ISSN 2347 - 3983
Volume 8. No. 8, August 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter06882020.pdf
https://doi.org/10.30534/ijeter/2020/06882020

A Comparative Study for Sentiment Analysis: LDA and LDA2Vec

Prerna Mishra[1], Ranjana Rajnish[2], Pankaj Kumar[3]
[1]Amity Institute of Information Technology, Amity University Uttar Pradesh, Lucknow Campus, Lucknow (India), India, prerna21.mishra@gmail.com
[2] Amity Institute of Information Technology, Amity University Uttar Pradesh, Lucknow Campus, Lucknow (India), India, rrajnish @lko.amity.edu
[3] Department of Computer Science, Sri Ramswaroop College of Engineering & Technology, Lucknow, India, pk79jan@gmail.com

ABSTRACT

Internet technology developed a new era of data mining by sharing their reviews, thoughts and ideas regarding any topic. Web 2.0 emerged as a new generation of Internet Services and created various social medium like Twitter, Facebook and creative blogs a data warehouse for exchanging and expressing their views. Nowadays enormous amount of data is accessible linked to any topic whether it will be Political agenda, Business enterprise, Survey organizations or different advertising firms. Customer contentment is the major component for various survey companies so as to improve their services. LDA2Vec is new approach proposed by Chris Moody and much research related to this approach is still not done. In this paper we used Latent Dirichet allocation (LDA) and LDA2Vec model for Sentiment classification. We evaluated performance of both the models by using corpus of 1000 records. After execution of model, we can easily demonstrate by our experimental results that hybrid approach of LDA2Vec (LDA and Word2Vec) performed better in comparison to LDA approach.

Key words: Latent Dirichlet Allocation, LDA2Vec, Sentiment Analysis.

1. INTRODUCTION

Sentiment Analysis is a classification problem which helps in mining or fetching sentiments/views or opinions of various end users regarding any topic. Various social networking sites like Facebook, Twitter etc are the sources of sharing the people's sentiments, thoughts and ideas which help in different aspects like from the government perspective it helps in improving their governance policies, understand their views and attitude etc. Internet has huge amount of data and thus sentiment analysis is a great topic of interest among researchers and Data Scientist [1]. There is almost 80% of data available on Internet which is unstructured and irrelevant. Sentiment analysis is a sub-field of data mining which helps in converting this unstructured data to structured data that is data which we can use for getting meaningful information. Sentiment analysis is process involved in correct brainstorming done on any raw data which further leads to understand the mood of the audience/customers [2]. Most of the unstructured data is hidden in various data patterns and fetched meaningful information whether it is positive or negative [3].

Topic modelling is a methodology for fetching structured data in field of text mining. It automatically identifies the various topics present in a text and helps in better decision making. The LDA algorithm (Blei et al.,2003) is one of the best successful topic models in a documents collection. LDA presumes that a document is a collection of various topics. LDA2Vec is hybrid approach of LDA and Word2Vec approach implemented by Chris Moody[4].LDA2Vec implements both words and topics into single framework. It improves the quality of topic modelling up to some extent in comparison to LDA.

In this paper we used the corpus of 1000 records of a product company which classifies the sentiments according to different categories using LDA and LDA2Vec. Some of the topic categories used for this Survey-Response is Customers, Products/Services and Technology.

This paper is divided into 6 Sections. In Section 2 we discussed about the related work done in field of sentiment analysis using LDA and Hybrid approach of LDA and Word2Vec. Section 3 gives a brief overview about models for Sentiment Classification. Section 4 discusses about the experiments done by models and processing steps for training the model. In Section 5 Results of model performance are presented in tabular format. Paper is concluded in Section 6 by discussing results and future scope in field of sentiment Classification.

2. RELATED WORK

Chris Moody [4] proposed a modified LDA2Vec model which is extension of LDA and Word2Vec Model. This model creates a context vector by combining features of LDA and Word2Vec model.

In [5] author proposed a three topical word Embeddings models which are combination of contextual word embedding and document embedding. Experimental results show that TWE-1 performed well in comparison to other models for text classification.

In this paper[6]author discussed about the text corpora modelling problem and discrete data. Author presented a convexity-based approach for inference which yields in better performance.Author proposes a bidirectional gated recurrent unit neural network model (BiGRULA) for analysis of sentiments by LDA2Vec model and attention mechanism. Adding to this author evaluated his model performance by sentiment classification of Hotel reviews data, Results demonstrate model achieve good performance in classification of sentiments[7]. Extraction of various aspects from the reviews is proposed by unsupervised Topic Modeling technique[8]. Author proposed a technique which mainly focuses on Sentence level analysis and different topics. They presented a method for automatic retrieval of polarity classification (Positive, negative and Neutral) on reviews in place of manually dictionary matching on seed words.

