TWITTER TRENDS ANALYSIS USING STRUCTURAL TOPIC MODELLING

**BY**

**ALEX MUIRURI MWANGI**

**A RESEARCH PROPOSAL** **SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF MASTER OF SCIENCE DATA ANALYTICS, KCA UNIVERSITY**

**MARCH, 2019**

**DECLARATION**

I declare that this research proposal is my original work and has not been presented for degree in any other university.

Sign: ……………………………………….. Date: ………………………

**ALEX MWANGI**

This research proposal has been submitted for examination with my approval as university supervisor

Sign: …………………………………… Date: ………………………….

**Supervisor**

**KCA UNIVERSITY**

# TABLE OF CONTENTS

[**DECLARATION** i](#_Toc4408697)

[TABLE OF CONTENTS ii](#_Toc4408698)

[LIST OF FIGURES iv](#_Toc4408699)

[LIST OF ACRONYMS AND ABBREVIATIONS vi](#_Toc4408700)

[CHAPTER ONE 1](#_Toc4408711)

[INTRODUCTION 1](#_Toc4408712)

[1.1 Background of the Study 1](#_Toc4408713)

[1.2 Statement of the Problem 3](#_Toc4408714)

[1.3 Motivation of the Study 4](#_Toc4408715)

[1.4 Main Objective 4](#_Toc4408716)

[1.5 Specific Objectives 5](#_Toc4408717)

[1.6 Research Questions 5](#_Toc4408718)

[1.7 Significance of the Study 5](#_Toc4408719)

[1.7.1 Social Scientists 5](#_Toc4408720)

[1.7.2 Companies 6](#_Toc4408721)

[1.7.3 Government 6](#_Toc4408722)

[1.7.4 Other Researchers 6](#_Toc4408723)

[1.8 Scope of the Study 6](#_Toc4408724)

[CHAPTER TWO 7](#_Toc4408725)

[LITERATURE REVIEW 7](#_Toc4408726)

[2.1 Introduction 7](#_Toc4408727)

[2.2 Theoretical review 7](#_Toc4408728)

[2.2.1 Evolution of Topic Models 7](#_Toc4408729)

[2.2.2 LDA 7](#_Toc4408730)

[2.2.3 Computational Methods. 9](#_Toc4408731)

[2.2.4 Extensions to LDA 9](#_Toc4408732)

[2.2.5 The correlated topic model 10](#_Toc4408733)

[2.2.6 The dynamic topic model 10](#_Toc4408734)

[2.2.7 Author-topic model 11](#_Toc4408735)

[2.2.8 Structural topic model 11](#_Toc4408736)

[2.3 Empirical Review 11](#_Toc4408737)

[2.4 Conceptual Framework 13](#_Toc4408738)

[CHAPTER THREE 15](#_Toc4408739)

[RESEARCH METHODOLOGY 15](#_Toc4408740)

[3.1 Introduction 15](#_Toc4408741)

[3.2 Research Design 15](#_Toc4408742)

[3.3 Data preprocessing 16](#_Toc4408743)

[3.3.1 Stopword Removal 16](#_Toc4408744)

[3.3.2 Stemming and Lemmatization 16](#_Toc4408745)

[3.3.3 Tokenization 17](#_Toc4408746)

[3.3.4 Converting to Lowercase 17](#_Toc4408747)

[3.4 Choosing Parameters 17](#_Toc4408748)

[3.5 Data Analysis 17](#_Toc4408749)

[3.6 Model Evaluation 19](#_Toc4408750)

[REFRENCES 19](#_Toc4408751)

[APPENDICES 22](#_Toc4408752)

[Appendix I: Budget 22](#_Toc4408753)

[Appendix II: Work Plan 23](#_Toc4408754)

# LIST OF FIGURES

[Figure 1 : Conceptual Framework 15](#_Toc3897839)

[Figure 2: Research Design 16](#_Toc3897840)

[Figure 3: Functions within the stm package form Roberts et al. (2015) 20](#_Toc3897841)

