Using Latent Dirichlet Allocation and Word embedding on Twitter data to gain insight in mindsets of young people.

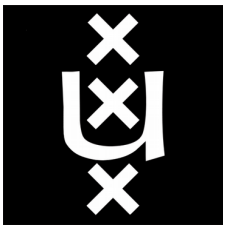
SUBMITTED IN PARTIAL FULLFILLMENT FOR THE DEGREE OF MASTER OF SCIENCE

Sije van der Veen
11422688

MASTER INFORMATION STUDIES
Data Science

FACULTY OF SCIENCE
UNIVERSITY OF AMSTERDAM

Date of defence



Academic Supervisor
Dr. Maarten Marx
UvA, FNWI, IvI

Industry Supervisor
Dr. Kyle Snyder
Radio Netherlands Worldwide Media

Using Latent Dirichlet Allocation and Word embedding on Twitter data to gain insight in mindsets of young people.

SIJE VAN DER VEEN, University of Amsterdam

Generating new topics to write about is not always easy for blog-writers. The linked social media data of the blog can be a solution. This paper describes a methodology to detect word topics within Twitter messages. Latent Dirichlet Allocation (LDA) is used to detect word topics, which is novel. Because it is has not been applied on 280 characters long Twitter messages. In two sub questions are the performances shown of LDA on Twitter messages. In the last experiment the word topics are projected using a combination of Word2Vec with Principal component analysis (PCA) or t-distributed Stochastic Neighbor Embedding (t-SNE). These projections are showing, in 2-dimensional way, the links between words within a topic. These links can be used by companies and organizations that wants to have insights in there followers thoughts.

CCS Concepts: • **Topic modeling → Latent Dirichlet Allocation**; • **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 Media (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 having young people consume this information and discuss it [1]. Since RNW Media wants to create content that 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 aims to tackle these problems by a data-driven approach.

Latent Dirichlet Allocation (LDA) is a well studied topic modeling method [19]. 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 contribute to 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:

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

(1) What is the optimal number of topics that can be generated? In the first sub-question, the number of topics generated by LDA is measured.
(2) Are hash-tags found in the same LDA output topic? In the second 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 similar output group?
(3) Can the relationship between words within a topic be visualized? In the last sub-question an word embedding will be creating based on the sentences from each topic. The word embedding will be used to create PCA, t-SNE and UMAP Projection of Word2Vec Model. In the projection the most frequent words will be shown.

The report will start with a related work section. What work has been done before, and/or what ideas and models have already been developed. 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.

2 RELATED WORK

2.1 Topic modeling with Latent Dirichlet Allocation

Since the introduction of LDA in 2000, this topic model method has been studied extensively [19]. This section will give 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 [2]. Different types of models were used for estimating continuous representation of words.

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 [15]. 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.[28] 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. [6] 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 [3]. In dynamic topic modeling the order of documents is retained.

2.2 Hash-tags clustering

In the second research question hash-tags will be used as true labels to verify the output of LDA model. In several studies hash-tags are clustered to analyze contextual semantics [23] [24] [16] [21]. For example in the study of Stilo and Velardi in which they created a clustering algorithm for hash-tags based on temporal mining [23]. They cluster hash-tags based on their temporal co-occurrence with other hash-tags. Tsur *et al.* created an algorithm which leverages users practice of adding tags to some messages by bootstrapping over virtual non sparse documents [24]. These non spare documents were then represented as vectors in the vector space model. Rosa et al. used hash-tag clusters to achieve topical clustering of tweets, where they compared the effects of expanding URLs found in tweets [21].

2.3 Word embeddings projections

In the last part of study, the application is introduced that is using Word2Vec on PCA, t-SNE and UMAP. Word embeddings are extensively studied on different text sources, from Web search [17] till Curriculum Vitae [7]. A combination of PCA projection using Word2Vec and Twitter messages is used in the tutorial of Leightley [13]. Another dimensional reduction method used in this tutorial is t-SNE which stands for t-distributed Stochastic Neighbor Embedding. t-SNE is multidimensional scaling method developed by Laurens van der Maaten and is used for dimension reduction approaches [26] [25]. Andrej Karpathy also used projection using t-SNE on Twitter data but in combination with JavaScript [12].

3 METHODS

In this section the accessed methodologies in the study are explained. The topic model method LDA will be explained in the first subsection. Followed by the evaluation method perplexity, which will be used to test the LDA model on Twitter data. Followed by the explanation of the weighted average of pair-wise maximum f-score. Finally Word2Vec, PCA and t-SNE and UMAP will be explained, which are used for the word projections.

3.1 Latent Dirichlet Allocation

LDA is a generative probabilistic model for the collection of data types like a text corpora. The model is structure can be explained as a three-level hierarchical Bayesian model. Each item/document in a collection is treated as mixture over an underlying set of topics. Each topic is treated as a mixture over an underlying set of term probabilities. In other words each document is treated as a mixture of different topics. Each document is considered to have a set of topics that are allocated by LDA [19].

3.2 Evaluating the model using Perplexity

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

For LDA, a test set is a collection of unseen documents w_d, and the model is described by the topic matrix and the hyperparameter for topic-distribution of documents. The LDA parameters is not taken into consideration as it represents the topic-distributions for the documents of the training set, and can therefore be ignored to compute the likelihood of unseen documents. Therefore, we need to evaluate the log-likelihood [18]

$$L(w) = log(w|\phi, \alpha)\Sigma_d logp(w_d|\phi, \alpha) \tag{1}$$

The measure traditionally used for topic models is the *perplexity* of held-out documents w_d defined as shown in formula 2. Which is a decreasing function of the log-likelihood L(w) of the unseen documents w_d; the lower the perplexity, the better the model

$$perplexity(testsetw) = exp(\frac{L(w)}{countoftokens}) \tag{2}$$

3.3 Clustering comparison using f-scores for evaluation

The weighted average of pair-wise maximum f^a-score is used as evaluation method in subsection 2. To score is obtained by calculating for each hash-tag the f^a-score. The f^a-score is based on the precision and recall.

