International Journal of Research Studies in Science, Engineering and Technology
Volume 7, Issue 6, 2020, PP 08-17
ISSN 2349-476X

Statistical Topic Modeling for Afaan Oromo Document Clustering

[1]Fikadu Wayesa Gemeda, [2]Million Meshesha (PhD)

[1]*Information Technology Department, Wachemo University, School of Computing, Hossana, Ethiopia*
[2]*Information Science Department, Addis Ababa University, Addis Ababa, Ethiopia*

***Corresponding Author:** Fikadu Wayesa Gemeda, Information Technology Department, Wachemo University, School of Computing, Hossana, Ethiopia.*

ABSTRACT

In today's world, the plenty of digital data poses challenge to understand and utilize the overwhelming amount of information.The amount of information available in digital form is getting double which is leading to the information overload almost in all languages.Manually reading this large data for content analysis without any specific idea of what a reader want is inefficient. It may waste a lot of time trawling through this varieties of data.Topic model is the basis of uncountable applications and plays an important role in diverse areas, such as information retrieval, information extraction,text summarization and document clustering. This paper presents unsupervised topic model that performs document clustering for Afaan Oromo documents that learns a mixture of latent topic in probability distribution representation over vocabularies. We combined word embedding approach to capture semantic structure of words how they are semantically correlated to each other with LDA algorithm to improve the quality of extracted topics since the LDA suffers from the bag-of-model approach. We used standard Gibbs sampling method to estimate the topic and word distributions.Theperformance of our proposed system isconfirmed through the Perplexity, Topic Coherence and human judgments.

Keywords: *Afaan Oromo, Topic Modeling, LDA, word embedding, Word Representation, Topic Extraction.*

INTRODUCTION

We live in a world where ideas and feelings are registered as a tool of human communication. In the digital age, the rise of multimedia data such as text files, video, audio and animation are continuously generated every day and the amount of data is expanding and increasing alarmingly which can be the basis for numerous analyses. Extracting quickly desired information in an unstructured data may be tedious and time consuming.

Topic modeling is an important tool in several applications in different fields such as Information Retrieval, Information extraction, Text Summarization, and Document clustering. In order to extract hidden patterns and analyze large digital datasets consists of many attributes, topic modeling is a popular for analyzing text content (Dean, 2014) (Zaki, 2014)(Jianguang Duy, 2015).

Topic models allow for grouping documents in a corpus based on their corresponding theme that provide an abstract view about a set of subjects. It helps in discovering latent or hidden topics that are present across the collection, annotating documents according to these topics and by using these annotations to organize, search and summarize texts.

The development of Topic Modeling for any language, to learn relevant information for further analysis is an important. This article aimed to make statistical topic modeling using Latent Dirichlet Allocation (LDA) Model for analyzing and organizing large document collections for Afaan Oromo text articles and cluster the documents according to the extracted topics.

The rest of this paper is organized as follows; the next section describes work related to our work.In Section 3, we present our proposed model. Section 4 describes the experimental results of this study. Finally, ourconclusion and some future research directions aredrawn in Section 5.

RELATED WORK

Statistical topic models, such as the Latent Dirichlet Allocation (LDA) (Blei D, 2010) have been verified to be effective and widely applied in various text analysis tasks. The nature of

these topic models is that they are unsupervised and entirely probabilistic; therefore, they do not exploit any prior knowledge in the models. Recently, numerous works and schemes have used to extract hidden topics from text using this statistical algorithm for text analysis.

Jonathan and David (Blei) developed the relational topic model, a model of documents and the links between them using LDA model and by the model they summarized a network of documents, to predict links between them, and predict words within them by derive efficient inference based on variational methods and evaluate the predictive performance of the model for large networks of scientific abstracts and web documents.

Shi Jian-hong et al. (Shi Jian-hong, 2014) applied LDA topic model to Chinese micro blog topic and carried out better micro blog topic discoveries. Li Wen-bo et al. (Li Wen-bo, 2008) used a labeled LDA topic model by adding text class information to the LDA topic model, which calculated the distribution of hidden topics in each class and raised the classification ability of the traditional LDA model.

