Using Latent Dirichlet Allocation on Twitter data to gain insight in the behaviour and mindsets of young people in restrictive settings.

Sije van der Veen

***Abstract*—**

A. KEYWORDS

Topic Modeling, Latent Dirichlet Allocation, Social media, Conversational data, Twitter data.

I. INTRODUCTION

This project was created in collaboration with Radio Netherlands Worldwide Media (RNW Media). RNW Media creates communities online to contribute to social change by writing about important and sensitive topics which young people in restrictive settings care about, and having young people consume this information and discuss it [1]. One problem they are dealing with is language gap between they organization and their audience. Since RNW Media wants to create content which speaks to young people, one of RNW Media's challenges is the language gap between the (technical or institutional) language of the organization and their audiences. Another problem they are dealing with is the costly supervised topic modeling for RNW media makers, which also introduce bias. Unseen topics by media makers should also be provided by content on the different platforms. This project is approach to tackle these problems by data-driven approach Latent Dirichlet Allocation (LDA).

LDA is well studied topic modeling method [7]. LDA learns the connection between words, topics and documents by assuming documents are generated by a probabilistic model. The applicability of topic modeling across languages makes the approach suitable for RNW Media since they work in Arabic, Chinese, Hindi, French and English. In general NGOs are working with many complex languages and therefore this study can good contribution for this field.

The goal of this thesis is to gain insight in the mindset and behaviour of young people from fragile countries to eventually positively influence the societies. We aim to do this by applying topic modelling on Twitter messages of the RNW Burundi Twitter platform, a platform containing Twitter messages from Burundi citizen. The assumption is that the topics will largely categorize and summarize what people of this platform talk about, which can ultimately be analyzed to gain insight in the mindsets of the people who posted the messages. This thesis aims to answer the following questions:

1) How accurately can topic modelling be applied on Twitter data? The first research question estimate how accurately topic modeling can be applied to Twitter data. Perplexity and coherence score will be used as evaluation on subs-set of the Burundi Twitter data-set. [2] [3] [4].
2) What is the optimal number of topics that can be generated? In the second sub-question, the influence the number of topics generated by LDA is measured. Also the perplexity and coherence score will used as evaluation, by testing the LDA model with different output settings on the data.
3) Can LDA distinguish between Twitter message using hash-tags as true labels? In the third sub-question the LDA output will be compared with the hash-tags which are labeled with each twitter message. Are the message with in common hash-tag assigned to the same topic group? For the this question the accuracy score will be used for evaluation [5] [6]. Since precision and recall necessarily depend on the notion of true classes. The hash-tags will be used as true classes. A LDA topic will be assigned to a hash-tag based on the biggest number labeled hash-tags within LDA topic. From there the true positives, true negatives, false positives and false negatives can be calculated.

The report will start with a related work section, in which previous studies about topic modeling will be discussed. Section 3 outlines methodologies used in the study. Subsequently, Section 4 outlines the experimental setup, explaining the data and implementation of the models. The results of the performed experiments are discussed in Section 5.

II. RELATED WORK

Since the introduction of LDA in 2000, this topic model method has been studied extensively [7]. This section will give an brief overview of how topic modeling algorithms can be adapted to many kinds of data. Among other applications, they have been used to find patterns in genetic data, images, and social networks [9]. Different types of models were used for estimating continuous representation of words. Two well known models are Latent Semantic Analysis (LSA) and LDA.