# LIST OF ACRONYMS AND ABBREVIATIONS

# TM Topic Modelling

# NLP Natural Language Processing

# BoW Bag of Words

# LDA Latent Dirichlet Allocation

# DTM Document Term Matrix

# TF-IDF Term Frequency – Inverse Document Frequency

# SVD Singular Value Decomposition

# LSA Latent Semantic Analysis

# CTM Correlated Topic Model

# STM Structural Topic Model

# CHAPTER ONE

# INTRODUCTION

## 1.1 Background of the Study

Since the invention of the internet over two decades ago, the world has experienced an unprecedented increase in the amount of information generated. Most of this information is in text format. Exploring this data manually would require much effort and would be too time consuming. This has led to a great deal of research in the domain of text mining to assist users in gaining insights from the ever growing textual data. Topic Modelling (TM) is a text mining method that has gained prominence in recent years. Topic modelling has shown to be able to give insights on huge corpus of textual data and hence improving exploration of unknown data, as an alternative to a traditional search engine.

Topic modelling was developed as an alternative to keyword search to enhance the exploration of text data collections (Deerwester, Furnas, Landauer, & Harshman, 1990). TM derives latent topics and patterns from textual data.TM has proven effective in summarizing large amounts of information and has been proposed as a solution to make long conversations, like microblogs, more approachable. This statistical model consists of the topics that appear in the data presented as a group of keywords sorted in their influence in forming the topic. TM also contain probabilities of each topic occurring in each of the document that can be used to filter all posts containing a particular topic.

Social Networking Sites (SNS) like Facebook and Twitter are a recent phenomenon that has transformed many aspects of our daily lives. Every second an average of 6000 Tweets are posted; this translates to about 500M Tweets per day1.[Internet Live Stats](https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/)

Twitter has become an indispensable source of news and information for a wide variety of users. People use Twitter to gather real-time news, follow events of interest and read updates by people they follow. The social networking site has become important for broadcasting breaking news and eyewitness accounts. Users also utilize the site for product marketing, product reviews and dissemination of news.

Twitter users have are used to receiving real-time updates on important local and global events. For example, Twitter was utilized to distribute information in many emergency and crisis situations such as the terrorists attack on the Dusit Complex in Kenya and the aftermath of the Kenyan elections in 2017. In addition, many companies and famous personalities use their Twitter accounts to keep in touch with clients and fans.

Being a mainstream Social Networking Site, Twitter offers researchers numerous prospects for research in text mining and Natural Language Processing (NLP) (Benhardus & Kalita, 2013). One such aspect is trending topics, highlighted on Twitter’s home page. They represent what is currently popular in users’ tweets. Studying the characteristics and content of these tweets is important to aid in important research such as detection of breaking news, recommending personalized messages, recommending friends, sentiment analysis among others.

It is important to analyze the huge amount of social media data generated daily to obtain meaningful information especially during any crisis and emergency situations. Topic Models are adept in summarizing, exploring and indexing large text document collections and can be used for this purpose (Manna, S., & Phongpanangam, O. (2018)).

## 

## 1.2 Statement of the Problem

Microblogging sites such as Twitter restrict the number of characters that a post can contain. For example, Twitter restricts the length of a post to between 140 and 280 characters. Due to this, microblogged messages have unconventional syntax and structure. The magnitude Twitter means that it can create a dynamic corpus. Due to these factors Twitter data presents several challenges not present in traditional analysis. Twitter messages are often characterized by the use of obscure Language and grammar because users often omit proper punctuation and use improper grammar in their posts. Tweets usually include shortened words and URLs, abbreviations and informal lingo such as “IRL” for “in real life”. The messages are also short and hence they contain very little grammatical structure. In addition the messages usually allude to diverse and specific events and locations and thus pre-defined entity recognition methods cannot be used.

In this study, I will investigate the use of Topic models for analyzing Twitter data. I propose that topic models are especially appropriate for analyzing Twitter data for various reasons. Firstly, topic models rely on the “bag-of-words” assumption. This means that they discard the word order and syntactic structure in the language. This makes them particularly suited to handle the improper grammar and obscure language contained in Twitter posts. Topic models can also infer latent (hidden) meanings in the data. This makes them sturdier in handling acronyms, slang and other idiosyncrasies in Twitter posts. The output of topic models are numerical vectors such as probability distributions. This makes them suitable for analysis, visualization as well advanced machine learning like clustering. Lastly, topic models are unsupervised algorithms. This makes them easily retrain able on other text data for a particular domain.

