USING TOPIC MODELING AND WORD EMBEDDING ON SOCIAL MEDIA DATA TO GAIN INSIGHT INTO THE MINDSETS OF YOUNG PEOPLE

SUBMITTED IN PARTIAL FULFILLMENT FOR THE DEGREE OF MASTER OF SCIENCE

Sije van der Veen
11422688

MASTER INFORMATION STUDIES
DATA SCIENCE
FACULTY OF SCIENCE
UNIVERSITY OF AMSTERDAM

2020-07-01

	Internal Supervisor	**External Supervisor**
Title, Name	Dr Maarten Marx	Kyle Snyder
Affiliation	UvA, FNWI, IvI	RNW Media
Email	maartenmarx@uva.nl	kyle.snyder@rnw.org

UvA

rnw media

SIJE VAN DER VEEN, University of Amsterdam

Generating new topics to write about is not always easy for blog-writers, but the social media data connected to the blog can provide a solution. The connected social media data can contain more timely relevant topics for future blogs. This paper describes a methodology for detecting word topics (combination of words that together form a topic) within Twitter and Facebook messages. Latent Dirichlet allocation (LDA) and Latent semantic indexing are used to detect word topics in this research. Which is a novel approach because it has not been applied on 280-character-long Twitter messages. Through two sub-questions, the performance of Topic models on Twitter messages is assessed. In the last experiment, the word topics are projected using a combination of Word2Vec with principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE) or uniform manifold approximation and projection (UMAP). These projections show, in a two-dimensional way, the links between words within a topic. The UMAP projection resulted in the most informative visualization. The visualization can be used by blog-writers, but also by companies and organizations that want to gain insights into their followers' thoughts.

CCS Concepts: • **Topic modeling → Latent Dirichlet Allocation**; **Latent Semantic indexing**; • **Word embedding** → *Word2Vec*; • **Dimension reduction** → Principal component analysis; t-distributed stochastic neighbor embedding; uniform manifold approximation and projection.

1 INTRODUCTION

This project was created in collaboration with Radio Netherlands Worldwide (RNW) Media. RNW Media creates communities online to contribute to social change by writing about important and sensitive topics that young people in restrictive settings care about, and giving them a place to discuss [1]. Since RNW Media wants to create content that targets young people, one of the organization's challenges is the gap between the (technical or institutional) language of the organization and that of their audiences. Another problem that the organization has encountered is the costly supervised topic modeling for RNW Media makers, which also introduces bias. Potentially mediamakers should also provided by topics with being demand which are not met. unseen topics by media makers should also be provided by content on different platforms. This project aims to tackle these problems through a data-driven approach. LDA and LSI are well-studied topic modeling methods [7] [28]. LDA learns the connections between words, topics, and documents by assuming that a probabilistic model generates documents. LSI learns latent topics by performing matrix decomposition (SVD) on term-document matrix. The applicability of topic modeling across languages makes the approach suitable for RNW Media since it publishes articles in Arabic, Chinese, Hindi, French, and English. In general, NGOs work with many complex languages and, therefore, this study can contribute to this field. The goal of this thesis is to gain insights into the mindset and behavior of young people from fragile countries and eventually positively influence these societies. We aim to do this by applying topic modeling to Twitter messages of the RNW Burundi Twitter platform, which contains Twitter messages from Burundi citizens. In addition, the same steps will be applied to Facebook messages and comments. The assumption is that the topics will largely categorize and summarize what people of this platform are discussing, which can ultimately be analyzed to gain insights into the mindsets of the people who posted the messages. This thesis aims to answer the following questions:

(1) What is the optimal number of topics that can be generated on a set of Twitter messages using a topic model?
(2) Are hashtags found in the same LDA output topic? In the second research question, hashtags are used as true labels (ground truth) to verify the output of the LDA model. Are the messages within common hashtags assigned to similar output groups?
(3) Can the relationship between words within a topic be visualized? In the last sub-question, a word embedding is created based on the sentences from each topic. The word embedding is used to create PCA, t-SNE, and UMAP of a Word2Vec model. In the projection, the most frequent words are shown.

The report commences by describing the related work that has been done and the ideas and models that have already been developed. Section 3 outlines the methodologies used in the study, after which 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.

2 RELATED WORK

2.1 Topic Modeling on social media data

This section provides a 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 [3]. Different types of models have been used to estimate the continuous representation of words. The four methods which can be considered as topic models are LSA, LSI, Probabilistic latent semantic analysis (pLSA) and Correlated topic model (CTM) [19]. Topic modeling has been applied for different purposes on social media data. In the study of Hu and Ester, topic modeling on Twitter data and spatial information are used to predict future users' locations [10]. Naskar *et al.* used LDA in their study to find topic distribution in mainly conversational Twitter data [23]. Topic modeling is also used for product opportunity planning in combination with sentiment analysis. In the study of Jeong *et al.* topic models were trained on social media data and in combination with sentiment analysis used to identify product opportunities [11]. Students' social media content are used for topic modeling in combination with visual features to optimize future education [26].

2.2 HashTag Clustering

In the second research question, hashtags are used as true labels to verify the output of the LDA model. A hashtag is a type of metadata tag used on social networks such as Twitter that allows users to apply dynamic, user-generated tagging, which makes it possible for others to easily find messages with a specific theme or content. A hash-tag archive is consequently collected into a single stream under the same hashtag [6] [37]. In several studies, hashtags have been clustered to analyze contextual semantics [21] [30] [32] [33],

Author's address: Sije van der Veen, University of Amsterdam.

such as in a study by Stilo and Velardi in which they created a clustering algorithm for hashtags based on temporal mining [32]. The researchers clustered hashtags based on their temporal co-occurrence with other hashtags. Tsur *et al.* created an algorithm that leverages users' practice of adding hashtags to some messages by bootstrapping over virtual non-sparse documents [33]. These non-sparse documents were then represented as vectors in the vector space model. Rosa et al. used hash-tag clusters to achieve the topical clustering of tweets, whereby they compared these effects with each other [30].

2.3 Word Embeddings Projections

In the final part of this study, the application that uses Word2Vec on PCA, t-SNE, and UMAP is introduced. Word embeddings are extensively studied on different text sources from Web search engines to sets of curriculum vitaes [22] [8]. A combination of PCA projection using Word2Vec and Twitter messages is used in the tutorial of Leightley [17]. Another dimensional reduction method used in this tutorial is t-SNE. t-SNE is a multidimensional scaling method developed by Laurens van der Maaten and is used in dimension reduction approaches [34] [35]. Andrej Karpathy also achieved projection using t-SNE on Twitter data but in combination with JavaScript [14]. UMAP is noval approach for dimension reduction for topic modeling with social media data [24]. In the tutorial of Joshi, UMAP is used in combination of LSI on '20 Newsgroup' dataset [13].

3 METHODS

In this section, the methodologies applied in the study are explained. LDA and LSI are explained in the first subsections, followed by the perplexity evaluation and coherence methods, which is used to test the topic models on Twitter data. An explanation of the weighted average of pairwise maximum f -score is then provided. Finally, Word2Vec, PCA, t-SNE, and UMAP, which are used for the word projections, are explained.

3.1 Latent Dirichlet Allocation

LDA is a generative probabilistic model for the collection of data types, such as text corpora. The model's structure can be explained as a three-level hierarchical Bayesian model. Each item or document in a collection is treated as a mixture of an underlying set of topics. Each topic is treated as a mixture of an underlying set of term probabilities. In other words, each document is treated as a mixture of different topics. A document is considered to have a set of topics that are allocated by LDA [20] [28].

