TWITTER TRENDS ANALYSIS USING 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](#_Toc3898513)

[TABLE OF CONTENTS ii](#_Toc3898514)

[LIST OF FIGURES iv](#_Toc3898515)

[LIST OF ACRONYMS AND ABBREVIATIONS vi](#_Toc3898516)

[CHAPTER ONE 1](#_Toc3898527)

[INTRODUCTION 1](#_Toc3898528)

[1.1 Background of the Study 1](#_Toc3898529)

[1.2 Statement of the Problem 2](#_Toc3898530)

[1.3 Motivation of the Study 4](#_Toc3898531)

[1.4 Main Objective 4](#_Toc3898532)

[1.4.1 Specific Objectives 5](#_Toc3898533)

[1.4.2 Research Questions 5](#_Toc3898534)

[1.5 Significance of the Study 5](#_Toc3898535)

[1.5.1 Social Scientists 5](#_Toc3898536)

[1.5.2 Companies 6](#_Toc3898537)

[1.5.3 Government 6](#_Toc3898538)

[1.5.4 Other Researchers 6](#_Toc3898539)

[1.6 Scope of the Study 6](#_Toc3898540)

[CHAPTER TWO 7](#_Toc3898541)

[LITERATURE REVIEW 7](#_Toc3898542)

[2.1 Introduction 7](#_Toc3898543)

[2.2 Theoretical review 7](#_Toc3898544)

[2.2.1 Evolution of Topic Models 7](#_Toc3898545)

[2.2.2 LDA 7](#_Toc3898546)

[2.2.3 Computational Methods. 9](#_Toc3898547)

[2.2.4 Extensions to LDA 10](#_Toc3898548)

[2.2.5 The correlated topic model 10](#_Toc3898549)

[2.2.6 The dynamic topic model 10](#_Toc3898550)

[2.2.7 Author-topic model 11](#_Toc3898551)

[2.2.8 Structural topic model 11](#_Toc3898552)

[2.3 Empirical Review 11](#_Toc3898553)

[2.4 Conceptual Framework 14](#_Toc3898554)

[CHAPTER THREE 15](#_Toc3898555)

[RESEARCH METHODOLOGY 15](#_Toc3898556)

[3.1 Introduction 15](#_Toc3898557)

[3.2 Research Design 15](#_Toc3898558)

[3.3 Data preprocessing 16](#_Toc3898559)

[3.3.1 Stopword Removal 16](#_Toc3898560)

[3.3.2 Stemming and Lemmatization 16](#_Toc3898561)

[3.3.3 Tokenization 17](#_Toc3898562)

[3.3.4 Converting to Lowercase 17](#_Toc3898563)

[3.4 Choosing Parameters 17](#_Toc3898564)

[3.5 Data Analysis 18](#_Toc3898565)

[3.6 Model Evaluation 19](#_Toc3898566)

[REFRENCES 20](#_Toc3898567)

[APPENDICES 23](#_Toc3898568)

[Appendix I: Budget 23](#_Toc3898569)

[Appendix II: Work Plan 24](#_Toc3898570)

# LIST OF FIGURES

[Figure 1 : Conceptual Framework 14](#_Toc3897839)

[Figure 2: Research Design 15](#_Toc3897840)

[Figure 3: Functions within the stm package form Roberts et al. (2015) 19](#_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 lot of research in the field 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, on average, users post around 6000 Tweets; this translates to 350,000 tweets per minute and 500 million tweets per day1.[Internet Live Stats](https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/)

Twitter has become an important source of information for a diverse spectrum of users. People use Twitter to gather real-time news, follow events of interest and read updates by people they follow. Twitter is increasingly used for communicating breaking news, eyewitness accounts, product marketing, product reviews, dissemination of news and organizing large groups of people.

Twitter users have become accustomed to receiving timely updates on important events of global or local value. For instance, Twitter was used to propagate information in many crisis events such as the terrorists attack on the Dusit Complex in Kenya and the aftermath of the Kenyan elections in 2017. In addition, many organizations and celebrities use their Twitter accounts to connect to customers and fans.

Being a popular microblogging and Social Networking Site, Twitter presents many opportunities for research in Natural Language processing (NLP) and machine learning (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 has become important for a number of things such as breaking news detection, personalized message recommendation, friends recommendation, sentiment analysis and others.

It is important to sieve through a huge amount of social media texts to extract meaningful information especially during any crisis and emergency situations. Topic Models have the ability to summarize, explore and index large text document collections and can be used for this purpose (Manna, S., & Phongpanangam, O. (2018)).