## 1.3 Motivation of the Study

Twitter has become an essential communication tool for diverse people across the world. Important events in the society are increasingly found in the timelines of individual people in Social Networking Sites such as Twitter. Trending topics utilize social media to provide a snapshot of topics and issues currently popular with users in the online community.

Researchers wish to use social media to infer users’ interests, model complex social relationships, follow news stories and identify developing topics. Companies want to use the messages posted to gain a competitive advantage, learn from their customers, better target marketing efforts and infer customers’ sentiment. Topic models are powerful algorithms to understand hidden patterns in the messages.

## 

## 1.4 Main Objective

The research proposes to implement a topic model of Twitter trending topics.

## 1.5 Specific Objectives

* To investigate how topic models can be applied to analyze Twitter trending topics.
* To establish the appropriate text preprocessing techniques and topic modelling algorithms to apply to Tweets.
* To design and develop a topic model of Twitter trending topics.
* To test and validate the effectiveness of the topic model for extracting the relevant topics from Twitter trending topics.

## 1.6 Research Questions

* How can topic models be applied to analyze Twitter trending topics?
* What are the appropriate text preprocessing and topic modelling algorithms that can be applied to Tweets?
* Which is the appropriate topic model for Twitter trending topics.
* How effective is the topic model in identifying the relevant topics.

## 1.7 Significance of the Study

The findings made in this study will be significant to several stakeholders:-

### 1.7.1 Social Scientists

Traditionally, most social scientists have used either human coding or dictionary methods that require high levels of pre-analysis making them very expensive. This problem is aggravated by the ever increasing volume and variety of unstructured text. Adopting computer assisted methods like topic modelling would augment and amplify their social science analysis.

### 1.7.2 Companies

Companies can benefit greatly from this study because their customers usually use Twitter to voice their sentiments as in the case of product launches and complaints or compliments about a product. Topic modelling can help them sieve through a huge amount of posts in a short time to discover which topics their customers are discussing.

### 1.7.3 Government

People use Twitter to get updates and post information during crisis situations and events of national significance. The user posts contain valuable information about events, places and people of interest. This research can help the government extract this valuable information in a short time which is critical especially in emergency situations.

### 1.7.4 Other Researchers

The study will benefit other scholars interested in topic modelling of short texts. They will know the best preprocessing techniques and the optimal algorithms to apply in these situations.

## 1.8 Scope of the Study

The motivation of this study is to derive topics from Twitter posts on a trending topic. The data will be obtained from tweets sampled from the Twitter search API.

# CHAPTER TWO

# LITERATURE REVIEW

## 2.1 Introduction

This chapter evaluates previous literature produced in the area of study. The review will be conducted in reference to the specific objectives of the study. The major topics that will be covered in this review are the theoretical review, empirical review and the conceptual framework.

**2.2 Theoretical review**

### 2.2.1 Evolution of Topic Models

Topic models find hidden topics within a collection of text documents. They are probabilistic models for discovering the hidden structure of a corpus of documents based on a Bayesian analysis of the documents(Griffiths & Steyvers, 2004). Latent Dirichlet Allocation (LDA) (D. M. Blei, Ng, & Jordan, 2003) is one of the widely used topic model today.

### 2.2.2 LDA

There are two main methods of computerized text analysis: statistical-based methods such as topic models (Hofmann, 2001) and natural language processing (NLP). NLP models perform Part-Of-Speech (POS) tagging, Named Entity Recognition (NER) and semantic labelling of documents. In contrast, statistical-based method use the “bag of words” (BoW) assumption. In doing this, word order and semantic structure in documents are ignored.

By ignoring the word order in documents, BoW models do not perform well short texts like question-and-answer. However, for large text corpus, the BoW assumption provides a wider range of statistical algorithms by the assumption exchangeability, i.e., the word order does not influence the outcome of the model(Blei & Lafferty, 2009.). This assumption aids statistical-based methods in identifying semantic themes in collections of related documents.

One of the earliest applications of topic models was reducing dimensionality of large text corpora. (Deerwester et al., 1990) came up with a seminal model, latent semantic indexing (LSI). They applied singular value decomposition (SVD) to summarize a document term matrix to its latent factors. (Landauer & Dutnais, 1997) improved the LSI model creating the latent semantic analysis model (LSA). They made the assumption that words with close meaning occur in similar documents. (Hofmann, 1999) extended the LSI model by incorporating a generalized Expectation Maximization algorithm with pLSI approach. The model could deal with polysemy and synonymy unlike LSA and LSI.