3.2 Latent Semantic Indexing

LSI learns latent topics by performing matrix decomposition (SVD) on term-document matrix. It assumes that a dataset has a Gaussian distribution. Compared to LDA, LSI is faster to train, but in general has a lower accuracy. It's a linear model, which performs well on datasets with linear dependencies [7].

3.3 Evaluating the Model Using Perplexity and Topic Coherence

Perplexity indicates how well the model describes a set of documents. In this research question, the documents are Twitter and Facebook messages. Perplexity is an intrinsic evaluation metric and is widely used for language model evaluation [4] [27]. It captures how surprised a model is by new data that it has not seen before and is measured as the normalized log-likelihood of a held-out test set.

This is a decreasing function of the log-likelihood L(w) of the unseen documents w_d, normalized by the number of words in the set N. The lower the perplexity, the better the model [16].

$$perplexity(D_{test}) = exp(\frac{\Sigma_{d=1}^{M} log \quad p(W_d)}{\Sigma_{d=1}^{M} N_d}) \tag{1}$$

D_{test} = data used for training and testing the model.
w_d = a test set of Twitter messages, which are not used in the training of the model.
M = numbers of documents.
N = numbers of words in the data set.

In topic coherence, the score for a single topic (is the degree of semantic similarity). Which is measured between high scoring words in the topic. This method can help to distinguish between topics, which are semantically identical topics and topics that are artifacts of statistical inference [29]. The score is defined as pairwise scores on the words $w_1, ..., w_n$.. The words used to describe the topic, usually the top n words by frequency $p(w|k)$. This measure is the sum of all edges on a complete graph [24].

The UCI scoring method is used to calculate the coherence. The UCI measure uses as pairwise score function the Pointwise Mutual Information (PMI). In UCI, $p(w)$ represents the probability of seeing w_i in a random document, and $p(w_i, w_j)$ the probability of seeing both w_i and w_j co-occurring in a random document [16].

$$CoherenceScoreUCI(w_i, w_j) = log(\frac{p(w_i, w_j)}{p(w_i)p(w_j)}) \tag{2}$$

3.4 Clustering Comparison using f-scores for Evaluation

In the second research question, hashtags are used as true labels (ground truth) to verify the output of the LDA model. The weighted average of the pairwise maximum f^a-score is used as evaluation method in the second research question. The steps explained in the report of Vermont university and Javed are used [36]. The score is obtained by calculating the f^a-score for each hashtag. The f^a-score is based on the metrics precision (formula 3) and recall (formula 4).

Precision is the ratio of correct hashtags in the cluster compared to the total number of hashtags in the cluster (i.e., the ground truth cluster). The precision for cluster C_i is calculated as follows, considering ground truth cluster G_j:

$$precision(C_i, G_j) = \frac{|G_j \cap C_i|}{|C_i|} \tag{3}$$

G_j = the ground truth cluster, which is a set of messages, with the same hash-tag attached.

C_i = the LDA cluster, which is a set of messages, assigned to the same LDA topic.

The recall score for a cluster is based on the number of correct hashtags in the cluster compared to the number of correct hashtags that should be in the cluster. The recall of a cluster C_i is calculated as follows, considering the ground truth cluster G_j:

$$recall(C_i, G_j) = \frac{|G_j \cap C_i|}{|G_j|} \tag{4}$$

The f-score is defined as:

$$f(C_i, G_j) = 2 * \frac{recall(G_j C_i) * precision(G_j C_i)}{recall(G_j C_i) + precision(G_j C_i)} \tag{5}$$

Depending on the purpose of the evaluation, the final f-score can be calculated in two ways. In this study, the overall accuracy of all clusters is calculated. For the calculation, the weighted average of pairwise maximum f-score is used (each pairwise maximum f-score by the size of the matching ground truth cluster). The f^a-score is defined as:

$$f^a(C, G) = \frac{\Sigma_{i=1}^{n}(f - score(C_i, G_i^m) * |G_i^m|)}{\Sigma_{i=1}^{n}|G_i^m|} \tag{6}$$

$G_i^m (\in G)$ is the ground truth cluster that gives the maximum f-score matching with the cluster $C_i (\in C)$.

3.5 Word2Vec

Word2Vec is a two-layer neural network that is designed to process text. Its input is a text corpus and its output is a set of feature vectors for words in that corpus. Word2Vec converts text into a numerical format that can be understood by a machine [25].

3.6 Principal component analysis

PCA is mathematically defined as an orthogonal linear transformation that transforms the data into a new coordinate system such that the greatest variance of projected data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on [12].

3.7 T-distributed stochastic neighbor embedding

The goal of t-SNE is to embed high-dimensional data in low dimensions in a way that respects similarities between data points. Nearby points in the high-dimensional space correspond to nearby embedded low-dimensional points, and distant points in high-dimensional space correspond to distant embedded low-dimensional points [17]. t-SNE does not preserve the global data structure, meaning that only distances within clusters are meaningful while similarities between clusters are not guaranteed.

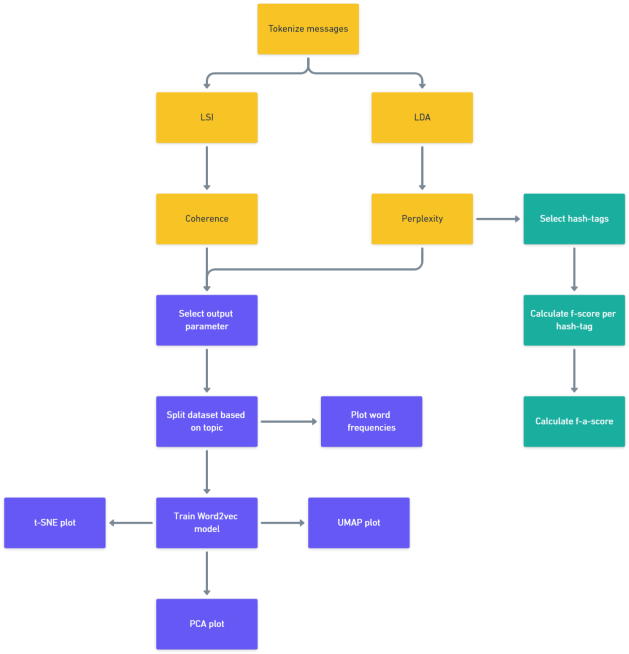

Fig. 1. Flowchart depicting the connections between the three sub-questions of this study. The yellow-colored blocks indicate the steps taken in sub-question 1, which concerns finding the optimal output topic number using a Topic model on Twitter data. The green-colored blocks indicate the steps taken for sub-question 2, which involves determining whether hashtags are found in the same LDA output topic. The blue colored blocks indicate the steps taken in sub-question 3, which regards visualizing word relationships in Word2Vec embedding. In the third sub-question, the RNW Media Burundi Twitter dataset and the RNW Media Burundi Facebook dataset are used.

3.8 Uniform Manifold Approximation and Projection

UMAP is a novel manifold learning technique for dimension reduction. UMAP can be used to visualize general non-linear dimension reductions. The algorithm assumes three characteristics about the data: (1) the data is uniformly distributed on a Riemannian manifold; (2) the Riemannian metric is locally constant; and (3) the manifold is locally connected. If the data has these characteristics, the manifold can be modeled with a blurry topical structure. The embedding is found by searching for a low-dimensional projection of the data [24].

4 EXPERIMENTAL SETUP

This section outlines the experimental setup, explaining the data and implementation of models. In the first section, the dataset is described, followed by the setups of the three sub-questions. Figure 1 presents a flowchart of the connections between the three sub-questions of this study. The blocks represent the steps taken per question.

