Application of Latent Dirichlet Allocation (LDA) for clustering financial tweets

SIFI Fatima-Zahrae[*], *SABBAR Wafae*[1], and *ELMZABI Amal*[2]

[1]Laboratoire Informatique de Mohammedia, Faculté des Sciences et Techniques Mohammedia, Maroc
[2]Laboratoire Performance Economique et Logistique, Faculté des Sciences Juridiques, Economiques et Sociales Mohammedia, Maroc

Abstract. Sentiment classification is one of the hottest research areas among the Natural Language Processing (NLP) topics. While it aims to detect sentiment polarity and classification of the given opinion, requires a large number of aspect extractions. However, extracting aspect takes human effort and long time. To reduce this, Latent Dirichlet Allocation (LDA) method have come out recently to deal with this issue.In this paper, an efficient preprocessing method for sentiment classification is presented and will be used for analyzing user's comments on Twitter social network. For this purpose, different text preprocessing techniques have been used on the dataset to achieve an acceptable standard text. Latent Dirichlet Allocation has been applied on the obtained data after this fast and accurate preprocessing phase. The implementation of different sentiment analysis methods and the results of these implementations have been compared and evaluated. The experimental results show that the combined uses of the preprocessing method of this paper and Latent Dirichlet Allocation have an acceptable results compared to other basic methods.

[*] fatimazahraesifi@gmail.com

1 Introduction

With the emergence and rapid development of Web 2.0, moreand more people begin to express their feelings, opinion andattitude over Internet, which increase the amount of usergeneratedreviews containing rich opinion and sentimentinformation. Sentiment analysis (SA), as a technique to automatic detection of opinionsembodied in text, is becoming a hotspot in many research fields, including natural language processing (NLP), with a number ofapplications including recommender and advertising systems,product feedback analysis and customer decision making.

However, one big problem is to find aspects that users evaluate in reviews. From the perspective of a user reading the reviews to get information about a product, the evaluations of the specific aspects are just as important as the overall rating of the product. Although sometimes the aspect information is available, it isunlikely to be a comprehensive set of all aspects that areevaluated in the reviews. Another important task in reviewanalysis is discovering how opinions and sentiments for different aspects are expressed. 'The company is in financial **difficulties**', and 'She organizes her financial affairs very **efficiently**'. These are sentiment words at the level of theaspect. We tackle these two problems at once with a unified generative model of aspect and sentiment.

The main intervention in this work can be listed as follows:

- Normalization, Removing stop words and Tokenization are three steps of pre-processing used to increase the quality of classification algorithms.
- Specialized and up-to-date libraries of Python programming language are also exploited and used in implementation of sentiment classification.
- The effectiveness of LDA algorithms will be applicated and evaluated in order to classify opinions of users on the Twitter social network.

The remaining sections of this paper are organized as follows. A literature review of commonand well-known methods and approaches of LDA is represented in section 2. The proposed methodology of this paper is introduced in section 3.
The experimental results and their analysis are represented in section 4, and finally, the work is concluded in finalsection 5.

2 Related Work

There are a lot of methods [1] [2] [3], to extract user reviews from the tweets in which different aspects are detected and reviews are grouped according to these aspects. For extracting product aspects in aspect based sentiment analysis, different studies were done. Also, same features of a product can be expressed with different words, these can be synonyms or not. For instance, coin and cash have same meaning for money although appearance and design do not have the same meaning, they can be used for same aspects of the product. So for effective summary, extracted aspect words should be grouped. However grouping manually is time consuming and difficult since there are so many feature expressions in a text corpus [4]. Topic modelling approaches as a clustering algorithm can be used for this purpose. One of the most popular topic modelling methods is LDA.

The LDA model is suitable for the following reasons: First, it provides an unsupervised way of discovering topics from documents (or aspects from reviews), and second, it results in language models that explain how much a word is related to each topic and possibly to a sentiment. We take LDA model and adapt it to match the granularity of the discovered financial topic to the details of the reviews. In this paper, we provide a survey study of set techniques about the application of LDA in financial domain and raising the results of experimental demonstration.

Weng et al. combined Twitter messages posted by the same user [5]. Nimala et al. did hashtag- based tweet aggregation strategy in their study [6]. Besides, there are many other studies for the adaptation of LDA to use on short texts. Lin et al. proposed the Joint Sentiment/Topic Model (JST) for aspect and sentiment extraction from user reviews [7].

