Topic Modelling: A Comparative Study for Short Text

Sara Lasri and El Habib Nfaoui
LISAC Laboratory Sidi Mohammed Ben Abdellah, University Fez, Morocco

Keywords: Topic Modelling, Latent Dirichlet Allocation, Biterm Model, LDA2Vec, WNTM.

Abstract: Massive amounts of short text collected every day. Therefore, the challenging goal is to find the information we are looking for, so we need to organize, search, classify and understand this large quantity of data. Topic modelling is a better performing technique to solve this problem. Topic modelling provides us with methods to organize, understand and summarize the short categorical text.TM is an intuitive approach to extract the most essential topics detection in a short text.

1 INTRODUCTION

Topic modelling is the task of identifying which underlying concepts are discussed within a collection of documents and determining which topics each document is addressing (Andra, Pietsch, Stefan, 2019).

Topic modelling is a method to find out the hidden semantic topics (Political, sports, or business, etc.) from the observed documents in the text corpus (Chris Bail, 2012).

Topic modelling provides methods for automatically organizing, understanding, searching, and summarizing corpus (Bhagyashree Vyankatrao Barde, A. M. Bain wad. 2017)

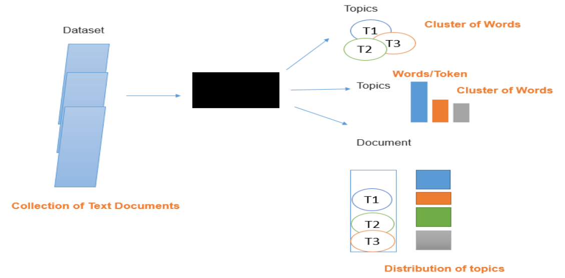

Figure 1: Topic Modelling.

In general, documents modelled as mixtures of subjects, where the subject is a probability distribution over Words (Hamed, Yongli, Chi, Xia Xinhua, Yanchao, Liang, 2019). Statistical techniques are then utilized to learn the topic components and mixture coefficients of each Document (Hamed, Yongli, Chi, Xia Xinhua, Yanchao, Liang, 2019).

Detection of the topics within short texts, such as tweets, has become a challenge. However, directly applying conventional topic models. (Hamed, Yongli, Chi, Xia Xinhua, Yanchao, Liang, 2019).

In this paper, we present different methods for topic modelling, and we compare them to find the most efficient for uncovering the hidden themes in the tweet.

2 TOPIC MODELLING METHODS

2.1 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is an unsupervised generative probabilistic method; it is the most popular topic modelling (Hamed, Yongli, Chi, Xia Xinhua, Yanchao, Liang, 2019). The basic idea is that documents represent random mixtures over latent topics, where each subject characterizes by a distribution over words (Hamed, Yongli, Chi, Xia Xinhua, Yanchao, Liang, 2019).

Lasri, S. and Nfaoui, E.
Topic Modelling: A Comparative Study for Short Text.
DOI: 10.5220/0010737000003101
In *Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning (BML 2021)*, pages 479-482
ISBN: 978-989-758-559-3

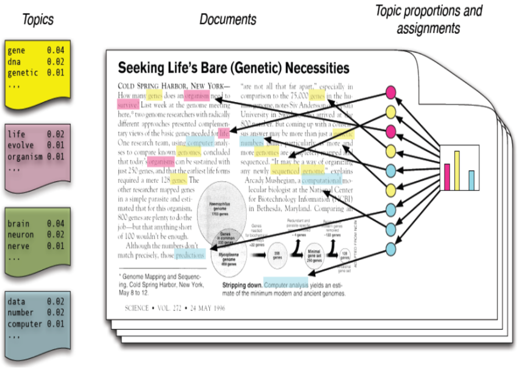

Figure 2: Latent Dirichlet Allocation (LDA) ((David M. Blei, 2010)

- Each topic is distributed over words (David M. Blei, 2010).
- Each document is a mixture of corpus-wide topics (David M. Blei, 2010).
- Each word is drawn from one of these topics (David M. Blei, 2010).

2.2 Biterm Topic Model (BTM)

Biterm proposed by, (Xiaohui, Jiafeng, Yanyan, Xueqi, 2013), is a topic model based on forming biterms from a corpus. It is the best method for detecting the topic in the short text (Xiaohui, Jiafeng, Yanyan, Xueqi, 2013).

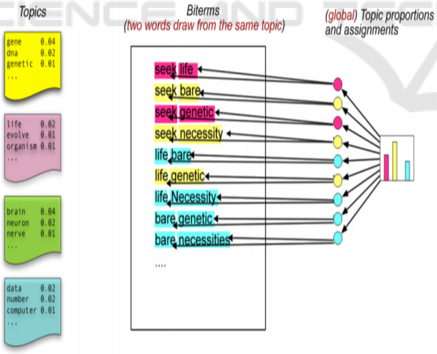

Figure 3. Biterm Topic Model (BTM) (Xiaohui, Jiafeng, Yanyan, 2013).