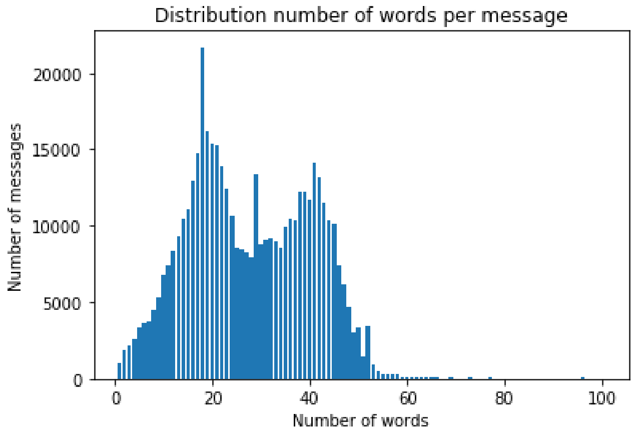

Fig. 2. The distribution of words per message. The largest group of message sizes is of 18 words. Most messages contain between 13 and 44 words. The figure clearly demonstrates that the message size is overall small, compared to documents where LDA have been applied on.

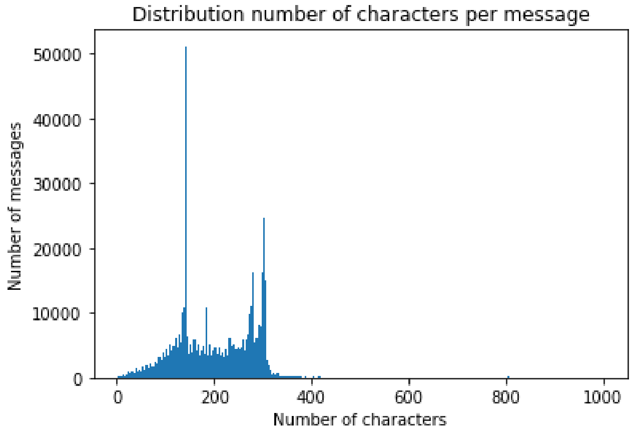

Fig. 3. The distribution of characters per message. Most messages contain less than 280 characters. Twitter's founders created Twitter messages to have a maximum length of 160 characters, which permitted a message of 140 characters and 20 characters for a username. The results exhibit a peak at around 160 characters, which could be related to this initial character limit. In 2018, the maximum character limit was changed to 280 characters [37].

4.1 RNW Media Burundi Twitter Dataset

The dataset that is used in all sub-questions is the RNW Media Burundi Twitter dataset, which comprises Twitter messages of the RNW Media Burundi Twitter platform. The messages are mainly written in French, English, and African languages. The dataset totals 464,845 Twitter messages. For this study five subset 20.000 Twitter messages. The Twitter message is presented in the first column and the hashtags for that message are shown in a second column. One message can have multiple hashtags. The Twitter messages were

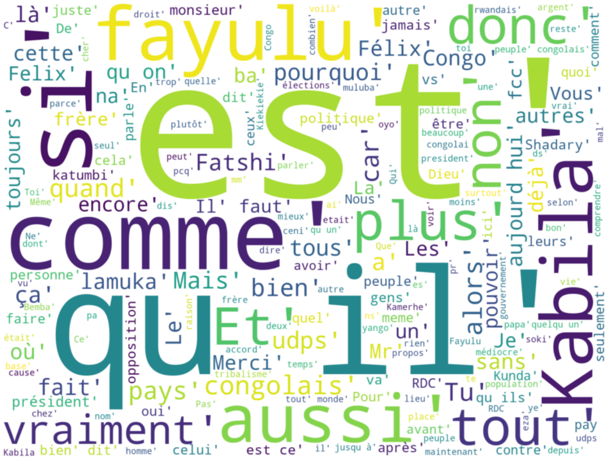

Fig. 4. Word cloud which shows the common words in RNW Media DRC Facebook dataset. The world cloud is generated before the English, French, Spanish and African stop words were removed. Which can be seen by the many French words like comme (as), aussi (also) and donc (therefore). Also, decision makers are in this figure like Fayulu (businessman and lawmaker from DRC) and Kabila (President of the DRC from 2001 up to 2018).

posted between January 2018 and May 2019. The dataset is not missing any values. In Figure 2, the distribution is shown for the number of words per message. From the figure, it can be clearly observed that the message size is small. Figure 3 presents the distribution of the number of characters per message. Most messages contain less than 280 characters. Twitter's founders created Twitter messages to have a maximum length of 160 characters, which permitted a message of 140 characters and 20 characters for a user name. The results exhibit a peak at around 160 characters, which could be related to this initial character limit. In 2018, the maximum character limit was changed to 280 characters [28]. In 2018, the maximum character limit was changed to 280 characters, which is also evident in Figure 3. In Figure 5, a word cloud of the 30 most commonly occurring hashtags in the dataset is presented. The hashtags are mainly related to countries, political parties and political events. In Table 1, the most frequent hashtags in the dataset are explained. A selection of these hashtags is used as true labels in sub-question 2.

4.2 RNW Media DRC Facebook Dataset

In the last subsection, the RNW Media Burundi Facebook dataset is used. This dataset consists of Facebook posts and messages from the RNW Media Facebook platform, which is active in the Democratic Republic of the Congo (DRC). The 91,579 Facebook posts and messages were posted between January 2018 and May 2019. For this study, a subset of 20.000 Facebook messages are used. The word cloud shown in Figure 4 displays the most common words in the dataset, among which many French words and the names of decision-makers, such as a president and a lawmaker from the DRC, are present.

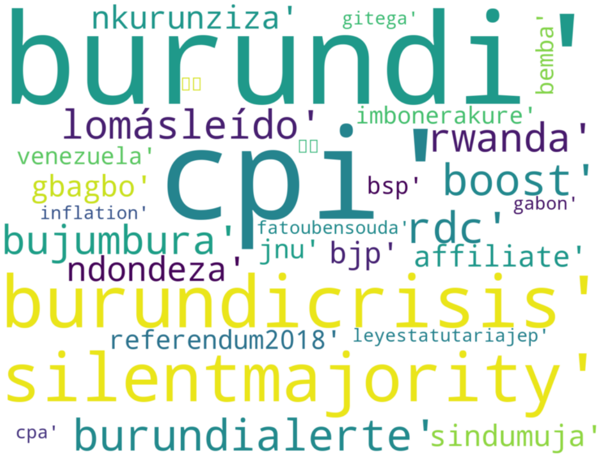

Fig. 5. Word cloud with the 30 most common hashtags in the dataset. The size of the word is equivalent on the number of occurrences in the dataset. The hashtags are mainly related to country names and political parties. Explanations of the hashtags can be found in table 1

4.3 Finding Optimal Output Topic Number

The goal of this section is to determine the parameter with which LDA, and LSI achieves the best results for the Twitter dataset. One setting is the number of output topics. To evaluate the LDA model applied in this study, perplexity (formula 1) is used, which is the common evaluation method for LDA [4] [29] [31]. Coherence (formula 2) is used for LSI [24]. Testing the total dataset is costly, instead five subsets are selected. Each subset contains 20.000 unique Twitter messages of the RNW Media Burundi Twitter dataset. From these subset common stop words English, French, Spanish, and African stop words are removed. Also common Twitter messages are filtered out. For each subset the perplexity is calculated in a range of one till thirty topics using LDA. The same setup is used to test LSI for different output topics.

4.4 Are Hashtags found in the same LDA Output Topic?

Can LDA distinguish between Twitter messages using hashtags as true labels? The hypothesis is that messages with common hashtags have a higher chance of being assigned to the same topic group. For this question, the f^a-score (formula 6) is used as a means of evaluation [36]. This score is based on precision (formula 3) and recall (formula 4). Since precision and recall necessarily depend on the notion of true classes, a selection of hashtags is used as true classes. An LDA topic is assigned to a hashtag based on the largest number of labeled hashtags within that LDA topic. Recall and precision are used to calculate the f-score, which is only for one specific hashtag (cluster). To estimate the score of all hashtags (clusters), the f^a-score is used. Explanations of the selected hashtags are provided in Table 1.