LDA is a well known topic modeling tool for exploiting the hidden thematic structure in large archives of text. LDA can become computationally expensive on larger data sets [8]. Since the introduction of LDA it is used for complex goals and tasks. One assumption of LDA is that the order words is not set a specific order, because it based on bag of words. To make LDA robust to non-exchange order of words, Wallach et all.[10] developed a topic model that relaxes the bag of words assumption by assuming that the topics generate words conditional on the previous word. Griffiths et al. [11] developed a topic model that switches between LDA and a standard Hidden Markov model (HMM). These models expand the parameter space significantly but show improved language modeling performance. A second assumption of LDA is that the order of documents is not set in a specific order. This assumption will not work perfect when the documents are over a wide range of time. In many situations topics change over time. From this conscience dynamic topic modeling is developed [12]. In dynamic topic modeling the order of documents is retained. A third assumption of LDA is that the number of number topics per document is known and equal. This assumption is fixed by the Bayesian nonparametric topic model [13]. This model determines the number of topics with posterior inference. Also previous unseen topics can be shown by new documents. This method is extended to analyze the hierarchy of topics. The representation is a tree, with general topics higher on the tree, and concrete topics lower on the tree [14]. In many cases the document includes pieces of meta-data, like a the name of writer, resources and geographical information. In several papers the integration of meta-data has been shown. The relational topic model is developed by Chang and Blei [15]. In their study the assumption is made that links between documents depends on the in-common shared topics. This is a combination of a new topic and network model.

III. Method

In this section the accessed methodologies in the study are explained. The main used method LDA will be explained in the first sub-section. Followed by the evaluation methods perplexity, coherence score and accuracy.

A. Latent Dirichlet Allocation

B. Perplexity

C. Coherence Score

D. Accuracy

Precision
Recall

IV. Experimental Setup

This section outlines the experimental setup, explaining the data and implementation of models. In the first section the data set will be described. Followed by the data processing steps.

A. Twitter data set

The data set which will be used in this project consist out Twitter messages of the RNW Burundi Twitter platform. The messages are mainly written in French, English and Rundi (language spoken in Burundi). The data-set is filled with 464845 Twitter messages and in a separate column the used hash-tags. The Twitter messages are posted between January 2018 and May 2019. The data-set did not contain missing values. In figure 1 the distribution is shown for the number of words per message. From the figure can be clearly seen that the size of the message is small. Most messages contains 18 words (4.66 percent). The most messages contains between 13 and 44 words.

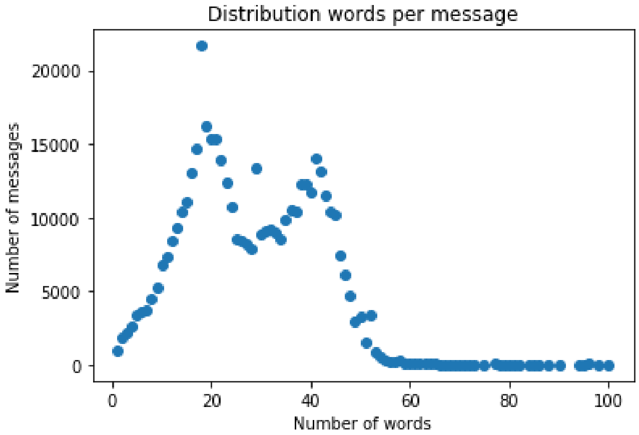

Fig. 1. The distribution of words per message is shown. Most messages contains 18 words (4.66 percent). Most messages contains between 13 and 44 words.

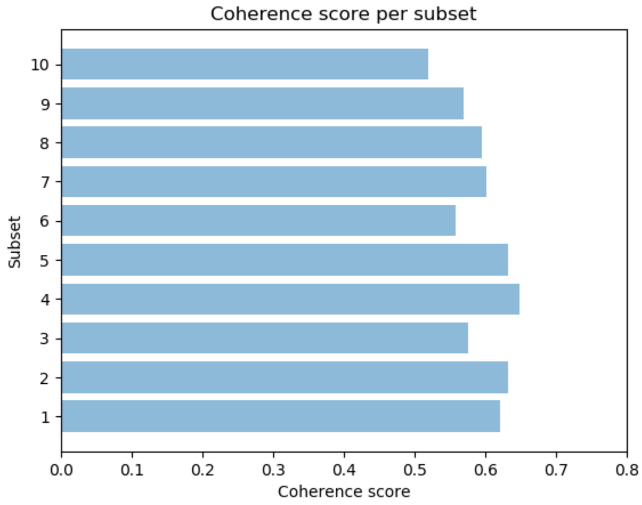

Fig. 2. For 10 subsets of the data set the perplexity is score is measured. The scores varies between 0.54 and 0.65.