## 1.2 Statement of the Problem

Length restrictions of microblogged messages creates syntactic and structural conventions not seen in traditional corpora. The size of the Twitter network also produces a continuously changing and dynamic corpus. Twitter data presents several challenges compared to traditional natural language processing technologies such as:

* **Obscure Language and grammar:** Twitter messages often lack proper punctuation and grammar. Tweets usually include shortened words and URLs. They also contain abbreviations including internet slang such as “IRL” for “in real life”.
* **Message length:** The short messages contain very little grammatical redundancy.
* **Locale-specific references:** Messages usually refer to specific locations, events and other named entities. Thus, one cannot rely on pre-defined entity lists or complex named entity recognition methods.

In this research, I will explore the use of Topics models for processing Twitter data. Topic models were originally developed for analyzing the semantic content of large document corpora. I propose that the family of topic models is particularly suited for analyzing Twitter data for the following reasons:

* **Bag of words:** Topic Models are usually “bag-of-words” models, meaning that they do not rely on the syntactic structure or word order in the language. This makes them especially suited to handle the obscure language and irregular grammar contained in Twitter posts.
* **Latent variables:** Topic models are able to infer latent (or hidden) connections between elements in data. This makes them sturdier in handling misspellings, acronyms and other idiosyncrasies in Twitter posts.
* **Representation:** Topic models represent statistical knowledge as homogeneous numerical vectors such as probability distributions. This makes them suitable for comparisons, visualization as well a complex modelling such as clustering.
* **Adaptability:** Because of their unsupervised nature, topic models can be easily retrained on a text corpus for a particular domain using other available text.

## 1.3 Motivation of the Study

Microblogging sites such as Twitter have become important communication tools for people across the globe. As these sites continue to grow, important events in the society are increasingly found not in headlines of traditional media institutions, but in the activity of ordinary individuals. The identification of trending topics utilizes social media to provide an overview of topics and issues currently popular within user posts in the online community.

Researchers wish to use messages posted by users to infer users’ interests, model social relationships, track news stories and identify emerging 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 tools to understand latent text patterns in the messages.

## 1.4 Main Objective

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

## 1.4.1 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.

## 1.4.2 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.

## 1.5 Significance of the Study

The findings of this research will be of great significance to several stakeholders:-

### 1.5.1 Social Scientists

Historically, most social science has focused on 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.

### 1.5.2 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.

### 1.5.3 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.

### 1.5.4 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.

## 1.6 Scope of the Study

The purpose 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 focuses on the previous studies done in the area of study. Literature will be reviewed in relation to the specific objectives of the study. The review is divided into theoretical and empirical review.

## 2.2 Theoretical review

### 2.2.1 Evolution of Topic Models

Topic models identify hidden topics within a corpus of text documents. Topic models are called generative models because they assume that observable data is generated by the joint probabilities of variables that are identified as the topics. LDA, one of the most popular topic model, is a Bayesian two-tiered mixture model that identifies word co-occurrence patterns that are identified as the topics.

### 2.2.2 LDA

In general, there are two approaches to computer assisted text analysis: natural language processing (NLP) and statistical-based methods such as topic models (Hofmann, 2001). Unlike NLP models that tag parts-of-speech and grammatical structure, statistical based models are largely based on the “bag-of-words” (BoW) assumption. In BoW models, collections of text documents are cast into a document term matrix (DTM) that counts the occurrence of each word (columns) for each document (rows).

The BoW approach prioritizes simplicity and statistical properties at the expense of word order. Without accounting for word order, BoW models perform poorly short texts like question-and-answer. However, for large text corpus, the BoW assumption provides a richer set of statistical algorithms by the assumption that words are exchangeable within each document, i.e., their order does not affect their probability under the model(Blei & Lafferty, 2009.). This assumption is a simplification that aids statistical-based methods in identifying semantic themes in a large enough collection of inter-related documents.

An early application of topic models was dimensionality reduction in large text corpora. (Deerwester et al., 1990) presented one of the earliest predecessor models, latent semantic indexing (LSI), by applying singular value decomposition (SVD), a linear algebra dimensionality reduction technique, to reduce the document term matrix to its latent factors. (Landauer & Dutnais, 1997) extended the LSI model to create the latent semantic analysis model (LSA). These models can be improved further by substituting term-frequency inverse document frequencies (TF-IDF) in place of raw term counts (Wesslen, 2018). However, (Hofmann, 2001) showed that LSA, which stem from linear algebra and performs singular value decomposition of co-occurrence tables, relied on the Gaussian distribution which could not be justified for word count. LSA could also not account for polsemy, the multiple uses of words in different contexts.