Zrigui et al (M. Zrigui, 2012) have proposed a new hybrid algorithm for Arabic topic identification named LDA-SVM. This algorithm is based on the combination of LDA and SVM. The LDA method is used to classify documents. Then the SVM method is employed to attach class label. The idea of this combination is to reduce the feature dimension by LDA before applying the SVM method. Pennington et al. (2014) proposed a Global Vectors (GloVe) to build a words representations model, GloVeuses the global statistics of word-word co-occurrence to build co-occurrence matrix *M*. Then, *M* is used to calculate the probability of word *wi*to appear inthe context of another word *wj*, this probability *P*(*i*/*j*) represents the relationship between words.

MODEL DESCRIPTION

Model Used

CBOW Parameter Estimation

This paper presents a probabilistic model using LDA model for news articles written in Afaan Oromo. Text documents are represented in NLP as a bag-of-words that represented as a fixed-length vector size. This model representation doesn't consider the semantic relation between words (Minglai Shao, 2014). Neighbor words in a document are useful for figuring out semantic relatedness. This is very good approach for LDA algorithm to capture the quality topics during learning to predict the context of words. We can handle this through word embedding to improve model's performance. There are many forms of word embedding. The most popular are word2vec, GoVe and FastText. From this we used word2vec to train our model by considering semantic structure. Word2vec is highly popular word embedding model, developed by Mikolov et al.

In (Mikolov et al., 2013a), all the methods (Collobert and Weston, 2008), (Turian et al., 2010), (Mnih and Hinton, 2009), (Mikolov et al., 2013c) have been evaluated and compared, and they show that CBOW and SKIP-G are significantly faster to train with better accuracy compared to these techniques. For this reason, we have used the CBOW word representations for Arabic model1 proposed by Zahran et al. (Zahran et al., 2015).

Training the Afaan Oromo CBOW model require choice of some parameters affecting the resulting vectors. All the parameters used by Zahran et al. (Zahran et al., 2015) are shown in Table 1.

Table1. *CBOW Model Parameters*

The Afaan Oromo CBOW Model Parameters	
Parameter	Value
Vector size	250
Window	2
Freq-threshold	10

Where: *Vector size: dimensionality of the word vectors, Window: number of words considered around the pivot word (context). Frequency threshold: threshold to discard less frequent words.*

Topic Model Parameter Estimation

In LDA parameters to improve our topics generated from the input text and the topic is used as a feature based on which we cluster our data according to their semantics. The parameter selection process for the LDA model can be provided by Gibbs sampling under fixed hyper-parameter.

Table2. *LDA Model Parameters*

The LDA Model Parameters	
Parameter	**Value**
α	2/K
β	0.1
K	10
Iteration	200

Where: *α (Alpha): Dirichlet prior on the per document topic distributions, β (Beta):Dirichlet prior on the per topic word distributions, K:Number of*

topics to be generated, Iteration: number of loop a topic is sampled for each word instance in the corpus.

Semantic Representation

Words are represented as fixed-length vectors or embedding for capturing semantic structure of documents. Words that occur in the same context are represented by vectors in close proximity to each other.

Then we interpreted data in a more general space, with fewer dimensions, to create an interesting the representation using word embedding. Similar words tend to have similar vectors. To identify the neighbor words (wi, wj) in figuring out the semantic relatedness this work integrates word embedding into LDA model using CBOW from Word2Vec approach.

The similaritybetween *wi*and *wj*is obtained by comparing theirvector representations *vi* and *vj*respectively. Thesimilarity between *vi*and *vj*can be evaluated using the cosine similarity, euclidean distance,Manhattan distance or any other similaritymeasure functions.