3. SENTIMENT CLASSIFICATION MODEL

3.1 LDA

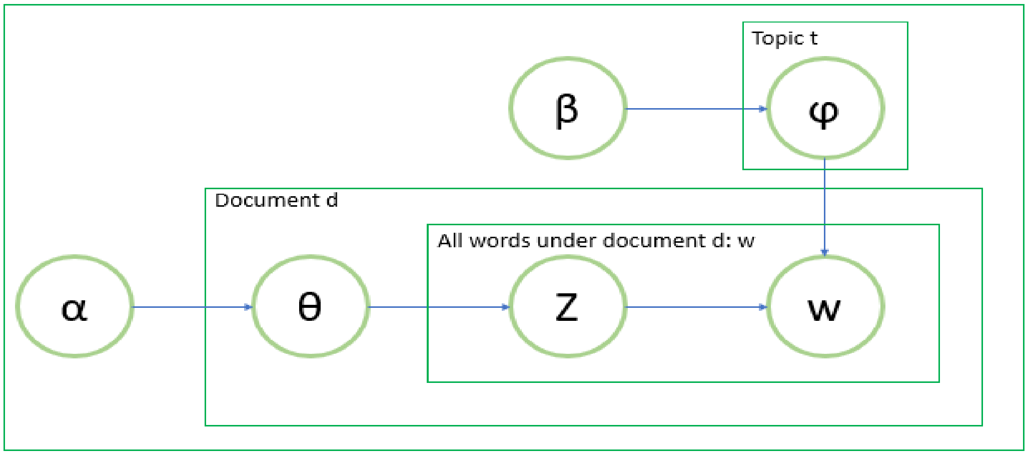
LDA is a stastical model in Natural Language Processing. Main role of LDA is Topic Modeling. LDA views each document as a collection of topics and each document has a set of certain topics. It is an example of Probabilistic Topic Model and can be easily interpreted by humans.

Important variables of LDA are -

- Alpha and Beta hyper parameters. Alpha parameter represents the weight of any topic with document whereas Beta parameter represents the weight of word with topic. So, we can conclude that higher the Alpha parameter documents contains large amount of topics and vice versa. By high-rise in Beta parameter, topics contain maximum words in a corpus and vice versa.
- Number of topics to be fetched from corpus.
- Number of topic terms that is number of terms which are present in a selected topic.
- Number of iterations allowed to LDA for convergence.

Extraction of Keywords from unstructured text simplifies the job of finding the relevant words. There are various types of stastical approach like TF-IDF (Term Frequency-Inverse Document Frequency) and RAKE (Rapid Automatic Keyword Extraction).TF-IDF uses the approach of calculating the word appearing in a text and compares it with the inverse document frequency.

LDA is a matrix factorization technique which transforms the corpus (collection of documents) in Document Term Matrix. LDA was proposed by David Blei in 2003 which is probability generation model[6]. This model is widely used in classification of text, processing of image and other fields etc[9]. Figure.1 represents the graphical representation of LDA model[10]



α: probability on the per-document topic distribution
β: probability on the per-topic word distribution
θ_m : the distribution for document d
φ_k : the word distribution for topic t
Z_{mn} : the topic for the n^{th} word in document d
W_{mn}: the specific word