B. Data processing

C. Sub-question 1

D. Sub-question 2

E. Sub-question 3

Explanation which hash-tags are selected.

V. Results

This section contains the results of experiments which has been done to answer the research questions. The study is divided into three sub-questions which will be answered separately in sub-sections.

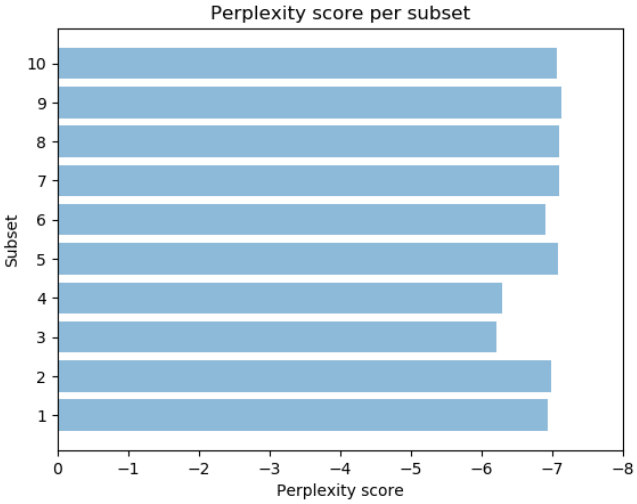

Fig. 3. For 10 subsets of the data set the perplexity is score is measured. The scores varies between -6.21 and -7.09.

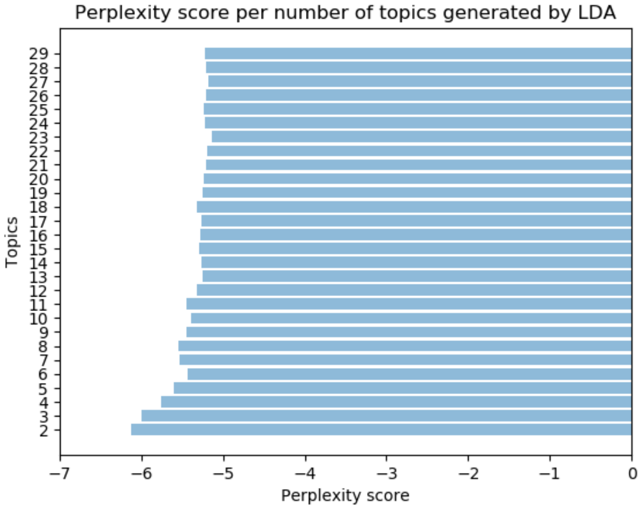

Fig. 4. Perplexity per topic. a lower perplexity suggests a better fit. The lowest perplexity is achieved using k = 2. Which results in perplexity score of -6.13.

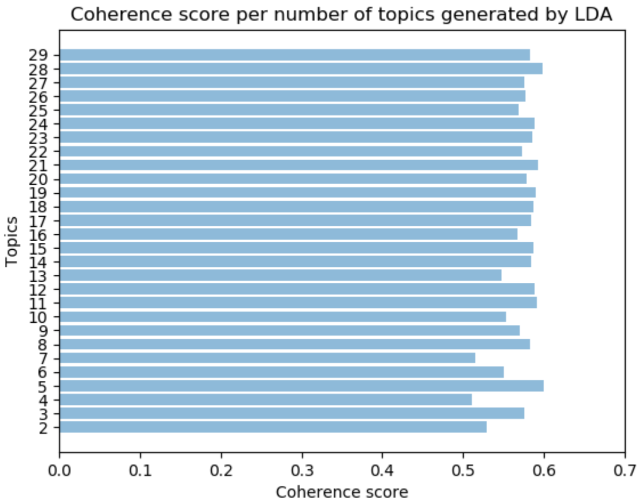

Fig. 5. Coherence per topics generated by LDA. A higher coherence score indicates a better fit. The highest score is generated with k = 5 and coherence score of 0.6.