For Example, Let ***Ogeessa (expert)***,***mana (house)*** and ***fayyaa (health)*** be three words, the similarity betweenthem is measured by computing cosine similarity between their vectors as follows;

Sim(ogeessa,mana)=0.1

Sim(ogeessa,fayyaa)=0.7

This means that, ***ogeessa (expert)*** and ***fayyaa(health)*** are semantically related to each other rather than ***ogeessa(expert)*** and ***mana(house).***

Topic Modeling

LDA

LDA is able to find out the topics and their relative proportions, which are distributed as a Latent Dirichlet random variable. Those topics then generate words based on their probability distribution(Blei D, 2010). It forms models in unsupervised mode, i.e., does not need labeled training data. The algorithm's performance can be managed though assumptions on the word and topic distributions. We used CBOW to capture a semantic structure of our corpus as a document-term matrix in vector space. The following matrix shows a corpus of N documents, D1, D2, D3 … Dn and vocabulary size of M words W1, W2 ... Wm. The value of i,j cell gives the frequency count of word Wj in Document Di.

Table3. *Document Term Matrix*

	w1	w2	w3	.....	wm
D1	0	2	1	....	3
D2	1	4	0	....	0
D3	0	2	3	....	1
.....	.....	.....	.....	.....	.....
Dn	1	1	3	....	0

Document-Term Matrix converted into two lower dimensional matrices – M1 and M2. M1 is a document-topics matrix and M2 is a topic – terms matrix with dimensions (N, K) and (K, M) respectively, where N is the number of documents, K is the number of topics and M is the vocabulary size.

Table3. *Document Topic Matrix and Topic word Distribution*

	K1	K2	K3	.....	Km
D1	1	0	1	....	1
D2	1	1	0	....	0
D3	1	0	1	....	1
.....	.....	.....	.....	.....	.....
Dn	1	0	0	....	0

	W1	W2	W3	.....	Wm
K1	0	1	1	....	1
K2	1	1	1	....	0
K3	1	0	0	....	1
.....	.....	.....	.....	.....	.....
Kn	1	1	0	....	0

The two matrices provide topic word and document topic distributions, and the distribution needs to be improved using sampling techniques in LDA by iterating through each word 'w' for each document 'd' and tries to adjust the current topic – word assignment with a new assignment. A new topic "k" is assigned to word 'w' with a probability P which is a product of two probabilities, p1 and p2 calculated for each topic.

P1 – p (topic t / document d) = the proportion of words in document d that are currently assigned to topic t.

P2 – p (word w / topic t) = the proportion of assignments to topic t over all documents that come from this word w.

The current topic – word assignment is updated with a new topic with the probability, product of p1 and p2. In this step, the model assumes that all the existing word – topic assignments except the current word are correct. Essentially this is the probability that topic t generated word w, so it makes sense to adjust the current word's topic with new probability. After a number of iterations, a steady state is achieved where the document topic and topic term distributions are fairly good.

The topic we reassign the word to is based on the probabilities below.

P(document "likes" the topic)×P(topic "likes" the word w′)

$$\frac{ni,k+\alpha}{Ni-1+K\alpha} * \frac{mw',k+\beta}{\sum_{w\epsilon\beta} mw,k+V\beta} \quad (1)$$

Where:

- ni,k - number of word assignments to topic k in document i
- α - smoothing parameter (hyper parameter - make sure probability is never 0)
- Ni - number of words in document i and −1- don't count the current word you're on
- K - total number of topics
- mw′,k - number of assignments, corpus wide, of word w′ to topic k
- β - smoothing parameter (hyper parameter - make sure probability is never 0)
- ∑w∈Vmw,k - sum over all words in vocabulary currently assigned to topic k
- V size of vocabulary i.e. number of distinct words corpus wide

LDA generative topic model approach that generates mixtures of topics based on word frequency from sets of documents. Topics Y and documents X Jointly P (Y, X) where the topics Y with highest joint probability given X has the highest conditional probability. That means when we put with equation:

$$p(Y|X = \max P(Y,X) \quad (2)$$

In generative process, assume we have X words and Y topics in a given document, as shown below:

Table4. *Word Topic Counters*

X1=Barattoonni mana barumsa fayyaa ispoortii atleetiksii kubbaa miilaa qilleensaa dhibee fayyaa midhamaa

Y1= 0 0 0 1 2 2 2 2 0 3 1 1

	Topic 0	Topic 1	Topic 2	Topic 3
Document i	4	3	4	1

Now, we can get matrix that shows the overall counter of words versus topics in the whole collection as shown below in table 5.