Precision is defined for one cluster the ratio of correct hash-tags in the cluster compared to the total number of hash-tags in the cluster. In the precision formula is G_j. The ground truth cluster and C_i

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

The recall score for a cluster is based on the number of correct hash-tags in the cluster compared to the number of correct hash-tags that should be in the cluster (the growth truth cluster). The precision of a cluster C_i is calculated as follows with taking in account ground truth cluster G_j:

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

The f^a-score is based on the precision and recall and is defined as following:

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

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 will be calculated. For the calculation the weighted average of pairwise maximum f-score are used (each pairwise maximum f-score is weighted by the size of the matching ground truth cluster). For a set of clusters C and a set of ground truth cluster set G, the f^a-score. Is formulated as following:

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

3.4 Word2Vec

Word2Vec is a two-layer neural network that is designed to processes text. It's input is a text corpus and its output is a set of vectors: feature vectors for words in that corpus. Word2Vec converts text into a numerical form that can be understood by a machine.

3.5 PCA

PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by some projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate and so on [11].

3.6 t-SNE

t-SNE stands for t-distributed Stochastic Neighbor Embedding. The goal 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 [13]. t-SNE does not preserve global data structure, meaning that only within cluster distances are meaningful while between cluster similarities are not guaranteed.

3.7 UMAP

UMAP (Uniform Manifold Approximation and Projection) is a novel manifold learning technique for dimension reduction. UMAP can be used to visualizing general non-linear dimension reductions. The algorithm assumes that 3 characteristics about the data. 1. The data is uniformly distributed on Riemannian manifold. 2. The Riemannian metric is locally constant. 3. The manifold is locally connected. If the data has these characteristics the manifold can be model with a blurry topical structure. The embedding is found by searching for a low dimensional projection of the data that has the most comparable blurry topological structure. [14]

4 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 setups of three sub-questions. In figure 1 is a Flowchart with the connection between the three sub-questions of this study. The blocks describes the steps taken per question.

4.1 RNW Media Burundi Twitter data-set

The data-set which will be used in all sub-questions is the RNW Media Burundi Twitter data-set. Which consist of Twitter messages of the RNW Media Burundi Twitter platform. The messages are mainly written in French, English and African languages. The data-set is filled with 464845 Twitter messages. The Twitter message is in the first column and in a second column the hash-tags for that message. One message can have multiple hash-tags. The Twitter