These pioneering models set the stage for (Blei et al., 2003) to build the LDA model. The main improvement in LDA was the a inclusion of a probabilistic model at the document level, thus assuming documents are a mixture of topics. In pLSI there is no generative probabilistic model for documents. This solved two major problems with the pLSI model (1) the number of parameters in the model grows linearly with the size of the corpus, which lead to serious problems with overfitting, and (2) it is not clear how to assign probability to a document outside of the training set.

The inclusion of a second probability component at the document level yielded a two-tiered model typical of the topic model framework. At the top, documents are a mixture of topics. At the bottom, topics are a mixture of words. Each topic is defined as a unique distribution 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 bast way to compute topic models.

### 2.2.3 Computational Methods.

In general, there are two main approaches to for topic model inference: sampling methods (e.g., Gibbs Sampling) and variational inference. Sampling based algorithms simulate samples of the posterior to approximate the true posterior. Gibbs Sampling was introduced by (Griffiths & Steyvers, 2004) and uses a Markov Chain Monte Carlo algorithm for inference of LDA.

Another approach to approximating the posterior are variational inference methods that transform the problem into an optimization problem. Variational methods posit a parameterized family of distributions over the hidden structure and then find the member of that family 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 minimize the estimated posterior to the true posterior.

### 2.2.4 Extensions to LDA

Since its introduction, LDA has been adapted and extended in many ways. LDA is defined by the statistical assumptions about the corpus. One active area of topic modelling research is how to relax and extend these assumptions to uncover more sophisticated structure in the texts (Blei, 2012).

### 2.2.5 The correlated topic model

One limitation of LDA is that it fails to model correlation between the occurrence of topics (Blei & Lafferty, 2009.). This limitation stems from the use of the Dirichlet distribution to model the variability among the topic proportions. It is natural to expect the occurrence of latent topics in most corpora will be highly correlated.

In the correlated topic model, (CTM) (Lda & Ctm, 2007) topic proportions are modelled with an alternative, more flexible distribution that allows for covariance structure among the components. This gives a more realistic model of the topic structure where the presence of one latent topic may be correlated with another.

The key to CTM is the logistic normal distribution that allows for a general pattern of variability between the components. The CTM is more expressive than LDA because the strong independence assumption imposed by the Dirichlet in LDA is not realistic when analyzing real document collections.

### 2.2.6 The dynamic topic model

LDA and CTM assume that words are exchangeable within each document. They further assume that documents are exchangeable within the corpus (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) captures the evolution of topics in a sequentially organized corpus of documents. In DTM, the data is dived by time slice, e.g., by year.

### 2.2.7 Author-topic model

(Rosen-Zvi, Griffiths, Steyvers, & Smyth, 2012) extended LDA to include authorship information. Each author is associated with a multinomial distribution over topics and each topic is associated with a multinomial distribution over words. By modelling the interests of the authors, it can be established which topics the author writes about, which authors may have written a similar observed document and which authors produce similar work. Using LDA, the only way to analyze the impact of the author was post-hoc by examining how topics compare relative to 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 estimate their relationship to the document metadata. Output of the model can be used to conduct hypothesis testing about these relationships. The model also introduces enhancements to the computational methods in order to make the model feasible for modelling and evaluation.

## 2.3 Empirical Review

(Lda & Ctm, 2007) applied the correlated topic model (CTM) to science articles from the JSTOR archive published from 1990 - 1999. The dataset comprised of 57M words. Their motivation was the limitation with LDA, which fails to directly model correlation between topics. In most science topics, it is natural to expect that subsets of the underlying topics will be highly correlated. They showed that it gives a better fit than LDA, as measured by the accuracy of predictive distributions over held out documents.

(Rosen-Zvi et al., 2012) applied the author-topic model on the NIPS data set that contains papers from the NIPS conferences between 1987 and 1999. The conference is characterized by contributions from a number of different research communities in the general area of learning algorithms. They illustrated the top 10 words most likely associated with a topic and the top 10 authors most likely to have generated a word associated with the topic. They showed that for each topic, the top 10 most likely authors are well known authors in terms of NIPS papers written on those topics.

(Roberts et al., 2014) used the structural topic model (STM) to analyze open-ended survey responses. Open-ended surveys have traditionally been considered more difficult to analyze than their closed counterparts as human coding is not always used. They showed that the model has a number of advantages over using human coders. First, it allows the researcher to discover topics from the data rather than assume them. Second, the model allows analysts to study how prevalence and content of topics changes with information that is particular to each respondent e.g., background demographic data.