Table5. *Sample Overall Word Topic distribution*

	Topic 0	Topic 1	Topic 2	Topic 3
Barattoonni	3	32	12	2
Mana	7	23	1	0
Barumsaa	10	5	4	1
Fayyaa	13	8	2	5
Ispoortii	6	11	16	2
Atleetiksii	1	0	7	23
Kubbaa	0	2	8	6
Miilaa	0	0	5	4
Qilleensaa	3	1	1	3
Dhibee	2	7	9	2
Miidhame	4	3	2	4

The Model Architecture

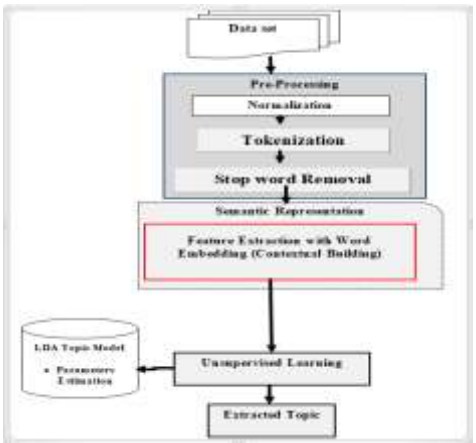

Figure1. *The Proposed Architecture*

EXPERIMENTS AND RESULTS

Sample Test

In its nature LDA is very effective for very large data collection. In fact, to measure the performance of our model we have used a dataset containing 16 documents that have been collected manually from 4 different domains (Education, Health, Weather and Sport categories). These documents were collected from different websites: Fana Broadcating Corporation (FBC), BBC Afaan Oromo and Oromia Broadcasting Network (OBN).

These instances consist of different sentence length and words, as shown in table 6.

Table6. *Distribution of data collected for training and testing*

News type	Sentences	Words
Education	350	3357
Health	256	2328
Sport	790	2348
Weather	420	5470
Total	**1816**	**13,503**

Pre-Processing

A set of data cleaning approaches have been applied to our dataset. The dataset went through the following preprocessing steps:

1. Punctuation mark apostrophe (‘) has replaced with character “h” or “y”.
2. Removing Punctuation marks
3. Splitting each sentences into individual tokens
4. Removing Stop words

Results

To evaluate the model, fitting parameters should be provided. The parameters are the suggested number of K topics that is done through perplexity. The gibbs sampling estimation sets for the model; $\alpha = 2/K$, $\beta = 0.01$ and $K=10$. After providing the parameters, the model produces output and the 7 topics with top terms were selected randomly from 10 generated topics.

Table7. *Topic Words Distributions*

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
Mana	Sammuu	Faalama	Itoophiyaa	Dhiibbaa	Atleetiksii	Oomisha
Barnootaa	Dhukkuba	Qilleensaa	istaadiyeemii	Lolaan	Dheeraa	Faalmni
Barattoota	Dhibee	Fayyaa	Fiigicha	Yaalaa	Isaayyaas	Bishaanii
Barattoonni	Kaansarii	sadarkaa	dorgommii	Lafa	Ijoollee	Shakkii
Barumsaa	yaalii	Addunyaa	Itoophiyaan	Hawaas	Riiyaal	MIDROC
Kuma	Qabxii	Qorannoo	xumuraaf	Nama	Tapha	Bishaan
Dandeettii	dhibamtoota	Biyyoota	Kirooshiyaan	Buna	Jiraattonni	Barattoonni
Seera	Qorannoon	Qilleensa	shaakala	Bunaa	Filannoo	Bulchaan
Fayyadamuu	Nageenya	Faalamni	sahaatii	Sabbataa	Kilabii	Warshichaa
Barnoota	Imaammata	atileetonni	Lama	Miidhaa	Jimmaa	Warqii