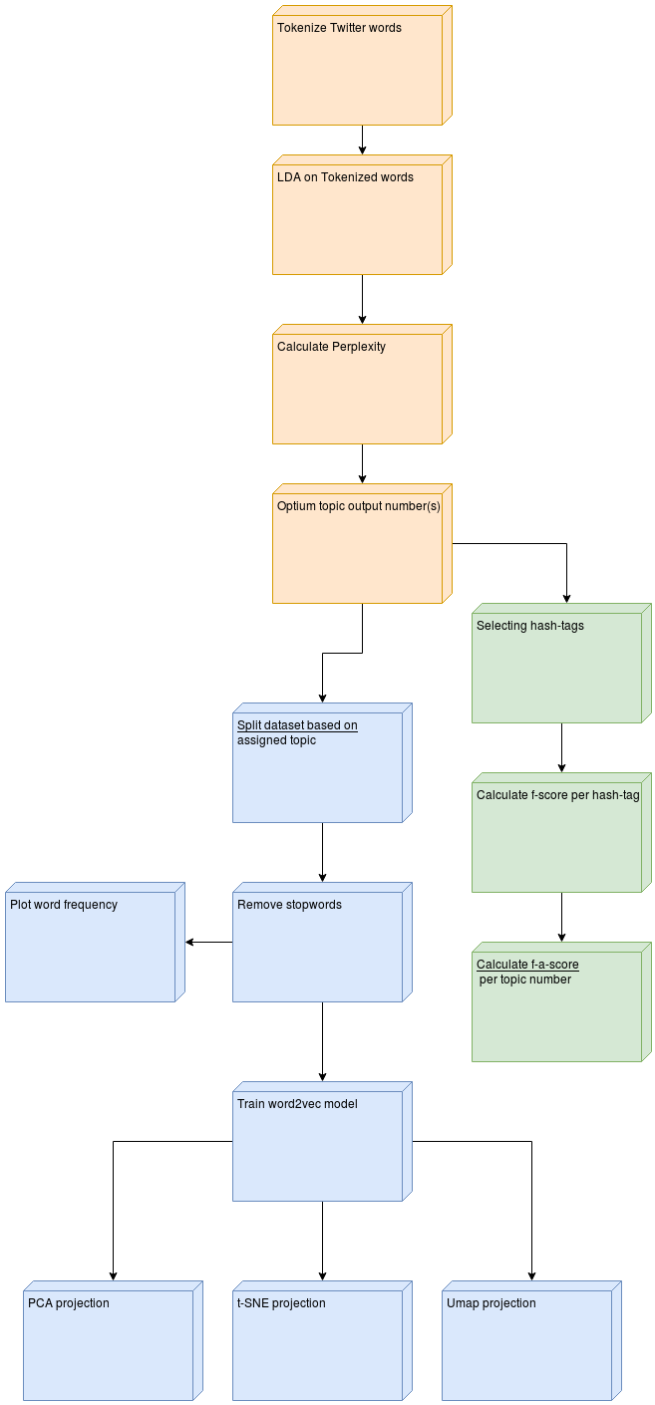


Fig. 1. Flowchart with the connection between the three sub-questions of this study. The yellow coloured blocks indicates the steps taken in sub-question 1: Finding optimal output topic number. The green coloured blocks indicates the steps taken for sub-question 2: Are hash-tags found in the same LDA output topic. The blue coloured blocks indicates the steps taken in sub-question 3: Visualizing word relationships in Word2Vec embedding. In third sub-question will the RNW Media Burundi Twitter data-set and the RNW Media Burundi Facebook data-set be used.

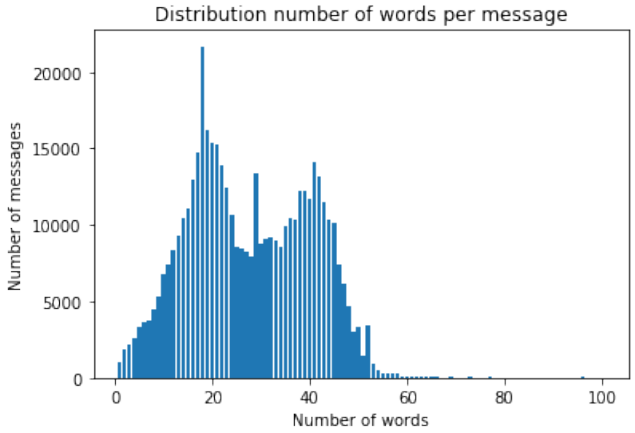


Fig. 2. The distribution of words per message is shown. Most messages contains 18 words (4.66 percent). Most messages contains between 13 and 44 words. From the figure can be clearly seen that the message size is small

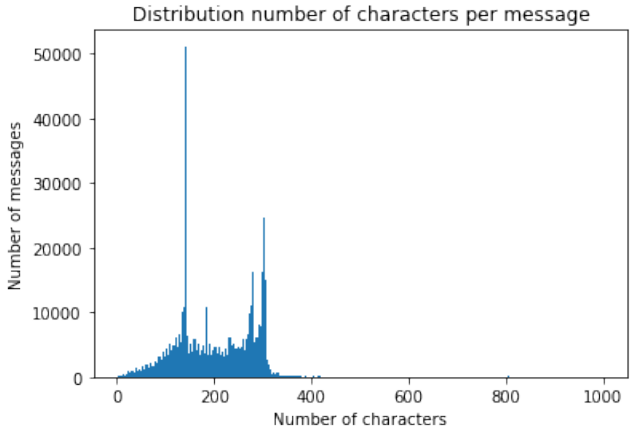


Fig. 3. The distribution of characters per message is shown. Most messages contains between the zero and 280 characters. Twitter founders created Twitter message with a size of 160 characters. So it could a message of 140 character and a 20 character for a user name. There is a pike around 160 characters, which could be related to previous text size of 160 characters. In 2018 the maximum message size is changed to 280 characters [27].

messages are posted between January 2018 and May 2019. The data-set did not contain missing values. In figure 2 the distribution is shown for the number of words per message. From the figure can be clearly seen that the message size is small. The most messages contains 18 words (4.66 percent) and the most messages contains between 13 and 44 words.

In figure 3 the distribution been seen for the number of characters per message. Most messages contain between zero and 280 characters. Twitter founders created Twitter message with a size of 160 characters. So it could be a message of 140 characters and a 20 character for a user name. There is a pike around 160 characters, which could be related to previous text size limit of 160 characters.

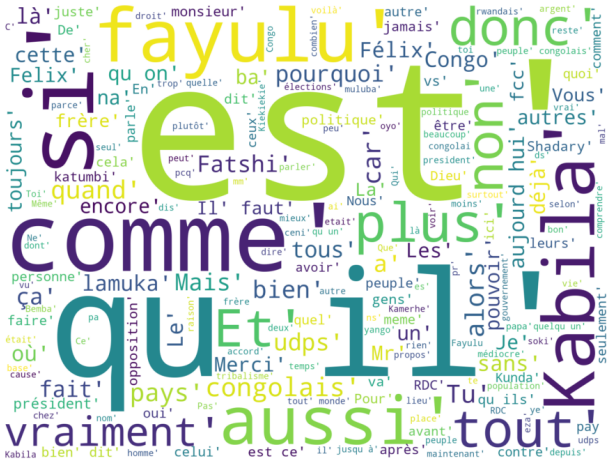


Fig. 4. Word cloud which shows the common words in RNW Media DRC Facebook data-set. The world cloud is generated before the English, French, Spanish and African stop words were removed. Which can been seen by the many French words like comme (as), aussi (also) and donc (therefore). Also decision makers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

In 2018 the maximum message size is changed to 280 characters [27]. The length changed to 280 characters can be seen in figure 3.

In figure 5 is a word cloud shown with the 20 most common occurring hash-tags in the data-set. The hash-tags are mainly related to countries, political parties and political events. In table 1 are the most frequent hash-tags in the data-set explained. A selection of this hash-tags will be used as true labels in sub-question 2.