Titov and McDonal [8] proposed the Multi-grain LatentDirichlet Allocation model (MG-LDA), since they are all based on the state-of-the-art topicmodel LDA. MG-LDA is argued to be more appropriate to build topics that are representative of table aspects of objects from eithera global topic or a local topic. Titov and McDonald [9]further proposed the Multi-Aspect Sentiment Model (MAS) byextending the MG-LDA framework. The major improvement ofMAS is that it can aggregate sentiment texts for the sentimentsummary of each rating aspect extracted from the MG-LDA.

In order to model document sentiments, Lin et al. added an additional sentiment layer to LDA between the document and the topic layer. Jo et al. proposed Sentence-LDA and Aspect Sentiment Unification Model (ASUM), which is a similar method to JST [10]. They used electronicdevices and restaurant reviews datasets to evaluate Sentence-LDA and ASUM.
García-Pablos et al. proposed W2VLDA method for aspect extraction and sentiment polarity detection on user reviews [11]. They used LDA algorithm in combination with continuous word embeddings, word2vec and maximum entropy classifier.

3 Latent Dirichlet Allocation:

Latent Dirichlet Allocation is an algorithm that has the ability to classify texts without having label value. It takes up a large amount of unsupervised text and classifies them. This is basically the majority of text handling, since naturally the texts that will come to the algorithm will be untitled.

The LDA form is an abbreviation of words Latent Dirichlet Allocation Topic or sometimes called Topic Modeling, which is modelling subjects. We should avoid confusing it with Linear Discriminant Analysis which has nothing with NLP.

LDA method was on the thought of the German athlete John Dirichlet [12], and there is a statistical mathematical distribution in his name named Dirichlet Distribution, which was invented recently as 2003, while he was in the 19th century.

Before dealing with the LDA, we need know the basics:

- There are so-called topics latent which means the underlying topics within a particular article, an assigned article may contain 10 underlying topics.
- Similar articles use similar words, most articles on economics use similar words.
- There is a natural distribution of underlying topics, within the full article.

For example, the figure 1 below shows that the whole subject type 2 has the largest ratio: We have five subjects in document 1, the highest value belong to topic type 2 with a value of eight.

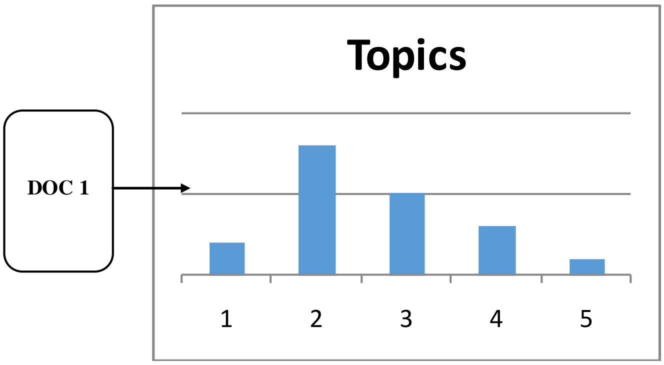


Fig.1. Proportion of topics according to subject in doc 1

In the second figure 2, we talk about document 2. In this example, the fourth topic type has the largest percentage, with a value of nine.

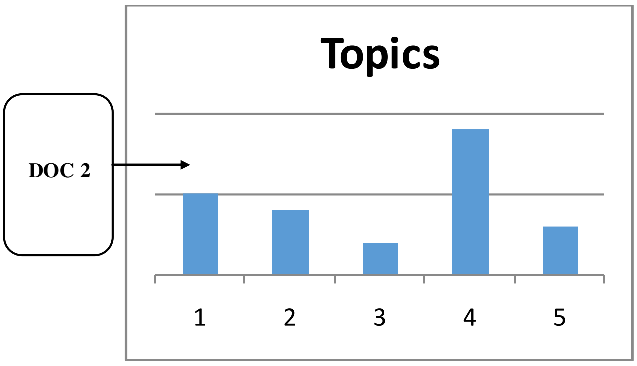


Fig.2.Proportion of topics according to subject in doc 2

Thus by reviewing the proportions of the topics within a particular article, we can conclude the type of the whole article.

The following figure 3 is about topic 1 of document 2, we can see that most words used are price, trading, million; we conclude that this article is about finance.