Hash-tag	Explanation	Frequency
Burundi	Country	111530
Cpi	Burundi political party	60839
Burundicrisis		11478
Burundialerte	Political tension	11478
Silentmajority	People who do not express their opinions publicly	10460
Rdc	Country	6160
Boost	Heroin-derived drug	4716
Bujumbura	Largest city and main port of Burundi	4458
Lomasleido	Political related	3337
Rwanda	Country	2928
Ndondeza	Burundi political party	2841
Nkurunziza	President of Burundi since 2005	2543
Sindumuja	Organization	2313
Gbagbo	Ivorian politician	1617
Bjp	Indian political party	1483
Affiliate	Marketing strategy	1456
Referendum2018	referendum about the president's power in Burundi	1454
Jnu	University in India	1417
Imbonerakure	Political group	1350
Bsp	Indian political party	1344
Bemba	Politician in Democratic Republic of the Congo	1310
Venezuela	Country	1305
Leyestatutariajep	Law in Venezuela	1270
Cpa	Certified Public Accountants	1250
Gitega	Gitega is one of the 18 provinces of Burundi.	1178
Gabon	Country on the west coast of Central Africa	1166
Inflation	Sustained increase in the general price level of goods and services in an economy over a period	931
Fatoubensouda	Gambian lawyer and international criminal law prosecutor	916

Table 1. Most common hashtags in the Burundi Twitter dataset with explanations and their frequency (number of messages with this hashtag). 65,927 messages did not have a hashtag attached to them in the dataset. The hashtags are mainly related to country names and political parties, but references to a university in India and the country of Venezuela and its laws also occur often in the dataset.

4.5 Visualizing word Relationships in Word2Vec Embedding

In this subsection, relationships between important words within a generated topic are visualized. This is achieved by using the Twitter and Facebook dataset. The experiment started with removing English, French, Spanish, and African stop words, and common Twitter messages are filtered out. For both datasets, the LDA algoritm is used. Based on the perplexity, the number of topics is selected. The most frequent words per topic are gathered and plotted. Subsequently the words of both topics are visualized in a projection by the Word2Vec model. Three different dimension reduction methods are used, which are PCA, t-SNE, and UMAP.

5 RESULTS AND DISCUSSION

This section contains the results of conducted experiments to answer the research questions. The study is divided into three sub-questions, which will be answered separately in sub-sections. In the fourth sub-section the further improvements are discussed.

5.1 Finding the Optimal Output Topic Number

What is the optimal number of topics that can be generated on a set of Twitter messages using LDA and LSI? The goal of this section is to determine the settings with which LDA, and LSI achieves the best results for the dataset. To evaluate the LDA model applied in this study, perplexity is used, which is the common evaluation method for LDA [4] [29] [31]. Coherence is used for LSI [24].

The number of topics that are generated by the LDA algorithm is written as num topics. To answer this research question, LDA is applied multiple times to five subsets of the dataset. In each run, the number of generated topics is set to a specific number num topics. If keywords are present in multiple topics, this indicates that num topics is too high.

In figure 6 the perplexity per topic number generated by LDA is shown. The *perplexity* of held-out documents w_d is defined as shown in Formula 3. This is a decreasing function of the log-likelihood L(w) of the unseen documents w_d; the lower the perplexity, the better the model. Therefore, a lower perplexity suggests a better fit [4]. A low perplexity is achieved using k = 1. Which results in perplexity score between -5.65 and -6.51. After topic number 12 the perplexity score is lowering for 4 subsets. Which indicates that for these subsets, the LDA model has a solid fit.

In figure 7 the coherence per topic number generated by LSI is shown. The Coherence is defined as shown in Formula 2. LSI is applied multiple times to five subsets of the dataset. In each run, the number of generated topics is set to a specific number num topics A higher coherence suggests a better fit. The best results for the subsets are achieved by using topic number between 4 and 6. For topic number larger than 20 the coherence seems to stabilize around 0.6.

5.2 Are HashTags Found in the Same LDA Output Topic?

Can LDA distinguish between Twitter messages using hashtags as true labels? Every Twitter message contains one or more hashtags. The hashtags can be used to find messages with a specific theme or content. The hypothesis for this question is, therefore, that messages that contain common hashtags have a higher chance of being assigned to the same LDA output group.

For this question, the f^a-score is used as evaluation metric. This score is calculated by a single cluster f-score. Since precision and recall necessarily depend on the notion of a true class, a selected group of hashtags is used as true classes. In Table 1, explanations can be found for the selected hashtags.

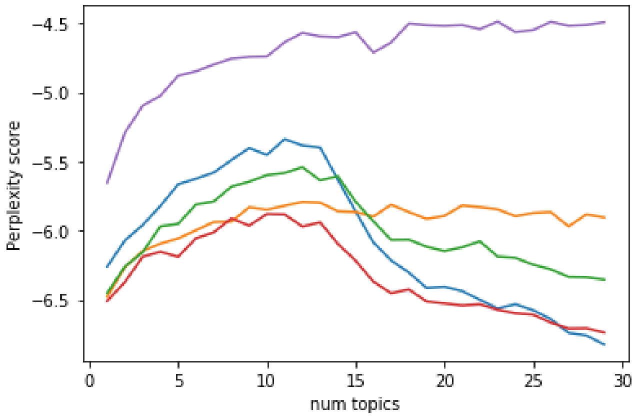

Fig. 6. Perplexity per topic number (num topics). The five different lines are the five different subsets. A lower perplexity suggests a better fit. A low perplexity is achieved using k = 1. Which results in perplexity score of between -5.65 and -6.51. After topic number 12 the perplexity score is lowering for 4 subsets. Which indicates that for these subsets, the LDA model has a solid fit.

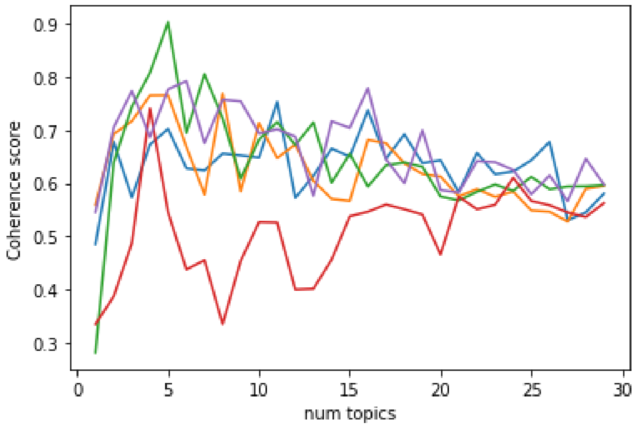

Fig. 7. Coherence per topic number (num topics). The five different lines are the five different subsets. A higher coherence suggests a better fit. The best results for the subsets are achieved by using topic number between 4 and 6. For topic number larger than 20 the coherence seems to stabilize around 0.6.

In Table 2, hashtags with a topic size of 2 and 3 are shown. For the topic size of 2, all f-scores are between 0.594 and 0.660. Hashtag "Jnu," used for a university in India, has the highest score. There is no direct link between this university and Burundi, which could explain this high score. For a topic output size of 3, the f-scores are smaller. For comparison, one hashtag is used. If a message has multiple hashtags, the first hashtag is selected.

In Table 3, hashtags with a topic size of 4 and 5 are shown. A lower number of hashtags are used to calculate the f^a-score. These scores are lower than the scores when using a topic number of 2 or 3.