These pioneering models paved the way for (Blei et al., 2003) who came up with LDA. The main improvement in LDA was the inclusion of a probabilistic model at the document level, by making the assumption that documents are a mixture of topics. In pLSI there is no probabilistic model for documents in a corpus. This solved two main problems of the pLSI model (1) the number of variables increase linearly as the size of the corpus grows, this may lead to overfitting. (2) it is not generalizable outside of the training set.

The inclusion of a second probability component at the document level in LDA introduced the two-tiered model typical of the topic model framework. In this model, documents are presumed to be a mixture of topics and topics a combination of words. Each topic unique occurrences of words.

The introduction of mixture components in LDA led to problems in estimating the optimal model due to exponentially large potential topic values. This led to questions of what was the best way to compute topic models.

### 2.2.3 Computational Methods.

There are two main methods of computing inference for topic models: sampling methods(e.g., Gibbs Sampling) and variational inference. Gibbs Sampling was introduced by (Griffiths & Steyvers, 2004) and uses a Markov Chain Monte Carlo algorithm for inference of LDA. The algorithm estimates the Dirichlet priors in order to approximate the true posterior.

Another approach to estimating the posterior are variational inference methods that are an optimization problem. Variational methods hypothesize a parameterized group of distributions over the latent structure and then find the one that is closest to the posterior (Blei, 2012). (Blei et al., 2003) introduced the Expectation Maximization (EM) algorithm for variational inference. EM uses Kullback-Leibler (KL) divergence to best estimate the posterior as close as possible to the true posterior.

### 2.2.4 Extensions to LDA

Since it was introduced, LDA has been adapted and extended in several ways. LDA is explained by the statistical assumptions it makes about the corpus. An active area of research in topic modelling is how to modify these assumptions to discover more advanced structures in the texts (Blei, 2012).

### 2.2.5 The correlated topic model

One constraint of LDA is that it does not represent correlation between the latent topics (Blei & Lafferty, 2009.). This constraint is because it uses the Dirichlet distribution to model the topic proportions. Naturally, the occurrence of latent topics in most corpora will be correlated.

In the correlated topic model (CTM), (D. Blei & John, 2007) topics are modelled using the logistic normal distribution. This is a more flexible distribution that accounts for covariance pattern among the proportions. This gives a more natural representation of the topics where one topic may be correlated with another.

### 2.2.6 The dynamic topic model

LDA and CTM ignore the word order within the documents. They further assume that the order of documents within the corpus does not matter (Blei & Lafferty, 2009.), this assumption is inappropriate. This is especially true when examining documents that span a long period.

The dynamic topic model (DTM) (Blei & Lafferty, 2006) models the evolution of topics in a serially catalogued corpus of documents. In DTM, the documents are divided by time, e.g., by month or year.

### 2.2.7 Author-topic model

(Rosen-Zvi, Griffiths, Steyvers, & Smyth, 2012) extended LDA to include document information about the author. Every author is linked to a multinomial distribution over the topics while each topic is a distribution over the words. By modelling the interests of the authors, it can be established which topics the author writes about. Documents by different authors can be compared to establish which authors create similar work. Using LDA, the only way to examine the effect of the author was by manually examining how topics change according to the author.

### 2.2.8 Structural topic model

The structural topic model (Margaret E. Roberts, Brandon M. Stewart, 2017) (STM) extends LDA by allowing users to include arbitrary metadata into the topic model. The goal of the STM is to discover topics and model their association to the document metadata. Hypothesis testing about these associations can be done using the model output. The model also introduces improvements to the model inference methods in order to render the model applicable to advanced modelling and evaluation.

## 

## 2.3 Empirical Review

(D. Blei & John, 2007) applied the correlated topic model (CTM) to science articles from the JSTOR archive published from 1990 - 1999. The dataset comprised of 57M words. In most science topics, we can presume that there will be a high correlation within the hidden topics. They showed that it estimated the topics better than LDA, as measured by the accuracy of predictive distributions over the test set documents.

(Rosen-Zvi et al., 2012) applied the author-topic model on the NIPS data set consisting of papers from NIPS conferences. NIPS is contributed to by researchers focused on learning algorithms and computational neuroscience. They illustrated the top 10 words most likely associated with a topic and the top 10 authors most probable to have written a word associated with the topic. They showed that for each topic, the top 10 most likely authors identified for each topic are popular for papers written on those topics.