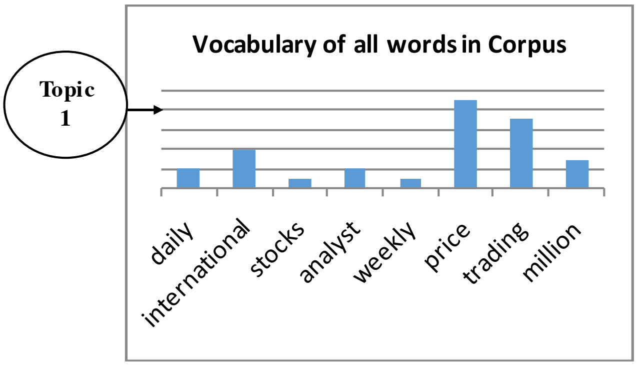


Fig.3. Words Proportion in topic 1 of the corpus.

Thus by reviewing the words proportions in a particular topic of the corpus, we can also conclude the type of the whole article.

The steps that the LDA does are as follows:

- LDA get into the article and examines the topics, then the words.
- Determine the total number of words in N topics.
- Know the distribution of proportions of topics in the entire article, it can be for example: 60 %business, 20 % Policy, this is for the highest number of subjects, for example 5 is the higher number.
- Then on the basis of distribution ratios, LDA does a clustering in an unsupervised manner, so that by dividing specific articles or sentences, for a number of sections those are given.

4 LDA implementation for clustering financial tweets:

In this section, we present the steps and details of our proposed work. First, we gathered financial data and applied preprocessing on them. We also applied manual labelling only for the methods that use them. After transforming text data into a set of feature vectors, LDA method is applied to determine the clusters type, as discussed later. Finally, the experimental result of the LDA application on financial data about clustering the product aspects is evaluated.

4.1 .Preprocessing

Applying text preprocessing before analyzing the tweets is very important for achieving good results. The purpose of preprocessing is to clarify the input data for next thorough analysis.

In this work, we exploit three methods of Natural Language Preprocessing, namely normalization, removing stop words, and tokenization, to make a standard dataset.

Normalization: The texts of the dataset should be normalized initially to be used in the next steps of opinion mining process. Even when some tweets are very similar, they may be considered as different opinions by sentiment analysis system, due to some superficial differences. Therefore, the proposed method attempts to remove such differences. To achieve this goal, normalization is conducted on opinions, before comparing their texts. This normalization leads to obtain more reliable results in comparison of the tweets. Removing accent, removing blank spaces, removing punctuation marks, and removing watermarks are the main operations conducted in the normalization phase of preprocessing. In every new load, the comments are normalized, and saved in the preprocessed data folder.

Removing stop words: Despite the frequent use of stop words, they are pragmatically insignificant. Although it is thought that only the linking words are stop words, many verbs, auxiliary verbs, nouns, adverbs, and adjectives can be stop words, too. In most text mining operations, the processing result is significantly improved by eliminating such words.

Tokenization: Opinion mining tools need to analyze their input tweets, lexically. In the process of lexical analysis, a sentence is converted into a sequence of tokens, which are meaningful parts of the text. This process is called tokenization and leads to the generation of a collection of independent meaningful parts, called tokens.

4.2 Dataset

In this study a sample dataset containing 1536 records of financial tweets is used (github.com). The goal is to explore the dataset and to understand what are the kinds of tweets in this dataset.

To do that, we used normalization, removing stop words and tokenization to clean and cluster the data. We also used our general knowledge of financial markets to analyze in human words the model's output.

This data is labeled completely. The dataset in this work is a collection of financial tweets in English language. The proposed method works on English language but with some little revisions, it can work on other languages as well.

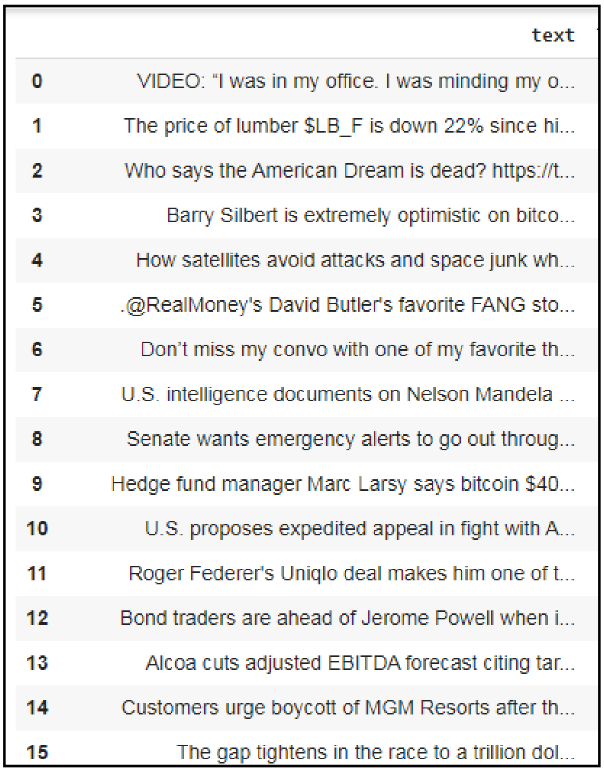
4.3 Experimental results:

