MACHINE LEARNING MODEL FOR CLASSIFYING FAKE NEWS IN KENYA.

SUBMITTED BY: OKUKU DENNIS DOME

REG NO: 17/02495

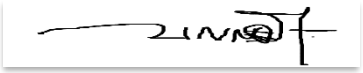
A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTER OF SCIENCE DATA ANALYTICS DEGREE IN THE SCHOOL OF TECHNOLOGY AT KCA UNIVERSITY.

November 2022

DECLARATION

I declare that this dissertation was my original work and had not been previously published or submitted elsewhere for the award of a degree. I also declare that this contains no material written or published by others except where due reference was made and the author duly acknowledged.

Student Name: Okuku Dennis Dome Reg No.: 17/02495

Sign:  **Date: 05/11/2022**

I did hereby confirm that I have examined the Master's proposal of

Okuku Dennis Dome

And has approved it for examination.

Sign: _____ Date: 05/11/2022

Name of Supervisor DR SIMON MWENDIA

ABSTRACT

The revolution in the digital age to the information age to the growth of social networks into go-to news sources and primary information pools has seen a change in the conventional approach to political information dissemination. This, unfortunately, also saw social media abuse through targeted mis- and dis-information to sway public opinion for political gain. Applied machine learning was a solution that bears promise. This research proposal explored the next level in Automated Machine learning to track and classify fake news in the Kenyan environment targeted on the Facebook Platform by applying Natural Language Processing at scale in renowned cloud computing frameworks.

This study would build multiple models and select the superior one for the final deployment of inaccurate word scenarios.

Keywords: adaptive boosting**,** machine learning, ensemble model, traditional machine learning, fake news, misinformation, disinformation.

ACKNOWLEDGEMENT

This is in appreciation of my research supervisor, Dr. Mwendia, the insights shared by the Kenya Police Service and support from the National Cohesion and Integration Commission, the Facebook Developer Community, and most importantly, my Family Grace, Leikana, Baraka, Mom and Dad for creating the right environment to focus on this work and complete it.

ACRONYMS

ANN - Artificial Neural Network

AU-ROC - Area under Receiver Operating Characteristic Curve

AutoML - Automated Machine Learning

BERT - Bidirectional Encoder Representations from Transformers

DTM/TDM – Document Term Matrix/Term Document Matrix

F1 - harmonic mean between TPR and FPR

FN - False Negative

FP - False Positive

FPR - False Positive Rate

IDF-TF/TF-IDF - Inverse Document Frequency - Term Frequency

MLP (Mlp) – Multi-Layer Perceptron

ML – Machine Learning

NCIC - National Cohesion and Integration Commission

NLTK - Natural Language Toolkit

ODK – Open Data Kit

RELU - Rectified linear unit

SGD (sgd) – Stochastic Gradient Descent

TF-IDF - Term Frequency - Inverse Document Frequency

TN - True Negative

TP - True Positive

TPOT - Tree-based Pipeline Optimization

TPR - True Positive Rate

VADER -Valence Aware Dictionary and Sentiment Reasoner

GLOSSARY

Artificial Neural Network - This is a commonly used and very effective neural network for various machine learning cases.

Bidirectional Encoder Representations from Transformers - It was a deep neural network modeling technique designed to pre-train two-way directional representations from the unlabeled text.

Natural Language Toolkit - This is a python library for natural language preprocessing, including tokenization, removal of stop words, and stemming of similar words.

PyCaret - This is a low-code ML library in Python, fast-tracking data pre-processing to deployment in minutes.

Rectified linear unit - This is a deep neural network activation function

Term Frequency - Inverse Document Frequency –This calculation assesses how relevant a word in a series or corpus was to a text.

Valence Aware Dictionary and Sentiment Reasoner - This is a python package for polarity sentiment analysis commonly used for classifying the sentiment of comments based on their polarity.

Mlp – This is a fully connected class of feedforward artificial neural network consisting of multiple layers of perceptrons.

SGD – This is one of the efficient approaches to fitting linear both binary and multi-class classifiers and regressor lines under convex loss functions. Commonly used for SVM's and Logistic Regression

TABLE OF CONTENTS

Contents

LIST OF FIGURES

LIST OF EQUATIONS

CHAPTER ONE: INTRODUCTION

1.1 Background of the Study

The digital age and improved internet connectivity have led to the scale-up in use and access to social networks. These social media networks were at the heart of communication strategies for governments and private entities. They continue to become the central platform for disseminating essential information on various topics and also our topic of interest - political information by (Abdulrauf, 2016) the information consumption and content creation was high before, during, and after both national and elections. This seasonality also strongly aligns with the five-year periods after every significant election and every time a constitutionally mandated reason for carrying out a by-election arises. They have been various computing approaches applied to the problem. Machine learning models have been widely applied to solve this problem. These have ranged from pure neural networks in deep learning to ensemble models, which improve on these traditional deep learning models. We explored the broad application of existing solutions and how this informed our study.

We focused on Facebook, which is not only a market leader in its category but the world's largest content distributor, yet not a news or broadcasting channel. Facebook, currently in Kenya, as of June 2021, had a market share of 56.16% (StatCounter GlobalStats, 2021). It widely proliferated political information and content to many Kenyan consumers in the country and across the globe. With its ability to effectively disseminate large volumes of political information that consisted of all multimedia forms to target individuals and groups alike had rendered it a target to many news agencies of a formal nature and also, unfortunately, peddlers of mis- and disinformation. In most extreme cases, these ill-intending entities were criminal organizations, think tanks bent on influencing Kenyans to take a particular political view, and most politicians who wanted to generate influence and shape opinions genuinely and at whatever cost and accuracy of the information shared. Though detecting fake news is challenging, avoiding mass disinformation and misinformation is essential.