4.2 RNW Media DRC Facebook data-set

In last sub-section the RNW Media Burundi Facebook data-set will be used. This data-set consists out Facebook posts and messages from the RNW Media Facebook platform which is active in the Democratic Republic of the Congo (DRC). The 91579 Facebook posts and messages are posted between January 2018 and May 2019. In the word cloud shown in figure 4 are the most common words shown in the data-set. In the word cloud many French words shown and also decision makers from DRC like a president and lawmaker.

4.3 Finding optimal output topic number

The goal of this section is determine the settings on which LDA achieves the best results on this data-set. One setting is the number of output topics, which should be determined. By testing the LDA model with different output settings on the data. Testing on the total data-set is costly and not necessarily. Therefore the different settings will be tested on similar sized subset. Topic output size in a range between 2 and 15 is selected, because the hypothesis is that the size of topic output is optimal when it is small. Perplexity score will used as evaluation.

Hash-tag	Explanation
Burundi	Country
Cpi	Burundi political party
burundicrisis	
Burundialerte	Political tension
Silentmajority	People who do not express their opinions publicly
Rdc	Country
Boost	Heroin-derived drug
Bujumbura	Largest city and main port of Burundi
Lomasleido	Political related
Rwanda	Country
Ndondeza	Burundi political party
Nkurunziza	President of Burundi since 2005
Sindumuja	Organization
Gbagbo	Ivorian politician
Bjp	Indian political party
Affiliate	Marketing strategy
Referendum2018	referendum about the presidents power in Burundi
Jnu	University in India
Imbonerakure	Political group
Bsp	Indian political party
Bemba	Politician in Democratic Republic of the Congo
Venezuela	Country
Leyestatutariajep	Law in Venezuela
Cpa	Certified Public Accountants
Gitega	Gitega is one of the 18 provinces of Burundi.
Gabon	Country on the west coast of Central Africa
Inflation	Sustained increase in the general price level of goods and services in an economy over a period of time
Fatoubensouda	Gambian lawyer and international criminal law prosecutor

Table 1. Most common hash-tags in the Burundi Twitter data-set with explanations. 65927 message did not have a hash-tag attached in the data-set. The hash-tags are mainly related to country names and political parties. But also a hash-tag of a university in India, the country Venezuela en law of this country occurs often in the data-set.

Fig. 5. Word cloud with the 30 most common hash-tags in the data-set. The size of the word is equivalent on the number of occurrence in the data-set. The hash-tags are mainly related to country names and political parties. Explanations of the hash-tags can be found in table 1

4.4 Are hash-tags found in the same LDA output topic

A hash-tag is a type of metadata tag used on social networks such as Twitter and other social media services, allowing 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 hash-tag. [27] [5] Can LDA distinguish between Twitter messages using hash-tags as true labels. The hypothesis is that messages with in common hash-tag have higher change to be assigned to the same topic group. For the this question the f^a-score will be used for evaluation [9] [10]. This score is based on the recall and precision. Since precision and recall necessarily depend on the notion of true classes, a selected group of 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.

The recall and precision are used to calculate the f-score, which is only for one specific hash-tag (cluster) To estimate the score of all hash-tags (clusters) the f^a-score is used. In table 1 explanations can be found of the selected hash-tags.

4.5 Visualizing word relationships in Word2Vec embedding

In this sub-section relationships between important words within a generated topic are visualized. This will be done by selecting the most frequent words per topic. These words will be visualized in projection of Word2Vec model. Three different dimension reduction methods will be used, which are PCA, t-SNE and UMAP. To obtain these words English, French, Spanish and African stop words has been removed. Also common Twitter messages has been filtered out.

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.

5.1 Finding optimal output topic number

This question should answer what number of topics is optimal. To evaluate our LDA model, perplexity is used, which is the common evaluation method for LDA [4] [22] [20].

The number of topics which will be generated by the LDA algorithm is written as N. 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 N. If key-words are in multiple topics, this can indicates that N is too high.

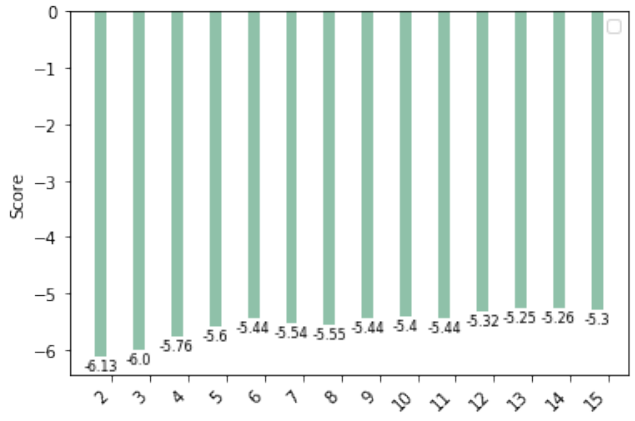


Fig. 6. Perplexity per topic number. a lower perplexity suggests a better fit. The lowest perplexity is achieved using k = 2. Which results in perplexity score of -6.13.

In figure 6 the perplexity per topic number generated by LDA is shown. *perplexity* of held-out documents w_d defined as shown in formula 2. That 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]. The lowest perplexity is achieved using N = 2. Which results in perplexity score of -6.13.

5.2 Are hash-tags found in the same LDA output topic

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

For this question the f^a-score will be used for evaluation [5] [6]. 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 hash-tags will be used as true classes. In table 1 explanations can be found of the selected hash-tags.