(Bhattacharya, Ganguly, Ghosh, Zafar, & Gummadi, 2014) developed a mechanism to infer topics of interest of individual users of Twitter. They observed that Twitter users generally follow experts in various topics of interest in order to acquire information on those topics. They checked the users a user of interest was following and then identified topics of interest of those users. Intuitively, if a user subscribes to tweets form several experts on a perticular topic, the user is likely to be interested in that topic.

(Reich, Tingley, Luis, Roberts, & Stewart, 2015) analyzed the use of massive open online courses (MOOC) course feedback to scale open-ended course responses. The 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) apply topic modelling to different types of documents such as fatwas and social media posts in order to facilitate comparison between countries.

(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 based on calculating the degree of semantic similarity between the words in the 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.

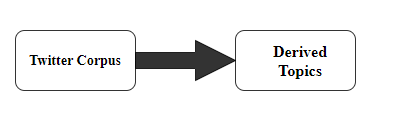


Figure 1 : Conceptual Framework

**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 will look at the research methods that will be employed in the study in order to achieve the objectives of the study. This chapter covers the research design to be adopted, population of study, sample size and sampling technique, data collection instrument, data collection procedures and data analysis methods.

## 3.2 Research Design

Research design is the general structure within which the research will be conducted. 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 a collection of relevant Tweets matching the specified 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.

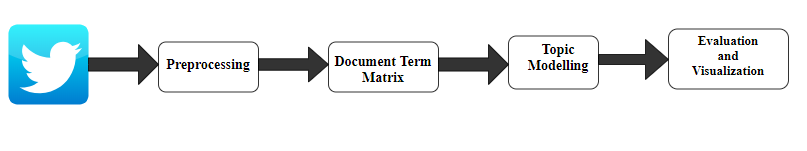


Figure 2: Research Design

Data will be collected via the Twitter API and 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 used to construct a document-term matrix (DTM). A DTM is a matrix where each row represents a document and each column represents a unique word. Following the “bag-of-words” assumption, the DTM format preserves information about how many times each word appears in a document while discarding information about the word order. 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 emoticons, hashtags, hyperlinks, user mentions, acronyms and slang.

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 frequently occurring function 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 removes the endings of conjugated verbs and plural nouns leaving just the “stem”, which is common to all forms of the word. Lemmatization is a more complicated algorithm that identifies the origin of the word, only returning the lemma, or common form of the word.

### 3.3.3 Tokenization

Tokenization is the process of breaking 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. The number of topics (K) needs to be specified, which indicates how many topics the model should classify the words in the documents into (Jacobi, Van Atteveldt, & Welbers, 2016). There is no default value for this parameter. The goal is to describe the documents with fewer topics than are actually present but with as little loss to relevant information as possible.

A second parameter is the *alpha* hyperparameter, which affects how many parameters a document can contain. A common default value for alpha is 50 divided by the number of topics. Essentially, a lower alpha leads to fewer topics within the documents.

## 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 for facilitating hypothesis testing.

To use the package first users ingest the data and prepare it for analysis. Next, a structural topic model is estimated. The package provides functions for evaluation, understanding and visualization of the results, as shown in the diagram below.

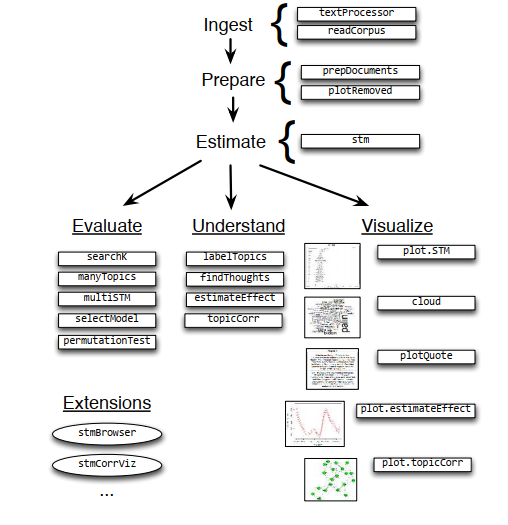


Figure 3: Functions within the stm package form Roberts et al. (2015)

## 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 (Kenya Shillings)** |
| Proposal development- Printing, stationery, internet costs | 10,000 |
| Data collection – Internet cost | 5,000 |
| Data analysis and report a) Printing and stationery  b) Hard cover Binding | 5,000 |
| 6,000 |
| Transport Fuel to Campus, airtime | 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 |  |  |  |  |  |