Fake news existed even before digital content became widely increased across the globe. Majorly fake news has various ill intents and typically takes one of two forms: - mis- or disinformation. At the same time, the latter means spreading false information, regardless of the intent to mislead. The former was the deliberate conscious effort to spread information that was not accurate and false with the intention to draw gains as a result of disbursing this information. While most vulnerable and targeted consumers were concerned with what was honest and accurate news on the internet, specifically on social media (Newman et al., 2020).

The very fabric of democracy is hinged on healthy politics and information sharing that typically consists of Facebook being at the forefront. The era of the digital age evolved to usher in social media and its ability to narrow the borders between counties in Kenya. When used for ulterior motives, there was a threat to democracy, as indicated by (Madowo, 2019). While regulation in news and journalism had been enhanced to reduce dis- and mis- information, the exception and weak areas in governance and application of these reforms were limited by the lack of robust regulations on Facebook in the country. The unhealthy competition rife in politics in Kenya before, during, and after every election cycle fueled the need to propagate falsehoods

for political gain or to curtail the popularity of a target competitor. Fake news was widely used to target the vulnerable. In the digitized world, information bias and asymmetry, even when the information was widely shared, enhanced these ills and motivated the entities at the forefront of spreading fake news in the political cycles of Kenya. The efforts to spread it were so well organized, involving a lot of resources and technical know-how from various parties who had commercialized several aspects of the processes and created a business model.

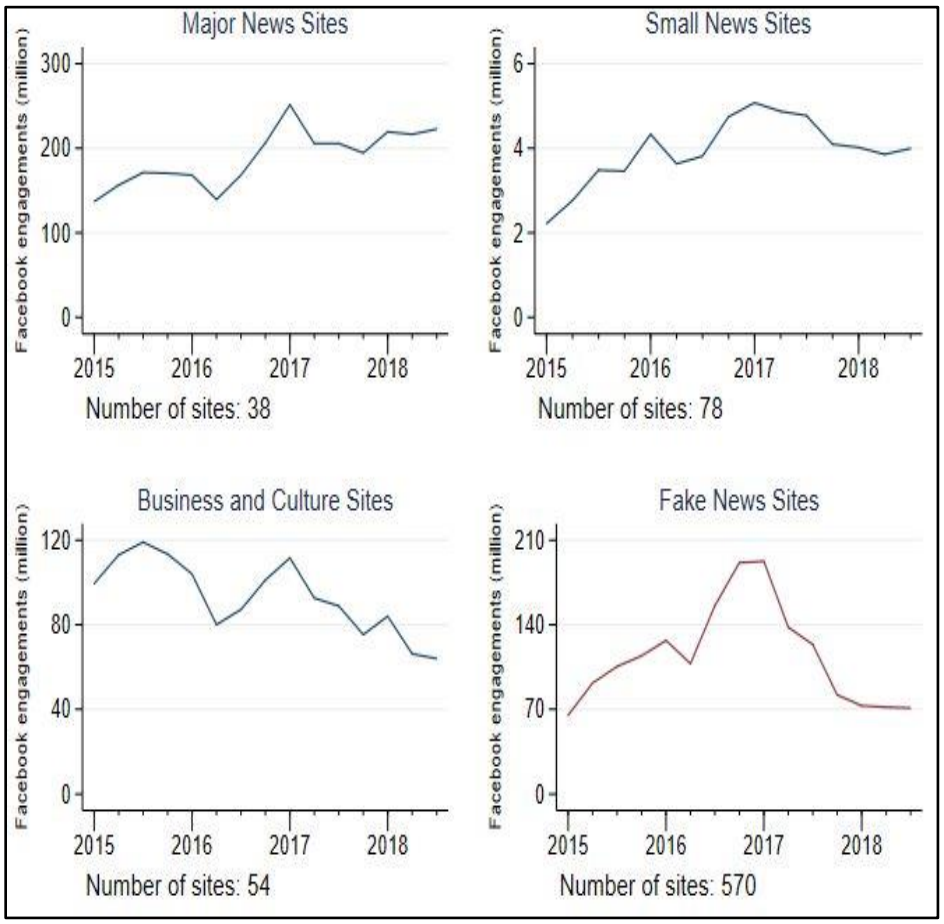
Facebook had tried self-policing and enhanced mitigation measures on propagating mis- and disinformation during primary elections for first-world countries. But, never in third-world or resource-constrained countries like Kenya where the violence was more disruptive to the economy and loss of lives was immediate and tragic. While working closely with Cognizant, it had become apparent that no form of or approach to content moderation or autonomous machine learning-driven system would ever be entirely infallible for these challenges and risks, while we recognize that those in power would be unrelenting as they continue to pursue durable solutions, for now, the local Kenya Government and citizenry must remain vigilant and stay on guard against both mis- and dis-information and stay skeptical of every aspect of what they read unless validated by multiple news sources pointing to the same facts for corroboration purposes. Machine learning and artificial intelligence algorithms had been deployed in a semi-autonomous design to work closely with content moderators in a man-in-the-middle approach. This was far from sufficient to tip the scales of catching and sieving all harmful forms of political mis- and disinformation in the Kenyan geography sense, let alone global scale. The challenge with multiple native languages and weak translation engines, mainly where the languages were not spoken in a pure sense, further complicated the problem (Facebook, 2020) I sought to solve.