In table 2 hash-tags are shown with a topic size of 2 and 3. For the topic size 2 all f-scores are between 0.594 and 0.660, which means that all f-score are above random (0.5). Hash-tag Jnu used for an University in India has the highest score. There is no direct link between this university and Burundi, which could explain this high score. For topic output size 3 the f-scores are smaller, also the random score is lower. For the comparison one hash-tag is used. If a message had multiple hash-tags, the first hash-tag was selected.

In table 4 are for topic sizes 2, 3, 4 and 5 the f^a-scores given. Which are compared with random score. The random score is the chance that a hash-tag is assigned to a cluster. For topic size 3 is the biggest difference between f^a-score and the random score. For topic size 2 is the smallest difference between f^a-score and the random score.

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	0.625	0.551

Table 2. For selected number of hash-tags are f-scores given using a topic number of 2 and 3. For the topic size 2 all scores are between 0.594 and 0.660, which means that all f-scores are above random (0.5). The f-score for topic size are lower, which is expected because score is on one more 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 selected number of hash-tags are f-score and the f^a-score is given using a topic number of 4 and 5. These scores are lower than the scores using a topic number of 2 or 3. But f^a-scores for topic 4 (0.406) is compared to the random score of (0.25) percentage beter scoring, because there is an increase of 62.0 percent

Topic size	f^a score	random score	percent > random
2	0.625	0.5	+0.25
3	0.551	0.34	+0.65
4	0.406	0.25	+0.62
5	0.285	0.20	+0.425

Table 4. For topic sizes 2, 3, 4 and 5 are the f^a-scores given. Which are compared with random score. The random score is the chance that a hash-tag is assigned to a cluster. For topic size 3 is the biggest difference between f^a-score and the random score. For topic size 2 is the smallest difference between f^a-score and the random score.

5.3 Visualizing word relationships in Word2Vec embedding

5.3.1 RNW Media Burundi Twitter data-set.

To find word relationship between words in the RNW Media Burundi Twitter data-set the most frequent words per topic is selected. In

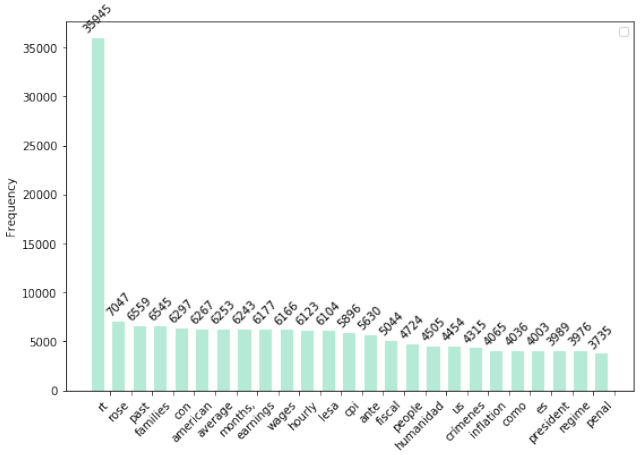


Fig. 7. Frequency of the most common words in Twitter message which are assigned to topic 2. These words are obtained to loop over all the Twitter messages which were assigned to topic 2. The 25 most occurring words are shown in this figure on the x-axis. The number of occurrence is shown on the y-axis. RT an Russian TV channel is again the most common word in this topic. This topic contains more French and Spanish words compared to the other topic. Crime and human related words are occurring often in this topic like crimes, militares, humanitaria, victimes, police, regimes and droits.

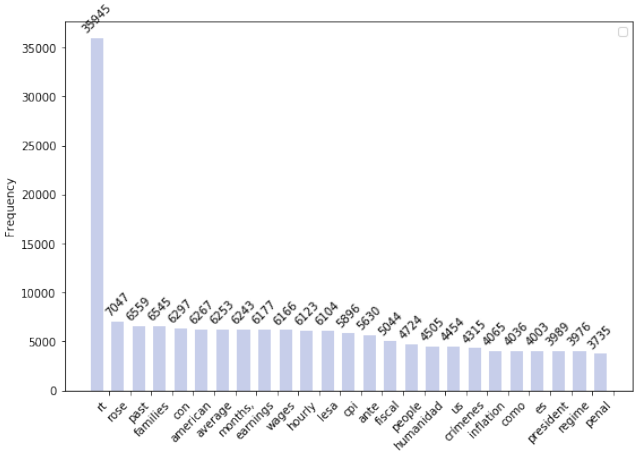


Fig. 8. Frequency of the most common words in Twitter message which are assigned to topic 1. These words are obtained to loop over all the Twitter messages which were assigned to topic 1. The 25 most occurring words are shown in this figure on the x-axis. The number of occurrence is shown on the y-axis. RT a Russian TV channel is the most common word in this topic. On the third position is the word rose. Besides that many finance related words are occurring often in this topic, like earnings, wages, inflation. Political words like president, regime, cpi and party can be found in this topic.

figure 7 are the frequencies of the 25 most common words in topic 1 shown. Many finance and political related words are in the top 25. For the second topic the same steps are taken. In figure 8 are frequency of the most common words shown for topic 2. Many of these words are related to crime and humans. Like crimes, militares (soldiers), humanitana (humanitarian), victimes (victims), police, regimes and droits (rights). Another observation is that topic 2 consist of more spanish and french words.

The most frequent words of both topics are shown in PCA projection of Word2Vec model in figure 9. In the left of the projection is a green cluster. Words like humanitana, victimes, commune, president and regime are close in space. In the bottom of the picture is a blue cluster with words as fiscal and hourly. In figure 10 is t-SNE projection shown. The most green words are occurring in the left side of the projection. The blue words are more spread out. The words police, regime and crimes are clustered together. The UMAP projection is shown in figure 11. Besides chambre (room) and politique (politics) are the green words clustered together. Humanitaria, victimes, aseguro (assure) and bloqueen (block) are close in position.