(Roberts et al., 2014) used the structural topic model (STM) to analyze open-ended survey responses. Open-ended surveys are more difficult to analyze than closed surveys since human coding is not always applied. They showed that the model was better than using human coding in a couple of ways. First, it permits the analyst to uncover topics from the data instead of guessing them. Second, the model allows analysts to study how frequency and content of topics changes with metadata that is associated with each respondent e.g., background demographic data.

(Bhattacharya, Ganguly, Ghosh, Zafar, & Gummadi, 2014) developed an approach to discover topics that individual users of Twitter are interested in. They observed that Twitter users generally follow experts in various topics in order to gain information on those topics. They observed the users a user they ware studying was following and then identified topics of interest of those users. They concluded that if a user follows several experts on a particular topic, they are likely to have a n interest in that topic

(Reich, Tingley, Luis, Roberts, & Stewart, 2015) analyzed the text generated by students of massive open online courses (MOOC) . They used the Structural Topic Model to map students’ self-reported motivations, identify themes in discussion forums and uncover patterns of feedback in course evaluations.

(Lucas et al., 2015) analyzed how the procedures for processing, managing, translating and analyzing textual data differ across languages. They then applied the structural topic model on different religious documents and social media posts to compare how they differ between countries. They compare how different news agencies cover news related to China by analyzing stories from Xinhua and Agence France Presse (AFP). They also compared different Muslim clerics depending on whether they were Jihadist or non-Jihadist. They showed differences topic proportions with regard to fighting and excommunication topics between the two.

(Sokolova et al., 2016) performed topic modelling and event identification from Twitter data. They worked on four datasets collected by the Umati project through Twitter’s streaming API: (1) The Gikomba Twitter data mainly covering a bombing incident in Nairobi’s Gikomba market. The dataset had 482 tweets. (2)The Mandera Twitter data that contained tweets on tribal clashes in Mandera town in Kenya. The data had 915 tweets in total. (3) The Makaburi dataset containing 20642 tweets. In those tweets, people were talking about the violent death of a controversial Muslim cleric, Sheikh Makaburi. (4) The Mpeketoni dataset containing 106348 tweets. In those Tweets, people discuss a terrorist attack that happened in Mpeketoni, a town in the coastal region of Kenya. They applied LDA for topic modelling after the initial steps of data pre-processing. They then analyzed the topics manually and by using topic coherence analysis. Topic coherence measures each topic by scoring it using the level of semantic similarity of words in a topic.

## 2.4 Conceptual Framework

As guided by the literature review, the following conceptual framework was developed. It shows how the Twitter corpus constructed from a trending topic will be used to derive topics users are discussing. Topic models are unsupervised algorithms, therefore, the variables will be inferred from the corpus. Consequently there are no independent and dependent variables.

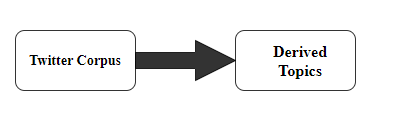


Figure 1 : Conceptual Framework

**2.5 Operationalization of Variables**

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | Sub-variables | Indicators | Values(data) |
| Tweet | Text | Twitter post text | Text |
| Created At | Date | Date |
| Source | Twitter Client | Twitter Web Client  TweetDeck  Twitter for Android  Twitter for Iphone  Others |

# CHAPTER THREE

# RESEARCH METHODOLOGY

## 3.1 Introduction

The chapter examine the research methods that will be used in order to achieve the objectives of the study. This chapter covers the research design to be utilized, data preprocessing methods, data analysis methods and model evaluation.

## 3.2 Research Design

Research design is the general framework within which the research will be carried out. It involves identifying the source and method of obtaining the relevant data and methods for analyzing the data. Data for this research will be obtained from the Twitter search API2.[Twitter API](https://developer.twitter.com/). The API returns Tweets matching the user defined query. Not all tweets matching a specified query are made available via the Search API. The search will include only Tweets in English and will exempt retweets.