1.1.1 Setting the context: Global and local perspective.

Social media provides the primary medium for consuming information in sovereign states. The sentiment towards governments could also be improved, worsened, or completely polarized using social networks. The human right of freedom to information was at risk owing to the nature of the information and the apparent growing inaccuracies of the shared information at a vast global scale (Kertysova, 2018).

Facebook is the market leader in this space. It accounts for a vast amount of mis- and dis-information being propagated. This practice has been documented extensively. The scenario worsened because the content did not belong to the social network firms or service providers, so they could not be held accountable and charged legally (Watson, 2021). It was estimated at the global level that fake news on engagements leads and spurs more site visits than any category of news or information on Facebook, the target social media platform for our research. The numbers stand at 210 million engagements on FaceBook, which is close to the volume on major news sites (Allcott et al., 2018).

Figure 1: Facebook Engagement



The graphs above by (Allcott et al., 2018, pg8) show the impact of Fake news on engagements. Fake news sites increase Facebook engagements at the highest rate compared to business and culture sites, major news sites, and small news sites. The ideal reality should, however, be different. We expect to see more significant news sites leading and spurring engagements than fake news sites. This not only points to the extent of the problem but the challenge that fake news

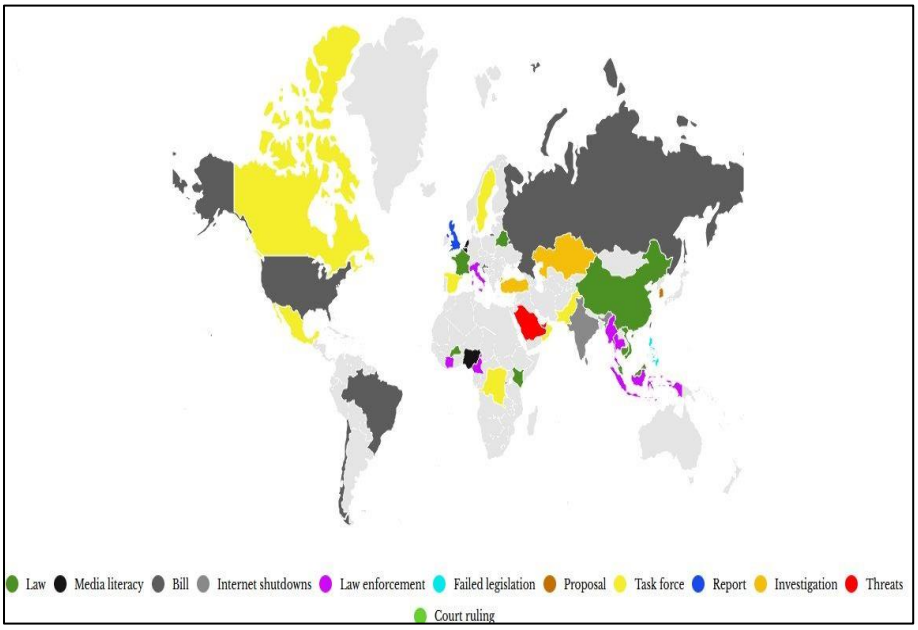
Fake news consisting of mis- and disinformation is a phenomenon that affects our social and civic participation in a significant way, especially with how geo-political issues affect everyone, not necessarily those in the immediate geography of the events. Fake news detection, identification, and classification need to multiply. However, it faces many difficulties because of the limited number of resources available, jurisdictions, different applications of regulations, and data policies for its immediate citizens. The primary objective of all proposed systems was to apply artificial intelligence to develop a very efficient system that could anticipate whether a piece of information was fake and label it by the extent to which it misleads based purely on its content. Thus stay effective in addressing the issue of dis- and mis- information from an NLP perspective. Important to note and observe that the use of automated machine learning to label and classify fake news was playing catchup to the new phenomenon of deepfakes - which uses neural networks to generate fake news in large volumes at scale. This study does not delve into

or focus on the evolving nature of fake news into deepfakes. We were, however, cognizant that mis- and disinformation had morphed further, and machine learning was being applied to create this new threat (Facebook, November 2020).

From a Kenyan perspective, there was a lack of legislation or a concerted effort by the Government to hold Facebook accountable. Victims of mis- and dis-information did not report this, nor did we have explicit regulations categorizing aspects of fake news as a crime or contravention of local laws. Though AI legislation conversations were ongoing, legislation on algorithms and ethics was not established. Global mapping of where governments had proactively decided against misinformation indicated little to no action in Kenya.

This was illustrated in the geo-chart below (Poynter, 2021)

Figure 2: Geo-map of fake news actions globally



The geo-map indicated little effort being put in place. Locally there were no third-party technology companies applying machine learning to resolve this challenge. The only effort was from the government during election periods monitoring hate speech at political events and on social media, where locals flagged them.

We move from the traditional approaches and assess the most advanced neural networks deployed and their inherent shortcomings. These include the following:

1. Simple multilayer Deep Neural Networks with highly optimized hyper-parameters. Riedel et al., 2017; built a single, robust system with an end-to-end design approach that consisted of lexical and similarity features. The model was an MLP consisting of one hidden layer. They achieved 85-88% accuracy on deployment in the FNC. The approach is limited in that it is heavily focused on one element of dis- and misinformation referred to as stance detection. This pre-processing technique is a first step towards identifying fake news that entails assessing what other media houses or news organizations say about the topic or issue of interest. For relatively new topics or issues, detection levels are deficient across media channels; hence, this will handicap the perceptron model in totality. Though the accuracy is very high, between 85-88%, this does not indicate superior performance compared to other approaches. This is because, for media posts that are very current, it's very likely that other media houses will not pick it up, negatively affecting the classification accuracy of this approach that is dependent on this aspect.