Figure 1: LDA Model Representation

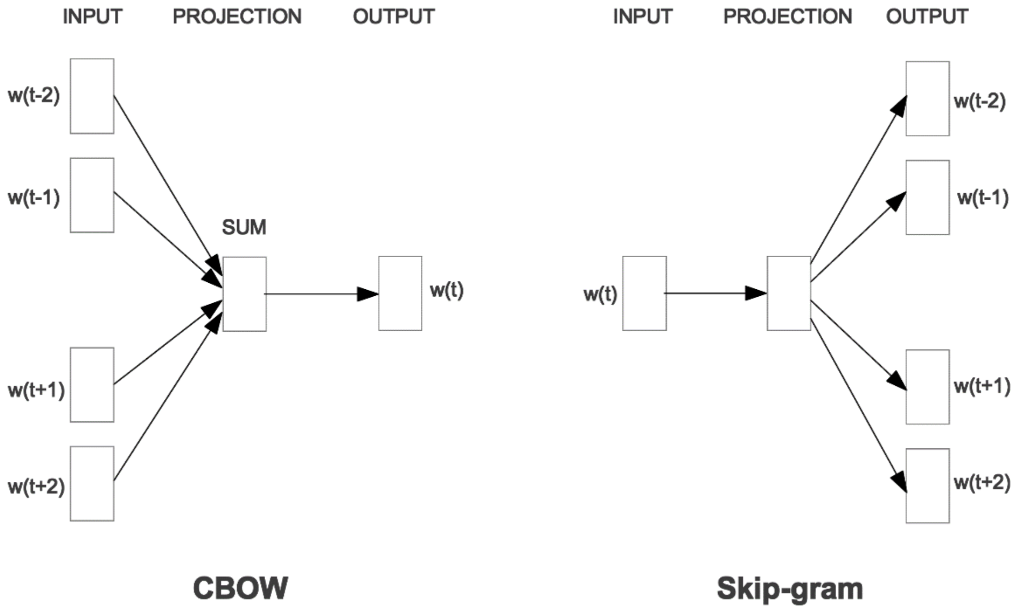


Figure 2: Word2Vec Model Representation

3.2 Word2Vec

Word2Vec is a Vector representation model. This model was published by Tomas Mikolv in 2013. For continuous representation of words, it uses a Recurrent Neural Network model which is a subfield of artificial neural networks. It is a word embedding model in which representation of words is done in numerical form.

This model is composed of two learning models which are CBOW (Continuous Bag of Words) and Skip-Gram. Figure 2 represents the graphical representation of model[11]. In CBOW model, each word in context is given as input and it attempts to forecast the word corresponding to text in middle[11]. This model gives better performance for bigger datasets.

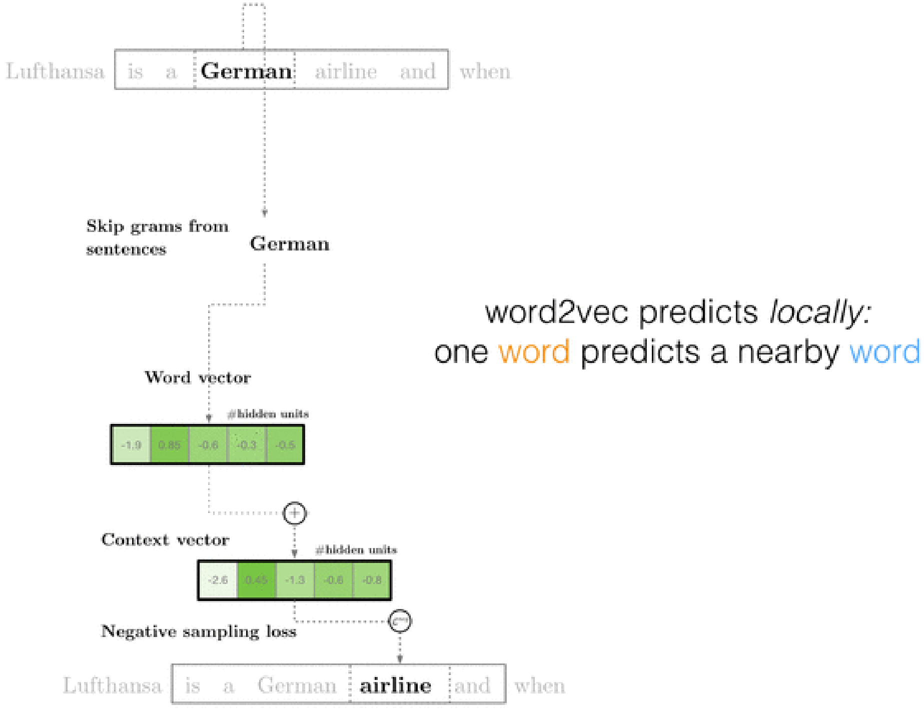

Lufthansa is a German airline and when
Skip grams from sentences
German
Word vector
#hidden units
-1.9 0.85 -0.6 -0.3 -0.5
+
Context vector
#hidden units
-2.6 0.45 -1.3 -0.6 -0.8
Negative sampling loss
Lufthansa is a German airline and when
word2vec predicts *locally:*
one word predicts a nearby word

Figure 3: Architecture of LDA2vec Hybrid Model

Table 1: Comparative Study of Topics Extracted by LDA and LDA2Vec

Topics	LDA	LDA2Vec
Topic 0	Goods, sense, choice, easier, you, could, enough	Better, service, experience, efficient, rubbish, left
Topic 1	You, can, will, remaining, easily, all, experience, bad, improve	Range, financially, money, super, facility, poor, excellent
Topic 2	Panel, up	Upgrade, will, better, updates
Topic 3	Is, little bit, making, enough	Next time, not upto, information, most
Topic 4	Usually, more, few, important	Bad service, excellent
Topic 0	Goods, sense, choice, easier, you, could, enough	Better, service, experience, efficient, rubbish, left

In Skip-Gram model the prediction of context words or surrounding words is done corresponding to target words [11]. This model is just reverse of CBOW model.