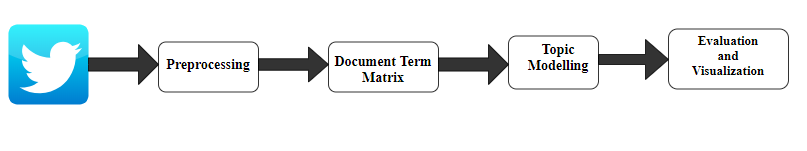


Figure 2: Research Design

Data collected via the Twitter API will be stored in a MySQL database. Common text preprocessing methods will then be applied to the datasets. Once all preprocessing is completed, the remaining terms will be converted into a document-term matrix (DTM). A DTM is a data structure where a document is represented by a row on the matrix and each unique word is represented by a column. It uses the “bag-of-words” approach where word order is discarded but the number of times each word occurs is recorded. Automated content techniques such as LDA or STM can then be applied to this matrix to learn the topics.

## 3.3 Data preprocessing

Twitter data is considered more challenging than other social media data due to character limit, misspellings and slang (Eisenstein, 2013). In this study, Twitter specific text preprocessing will be performed. This will include removal of Twitter-specific features like hashtags, hyperlinks, emoticons, user mentions, slang and acronyms.

Other common text pre-processing procedures that will be performed include stop word removal, stemming, lemmatization, Tokenization and converting to lowercase.

### 3.3.1 Stopword Removal

This involves removing recurring words such as “and” and “the” to aid in model performance and interpretation. These words do not contribute to the meaning of the document. The most common way to remove them is to use a fixed list. Such lists are available in many languages and typically take care of most stopwords.

### 3.3.2 Stemming and Lemmatization

Stemming gets rid of the letters added to conjugated verbs and nouns leaving just the base form of the word. Lemmatization is an advanced algorithm that returns the *lemma*, or canonical form of a word.

### 3.3.3 Tokenization

Tokenization is the process of separating a string of text into its constituent words. For English, whitespace and punctuation are usually used to detect word boundaries. Tokenization is the first step before creating a document term matrix.

### 3.3.4 Converting to Lowercase

For a particular concept there can be many character strings representing it depending on the case. For example, “Topic”, “TOPIC” and “topic” refer to the same underlying concept. Converting to lowercase will convert them to the same word for purposes of topic modelling.

## 3.4 Choosing Parameters

After data pre-processing and creation of the document-term matrix. There are a number of parameters the researcher needs to choose before running the model. First is the number of topics (K) which indicates the number of topics that should be identified by the model (Jacobi, Van Atteveldt, & Welbers, 2016). There is no default value for this parameter. The aim is to represent the documents with fewer than the actual number of topics present but with as little loss to relevant information as possible.

Second is a hyperparameter, *alpha,* which affects the number of topics to be identified within the documents. A widely used default is 50 divided by K.

## 3.5 Data Analysis

The data will be analyzed using the R language. R has many open source implementations for topic modelling. In this case I will use the stm R package (Margaret E. Roberts, Brandon M. Stewart, 2017). The choice of this package is informed by the fact that it allows modelling of document metadata and provides a statistical-based framework to enable hypothesis testing.

To use the package, first the data is read and prepared for analysis. Then a structural topic model is approximated. The package provides functions for evaluation, understanding and visualization of the results, as shown in the diagram below.

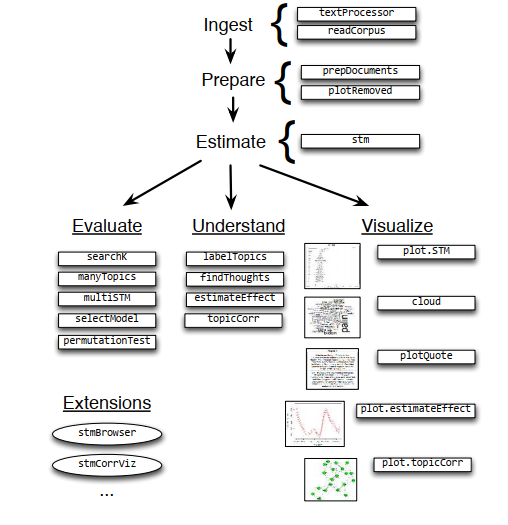


Figure 3: Functions within the stm package form Roberts et al. (2015)

## 3.6 Model Evaluation

The results of the topic model will be evaluated using perplexity and manual analysis of each topic. Manual analysis will be done by examining each topic and the top words closely. Perplexity will also be used to evaluate the topic model. Perplexity can be considered as a measurement of how well a probability distribution predicts a sample. If a topic model has low perplexity, then it is considered more generalized, compared to the one that has high perplexity.