2. This is an ensemble model (Largent, 2017); the model employed a 50/50 split on a weighted average between a standard tree-based model known as the gradient boost model and a DNN. They used handmade optimized features. The gradient-boosted classifier, combined with the DNN, improved the model's performance in tackling the fake news classification problem. This model treats the news headline or post as different from the topic, creates features, and compares the two. The headline is run through a pre-trained google vector, while the body is through XGB. Three hidden layers are finally created, and their predictive accuracy is reasonably standard at 57.99%. The model can then be regularized and does not need to employ normalization.

3. Chopra and Jain, 2017 developed a model that used bidirectional LSTM in GRU computing architecture with modifications. They start by leveraging SVM, trained on TF-IDF cosine similarity features. The intention here is to discern whether a headline and article of a post's pairing are related. Where the classification indicates a relation, the next step is triggered. This is the use of LSTM neural networks to label the pairing using three labels -
   agree, disagree, or discuss. Ultimately, the best-performing neural network was the bidirectional conditionally encoded LSTM neural network using bidirectional global attention. The limitation of the model is that in both its train and test sets after implementation on a carefully selected dataset, the model was not generalized in a real-world setting; hence, the accuracy is based only on the training and test dataset. The real-world application was not thoroughly tested.

4. Thorne et al., 2017 developed a stacked model of 5 independent classifiers. The model comprises two layers that seek to improve from a lower layer of weaker classifiers within the two-level set. These five models were composed of complex encoders and neural networks as follows:- 1) vanilla CNNs; 2) independent Encoders; 3) conditional Encoder by Neel Rakholia, 4) multipass conditional encoders with attentive Readers and also containing weighted cross entropy function, and lastly, 5) Kurt Miller's bidirectional LSTMs. The

accuracy exceeded 90%. The only limitation is the lack of consistency in this score due to constant change in the K-Folds during the training phase.

5. Akshay et al. 2017 developed a supervised ML model known as the Siamese Regression model. They used one 1-layer for the model's homologous networks. The final hidden states for each subnetwork are output as the cumulative representation of the model's critical feature categories, consisting of the headline and body text of the posts or news in question. The structured data set used had a disproportionate mix of the three required labels for classification – unrelated (which assuages fakeness, agree, disagree, discuss). The performance accuracy was 52% on the validation data set, and the model achieved 86% on the training data set.

6. In this case, reinforcement learning is used to model the spread of news as a social learning game on a social network, Aymanns et al., 2017. Their model is a multi-agent deep reinforcement learning. They demonstrate the ability of agents to classify fake news after independently receiving signals. Their model is a recurrent deep neural network that applies training through Q-learning. Computationally the model was further illustrated but illustrated mathematically as foundations for model building on a computational network. Assessing the accuracy by the usual model accuracy measurement networks was thus not possible. This model is deemed unusable because of this limitation.

7. Karadzhov et al., 2017 built BiLSTMs (Bidirectional LSTMs that used text encoding to combine semantic kernels. With a very task-specific embedding that would encode a claim together with pieces of potentially relevant text fragments. The deep neural network model was effective in generalization, employed the principles of model robustness, was simple enough to replicate, and was reusable. The model had a performance of 80% from an initial performance of 66%. This model was implemented on a select structured dataset. Thus still limited in ease of deployment.

8. Ruchansky et al., 2017, built a recurrent neural network model. Their RNN ensures that temporal engagements are stored as vectors and are fed into the RNN. Their model is feature reach and includes more than text data and images. They take a universal approach to classifying fake news. Their feature set includes the number of likes, reactions, shares, and tags for each post assessed for genuineness or fakeness. They, however, limited their model-building algorithm use and only worked with SVM, random forest, and Logistic Regression to build their hybrid model.

1.2 Statement of the Problem

With the onset of journalism and the evolution of news reporting, regulation and editorial ethos had long governed and checked excesses of sensational and fact-lacking reporting. However, digital broadcasting regulatory bodies' role needed to evolve to monitor the increased information sources that could be abused. In the digital age, a lot of multinational organizations that don't own content but disseminate it equally as fast as how it's created pause a new challenge

to regulation, censorship, and bringing about punishment to content creators who dis- and/or mis-inform consumers (Ahmed, 2017). In Kenya's case, well-crafted information that was wholly false or contained falsehoods blended with some level of truth was not only generated by local entities but by actors of a foreign nature as well to influence certain political entities' election to offices of interest. In other cases, this influenced favorable legislation for certain political classes.

Efforts to use or employ machine learning to classify and label information and its affinity to carry the truth have been ongoing, as shown by (Ahmad, 2020). This had been ongoing since the US elections of 2016 (The Guardian, 2016), and EU exit polls of the UK showed just how extensive and influential fake news could be on the voting population if left without putting measures in place to reduce and mitigate its adverse effects (Madrigal, 2017). The predominantly negative impact of mis- and disinformation was civic unrest and majorly political unrest based on our focus on election periods, the radicalization of voters and disruption of election processes, and the break out of electoral violence in high-risk regions in Kenya. The vulnerable age of Facebook users likely to consume fake news and spread it accounts for 12% in Kenya, and the youngest voting population accounts for 70% (Tankovska, 2021). This problem led to the cessation of socio-economic activities, massive disruption of commercial activities, collapse or closure of vital services in the affected regions for prolonged periods, loss of both private and public property, and related criminal activity in close locations and neighborhoods. This, unfortunately, ends with the loss of lives.