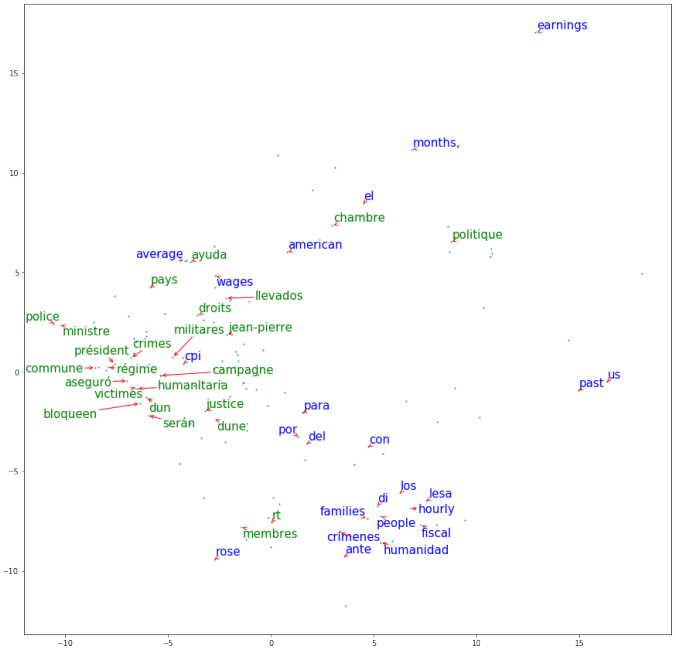


Fig. 9. PCA projection of Word2Vec Model. The 25 most occurring words per topic are highlighted in green for topic 1 and blue for topic 2. The position is shown with a red arrow. In left of the projection is a green cluster. Words like humanitana, victmes, commune, president and regime are close in space. In the bottom of the picture is a blue cluster with words as fiscal and hourly.

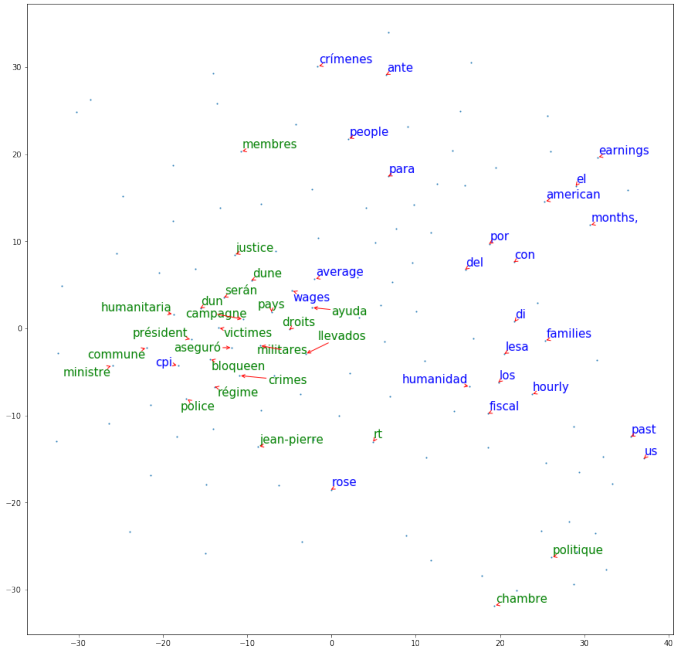


Fig. 10. t-SNE projection of Word2Vec Model. The 25 most occurring words per topic are highlighted in green for topic 1 and blue for topic 2. The position is shown with a red arrow. The most green words are occurring at the left side of the projection. The blue words are more spread out. The words police, regime and crimes are clustered together.

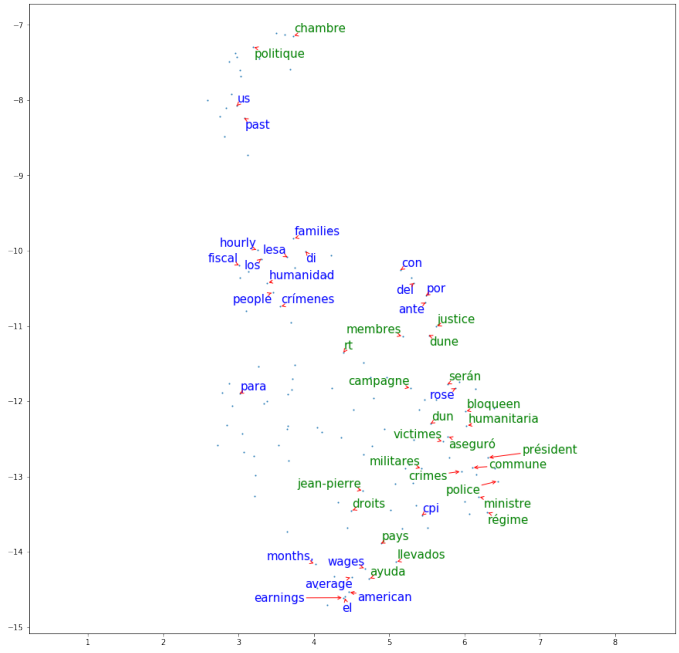


Fig. 11. UMAP projection of Word2Vec Model. The 25 most occurring words per topic are highlighted in green for topic 1 and blue for topic 2. The position is shown with a red arrow. The green coloured words are clustered, besides chambre and politique. Humanitaria, victimes, aseguro (assure) and bloqueen (block) are close positioned.