Hash-tag	Topic size 2	Topic size 3
Rwanda	0.594	0.506
Ndondeza	0.607	0.521
Nkurunziza	0.616	0.531
Sindumuja	0.600	0.541
Gbagbo	0.637	0.564
Bjp	0.622	0.554
Affiliate	0.611	0.554
Referendum2018	0.641	0.573
Jnu	0.660	0.556
Imbonerakure	0.657	0.600
Bsp	0.597	0.543
Bemba	0.596	0.584
Venezuela	0.658	0.569
f^a-score	0.619	0.547

Table 2. For the selected number of hashtags, f-scores are given using a topic number of 2 and 3. For the topic size of 2, all scores are between 0.594 and 0.660. The f-scores for topic size 3 are lower, which is expected because the score is applied to an additional cluster.

Hash-tag	Topic size 4	Topic size 5
Gbagbo	0.392	0.280
Bjp	0.386	0.273
Affiliate	0.421	0.283
Referendum2018	0.400	0.286
Jnu	0.409	0.286
Imbonerakure	0.418	0.294
Bsp	0.412	0.284
Venezuela	0.410	0.294
f^a-score	0.406	0.285

Table 3. For the selected number of hashtags, the f-score and f^a-score are given using a topic number of 4 and 5. These scores are lower than the scores when using a topic number of 2 or 3.

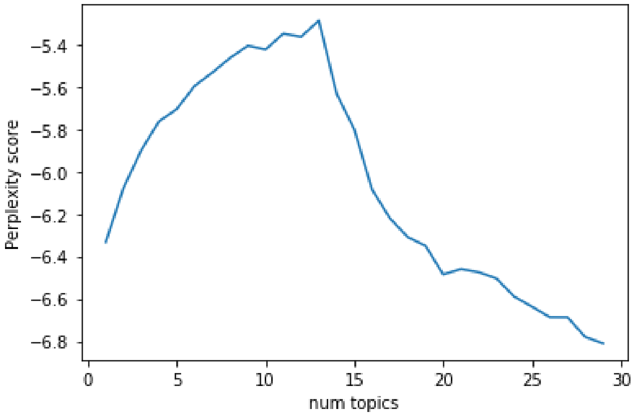

Fig. 8. Perplexity per topic number (num topics) applied on a subset of the Twitter messages. A lower perplexity suggests a better fit. In the figure can be seen that the lowest scores are achieved by a high topic number. Descent score is achieved using topic number of two.

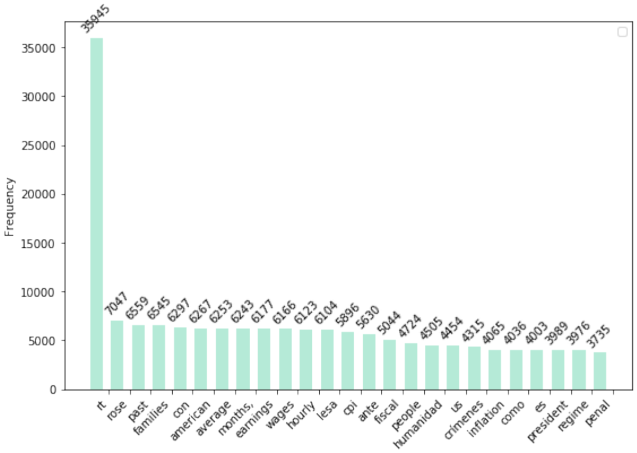

Fig. 9. Frequency of the most common words in Twitter messages that are assigned to topic 1. These words are selected from all the Twitter messages that are assigned to topic 1. The 25 most frequently occurring words are shown in this figure on the x-axis, while the number of occurrences is shown on the y-axis. *RT*, a Russian TV channel, is the most common word in this topic. In the second position is the word *rose*. Besides that, many finance-related words occur often in this topic, such as *earnings*, *wages*, and *inflation*. Political words, such as *president*, *regime*, *cpi*, and *party*, can also be found in this topic.

5.3 Visualizing Word Relationships in Word2Vec Embedding

5.3.1 *RNW Media Burundi Twitter Dataset.*

To identify relationships between the words in the RNW Media Burundi Twitter dataset, two topics are created. In figure 8 can be seen that perplexity score for two topics is lower, compared to other topic numbers. The most frequent words per topic are selected. In Figure 9, the frequencies of the 25 most common words in topic 1 are shown. Many financial and politics-related words are in the top 25. For the second topic, the same steps are taken. In Figure 10, the frequency of the most common words for topic 2 is shown. Many of these words are related to crime and humans, such as *crimes*, *militares* (soldiers), *humanitana* (humanitarian), *victimes* (victims), *police*, *regimes*, and *droits* (rights). Another observation is that topic 2 includes more Spanish and French words than topic 1. The most frequent words in both topics are shown in the PCA projection of the Word2Vec model in Figure 11. On the left of the projection is a green cluster; words such as *humanitana*, *victimes*, *commune*, *president*, and *regime* are grouped near one another. At the bottom of the picture is a blue cluster with words such as fiscal and hourly. In Figure 12, a t-SNE projection is shown. Green words occur on the left side of the projection, while the blue words are more spread out. The words *police*, *regime*, and *crimes* are clustered together. The UMAP projection is shown in Figure 13. Besides *chambre* (room)

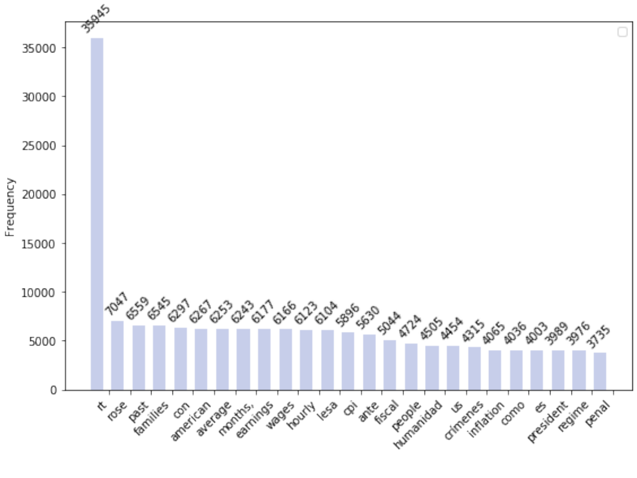

Fig. 10. Frequency of the most common words in Twitter messages that are assigned to topic 2. These words are selected from all the Twitter messages that are assigned to topic 2. The 25 most frequently occurring words are shown on the x-axis, while the number of occurrences is shown on the y-axis. *RT*, a Russian television channel, is the most common word in this topic. This topic contains more French and Spanish words compared to topic 1. Crime and human-related words occur often in this topic, such as *crimes militares*, *humanitaria*, *victimes*, *police*, *regimes*, and *droits*.

and *politique* (politics), the green words are clustered together. *Humanitaria*, *victimes*, *aseguro* (assure) and *bloqueen* (block) are closely grouped. Two clusters are formed for topic 1. The lower blue cluster consists of financial and work-related words like *earnings* and *wages*. The blue cluster in the center consists of humanity related words like *crimes*, *people*, *families* and *humanidad*.