Mis- and disinformation on Facebook significantly affect the reliability of the mediums (Poddar, 2019). Since Facebook was used by both regulated news agencies and unregulated information vendors, trust was broken to the extent that differentiating what was genuine was very hard; since there were ulterior motives of the peddlers of dis- and mis- information, the ultimate result was chaos and violence in the Country. Regulating a social media platform of Facebook's size needs a practical and durable solution that content moderators could not offer like Cognizant tried and later withdrew its services (BBC, 2019). This was further complicated by the number of Kenyan voting population with smartphones that could access the platform and the limited resources of security officials to monitor, track and prevent the negative impact, let alone anticipate the following challenges of dis- and mis- information during election periods.

The impact of fake news was devastating to civic order, and there has been a successive increase in solutions to tackle this challenge (Setiawan, 2021). Most of the solutions fell short in this attempt, and we still had limited information and gaps that needed to be comprehensively addressed. These include the following; 1) the context, base language, and geographical regions for which these solutions were applied were primarily European and North American in grounding (Hakak, 2021). The pre-processing algorithms, corpus created, and dictionaries used were not optimized to deal with Kenyan-specific data sets. 2) there was little attempt to optimize and test the solutions when applied to the natural languages spoken in Kenya in pure or hybrid forms; this leads to classification techniques not optimized for language variations. Lastly, 3) there were gaps in connecting and correlating fake news to actual victims to draw and build personas. There were gaps in targeting recipients of fake news or creating classification solutions targeting the most affected.

While robust neural network solutions exist that apply hybrid or ensemble model-building approaches, these machine learning solutions are faced with various challenges and problems, predominantly the language and geography of application. These approaches have not been robustly applied in the Kenyan context with pidgin languages. The various challenges include

and are not limited to the following 1) applied on structured data sets with limited application in real-time use cases during the election period; 2) challenges to deployment as a result of generalization challenges; 3) applied on the particular pre-processing step and not the entire classification problem in fake news detections, for example, stance detection or level of agreement between the body and the title of a post, use of simple train-test split without due consideration for validation in multiple folds to reduce overfit; and last but not least, manually parameter optimization as opposed to automation against the more than 12 activation functions and various ways of values adjustments. These are some critical notable problems and gaps that our approach will seek to address Tiwari, S. (2018) and Baheti, P. (2022, March 8).

GeoPoll and Portland (2017) find out that there is an alarming increase, spread, and consumption of fake news or mis- and dis-information in Kenya. They conducted research with a sample size of 2000 respondents and established that 90% of the respondents had seen and consumed false or inaccurate information. There were also 87% of these respondents who were able to conclude that this information was deliberately fake.

1.3 Main objective

The study's main objective was to develop a machine-learning model for classifying fake news on Facebook in Kenya.

1.4 Specific Objectives

1. To investigate and identify attributes that could be used to
   Classify fake news. We initially identified and investigated various variables, including the following:- Facebook posts, after pre-processing, Profiles of the individuals posting, Posts flagged in fact-checker websites, Posts with URLs flagged in fact-checker websites, Title, Text, Subject, Date, Entity, Likes, Category and Status.

2. To develop a machine learning model that uses the identified attributes to classify fake news in Kenya. The focus was on neural networks, with the specific use of automation and ensemble model building. We will focus on adaptive boosting models, and this is because of the inherent good performance they carry during classification tasks. The boosting models will be hybrid or ensembles owing to the robust nature of combining traditional models. The model-building process will use AutoML to ensure painstaking manual parameter optimization and values adjustment are enhanced and implemented quickly with high accuracy. This is further enhanced by cloud computing over Google's TPU to increase computing power and model building, especially with feature reach vast data sets. Specifically, the following algorithms will be experimented on:- a) Ensemble AutoML model building, including adaptive boosting models; b) Decision Tree, supervised model; c) LSTM, supervised neural network models; d) Naïve Bayes, supervised models; e) Random Forest, tree-based supervised models, and f) SVM, supervised neural network models.

3. To evaluate the accuracy and ease of deployment of the model.
   Accuracy and ease of deployment objectives were interrogated using industry-tried and tested approaches. We tested using K-Fold validation; the details and choice of the number of folds will be detailed in the model development section of this document.

1.5 Research Questions/hypothesis

1) Which attributes could be used to classify fake news?
2) Which machine learning model could use the identified attributes to classify news in Kenya?
3) What was the accuracy and ease of deployment of the model?

1.6 Significance of the Study

This research was intended to illustrate the efficacy of automated natural language processing solutions in tackling the triggers of violence. This would greatly assist and support local law enforcement in carrying out their duties during election periods and deter perpetrators of fake news. The study would also directly assist the targeted individuals in staying vigilant and actively taking steps to prevent, reduce and report fake news before many people view it.

1.7. Motivation of the Study

Mis- and Dis- information was only possible because consumers were vulnerable and lacked the tools to genuinely self-determine information on Facebook's platforms. With Facebook on countless occasions being summoned to explain why they hadn't taken decisive action that had ended up being harmful to voters and citizens from many nations, there was a need to propose solutions independent from their immediate and direct participation owing to their lack of motivation to take up this role as their platform grows even more prominent and as they acquire other social media platforms.