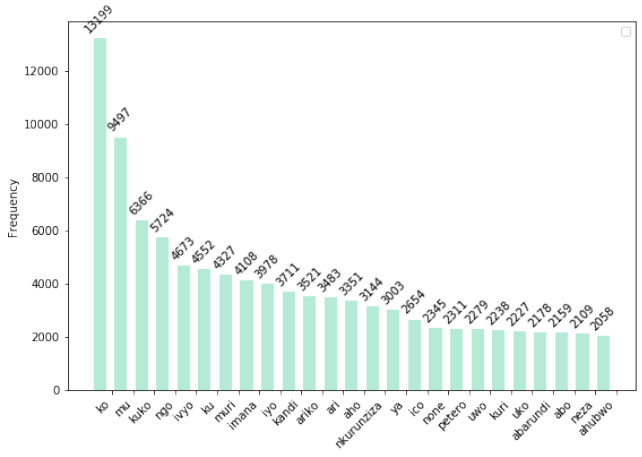


Fig. 12. Frequency of the most common words in Facebook message which are assigned to topic 1. These words are obtained to loop over all the Facebook messages which were assigned to topic 1. This topic contains mostly African words.

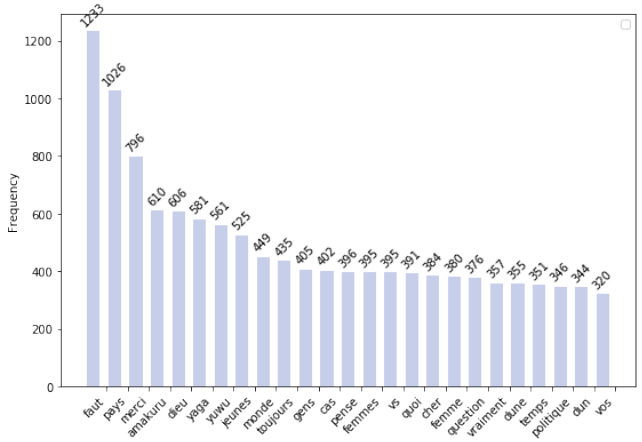


Table 5. Frequency of the most common words in Facebook message which are assigned to topic 2. These words are obtained to loop over all the Facebook messages which were assigned to topic 2. This contains many French words like femmes, chere and pense.

5.3.2 RNW Media DRC Facebook data-set. Besides a Twitter forum, RNW Media is also operating on Facebook in DRC. Posts and messages from this Facebook platform are stored in the RNW Media DRC Facebook data-set. For this data-set the same steps are taken to find word relationships. Two topics are created. The frequency of most common words for topic 1 are shown in figure 12 and for topic 2 in figure 5. Topic 1 contains many African words. Topic 2 contains more french words, like femmes (woman), chère (dear) and pense (thought). The most frequent words of both topics are shown in PCA projection of Word2Vec model in figure 13. In this projection is on the left side a cluster of words formed. In figure 14 t-SNE projection is shown. In this projection the word are spread over projection. The UMAP projection is shown in figure 15, which

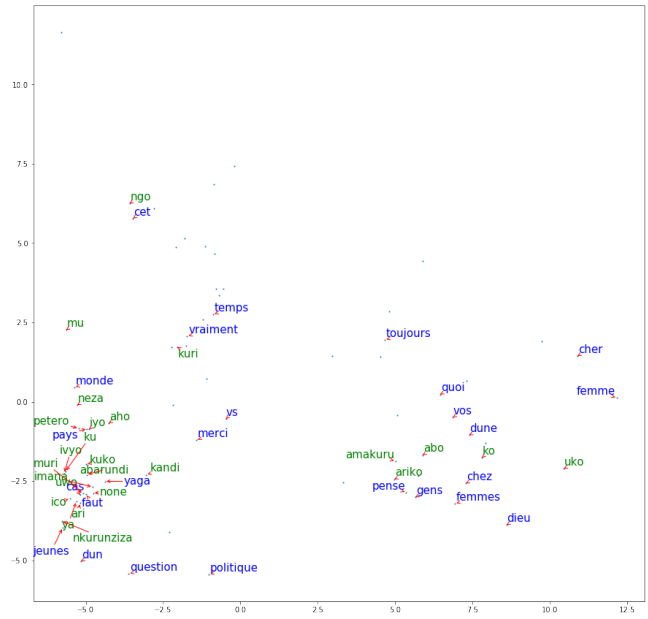


Fig. 13. Scatter Plot of PCA projection of Word2Vec Model. The 25 most occurring words per topic are highlighted in green for topic 1 and blue for topic 2. The position is shown with a red arrow. In the left side of the projection is a cluster formed with words like jeunes (youth), Yaga (RNW Media social media platform) and Nkurunziza (President of Burundi since 2005)

shows two clear separated clusters. The top cluster contains some political related words like, politique and Nkurunziza (President of Burundi since 2005).

5.3.3 Future improvements.
In further research different Twitter data-sets should be used for testing. On the Brurundi Twitter data-set, 2 topics would give solid results. It could be possible that on a larger data-set more topics could be generated by the LDA algorithm. In the Burundi Twitter data-set mostly English, French and Spanish words occur. These languages are easily to deal with. RNW Media is also working with languages with more complex structures like Arabic, Hindi and Rundi. A future study could be continued on data-sets which contains these languages.
In addition to a word embedding projection, the word similarity could also be visualized in a heat-map. The Word2Vec similarity function can used to obtain a score between words. On the x-axis and y-axis of the map are the words in similar order. The colour intensity will be the similarity intensity.