For a given word wj, Skip-gram modeling predicts the context of this word by wj−2wj−2, wj−1wj−1, wj+1wj+1, wj+2wj+2 where as CBOW is reverse of skip gram model approach it predicts the context by wj−2wj−2, wj−1wj−1, wj+1wj+1, wj+2wj+2.

3.3 LDA2Vec

Architecture of Word2Vec model is illustrated in Figure 3, in which document Vector depicts the percentile of various topics whereas Context vector is built by combination of various topic vectors which are represented in a document. LDA2Vec predicts the word globally and locally synchronously. It predicts the available words by both nearest words and globally available documents.

LDA2Vec model published by Chris Moody in 2016 as depicted in Figure 4. This model is extension of word2vec model and LDA model by incorporating topic and document vectors. Word Embedding and Topic models are the key factors of this approach.

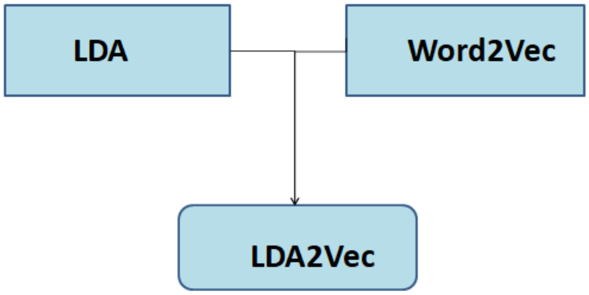


Figure 4: LDA2vec Hybrid Model

It is concluded from Table 1 that alone passive or active systems are not appropriate and sustainable due to increasing energy demand trend in space heating/cooling. It forces us to adopt suitable hybrid systems according to tailor made situations.

4. Experiments

In this section, I am analyzing the sentiments of Customer Survey corpus by probabilistic topic model. We are demonstrating performance by different model like LDA and LDA2vec.For this we used a corpus of 1000 records from a Survey firm which has response of customers availing different services.

For any classification problem there are 4 major steps to be followed for analysis which are as follows-

4.1 Identification of Feature

Feature identification is a preliminary step for data analysis. We will try to find out how many documents are there for a given category. We will define one specific entry per text column as a document according to category wise.

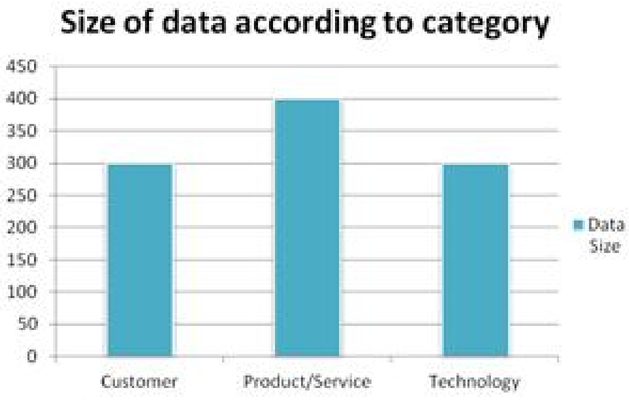


Figure 5: Category wise Data Size

Categories defined in our corpus are “Customer”, “Product/Service” and “Technology”. As we can see from Figure 5, Size of data for category “Computer” and “Technology” is almost same having documents size of 300.

4.2 Preprocessing Data

Data cleaning is the important step for further processing. For processing in classification algorithm data need to be structured which requires whitespace removal, stop words removal, punctuation removal etc.

To process any model following steps for cleaning the data are executed.

1. Tokenization-In this step we split the sentence into tokens for processing. For some Non-ASCII characters, we replaced it with some special character while processing with LDA2vec approach.
2. Stop Word Removal- Removal of noisy data makes corpus more structured as these words like prepositions (at, in, on), conjunctions (and, if, but) makes no sense for sentiment analysis process. We used NLTK stopwords removal corpora library for this step.
3. Stemming- The process of stemming includes removing of difficult words like “ing”, “ion” etc. For this we used NLTK library “porterstemmer”. It converts the word like “playing” with “play”. Vectorization of words-Model always takes input in form of vectors. For converting our cleaned data to vectors we used “Genism” corpora dictionary.

