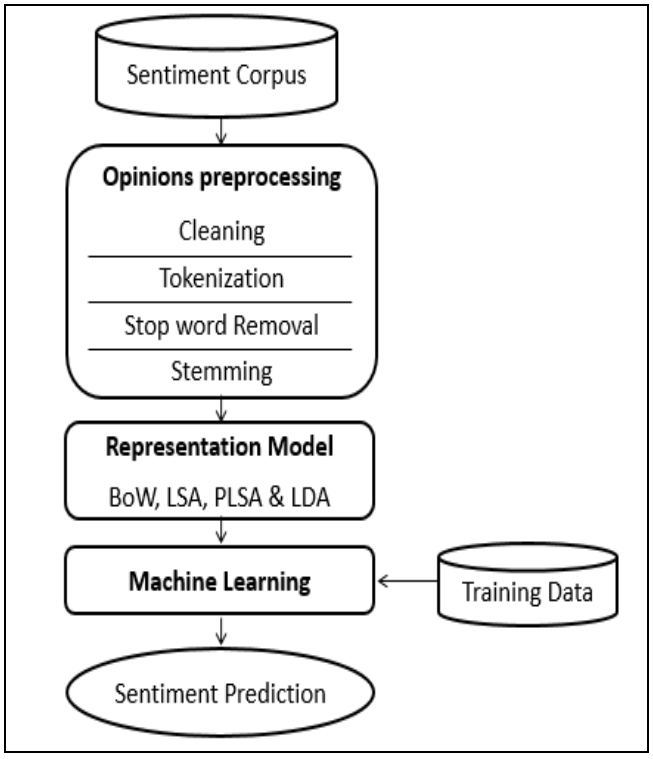

Figure 4: Arabic Sentiment Analysis System

In the first step, each opinions in the corpus will be cleaned by removing Arabic stop words, Latin words, and special characters like (/, #, $, etc.). After that, a stemmer algorithm may be used to generate the root/stem and eliminate the redundancy. Then in the second step, one of the representation models previously presented (BoW, LSA, PLSA and LDA) is chosen, to generate a corresponding corpus to each model.

Dataset Description

The opinions used in our experiments are collected from Twitter by the NodeXL Excel Template which is a freely Excel template that makes it super easier to collect Twitter network data [9]. The corpus contains 7086 marked tweets (3092 negative tweets and 3994 positive tweets) on various topics such as politics and art, …etc. These tweets include opinions in modern standard Arabic.

The Obtained Results

To assess the performance of the proposed system, a series of experiments has been conducted. The effectiveness of our system has been evaluated and compared in terms of the F1-measure using the Naïve Bayesian (NB) and Support Vector Machine (SVM) classifiers.
F1-measure can be calculated using Precision and Recall measures as follows:

$$F1\text{-measure} = 2 * ((Precision * Recall) / (Precision + Recall))$$

Precision and Recall both are defined, respectively, as follows:

$$Precision = TP / ((TP+FP))$$
$$Recall = TP / ((TP+FN))$$

where:

- True Positive (TP) refers to the set of tweets which are correctly assigned to the given category.
- False Positive (FP) refers to the set of tweets which are incorrectly assigned to the category.
- False Negative (FN) refers to the set of tweets which are incorrectly not assigned to the category.

The chosen technique for validation is cross validation over a 10-fold, considering the number of tweets in our collection; and for the model to learn as much as possible in the learning phase. The obtained results for Arabic sentiment analysis using Bow, LSA, and PLSA models are presented in table 1. Moreover, we want to test how changing the number of topics can affect the sentiment analysis results using LDA as a representation model. To this end, table 2 presents the obtained results for Arabic sentiment analysis using different numbers of topics (10, 20, 30, 40, 50 and 60).

Table 1: The obtained results for Arabic sentiment analysis using BoW, LSA and PLSA

	NB	SVM
BoW	0.939	0.972
LSA	0.662	0.67
PLSA	0.846	0.868

Table 2: The obtained results for Arabic sentiments analysis using LDA

	NB	SVM
10 Topics	0.715	0.743
20 Topics	0.734	0.719
30 Topics	0.779	0.755
40 Topics	0.795	0.81
50 Topics	0.809	0.826
60 Topics	0.802	0.834

We notice that BoW gives the best results compared to the other models, despite the high dimension of the matrix that represents this model. The reason behind this performance is that in sentiment analysis always we talk about senti-words that reflect the polarity in a tweet. Therefore, because of the volume of terms contained in the representative matrix of this model, it give remarkable results up to 97% accuracy.