From the above table, topic 1 is about Barnoota (Education). Topic 2 is about Fayyaa (Health). Topic 3 is about FaalamaQilleensa (pollution). Topic 4 is about ispoortii (Sport).The words are arranged in the order of highest probability of words distribution to topics. For the manual analysis of the extracted topics, we confirmed that the LDA results are able to identify and reveal relevant information. topic 1: “mana”, “barnoota”, “barattoota”, “barumsaa” give us an idea of ‘Barnoota’ (Education). In topic 2, the outliers are “sammuu”, “dhukkuba”, “dhibee”, “kaansarii”, “yaalii” lead us to Fayyaa (Health). In topic 3, “faalama”, “qilleensaa”, “fayyaa”, “sadarkaa” indicate the FaalamaQilleensaa (Air pollution). Not every word in a topic can be justified as to carry the semantic structure of topics. Topics are appeared with their keywords as their proportions contributed in the topic.

Table8. *Document Topics Distributions*

DocId	Top-topics	Topic contributions to Documents											
1	1	0.71	6	0.089	4	0.065							
2	1	0.403	6	0.202	7	0.121	5	0.097	4	0.089	3	0.065	
3	2	0.471	1	0.178	3	0.138	6	0.103	5	0.069			
4	7	0.312	1	0.201	5	0.134	6	0.114	4	0.109	2	0.089	
5	2	0.396	3	0.173	5	0.115	1	0.108	6	0.101	4	0.101	
6	4	0.327	3	0.245	6	0.163	5	0.102	7	0.082	2	0.061	
7	2	0.252	6	0.215	5	0.178	1	0.131	4	0.093	7	0.075	3	0.056
8	7	0.354	2	0.177	1	0.125	6	0.115	3	0.104	4	0.083		
9	4	0.647	3	0.141	2	0.082	5	0.059						
10	4	0.718	3	0.068	6	0.06								
11	6	0.355	2	0.211	3	0.184	4	0.132	5	0.066				
12	6	0.654	3	0.102	4	0.079	7	0.071						
13	5	0.4	3	0.179	2	0.105	1	0.084	7	0.079	4	0.079	6	0.074
14	5	0.655	4	0.084	7	0.059	6	0.059	2	0.059				
15	3	0.531	2	0.166	5	0.094	4	0.067						
16	7	0.544	5	0.133	3	0.095	6	0.076	1	0.057	4	0.051		

Table 8 shows how the topics are contributed in each document; the first document, **doc 1** is 71% of **topic 1**, 8.9% of **topic 6**, and 6.5% of **topic 4** and so on. This allows for a more comprehensive understanding of the themes present in each text. For each document to what extent the topics contributed is specified.

Table9. *Perplexity of the model*

Number of Latent Topics	Perplexity
2	-2.53
3	-4.91
4	-5.97
5	-6.5
6	-7.95
7	-9.72
8	-9.12
9	-9.74
10	-9.81
11	-5.77
12	-3.45
13	-2.33
14	-2.22
15	-3.44
16	-5.88
17	-4.22
18	-4.12
19	-6.67
20	-3.54

Table 9 shows that the perplexity decreases as the number of topics increases from topic 2 to 10, at topic 7 to 10 it registered some related constant perplexity value with lowest one. So, we inferred some meaningful semantic structure of topics between 7 to 10. Thereby, 10 latent number of topics was selected since this is lowest obtained perplexity value as we can see from the table. The number of topics, in an observed pair with the least perplexity difference, was considered as the optimal number of topics. The differences between the topic perplexities for the selected range of topics have been presented in Table 10. From this table, it can be observed that the difference in the perplexities is less when the number of topics is from 7 to 10.