4.3 Training Model

For training the model we used LDA2vec for topics extraction. At this stage we transformed our documents into features. For extracting features from data, we used LDA and word2Vec.We used sklearn “LabelEncoder” for classifying our categories into numbers. Now we train our data with LDA and LDA2Vec approach.

By executing LDA and LDA2vec, we got a topic for each document having relevant words about that topic. Table 1 represents the most relevant words fetched from both approaches. As we can see the words fetched from both approaches are not relative to each other. We can easily demonstrate that the topics fetched by LDA2Vec are most relevant according to human’s interpretation while LDA extraction is not that meaningful.

5. RESULTS

In this section we demonstrated the performance of both the model by execution of dataset from Survey Company. Our dataset contains 1000 reviews from different categories.

Table 2: Performance Measure of LDA and LDA2Vec

Topics	**LDA**		**LDA2Vec**	
	Precision	**Recall**	**Precision**	**Recall**
Computer	.56	.52	.72	.65
Service	.66	.64	.69	.55
Technology	.54	.44	.82	.78

We evaluated the dataset by LDA and LDA2Vec approach. From experiment results we can easily conclude that LDA2vec approach is better in comparison to LDA approach. LDA2Vec has maximum precision of .82 on Technology category while Service category has minimum Precision and Recall score in LDA2vec of .69 and .55.

6. CONCLUSION

Despite of growing research problems Sentiment analysis has been active research topics among researchers, Data scientist etc. Sentiment analysis helps in better understanding of customer relationship management which helps in analyzing the customer feedback about any services whether he is happy or unhappy about the services provided.

In this paper, we evaluated the performance of Hybrid approach of LDA and Word2Vec with LDA on corpus of 1000 records.

Our experimental results represent thatLDA2Vec approach gives better accuracy in comparison to traditional LDA approach.

In future, we will try to work upon larger dataset and can measure the effectiveness of Hybrid approach by implementing our model. Although much work has not been done on this approach by researchers so more experiments are required for evaluating performance of this approach on various dataset.

REFERENCES

[1] A. K. Mohamad, M. Jayakrishnan, and N. H. Nawi, “**Classification of twitter data by sentiment analysis in the malay language**,” *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 6, pp. 2730–2738, 2020, doi: 10.30534/ijeter/2020/83862020.

[2] P. Mishra, R. Rajnish, and P. N. Kumar, “**Sentiment analysis of Twitter data: Case study on digital India**,” *2016 Int. Conf. Inf. Technol. InCITe 2016 - Next Gener. IT Summit Theme - Internet Things Connect your Worlds*, pp. 148–153, 2017, doi: 10.1109/INCITE.2016.7857607.

[3] N. P. Kumar, J. K. R. Sastry, and K. R. S. Rao, "**Mining negative frequent regular itemsets from data streams**," *Int. J. Emerg. Trends Eng. Res.*, vol. 7, no. 8, pp. 85–98, 2019, doi: 10.30534/ijeter/2019/02782019.

[4] DataCamp, "**LDA2vec: Word Embeddings in {Topic} Models,**" *DataCamp Community*. 2017.

[5] Y. Liu, Z. Liu, T. S. Chua, and M. Sun, "**Topical word embeddings**," 2015.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "**Latent Dirichlet allocation**," *J. Mach. Learn. Res.*, 2003, doi: 10.1016/b978-0-12-411519-4.00006-9.

[7] Q. Li, S. Li, J. Hu, S. Zhang, and J. Hu, "**Tourism review sentiment classification using a bidirectional recurrent neural network with an attention mechanism and topic-enriched word vectors**," *Sustain.*, 2018, doi: 10.3390/su10093313.

[8] S. Brody and N. Elhadad, "**An unsupervised aspect-sentiment model for online reviews**," 2010.

[9] J. Ye, X. Jing, and J. Li, "**Sentiment Analysis Using Modified LDA**," in *Lecture Notes in Electrical Engineering*, 2018, vol. 473, pp. 205–212, doi: 10.1007/978-981-10-7521-6_25.

[10] "**Combing LDA and Word Embeddings for topic modeling** | by Edward Ma | Towards Data Science." https://towardsdatascience.com/combing-lda-and-word-embeddings-for-topic-modeling-fe4a1315a5b4 (accessed Aug. 23, 2020).

[11] T. Mikolov, Q. V. Le, and I. Sutskever, "**Exploiting Similarities among Languages for Machine Translation**," Sep. 2013, Accessed: Aug. 23, 2020. [Online]. Available: http://arxiv.org/abs/1309.4168.