Model the generation of biterms with latent topic structure (Xiaohui, Jiafeng, Yanyan, 2013)

- A topic is a probability distribution over words (Xiaohui, Jiafeng, Yanyan, 2013)
- A corpus is a mixture of topics (Xiaohui, Jiafeng, Yanyan, 2013)
- A biterm is two , i.e. sample drawn from one topic (Xiaohui, Jiafeng, Yanyan, 2013)

BTM bases on word co-occurrence that learns subjects by modelling word-word co-occurrences patterns, which calls biterms. (Xueqi, Xiaohui, Yanyan, 2014).

2.3 A Hybrid of LDA and Word2Vec (LDA2Vec)

LDA2Vec is a topic-modelling algorithm, is a hybrid of the two algorithms. (Xueqi, Xiaohui, Yanyan, 2014) LDA (latent Dirichlet allocation) and Word2Vec.

LDA2Vec, a model that learns dense word vectors jointly with Dirichlet-distributed latent document-level mixtures of topic vectors (Christopher, 2016).

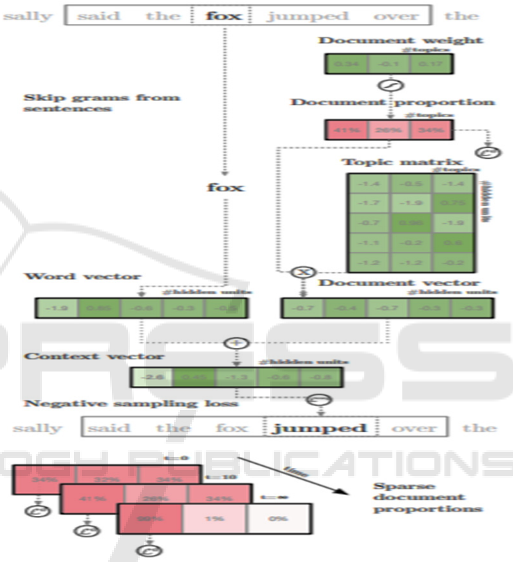

Figure 4: LDA and Word2Vec (LDA2Vec) (Christopher, 2016)

LDA2Vec embeds both words, document vectors into the same space, and trains both representations simultaneously by mixing word2vec's skip-gram architecture with Dirichlet-optimized sparse topic mixtures vectors (Christopher, 2016).

2.4 Word Network Topic Model (WNTM)

Word co-occurrence network based model named WNTM, WNTM models the distribution over topics for each word instead of learning topics for each document (Yuan, Jichang, KeXu, 201).

WNTM uncovers latent word groups in a word co-occurrence network. Here latent word-groups of the network are taken as topic components of a corpus space (Yuan, Jichang, KeXu, 2015).

WNTM learns to generate each word's adjacent word list in the network by using latent word groups and words belonging to that group's space (Yuan, Jichang, KeXu, 2015).

3 EXPERIMENTS AND RESULTS

3.1 Testing Corpus

To compare the four topic-modelling methods on the short text, we carried experiments on a standard short text collection of the tweet.

To compare these models' ability to provide a suitable thematic of a collection of documents (tweet), in this section. We present the difference between all these models. To find the most efficient model, we compare the probability distribution over the k topics for each tweet.

Table 1: Tweet Topic Detection

LDA	**Tweet 1**	0.00903	0.00121	0.01132
	Tweet 2	0.01142	0.00111	0.00945
	Tweet 3	0.00249	0.01754	0.00755
Biterm	**Tweet 1**	0.00956	0.00204	0.01177
	Tweet 2	0.01246	0.00239	0.00945
	Tweet 3	0.00567	0.01987	0.00653
LDA 2Vec	**Tweet 1**	0.00975	0.00298	0.01870
	Tweet 2	0.01543	0.00373	0.00975
	Tweet 3	0.00756	0.01994	0.00722
WNTM	**Tweet 1**	0.00998	0.00321	0.01934
	Tweet 2	0.01570	0.00311	0.00934
	Tweet 3	0.00623	0.01998	0.00683

Latent Dirichlet Allocation (LDA): is a probabilistic technique. For each document, the results give us a mix of topics that make up that document.

Biterm Topic Model (BTM): gets topics by modelling the generation of word co-occurrence patterns

LDA2Vec (LDA, Word2Vec): LDA2Vec is a hybrid of the two algorithms LDA and Word2Vec.