First, the code in Python starts by reading the data file whichis a huge file. About 25996 tweets, every tweet speaks of a particular topic, and there is similarity in topics, and we want to divide them in a number of homogeneous sections according to the presence of similar words.

First, we read articles on scikit-learn library to get word matrices. It is used to transform the given text into a vector on the basis of the frequency of each word that occurs in the entire text.

It preferred to define the parameters values during the work, which are: Maximum Document Frequency, maximum of a particular word presence, as well as Minimun Document Frequency until the exclusion of abnormal words.

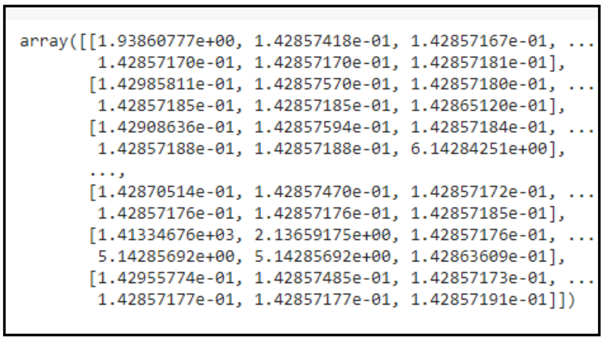
https://doi.org/10.1051/e3sconf/202129701071



	text
0	VIDEO: "I was in my office. I was minding my o...
1	The price of lumber $LB_F is down 22% since hi...
2	Who says the American Dream is dead? https://t...
3	Barry Silbert is extremely optimistic on bitco...
4	How satellites avoid attacks and space junk wh...
5	.@RealMoney's David Butler's favorite FANG sto...
6	Don't miss my convo with one of my favorite th...
7	U.S. intelligence documents on Nelson Mandela ...
8	Senate wants emergency alerts to go out throug...
9	Hedge fund manager Marc Larsy says bitcoin $40...
10	U.S. proposes expedited appeal in fight with A...
11	Roger Federer's Uniqlo deal makes him one of t...
12	Bond traders are ahead of Jerome Powell when i...
13	Alcoa cuts adjusted EBITDA forecast citing tar...
14	Customers urge boycott of MGM Resorts after th...
15	The gap tightens in the race to a trillion dol...

Cap.1.Text transformation to vecteurs

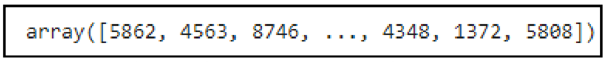
The selected words are viewed in the sum of all the words in 12764 tweets, with the repetition omitted. A number of words can be browse randomly. After, we divide the clusters or topics into seven sections and present the attribute components as a matrix of 7 rows, and 12764 columns, where each row is part of the cluster, and each column is the values of existing words.



```
array([[1.93860777e+00, 1.42857418e-01, 1.42857167e-01, ...,
         1.42857170e-01, 1.42857170e-01, 1.42857181e-01],
        [1.42985811e-01, 1.42857570e-01, 1.42857180e-01, ...,
         1.42857185e-01, 1.42857185e-01, 1.42865120e-01],
        [1.42908636e-01, 1.42857594e-01, 1.42857184e-01, ...,
         1.42857188e-01, 1.42857188e-01, 6.14284251e+00],
        ...,
        [1.42870514e-01, 1.42857470e-01, 1.42857172e-01, ...,
         1.42857176e-01, 1.42857176e-01, 1.42857185e-01],
        [1.41334676e+03, 2.13659175e+00, 1.42857176e-01, ...,
         5.14285692e+00, 5.14285692e+00, 1.42863609e-01],
        [1.42955774e-01, 1.42857485e-01, 1.42857173e-01, ...,
         1.42857177e-01, 1.42857177e-01, 1.42857191e-01]])
```