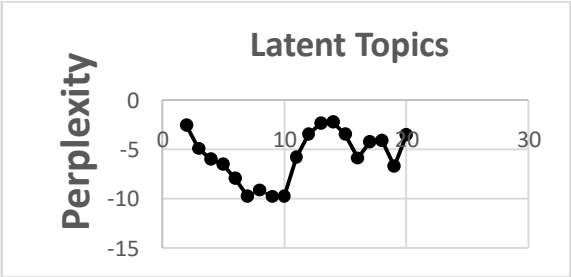

Figure2. *The Perplexity of LDA evaluated on the Corpus*

# of Topics	Average Perplexity
(2,3)	-3.72
(3,4)	-5.44
(4,5)	-6.235
(5,6)	-7.225
(6,7)	-8.835
(7,8)	-9.42
(8,9)	-9.43
(9,10)	-9.775

Table10. *Perplexity difference among the topics*

Table 10 shows the coherence score for the number of topics from 2 to 10. From topic 2 to topic 10 the coherence score decreases and increases from topic 11 to 20 so that we discard all perplexity value of this range. Now, choosing the number of topics would be with lowest perplexity from all average that is at 9 and 10.

We used PMI for the ranked ten words in eachtopic. Candidates for the best words were selectedby choosing the top-1, top-2 and top-3 ranked words.

We calculated the conditional likelihood of the words occurring with each. For instance, Topic 1**{manabarnootaaBarattootaabattoonnibarumsaa}** → [**{mana, barnootaa}, {mana, barattoota}, {mana,barumsaa},{barnoota, arattoota}**, etc] the confirmation measure of the topic coherence was calculated from. In this case, the conditional likelihood is obtained for the frequently co-occur words in the topic. We used two words conditional probability i.e. p(wi,wj)=number of documents that contain words wi and wj per total number of documents.

The nature of subjects influences the choice of words. Though the choice of top-m words for a topic affects its interpretability we just used top-10 words of a topic. That is the choice of 'm' for different topics may be different and not constant. PMI is computed for each pair of topic's words. We got average PMI of **0.525% or 52.5 %**. This only to quantify how much the extracted topics is interpretable.

In the topics, words that are semantically coherent in our corpus have been captured by an LDA with referenced labeled corpus according to the domain category based on the defined kinds of keywords in the domain for human readability. We put a thresh hold of 0.2 (the words that contributed less than 20% of average mean in a topic is considered as difficult to label otherwise the co-occurred word can be referenced to label.

For example, given a list of words mana, barumsaa, barnoota, barattoota, barataa, which is an Education topic, the first four words could be the most representative word related to education. This is because it is natural to think about the Education after seeing the words mana, barumsaa and barnoota individually. A good candidate for best word could be the word that has high average conditional probability given each of the other words.

In our experiment, we had 16 documents with 4 labels (ground truth). As our result shows, from 10 topics almost more than 5 of them perfectly matched with accurate labels.

The performance measure of the learnt topics is also done using human judgment evaluation metric. We selected nine participants (N=9) to rate the coherence of each topic and they were presented with top 10 term sets, each of which characterize a topic. We selected 3 masters' students, 3 Afaan Oromo expert and 3 others. Participants were asked to read the given topic and rate using a given standards. They asked to judge topic on two scale; Relevant (topic is coherent and interpretable if can easily assign predefined categories) and Irrelevant (words appear random and unrelated to each other). For our purposes, the usefulness of a topic can be thought ofas whether one could categorize to one of the four topical areas (Education, Health, Sport and Weather condition) as particular to describe a topic.