In trying to label and flag fake news at scale, data science provides the most reliable approach through machine learning, enabling us to build a model that learns from multiple features from enormous datasets. Machine learning was practical. Applying natural language processing for local dialects was key to solving this problem of flagging and labeling fake news in Kenya's specific case. I propose this homegrown solution to tackle this challenge in our local context. The level of mis- and dis-information witnessed in Kenya circulates predominantly during political periods due to the campaign approaches pursued by the candidates and the supporting public. The highly heightened media coverage of fake news in Kenya during these election periods has always impacted the audience's perception of what's real from what is not. Therefore influencing their decisions for some of the population, as advanced by Oloo, 2021.

1.8 Scope of the Study

- The study targets the Kenyan population in the country. Fake news has been documented during election periods in Kenya (The Wire, 2017). The target for our research was Facebook

users who reside in Kenya and not in the diaspora. We ensure age and gender consideration. We target the male population between the ages of 20 to 65 years.

- Since we would be building profiles of the vulnerable or targeted population, we would visit up to 5 police stations, 1 in Nairobi, 1 in Kiambu, 1 in Machakos, and 2 in Nakuru, to collect data on cases reported on fake news or its harms as documented. This would be necessary to ensure we could validate the profile of the targeted population so that when we start the crawling process to collect the data, we could effectively target and not crawl in general.

- We would seek fake news between 2017 and 2018. This was because the 2017 election year was the official period before being annulled by the courts and set in 2018.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

We reviewed guiding and inspiring literature of previous studies on creating and implementing both classical and automated machine learning approaches to classifying and labeling fake news specific to election or political information. There were limited Kenyan studies on this topic. We would borrow from extensive research in first-world countries of a similar nature. To synthesize representative and existing literature on automated machine learning for fake news classification in an integrated way so that new perspectives on the context-aware application could be made in the Kenya election news as to whether it's fake or genuine.

In our theoretical review, we highlight key attributes necessary for ensuring the classification problem can be solved, and a practical model can be built. We review and distinguish key highlights existing in natural language expressed in Facebook posts, such as titles, headers, and body messaging or source information of Facebook posts that were key to classifying fake and genuine posts. We also assess some key and effective machine learning techniques that would inspire our approach and provide key direction to implement the proposed approach. This was all illustrated in our conceptual framework and demonstrated how the identified variables would be operationalized in the model-building exercise.

2.2 Theoretical Review

There were various practical applications of techniques for classifying fake news whose used cases could be adjusted or enhanced to solve the problem as identified in our study. Automated machine learning to classify and label fake and genuine news at scale primarily draws its foundation from natural language processing theory and practices. This entails the identification of authentic and fake news data sets, pre-processing both, then splitting into test and train data sets before training and validation, and then later deployment on real-world data. It highlighted below some salient attributes and techniques from previous studies of near similar nature.

2.2.1 Attributes that could be used to classify fake news

The attributes distinctly used for this problem were not different from attributes used in other countries globally. The difference was in the combination of these attributes; while most researchers would use one or two, I am including all three in solving this problem. Fake news classification relies on salient features that act as flags in this classic binary classification problem. To classify fake news, reliance on attributes commonly used in natural language processing problems was commonly applied. These attributes include the following: -

- Assessed and mapped Facebook post sources. This attribute helped assess and compare posts' sources if they emanate from well-known and genuine references or sources (Conner-Simmons, 2018). This was an initial step, and key for posts shared concerning news media official pages or links. Media bias initial detection

and training dataset or dictionary building could effectively be implemented by applying human-in-middle approaches in machine learning model building validation. This was through human data review and improved approaches. The resultant dataset was then used to build a database of classified fake and non-fake news, effectively training the model based on the type of news sources. Such key training datasets were adequate for traditional primary aspects of model building (Media Bias/Fact Check, 2021). Analyzing and recognizing the role of Facebook sources remains critical to modeling and classifying how genuine or untrustworthy the Facebook post is and ensures we accurately create a very effective model (Bharadwaj et al., 2020, pg2).

- The polarity of the sentiment in the Facebook post was a key attribute used to assess and classify Fake posts. Usually, a word repeated throughout various Facebook posts would naturally imply it had a high degree of importance, and we verify this after we normalize the occurrence of that word with the size of the document. The focus on building TF-IDF was on Facebook post sentiments during the political periods of interest. The number of times keywords appear carrying fundamental sentiment and polarity of defined ranges helps in classification. The high proportional score in this attribute could be used in comparing critical terms used in fake or genuine news articles depending on the training data set for the model. Used of a critical frequency measuring statistic known as TF-IDF would enable us to interpret text-based data with accuracy after transforming it into a numerical representation. The TF-IDF vectorization was crucial in assessing the importance of sentiment inherent in Facebook posts when repeated or not repeated across multiple posts before applying the model classification algorithms and selecting the best-performing model (Palriwala, 2020).

- Profiles of account holders include bio-data that renders users identifiable as well as geo-data that would help pin the location of individuals posting and validate as information triangulation sources on whether the posting entity was genuine or a fictitious creation by the perpetrators of fake news (Rao et al., 2020, 95-100). Account holder profile was a critical complementary group of data points necessary for fake news detection and classifying their Facebook posts (Boshmaf et al., 2016). These profile attributes explored would include the following:- screen name of the Facebook account user who posted messages of interest for the political periods stated, key creation dates for the posts, count of the number of posts, number of friends following or with activity on the posts, status posts, followers posts, number of groups followed, the actual URL link of the Facebook account, time zone of the posts was critical, the actual location of the account

holder when posting and geographic positioning data if enabled for the user. These were some of the critical attributes to be explored further.