Cap.2.The attributes components representation

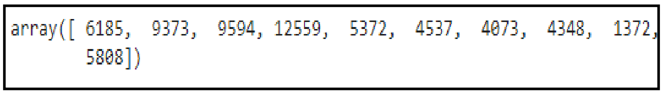
For example, we applied the Argsort command for the first topic, which shows in order the Index values for the words less prevalence and the most spread one,.
In our work, the word 5862 is less likely, and the word 5808 is the most widespread.



```
array([5862, 4563, 8746, ..., 4348, 1372, 5808])
```

Cap.3.Exemple of values from less prevalence to most spread

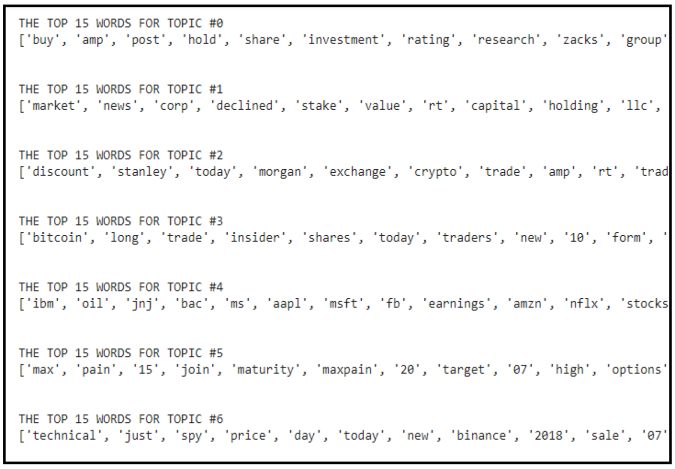
If we want to get the spread 10 words in the first cluster, we show them through the use of get feature names.



```
array([ 6185,  9373,  9594, 12559,  5372,  4537,  4073,  4348,  1372,
        5808])
```

Cap. 4.Exemple of 10 spread words in the first cluster

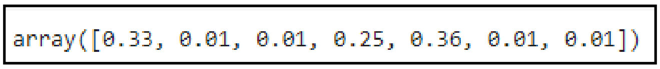
It shows clearly that the first cluster is purchase. We can view the most important words in all clusters.



```
THE TOP 15 WORDS FOR TOPIC #0
['buy', 'amp', 'post', 'hold', 'share', 'investment', 'rating', 'research', 'zacks', 'group'


THE TOP 15 WORDS FOR TOPIC #1
['market', 'news', 'corp', 'declined', 'stake', 'value', 'rt', 'capital', 'holding', 'llc',


THE TOP 15 WORDS FOR TOPIC #2
['discount', 'stanley', 'today', 'morgan', 'exchange', 'crypto', 'trade', 'amp', 'rt', 'trad


THE TOP 15 WORDS FOR TOPIC #3
['bitcoin', 'long', 'trade', 'insider', 'shares', 'today', 'traders', 'new', '10', 'form', '


THE TOP 15 WORDS FOR TOPIC #4
['ibm', 'oil', 'jnj', 'bac', 'ms', 'aapl', 'msft', 'fb', 'earnings', 'amzn', 'nflx', 'stocks


THE TOP 15 WORDS FOR TOPIC #5
['max', 'pain', '15', 'join', 'maturity', 'maxpain', '20', 'target', '07', 'high', 'options'


THE TOP 15 WORDS FOR TOPIC #6
['technical', 'just', 'spy', 'price', 'day', 'today', 'new', 'binance', '2018', 'sale', '07'
```

Cap. 5. Most important words in all clusters

Now, we apply the partition on all 25995 sentences, via LDATransform. The result of the transformer is a matrix of 25995 rows in 7 columns and the percentage of each section is considered to be the proportion of this text. For example, the first text is 33% from the second section and 36% in the fifth section.