LSA gives very poor results compared to other models. It can be concluded that LSA does not deal correctly with short text. Although, PLSA really improves the results of LSA while replacing the text / words matrix by text / topic. Likewise, for LDA, the representation according to the topics allows to improve the obtained results for Arabic sentiment analysis while enlarging the number of topics.

5. CONCLUSION

With the explosion use of social networks all over the world, these users express their opinions about different products and services. As a result, there is a great need to analyze and classify these opinions. Sentiment analysis or opinion mining is relatively a new field of research for the Arabic language compared to English or some other European languages. In this paper, we implemented four systems based on the four representation models, namely BoW, LSA, PLSA and LDA. The obtained results showed that the most efficient system is the one that uses BoW as representation model, also PLSA and LDA.

The future work will consist of testing our systems with different corpora from different domains, and using these results to attack other lines of research such as recommendation systems, detection of communities as well as security on the Internet.

REFERENCES

1. Chen, M., Jin, X., Shen, D**. Short text classification improved by learning multigranularity topics.** In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence-Volume, vol. 3, pp. 1776–1781. AAAI Press (2011)
2. Scott Deerwester, Susan Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman**. Indexing by Latent Semantic Analysis.** Journal of the Society for Information Science, vol. 41,no 6, 1990, p. 391-407
3. Thomas Hofmann**. Probabilistic Latent Semantic Indexing.** Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99), 1999 https://doi.org/10.1145/312624.312649
4. D. M. Blei, A. Y. Ng, and M. I. Jordan. **Latent dirichlet allocation**. Journal of Machine Learning Research, 3:993–1022, 2003
5. Alahmary RM, Al-Dossari HZ, Emam AZ (2019)**. Sentiment analysis of Saudi dialect using deep learning techniques.** In: 2019 international conference on electronics, information, and communication (ICEIC). IEEE, pp 1–6
6. Alharbi, F.R. & Khan, M.B**. Identifying comparative opinions in Arabic text in social media using machine learning techniques.** SN Appl. Sci. (2019) 1: 213. https://doi.org/10.1007/s42452-019-0183-3
7. Gamal D, Alfonse M, El-Horbaty ESM, Salem ABM (2019)**. Twitter benchmark dataset for Arabic sentiment analysis.** Int J Mod Educ Comput Sci 11(1):33 https://doi.org/10.5815/ijmecs.2019.01.04
8. Sebastiani F. "**A Tutorial on Automated Text Categorisation**". Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence. PP 7-35. 1999.
9. Lamberson PJ. Collecting and visualizing twitter network data with NodeXl and Gephi. http://social-dynamics.org/twitter-network data/. Accessed Dec 2019
10. Mohammed Bekkali, Abdelmonaime Lachkar. (2019)**. Arabic Sentiment Analysis based on Topic Modeling. Analysis based on Topic Modeling.** In Proceedings of ACM SMC conference (SMC '19). ACM, Kenitra, Morocco https://doi.org/10.1145/3314074.3314091
11. Alayba AM, Palade V, England M, Iqbal R (2018)**. Improving sentiment analysis in Arabic using word representation.** In: 2018 IEEE 2nd international workshop on Arabic and derived script analysis and recognition (ASAR). IEEE, pp 13–18
12. Al-Azani S, El-Alfy ESM (2017)**. Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short Arabic text.** Procedia Comput Sci 109:359–366 https://doi.org/10.1016/j.procs.2017.05.365
13. Abdul-Mageed, M., Diab, M., & Kübler, S. (2014)**. SAMAR: Subjectivity and sentiment analysis for Arabic social media.** Computer Speech & Language, 28(1), 20–37
14. Abdul Karim Mohamad, Mailasan Jayakrishnan, Nurnajwa Hazwani Nawi (2020)**. Classification of Twitter Data by Sentiment Analysis in the Malay Language.** Int. J. Emerg. Trends Eng. Res., vol. 8, no. 6, June 2020. https://doi.org/10.30534/ijeter/2020/83862020
15. V. Shailaja**. Predictive Analytics of Performance of India in the Olympics using Machine Learning Algorithms.** Int. J. Emerg. Trends Eng. Res., vol. 8, no. 5, May 2020. https://doi.org/10.30534/ijeter/2020/57852020
16. John Paul P. Miranda, Jonar T. Martin**. Topic Modeling and Sentiment Analysis of Martial Arts Learning Textual Feedback on YouTube.** Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, no. 3, June 2020. https://doi.org/10.30534/ijatcse/2020/35932020