2.2.2 Classification Machine Learning Techniques

Other researchers have used various machine learning techniques to solve this fake news classification problem in the past. Though AutoML was at the heart of our model building and deployment, the techniques used would align and borrow from the following reviewed approaches:

2.2.2.1 Supervised machine learning techniques

This technique entails training a fake news classification machine learning model using well-labeled and categorized data sets as input with an expectation of the output to be either fake Facebook posts or genuine Facebook posts. The commonly used algorithms in this approach, as observed from our reviewed literature, include the following:-

- **Decision Trees**
  This was a classical and standard classification machine learning model that used a leaf node, which was assigned a class label. The other non-terminal nodes, including the root and other internal nodes, contain attribute test conditions that segregate records of different characteristics (University of Minnesota CSE, 2021, pg 150).
- **Random Forests**
  These were more effective when the data, in this case, text data of political period posts on Facebook, had very high or large dimensions, and there was a need to reduce the white noise or non-intuitive aspect of the text data in the model. It mitigates the inherent challenges (Islam et al., 2019).
- **Simple Vector Machine (SVM)**
  The simple vector machine (SVM) could be effectively used for binary or multiclass classifications. SVMs were known to efficiently perform a non-linear classification process while effectively applying a kernel trick. This ensures it consistently maps inputs into high-dimensional feature space (Bedi, 2018).

2.2.2.2 Unsupervised machine learning techniques

This technique entails using deep neural networks in which our outcomes were known, another effectively used approach to resolving this problem. Used of ANN with a particular focus on RELU as the activation function was vital for the fake news classification of specific models as reviewed for this study. One of the critical approaches was built on python on the backend and flask as the front-end user interface. The python backend primarily uses VADER and pickle

packages to manage the supervised machine learning model and allow it to conduct sentiment analysis and allocate polarity scores based on the comments of news or media pieces from every Facebook post one reads that could be classified in more than binary categories, with extreme ends of our classification ends being the fake, not fake and varying degree of fakeness.

- **K-Means Clustering**
  This was an unsupervised machine learning model in which we randomly initialize the K starting centroid at the beginning of the process. After that, each key Facebook post, after pre-processing, was assigned to its nearest centroid and arrived at arbitrarily or through a random walk probabilistic approach. The centroids were then recomputed as the mean of the data points assigned to the respective cluster. This recurs until we trigger our stopping criteria. The aim was to cluster the text around key centroids that determine fake or genuine posts (Foley, 2019).

2.2.2.3 Ensemble machine learning

Ensemble model building effectively applied the various algorithms' strengths into one effective model. This was an approach in which we implemented Facebook post classification modeling, which inherently combines multiple models to get better results. We explore several models that create hybrid deep learning models. This is at the core of our proposed approach. Most models that prove effective, as discussed in the gaps and background of the study, cover work by Riedel et al.,2017, Largent 2017, Chopra and Jain 2017, Thorne 2017, Akshay 2017Aymanns 2017, Karadzhov 2017, and Ruchansky 2017. These models consist of MLP use; XGB infused with DNN; the hybrid mix of CNNs, encoders, and LSTM; regression, while using 1-layer Siamese approaches; multi-agent deep reinforcement learning with Q-learning, Bidirectional LSTM; and effective combination of SVM, random forest and Logistic Regression. Specific models of interest in applying ensemble modeling of neural networks to our problem focus on adaptive boosting with automated machine learning as follows:-

- **Ada boost machine learning model**
  This could be applied and implemented on multiple types of machine learning text classifiers to learn and improve the downsides or shortcomings. This gives it the common phrase “best out-of-the-box classifier” name in most circles. For example, ada boost could be used in decision trees when classifications were wrong or least accurate to give weights until the classification improves on the training data and when tested with the validation data set. Therefore more weight was assigned to the incorrectly classified sample to improve the classification goal in the next cycle (Chengsheng et al., 2017).

This would be achieved on AutoML platforms by applying the following AutoML pipelines and selecting the most effective: - (i) TPOT enables the exploration of many combinations and shows partial results as it runs. It nicely integrates with Dask to parallelize the training tasks. (Moore, 2021). (ii) Snorkel - it ensures we could be considerate of the nuances in the sentiments in the data set as we attempt to classify fake from genuine news specific to political information in Kenya. This approach was also beneficial for classification tasks that started with incomplete data or a complete lack of target labels. Our classification problem may be nuanced for more complex posts (Ratner, 2020). (iii) Used of H20's AutoML pipeline allows us to automate with pidgin or multiple language mix considerations from the data source we would be working with (Pandey, 2019).