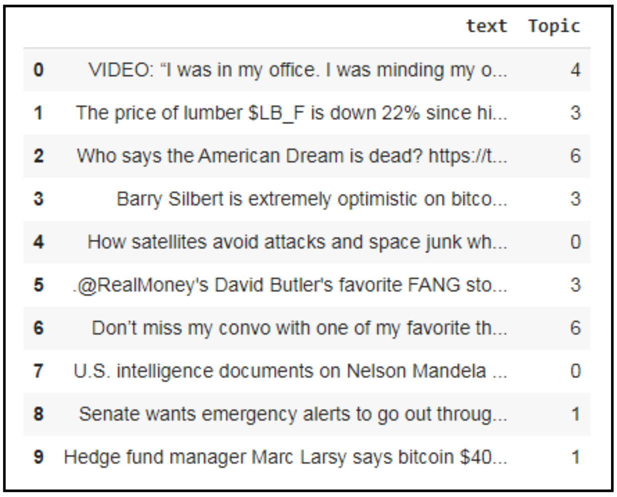


```
array([0.33, 0.01, 0.01, 0.25, 0.36, 0.01, 0.01])
```

Cap. 6. The percentage of each topic in the text

After we determine the order of the biggest value, and give the full matrix of this work.

In the end, a new column can be added to the basic table.



	text	Topic
0	VIDEO: "I was in my office. I was minding my o...	4
1	The price of lumber $LB_F is down 22% since hi...	3
2	Who says the American Dream is dead? https://t...	6
3	Barry Silbert is extremely optimistic on bitco...	3
4	How satellites avoid attacks and space junk wh...	0
5	.@RealMoney's David Butler's favorite FANG sto...	3
6	Don't miss my convo with one of my favorite th...	6
7	U.S. intelligence documents on Nelson Mandela ...	0
8	Senate wants emergency alerts to go out throug...	1
9	Hedge fund manager Marc Larsy says bitcoin $40...	1

Cap. 7. Text transformation according to Topics

5 Conclusion

In this work, we have described LDA, a topic modeling based method for aspect clustering for collection of data. LDA is based on a simple exchangeability assumption for the words and topics in a document. The dataset in this work is a collection of financial reviews in English language. This application works on English language with some little revisions, it can works on other languages as well. The experimental results show the application of LDA in financial domain is quite successful in clustering the product aspects.

Some improvements can be made on this work. We can also consider partially exchangeable models in which we condition on exogenous variables. Thus, for example, the topic distribution could be conditioned on features such as "paragraph" or "sentence," providing a more powerful text model that makes use of information obtained from a parser.

References

1. Lin F, Xiahou J, Xu Z. TCM clinic records data mining approaches based on weighted-LDA and multi-relationship LDA model. Multimed Tools Appl 75:14203–14232, (2016).
2. A. Etzioni. Extracting product features and opinions from reviews. In the Conference on Empirical Methods in Natural Language Processing (EMNLP).
3. S. H. L. L. Zhang, B.L. and E. O'Brien-Strain. Extracting and ranking product features in opinion documents. In the 23rd International Conferenceon Computational Linguistics: Posters, COLING '10.
4. H. X. P. J. ZhongwuZhai, B.L. Clustering product features for opinionmining. WSDM (2011).
5. Weng, J., Lim, E., Jiang, J., & He, Q. Twitterrank: finding topic-sensitive influential twitterers. In Proceedings of the third ACM international conference on Web search and data mining (WSDM '10), pp. 261–270 (2010).
6. Nimala, K., Magesh, S., &Arasan, R. T. Hash tag based topic modelling techniques for twitter by tweet aggregation strategy. Journal of Advanced Research in Dynamical and Control Systems, 3, 571–578 (2018).
7. Lin, C., & He, Y. Joint sentiment/topic model for sentiment analysis. In Proceedings of the 18th ACM conference on Information and knowledge management (CIKM '09), pp. 375–384 (2009).
8. I. Titov, R. McDonald, Modeling online reviews with multi-grain topic models, in: Proceeding of the 17th International Conference on World Wide Web.Publishing, pp. 111–120, Beijing, China, (2008).
9. I. Titov, R. McDonald, A Joint Model of Text and Aspect Ratings for Sentiment Summarization, ACL'08. Publishing, pp. 308–316 (2008),
10. Jo, Y., & Oh, A. Aspect and sentiment unification model for online review analysis. In Proceedings of the fourth ACM international conference on Web search and data mining (WSDM '11), pp. 815–824 (2011).
11. García-Pablos, A. Cuadros, M., &Rigau, G. W2VLDA: Almost unsupervised system for aspect based sentiment analysis. Expert Systems with Applications, 91, pp.127–137, (2018).
12. Blei, D. M., Ng, A. Y., & Jordan, M. I. Latent dirichlet allocation. The Journal of Machine Learning Research, 3, pp.993–1022 (2003).