We first judge our own judgment that means we set it as golden judgment and we compared the golden judgment to the judgments made by the participants. Each participant was asked to judge 10 topics and the average response for each topic was calculated to measure the model as recall, precision, F-measure. If a given topic can falls under one of categories (Education, Health, Sport and Weather condition), participants put relevant, otherwise irrelevant. Table 10 shows the rating of 10 topics and human judgment.

	u1	u2	u3	u4	u5	u6	u7	u8	u9	Average Recall	Average Precision
Topic1	R	R	I	R	I	R	R	R	R		
Topic2	R	R	R	R	I	R	R	R	R		
Topic3	I	R	R	R	I	I	R	R	R		
Topic4	R	R	R	R	I	I	R	R	R		
Topic5	R	R	R	R	R	R	I	R	I		
Topic6	R	I	R	R	R	R	R	R	R		
Topic7	R	R	R	R	R	R	R	R	R		
Topic8	I	R	R	R	R	R	R	R	R		
Topic9	I	R	R	R	R	R	R	R	R		
Topic10	R	R	R	R	R	R	R	R	I		
Recall	0.7	0.9	0.6	0.8	0.7	0.7	0.9	0.6	0.8	**0.744**	
Precision	0.57	0.6	0.7	0.75	0.43	0.57	0.67	0.5	0.62		**0.605**
F-Measure											**0.66**
R-Relevant I-Irrelevant U-User											

Table11. *Overall human ratings*

Table 11 show that the model has an **average** score of **74.4% Recall, 60.5% precision and 66% F-measure** based on the human judgment. The challenge here is to judge a topic as relevant, human understanding with similar topic was varied. Everyone held a different opinion on a topic. Even with the same people, we can have a different outcome in a different time frame. It depends on many factors, word relatedness, and personal educations. This method is limited with the size of the experiment and may have different results in different people. But the idea of the method and the way to evaluate may be useful for inspiring other researchers.

Document Clustering and Exploration

We combine the clustering results and document similarity to assemble the document exploring system for our dataset. This enables user to choose one of the topics and look at the related documents of that topic. For each document in our data set we identify the topic index for which the probability is the largest, i.e., the main topic. Grouping by the topic index, counting, and sorting results in the counts of documents per topics.

Figure3. *Topic List*

By clicking certain topic, the related documents and it enables user to explore documents through topicsand documents according to a topic.

Figure4. *Top Ranked Documents Related to Topic 1*

If we click document 4 from the above link, we get how a specific topic contributed in the document. It shows us for this specific document topic 7 holds 31%, topic 1 20%, topic 5 13%, topic 6 11% and so on. The result looks the following figure.

Figure5. *Topic Distribution of certain Document*

CONCLUSION AND FUTURE WORK

In this article, we presented an unsupervised topic model that performs document clustering for Afaan Oromo documents that uses distributed representations of words. We used LDA modeling schemes in our model that uses word embedding approaches to capture the semantic structure of Afaan Oromo words. We performed document exploration based on the extracted topics how much they are related to a given topic.In our experimentation results we discussed and evaluate the performance of our model using three methods; Perplexity for estimating the optimal number of topics, automatic topic coherence and Human judgment.

As future work, we can make further study by conducting deeper grammatical analysis focusing on Part of Speech Tagging (POS) since Nouns are more representative of the topics than frequencies of features to acquire good topics.The application of topics-based search is efficient than keywords based So, our further study would be applying topic-based search to information retrieval.

REFERENCES

[1] J. Dean, "Big Data, Data Mining, and Machine Learning : Value Creation for Business Leaders and Practitioners," 2014.

[2] I. Biro, "Document classification with latent dirichlet allocation," Ph.D.dissertation," 2009.

[3] M. J. &. M. W. Zaki, "Data Mining and Analysis: Fundamental Concepts and Algorithms," Cambridge University Press, 2014.

[4] L. JianguangDuy, "Topic Modeling with Document Relative Similarities," Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, 2015.

[5] V. M. S. Suganya C, "Statistical topic Modeling for News Articles," International Journal of Engineering Trends and Technology (IJETT), vol. Volume 31, Number 5- January 2016.