5.3.2 *RNW Media DRC Facebook Dataset.*

Besides a Twitter forum, RNW Media also operates on Facebook in the DRC. Posts and messages from this Facebook platform are stored in the RNW Media DRC Facebook dataset. The steps described in section 4.5 are taken to identify word relationships. Two topics are created. The frequency of the most common words for topic 1 are shown in Figure 15 and for topic 2 in Figure 16. Topic 1 contains many African words, while topic 2 contains more French words, such as *femmes* (woman), *chère* (dear), and *pense* (thought). The most frequent words of both topics are shown in the PCA projection of a Word2Vec model in Figure 17. In this projection, words are clustered on the left side. In this figure, there is a relationship between the words from both topics which contain both different languages. For example, there is a close relationship between *amakuru* (major) and *pense* (thought). In Figure 18, a t-SNE projection is shown. In this projection, the words are spread over the projection. The UMAP projection is shown in Figure 19, which shows two clearly separated clusters. The upper cluster contains some politics-related words, such as *politique* and *Nkurunziza* (President of Burundi since 2005). In all three plots the words per topic are mixed in the clusters. In the Figure 14 can be seen that the low Perplexity scores are achieved by low topic numbers. Perhaps a better clustering could be achieved using more topics as output.

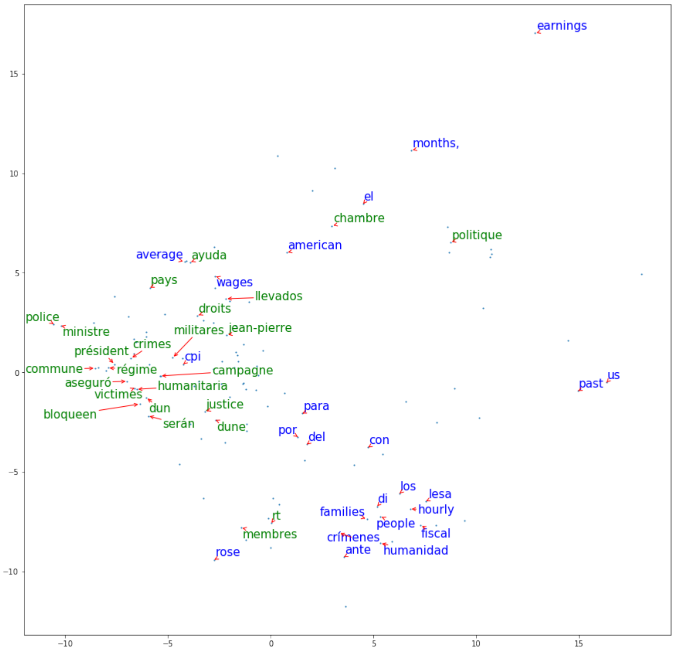

Fig. 11. PCA projection of a Word2Vec model. The 25 most frequently occurring words per topic are highlighted in green for topic 1 and blue for topic 2. The positions are shown with red arrows.

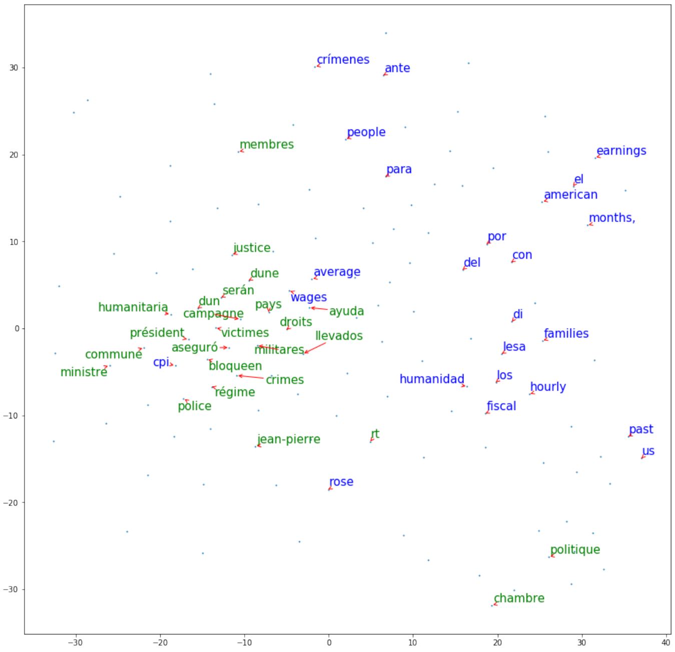

Fig. 12. t-SNE projection of a Word2Vec model. The 25 most frequently occurring words per topic are highlighted in green for topic 1 and blue for topic 2. The positions are shown with red arrows.

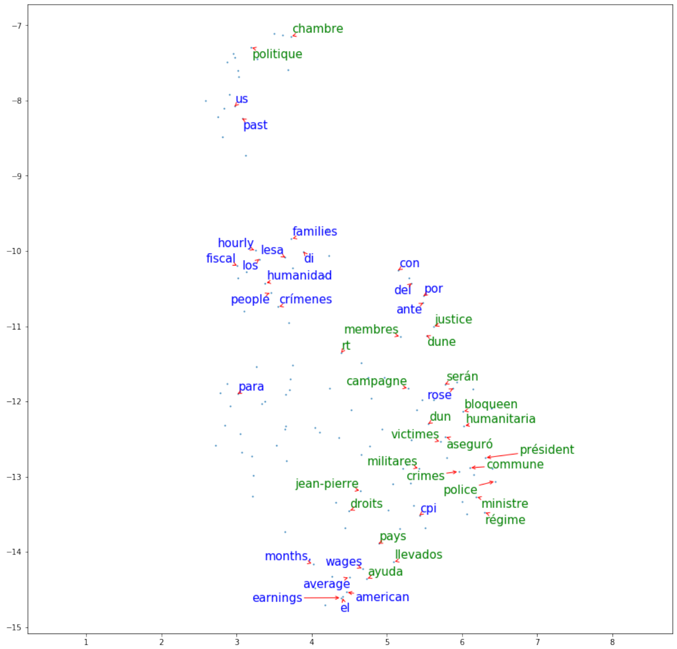

Fig. 13. UMAP projection of a Word2Vec model. The 25 most frequently occurring words per topic are highlighted in green for topic 1 and blue for topic 2. The positions are shown with red arrows.

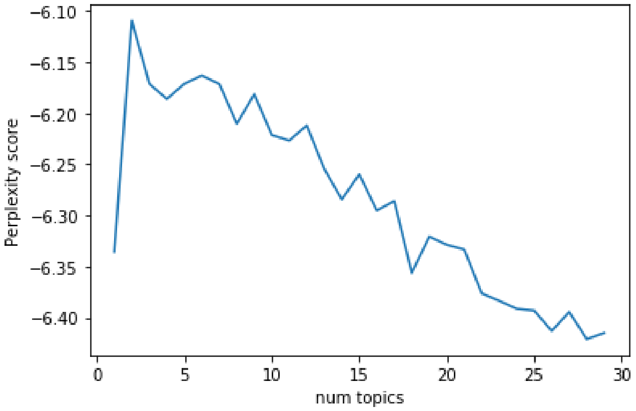

Fig. 14. Perplexity per topic number (num topics) applied on Facebook dataset. A lower perplexity suggests a better fit. In the figure can be seen that the highest scores are achieved by low topic numbers.

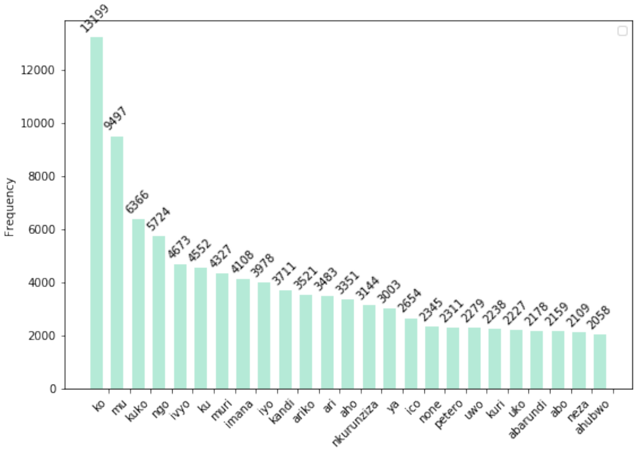

Fig. 15. Frequency of the most common words in Facebook messages that are assigned to topic 1. These words are selected from all the Twitter messages that are assigned to topic 1. This topic contains mostly African words.