A. Sub-question 1

How accurately topic modeling can be applied to conversational data? To answer this question LDA has been applied to 10 sub-sets of the Twitter data-set. To evaluate a LDA model, the model perplexity and model coherence are common used evaluation methods [2] [3] [4].

For each subset the perplexity score is shown in figure 3. Between the perplexity scores is a variance of 8.25 and standard deviation of 2.87. The performance per data set is slightly different, but stable. Another evaluation method applied to subsets of the Twitter dataset is the coherence score. For each subset the coherence score is shown in figure 2. Between the coherence scores is a variance of 0.0014 and standard deviation of 0.038. Also the coherence score per data set is slightly different, but stable. Both evaluation methods shows slightly different results between sub-sets, but overall performs stable. How accurately can topic modeling be applied on Twitter data? By using a LDA model with 10 output topics, perplexity scores can achieved between -6.21 and -7.09 and a coherence score between 0.54 and 0.65.

B. Sub-question 2

This question should answer what number of topics is optimal. The number of topics is written as k. To answer this question LDA is applied multiple times on the entire data set. Each run the number of generated topics has been set to a specific number k. If key-words are in multiple topics, this can indicates that k is too high.

In figure 4 the perplexity score per topic number generated by LDA are shown. Perplexity indicates how well the model describes a set of documents. Perplexity is equivalent to the inverse of the geometric mean a lower perplexity implies data is more likely. As such, as the number of topics increase, the perplexity of the model should decrease. Therefore a lower perplexity suggests a better fit [2]. The lowest perplexity is achieved using k = 2. Which results in perplexity score of -6.13. In figure 5 the coherence score per topic number generated by LDA are shown. A higher coherence score indicates a better fit. The highest score is generated with k = 5. In the following sub-question the optimal number for k should be selected. The optimal number of k is different between perplexity and coherence score. The following question can be performed with value for k which is in between 2 and 5.

C. Sub-question 3

VI. Conclusion

References

[1] RNW media website: https://www.rnw.org/about-us/
[2] Blei D.M.; Ng A.Y.; Jordan M.I. Latent Dirichlet Allocation (2003)
[3] Stevens K.; Kegelmeyer P.; Andrzejewski D.; Buttler D. Exploring Topic Coherence over many models many topics.

[4] RÃűder M.; Both A.; Hinneburg A. Exploring the Space of Topic Coherence Measures

[5] Javed A.; Byung S.U.Sense-Level Semantic Clustering of Hashtags in Social Media.

[6] Javed A. A Hybrid Approach to Semantic Hash-tag Clustering in Social Media.

[7] Pritchard, J. K.; Stephens, M.; Donnelly, P. (June 2000). "Inference of population structure using multilocus genotype data". Genetics. 155 (2): pp. 945âĂŞ959. ISSN 0016-6731.

[8] Mikolov T.; Sutsekever I.; Chen K.; Carrdo G.; Dean J. Distributed representations of words and phrases and their compositionality. NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, pages 3111-3119

[9] Blei D.M., Probabilistic topic models

[10] Wallach, h. topic modeling: beyond bag of words. in Proceedings of the 23rd International Conference on Machine Learning(2006).

[11] Griffiths, t.; Steyvers, M., Blei, D.; Tenenbaum, J. integrating topics and syntax. Advances in Neural Information Processing Systems 17. l. K. saul, y. weiss, and l. bottou, eds. mit Press, cambridge, ma, 2005,

[12] Blei, D.; Lafferty J. Dynamic topic models. in International Conference on Machine Learning (2006), acm, new york, ny, usa,

[13] Teh, Y.; Jordan M.; Beal M.; Blei D. hierarchical Dirichlet processes. J. Am. Stat. Assoc. 101, 476 (2006)

[14] Blei, D.; Griffiths T.; Jordan m. the nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. J. ACM5 7,2 (2010)

[15] Chang, J.; Blei D. hierarchical relational models for document networks. Ann. Appl. Stat. 4, 1 (2010).