CONCLUSION

The goal of this thesis to gain deeper insight in the Twitter and Facebook post/comments data of social media platform of a humanitarian organisation. LDA was used to create grouped word topics, which is unique because it is not yet used on text sizes like Twitter messages. Output size is optimal when it is small, which meets the hypothesis. In addition is shown that messages which contain in

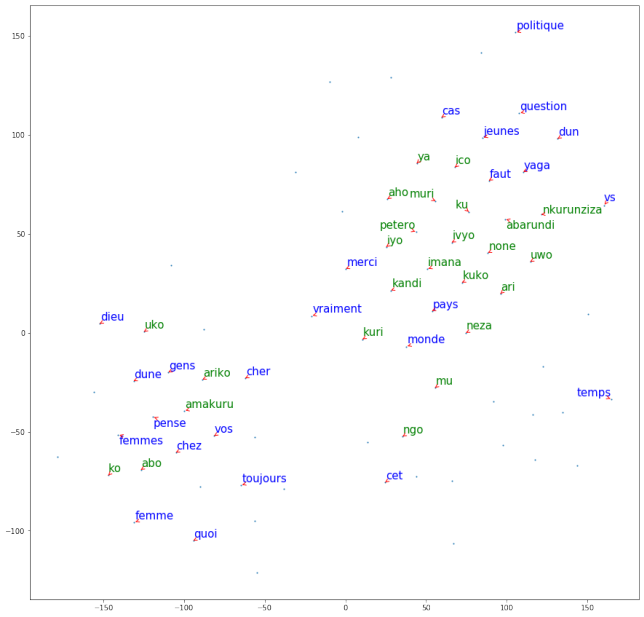


Fig. 14. Scatter Plot of t-SNE projection of Word2Vec Model. The 25 most occurring words per topic are highlighted in green for topic 1 and blue for topic 2. The position is shown with a red arrow. In this topic can been seen that there is relationship between the words from both topics, which contain both different languages. For example there is a close relationship between amakuru (major) and pense (thought).

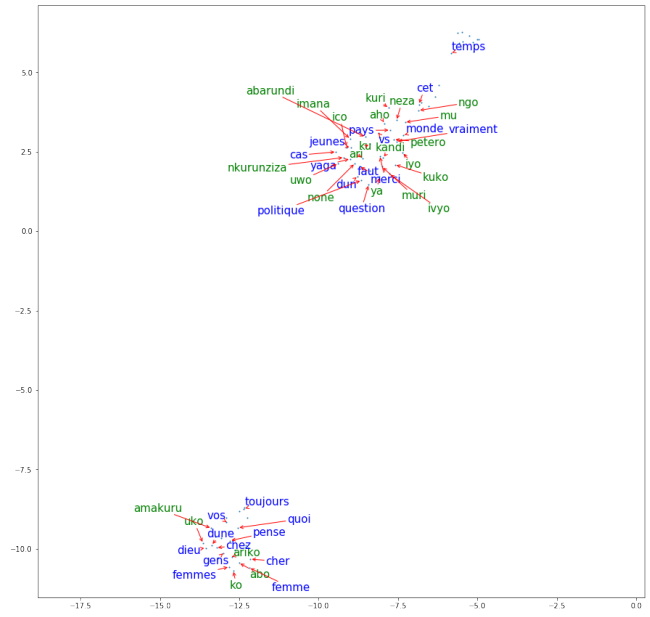


Fig. 15. Scatter Plot of UMAP projection of Word2Vec Model. The 25 most occurring words per topic are highlighted in green for topic 1 and blue for topic 2. The position is shown with a red arrow. The projection shows two clear clusters with a mixture of both topics. The top cluster contains some political related words like, politique and Nkurunziza (President of Burundi since 2005)

common hash-tags have higher change to be assigned to the same LDA output group. The word topics are subsequently projected using a combination of Word2Vec with PCA, t-SNE and UMAP. These projections are shown in a 2-dimensional way, whereby links between words from mixed languages within and between topics can be obtained. The UMAP projection resulted in the most informative visualizations.

REFERENCES

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

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

[3] 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

[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] 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/

[6] Griffiths, T. L., Steyvers, M., Blei, D. M., Tenenbaum, J. B. (2005). Integrating Topics and Syntax. Advances in Neural Information Processing Systems, 17, 537–544. Consulted from https://cocosci.princeton.edu/tom/papers/composite.pdf

[7] 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

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

[9] Javed, A., Lee, B. S. (2017). Sense-Level Semantic Clustering of Hashtags. Information Management and Big Data, 1–16. Consulted from https://arxiv.org/pdf/1802.03426v2.pdf

[10] Javed, A., Lee, B. S. (2018). Hybrid semantic clustering of hashtags. Online Social Networks and Media, 5, 23–36. https://doi.org/10.1016/j.osnem.2017.10.004

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

[12] 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/

[13] 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/

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

[15] 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

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

[17] 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

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

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

[20] 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

[21] 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
[22] 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
[23] Stilo, G., Velardi, P. (2014). Temporal Semantics: Time-Varying Hashtag Sense Clustering. Lecture Notes in Computer Science, 563–578.
[24] 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
[25] 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
[26] 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
[27] 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
[28] Wallach, H. M. (2006). Topic modeling. Proceedings of the 23rd international conference on Machine learning - ICML '06, 23, 977–984. https://doi.org/10.1145/1143844.1143967