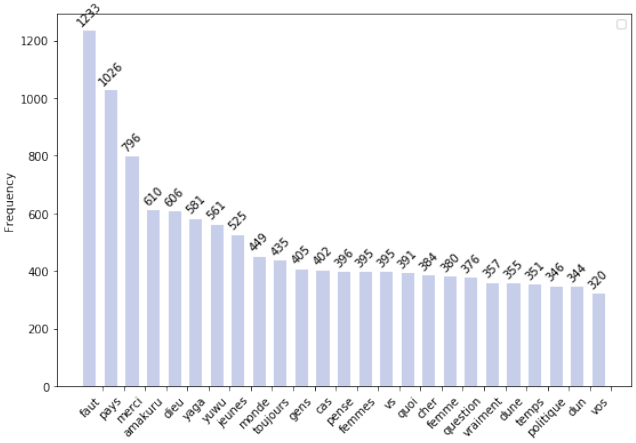

Fig. 16. Frequency of the most common words in Facebook messages that are assigned to topic 2. These words are selected from all the Twitter messages that are assigned to topic 2. This topic contains mostly French words

5.4 Future improvements

In further research, different social media datasets should be used for testing. It could be possible that in a larger dataset, more topics could be generated by the LDA algorithm. In the Burundi Twitter dataset, mostly English, French, and Spanish words occur. These languages are commonly used languages in topic modeling, for example using Python as programming language. RNW Media is also working with languages with more complex structures, such as Arabic, Hindi, and Rundi. A future study could be continued on datasets that contain these languages. Kelaiasa and Merouani applied LDA and K-means clustering on Arabic data. Their aim was to a to compare the influence of morpho-syntactic characteristics of Arabic language on both methods [25]. In the study of Alhawarat and Hegazi, LDA and K-means clustering was applied on an Arabic news dataset [2], which resulted in a combined method to cluster Arabic text documents.

In the second research question, hashtags are used as true labels (ground truth) to verify the output of the LDA model. The assumption is made that the first hashtag used by a user is most informative. For future investigation, the strength of the experiment could be increased by including other hashtags.

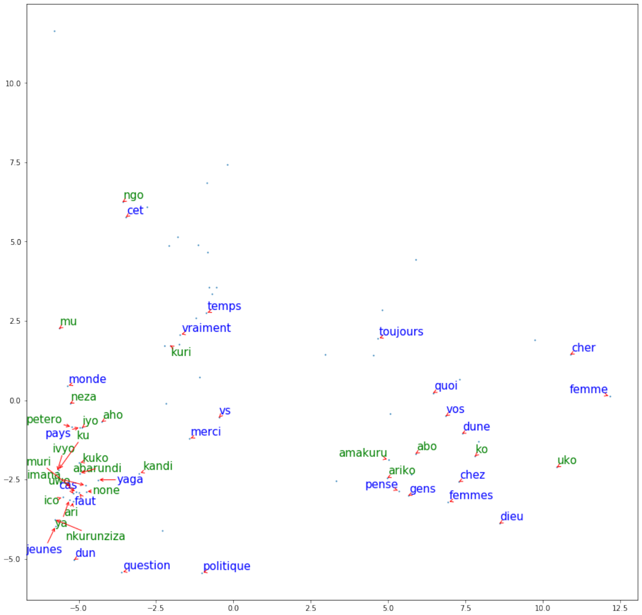

Fig. 17. Scatter plot of a PCA projection of a Word2Vec model. The 25 most frequently occurring words per topic are highlighted in green for topic 1 and blue for topic 2. The positions are shown with red arrows.)

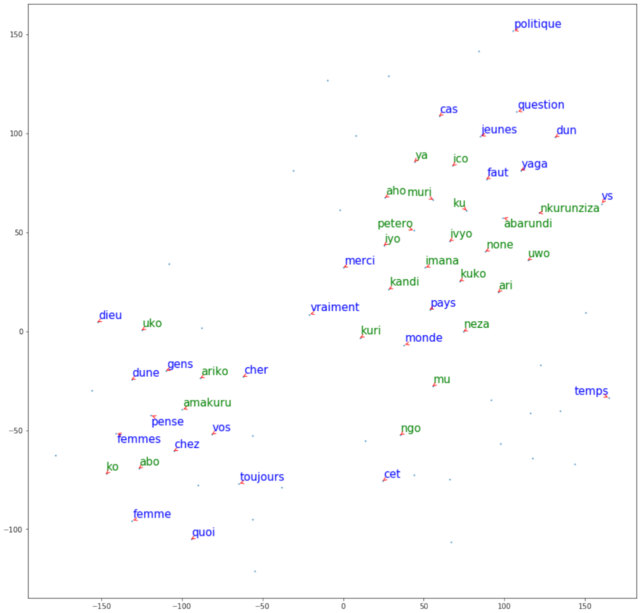

Fig. 18. Scatter plot of a t-SNE projection of a Word2Vec model. The 25 most frequently occurring words per topic are highlighted in green for topic 1 and blue for topic 2. The positions are shown with red arrows.

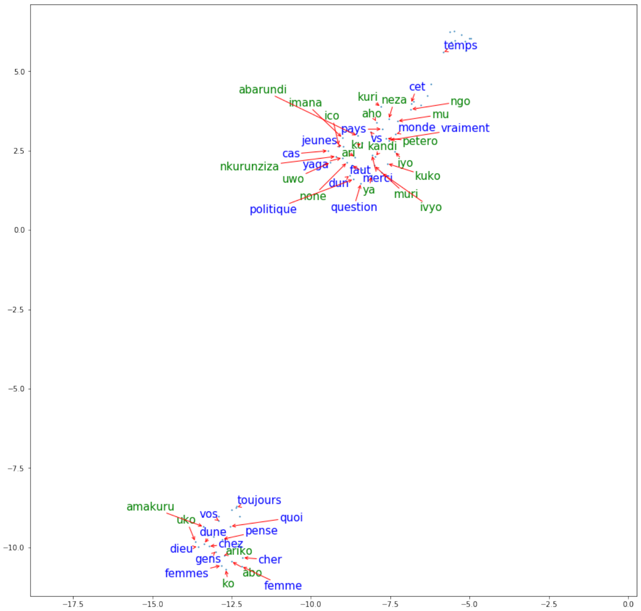

Fig. 19. Scatter plot of a UMAP projection of a Word2Vec model. The 25 most frequently occurring words per topic are highlighted in green for topic 1 and blue for topic 2. The positions are shown with red arrows.

Following up on the current investigation evolution models could be applied on the social media data. Word and topic co-occurrence mainly change over time. This has been seen in the cause of the study by Blei and Lafferty [29]. In their study a topic is considered as being associated with a continuous distribution over time.
In addition to a word embedding projection, the word similarity could also be visualized in a heat-map. The Word2Vec similarity function can be used to obtain a score between words. On the x-axis and y-axis of the map, the words in similar order are positioned. The color intensity would represent the similarity intensity.

CONCLUSION

The goal of this thesis was to gain deeper insights into the Twitter and Facebook post/comments data on the social media platform of a humanitarian organization. LDA and LSI were used to create grouped word topics. This approach is unique because it has not yet been used on text with character limits, like Twitter messages. Topic number was found to be optimal when it is larger, which confirms the hypothesis. In addition, it was shown that messages that contain common hashtags have a higher chance of being assigned to the same LDA output group. The word topics were subsequently projected using a combination of Word2Vec with PCA, t-SNE, and UMAP. These projections were presented in a two-dimensional way, whereby links between words from mixed languages within and between topics can be obtained. Abbreviations are an underestimated problem when social media data consists of different languages and dialects. The UMAP projection resulted in the most informative visualization.