# REFRENCES

Manna, S., & Phongpanangam, O. (2018). Exploring Topic Models on Short Texts: A Case Study with Crisis Data. 2018 Second IEEE International Conference on Robotic Computing (IRC), 377-382.

Benhardus, J., & Kalita, J. (2013). Streaming trend detection in Twitter. *International Journal of Web Based Communities*, *9*(1), 122. <https://doi.org/10.1504/IJWBC.2013.051298>

Bhattacharya, P., Ganguly, N., Ghosh, S., Zafar, M. B., & Gummadi, K. P. (2014). Inferring user interests in the Twitter social network, 357–360. <https://doi.org/10.1145/2645710.2645765>

Blei, D. M. (2012). Probabilistic Topic Models, 77–84.

Blei, D. M., Edu, B. B., Ng, A. Y., Edu, A. S., Jordan, M. I., & Edu, J. B. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022. <https://doi.org/10.1162/jmlr.2003.3.4-5.993>

Blei, D. M., & Lafferty, J. D. (2006). Dynamic Topic Models.

Blei, D. M., & Lafferty, J. D. (2009). Topic models.

Deerwester, S., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis.

Eisenstein, J. (2013). What to do about bad language on the internet. *NAACL HLT 2013 - 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Main Conference*, (June), 359–369. <https://doi.org/10.1016/j.febslet.2009.03.055>

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics.

Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis, 177–196.

Jacobi, C., Van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, *4*(1), 89–106. <https://doi.org/10.1080/21670811.2015.1093271>

Landauer, T. K., & Dutnais, S. T. (1997). A Solution to Plato ’ s Problem : The Latent Semantic Analysis Theory of Acquisition , Induction , and Representation of Knowledge, *1*(2), 211–240.

Lda, T., & Ctm, T. (2007). A correlated topic model of, *1*(1), 17–35. <https://doi.org/10.1214/07-AOAS114>

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis*, *23*(2), 254–277. <https://doi.org/10.1093/pan/mpu019>

Margaret E. Roberts, Brandon M. Stewart, D. T. (2017). stm: R Package for Structural Topic Models, *VV*(Ii). <https://doi.org/10.18637/jss.v000.i00>

Reich, J., Tingley, D., Luis, J. L., Roberts, M. E., & Stewart, B. M. (2015). Computer-Assisted Reading and Discovery for Student-Generated Text in Massive Open Online Courses 1 OVERVIEW, *2*, 156–184.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., … Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*(4), 1064–1082. <https://doi.org/10.1111/ajps.12103>

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2012). The Author-Topic Model for Authors and Documents, (January). Retrieved from <http://arxiv.org/abs/1207.4169>

Sokolova, M., Huang, K., Matwin, S., Ramisch, J., Sazonova, V., Black, R., Sambuli, N. (2016). Topic Modelling and Event Identification from Twitter Textual Data. *arXiv Preprint arXiv:1608.02519*, 17. Retrieved from <http://arxiv.org/abs/1608.02519>

Wesslen, R. (2018). Computer-Assisted Text Analysis for Social Science: Topic Models and Beyond. Retrieved from <http://arxiv.org/abs/1803.11045>

# APPENDICES

## Appendix I: Budget

|  |  |
| --- | --- |
| **Budget Items** | **Cost (KES)** |
| Proposal - Printing, binding and internet costs | 10,000 |
| Data collection – Internet cost | 5,000 |
| Data analysis and reporting a) Printing and stationery costs  b) Hard cover Binding | 5,000 |
| 6,000 |
| Transport to Campus, airtime costs | 5, 000 |
| Miscellaneous | 5,000 |
| **TOTAL BUDGET** | **KES 36,000** |

## Appendix II: Work Plan

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Timelines**  **Activity** | **Jan - March 2019** | **March**  **2019** | **March-April**  **2019** | **April**  **2019** | **May-June**  **2019** |
| Proposal Development |  |  |  |  |  |
| Proposal Defense |  |  |  |  |  |
| Data Collection |  |  |  |  |  |
| Data analysis and Report Writing |  |  |  |  |  |
| Project Defense and Final Report Submission |  |  |  |  |  |