Word Network Topic Model (WNTM): learns the topic for each word by uses global word co-occurrence

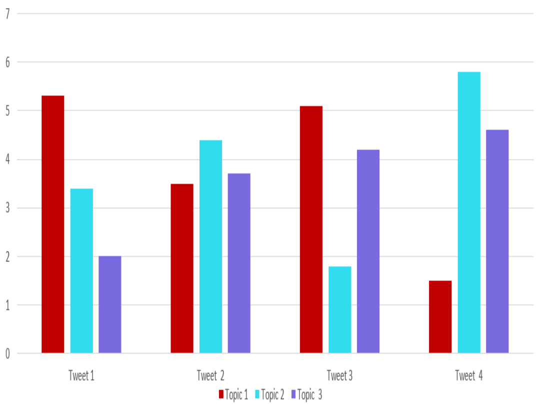

Figure 6: Tweet Topic-Detection

From the result, we can find which model is performant than another. However as shown the Table, in Figure 6, WNTM still gains a relatively better performance as compared to LDA and Biterm, the same thing for the LDA2Vec model. From Table 1 we can note that the probability distribution over the k topics for each tweet with LDA2Vec and WNTM is better than LDA and Biterm.

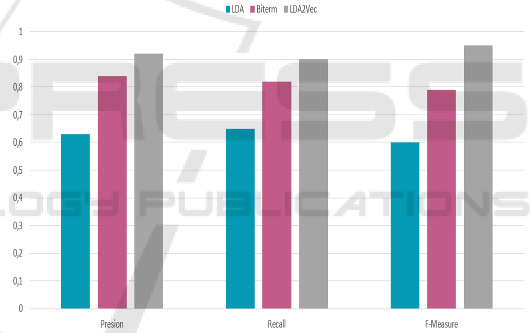

Figure 7: Comparison between LDA, Biterm, LDA2Vec.

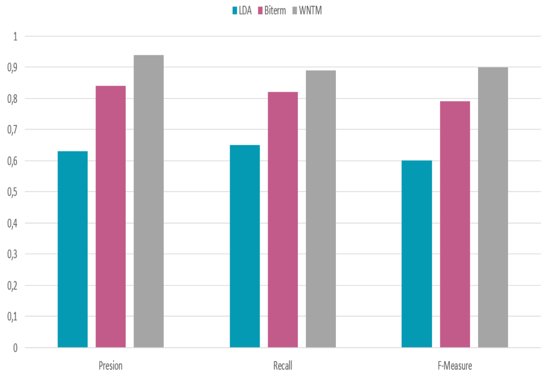

Figure 8: Comparison between LDA, Biterm, WNTM.

This result indicates that WNTM and LDA2Vec are more accurate than LDA and Biterm in selecting topics. The averaged precision, recall, and F-measure of Biterm are better than that of LDA. However, the averaged precision, recall, and F-measure of LDA2Vec is better than that of LDA and Biterm the same thing for the WNTM model. We can find that WNTM and LDA2Vec are among the models that allow predicting the themes of the tweet

4 CONCLUSIONS

Topic modelling has gained a trend in extracting the hidden themes of the short text. All the topic-modelling methods can be uncovering and detecting the themes, subjects, for a collection tweet.in this paper. First, we list the different techniques to identify the topic detection for each tweet. Further, we compare between themes. LDA2Vec and WNTM are the most performant than LDA and Biterm .in the following work; we will improve the result to find the best solution to uncovering the topic in a tweet collection.

REFERENCES

Andra-Selina, Pietsch, Stefan Lessmann, 2019 *Journal of Business Analytics* Topic modeling for analyzing

Bhagyashree Vyankatrao Barde, A. M. Bainwad. , 2017 *International Conference on Intelligent Computing and Control Systems (ICICCS)*.An overview of topic modeling Methods and tools.

Chris Bail (2012). Topic modeling https://cbail.github.io/SICSS_Topic_Modeling.html.

Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng

Xinhua Jiang, Yanchao, L, Liang Zhao, (2019).*Multimedia Tools and Applications volume 78* Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey.

David M. Blei, November 2010.Probabilistic Topic Models: A focus on graphical model design and applications to document and image analysis. (David M. Blei, 2010)

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Xueqi Cheng.A, May 13–17, 2013. *The International World Wide Web Conference Committee (IW3C2)*. Biterm Topic Model for Short Texts

Xiaohui, Jiafeng, Yanyan (2013). A Biterm Topic Model For Short texts. http://www.bigdatalab.ac.cn/~lanyanyan/slides/2013/WWW2013-Yan.pdf

Xueqi Cheng, Member, IEEE, Xiaohui Yan, Yanyan Lan, Member. IEEE, Jiafeng Guo,Member, *IEEE IEEE. Transactions on Knowledge and Data Engineering* .BTM: Topic Modeling over Short Texts

Christopher, Moody 6 May 2016. Mixing Dirichlet Topic Models and Word Embeddings.

Lars Hulstaert, (October 19th, 2017).LDA2vec: Word Embeddings in Topic Models https://www.datacamp.com/community/tutorials/lda2vec topic-model

Yuan Zuo, Jichang Zhao, KeXu Received. Aug24, 2015 Word Network Topic Model: A Simple but General Solution for Short and Imbalanced Texts