REFERENCES

[1] About us. (2019, 19 August). Accessed at 19 August 2019, van https://www.rnw.org/about-us/

[2] Alhawarat, M., Hegazi, M. (2018). Revisiting K-Means and Topic Modeling, a Comparison Study to Cluster Arabic Documents. IEEE Access, 6, 42740–42749. https://doi.org/10.1109/access.2018.2852648

[3] Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77. https://doi.org/10.1145/2133806.2133826

[4] Blei, D. M., NG, A., Jordan, M. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022. Consulted from http://jmlr.org/papers/volume3/blei03a/blei03a.pdf

[5] Blei, D. M., Lafferty, J. D. (2006). Dynamic topic models. Proceedings of the 23rd international conference on Machine learning - ICML '06, 113–120. https://doi.org/10.1145/1143844.1143859

[6] Doctor, V. (2012, 12 June). What Characters Can A Hashtag Include? Accessed at 4 May 2019, van https://www.hashtags.org/featured/what-characters-can-a-hashtag-include/

[7] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. (1990), Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci., 41: 391-407. doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

[8] Hansen, C., Tosik, M., Goossen, G., Li, C., Bayeva, L., Berbain, F., Rotaru, M. (2015). How to get the best word vectors for resume parsing. SNN Adaptive Intelligence/Symposium: Machine Learning. Consulted from https://lenabayeva.files.wordpress.com/2015/03/snn-textkernelposter-2015.pdf

[9] He, J. (2012). Exploring topic structure. ACM SIGIR Forum, 46(1), 84. https://doi.org/10.1145/2215676.2215690

[10] Hu, B., Ester, M. (2013). Spatial topic modeling in online social media for location recommendation. Proceedings of the 7th ACM conference on Recommender systems - RecSys '13, 25–32. https://doi.org/10.1145/2507157.2507174

[11] Jeong, B., Yoon, J., Lee, J.-M. (2019). Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. International Journal of Information Management, 48, 280–290. https://doi.org/10.1016/j.ijinfomgt.2017.09.009

[12] Jolliffe, I. T. (2006). Principal Component Analysis. x: Springer New York.

[13] Joshi, P. (2020, 7 mei). Text Mining 101: A Stepwise Introduction to Topic Modeling using Latent Semantic Analysis (using Python). Consulted from https://www.analyticsvidhya.com/blog/2018/10/stepwise-guide-topic-modeling-latent-semantic-analysis/

[14] Karpathy, A. (2014, 2 juli). Visualizing Top Tweeps with t-SNE, in Javascript. Accessed at 10 June 2019, van https://karpathy.github.io/2014/07/02/visualizing-top-tweeps-with-t-sne-in-Javascript/

[15] Kelaiaia, A., Merouani, H. F. (2013). Clustering with Probabilistic Topic Models on Arabic Texts. Modeling Approaches and Algorithms for Advanced Computer Applications, 65–74. https://doi.org/10.1007/978-3-319-00560-7-11

[16] Kim, A. (2018, 11 oktober). Perplexity Intuition (and its derivation). Geraadpleegd op 18 januari 2020, van https://towardsdatascience.com/perplexity-intuition-and-derivation-105dd481c8f3

[17] Leightley, D. (2018, 1 August). Visualizing Tweets with Word2Vec and t-SNE, in Python. Accessed at 5 May 2019, van https://leightley.com/visualizing-tweets-with-word2vec-and-t-sne-in-python/

[18] McInnes, L., Healy, J., Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Consulted from https://arxiv.org/pdf/1802.03426.pdf

[19] Martin, M. E., Schuurman, N. (2017). Area-Based Topic Modeling and Visualization of Social Media for Qualitative GIS. Annals of the American Association of Geographers, 107(5), 1028–1039. https://doi.org/10.1080/24694452.2017.1293499

[20] Mikolov, T., Chen, K., Corrado, G. S., Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. Advances in neural information processing systems, 1–9. Consulted from http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

[21] Muntean, C. I., Morar, G. A., Moldovan, D. (2012). Exploring the Meaning behind Twitter Hashtags through Clustering. Business Information Systems Workshops, 231–242.

[22] Nalisnick, E., Mitra, B., Craswell, N., Caruana, R. (2016). Improving document ranking with dual word embeddings. Proceedings of the 25th International Conference Companion on World Wide Web, 83–84. Consulted from https://www.microsoft.com/en-us/research/wp-content/uploads/2016/04/pp1291-Nalisnick.pdf

[23] Naskar, D., Mokaddem, S., Rebollo, M., Onaindia, E. (2016). Sentiment Analysis in Social Networks through Topic modeling. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 46–53. Consulted from https://www.aclweb.org/anthology/L16-1008.pdf

[24] Newman, D., Chemudugunta, C., Smyth, P. (2006). Statistical entity-topic models. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06, 680–686. https://doi.org/10.1145/1150402.1150487

[25] Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11), 559–572. https://doi.org/10.1080/14786440109462720

[26] Peng, J., Zhou, Y., Sun, X., Su, J., Ji, R. (2019). Social Media Based Topic Modeling for Smart Campus: A Deep Topical Correlation Analysis Method. IEEE Access, 7, 7555–7564. https://doi.org/10.1109/access.2018.2890091

[27] Pleplé, Q. (2013 June). Perplexity To Evaluate Topic Models. Accessed at 11 May 2019, van http://qpleple.com/perplexity-to-evaluate-topic-models/

[28] Pritchard, J. K., Stephens, M., Donnelly, P. (2000). Advances in neural information processing systems. Genetics, 155, 945–959.

[29] Röder, M., Both, A., Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15, 399–408. https://doi.org/10.1145/2684822.2685324

[30] Rosa, K. V., Shah, R., Lin, B., Gershman, A., Frederking, R. (2011). Topical Clustering of Tweets. Proceedings of the ACM SIGIR: SWSM 63. Consulted from http://www.cs.cmu.edu/ kdelaros/sigir-swsm-2011.pdf

[31] Stevens, K., Kegelmeyer, P., Andrzejewski, D., Buttler, D. (2012). Exploring Topic Coherence over many models and many topics. Association for Computational Linguistics, 952–961. Consulted from https://www.aclweb.org/anthology/D12-1087.pdf

[32] Stilo, G., Velardi, P. (2014). Temporal Semantics: Time-Varying Hashtag Sense Clustering. Lecture Notes in Computer Science, 563–578.

[33] Tsur, O., Littman, A., Rappoport, A. (2012). Scalable multi stage clustering of tagged micro-messages. Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion, 621–622. https://doi.org/10.1145/2187980.2188157

[34] van der Maaten, L. (2014). Accelerating t-SNE using Tree-Based Algorithms. Machine Learning, 15, 3221–3245. Consulted from http://www.jmlr.org/papers/volume15/vandermaaten14a/vandermaaten14a.pdf

[35] van der Maaten, L., Hinton, G. (2012). Visualizing non-metric similarities in multiple maps. Machine Learning, 87, 33–55. Consulted from https://link.springer.com/content/pdf/10.1007/s10994-011-5273-4.pdf

[36] Vermont University, Javed, A. (2016). A Hybrid Approach to Semantic Hashtag Clustering in Social Media. Geraadpleegd van https://scholarworks.uvm.edu/cgi/viewcontent.cgi?article=1622context=graddis.

[37] Wagner, K. (2017, 7 November). Twitter's 280 character tweets are now available for everyone. Accessed at 3 June 2019, van https://www.vox.com/2017/11/7/16615914/twitter-longer-tweets-280-characters-update-available-everyone