[6] Clint P. George, "A Machine Learning based Topic Exploration and Categorization on Surveys," International Conference on Machine Learning and Applications, 2012.

[7] David M. Blei, "Latent Dirichlet Allocation," Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[8] Blei D, "Probabilistic topic models," IEEE Signal Process Mag, 2010.

[9] D. M. Jonathan Chang, "Reading Tea Leaves:How Humans Interpret Topic Models," in Advances in neural information processing systems, p. 288–296, 2009.

[10] C. NitinSukhija, "Topic Modeling and Visualization for Big Data in Social Sciences," Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress, 2016.

[11] A. RubayyiAlghamdi, "A Survey of Topic Modeling in Text Mining," International Journal of Advanced Computer Science and Applications, Vols. Vol. 6, No. 1, 2015.

[12] Sarah ElShal, "Topic modeling of biomedical text," IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2016.

[13] M. Verberne, "Automatic thematic Classification of election Manifestos," Information Processing & Management, vol. 4, pp. 554-567, 2014.

[14] Mohamed A Zahran, Ahmed Magooda, Ashraf Y Mahgoub, HazemRaafat, Mohsen Rashwan, and AmirAtyia. 2015. Word representations in vector spaceand their applications for arabic. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 430–443. Springer.

[15] B. Daniel Maier, "Applying LDA topic modeling in communication research: Toward a valid and reliable methodology, Communication Methods and Measures," Communication Methods and Measures, 2018.

[16] Tomas Mikolov, Kai Chen, Greg Corrado, and JeffreyDean. 2013a. Efficient estimation of word representations in vector space. In *In: ICLR: Proceeding ofthe International Conference on Learning Representations Workshop Track*, pages 1301–3781.

[17] H. Merouani, "Clustering with probabilistic topic models on arabic texts," in Modeling Approaches and Algorithms for Advanced Computer Applications, ser. Studies in Computational IntelligenceSpringer International Publishing, vol. vol. 488, p. pp. 65–74, 2013.

[18] J. Zhao, "Topic modeling for cluster analysis of large biological and medical datasets," BMC bioinformatics, 2014.

[19] V. Jensen, An introduction to Bayesian networks, London: UCL press, 1996.

[20] M. Spruit, "Examining Topic Coherence Scores Using Latent Dirichlet Allocation," in The 4th IEEE International Conference on Data Science and Advanced Analytics, p. 165–174, 2017.

[21] M. K. Christidis, "Exploring Customer Preferences with Probabilistic Topics Models.," 2014.

[22] D. Berkani, "A Topic Identification Task for Modern Standard Arabic," In Proceedings of The 10th WseasInternationalConference On Computers, pp. 1145-1149, 2006.

[23] M. Zrigui, "Arabic text classificationframework based on latent dirichlet allocation," Journal of

Computing and Information Technology, vol. 3, pp. 125-140, 2012.

[24] H. Merouani, "Clustering with Probabilistic Topic Models on Arabic Texts," In Modeling Approaches and Algorithms for AdvancedComputer Applications, pp. 65-74, 2013.

[25] S. David Newman, "Analyzing Entities and Topics in News Articles using Statistical Topic Models," 2011.

[26] M. Hanna M. Wallach, "Evaluation Methods for Topic Models," in Proceedings of the 26th Annual International Conference on Machine Learning, New York, NY, USA, p. 1105–1112, 2009.

[27] N. Ralf Krestel, "Latent Dirichlet Allocation for Tag Recommendation," ACM, 2009.

[28] C. Muhammad Omar, "LDA Topics: Representation and Evaluation," Journal of Information Science, 2015.

[29] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig.2013c. Linguistic regularities in continuous spaceword representations. In *Hlt-naacl*, volume 13,pages 746–751.

Citation: *Fikadu Wayesa Gemeda, Million Meshesha. "Statistical Topic Modeling for Afaan Oromo Document Clustering" International Journal of Research Studies in Science, Engineering and Technology, 7(6), 2020, pp. 08-17.*