With the necessary libraries imported and configured appropriately, the next step was to call the AdaBoost model and optimize our parameters. The critical parameters of interest that were optimized included the following: -

a) n_estimator – The value used was 50 because, above which, no further improvement in model performance was experienced. This was the maximum number of estimators at which boosting was terminated in the proposed model. Usually, the goal was to achieve a good fit, and when we had a perfect fit, the model would stop the learning procedure through early stopping. We sought to avoid underfitting or overfitting. Fitting is an instance where we had errors in the estimators due to bias (Hansen 2021). While overfitting was where we had an error in the estimator due to variance (Hansen 2021). The initial value was 50, which was incremented while observing model performance.

b) Random state – Used 10 to ensure try to improve the reproducibility of the model again for further tuning and improvement while building on the immediate results of the initial model tested since I iterated on the findings as I sought to improve the model parameters from the primary model I began with from the outset. This controlled the random seed for each base estimator. The default was typical "none," initially set to 10, for this process (Hansen, 2021).

c) Base estimator –This was set to Decision Tree as the primary algorithm used before performing adaptive boosting, and it improved the weak learners. This was used for base weighting; when set to default, the weak learners weren't improved, and we would implement a decision tree classifier, which would not achieve our objective (Hansen, 2021).

d) Learning rate – The learning rate employed for this model was 0.001, with one being maximum but initialized much faster and traded off with some accuracy. The intent was to slow the model to improve the learning rate since we were dealing with weak learners. This was the absolute weight we applied to each classifier, then iterated during the adaptive boosting process. There was usually a trade-off between the learning_rate and n_estimators parameters as we sought to optimize our model and improve performance (Hansen, 2021).

e) Change in the activation function varied between sigmoid and softmax and compared the performance before settling on sigmoid as the activation function. ReLU was not ideal because of the computation demand required and the time to let it run on the data; the volume of our dataset consistently crashed the cloud computing environment being used.

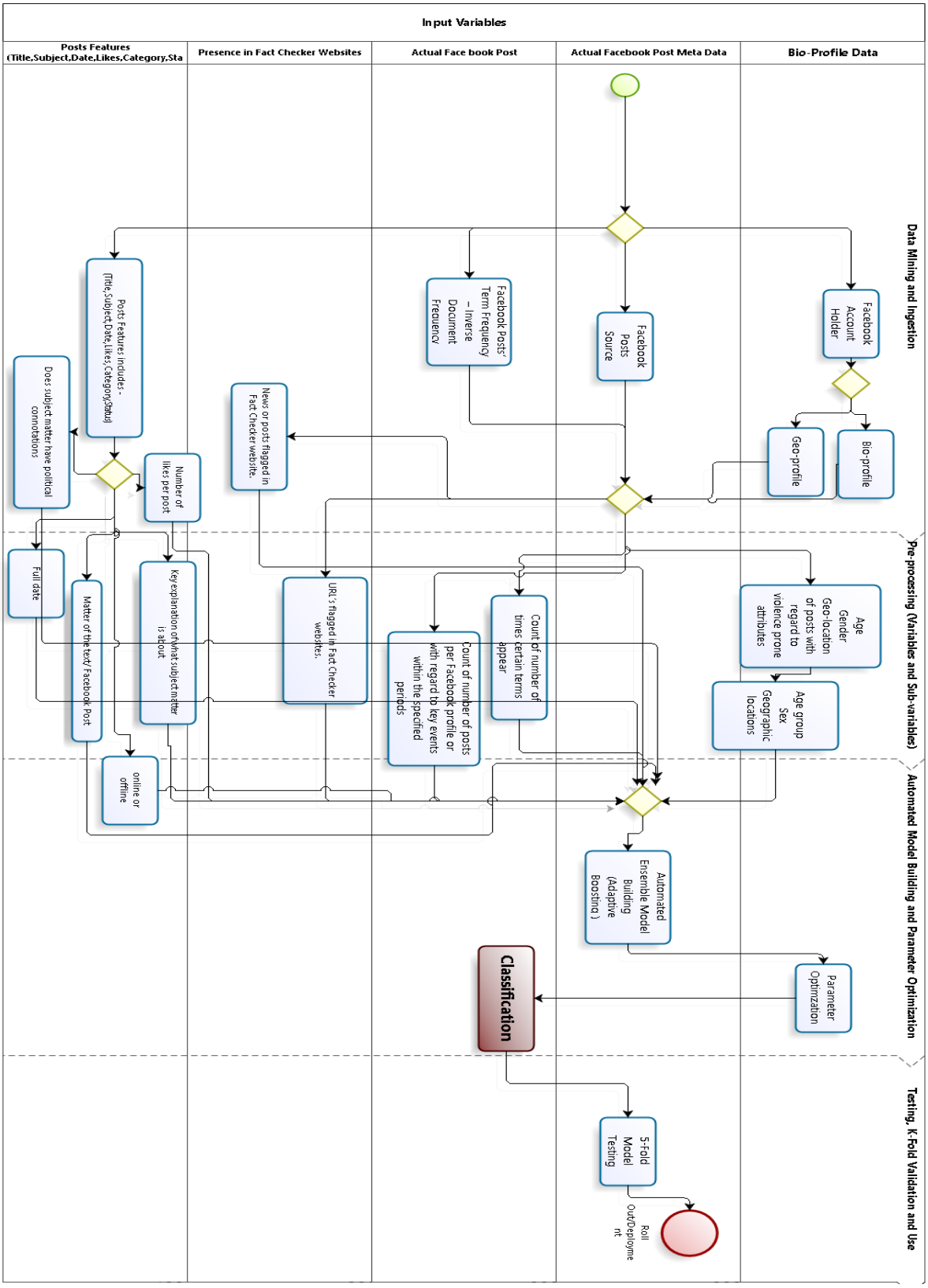
f) Amount of data used, we were using 45,263 rows and 13 columns/variables in total, this proved to be very large for ordinary ML model building, and the decision was to use Google Colab as a cloud computing engine.

g) Normalizing/Scaling data was implemented using TF-IDF as our vectorizer. The size of the matrix after condensing to a dense matrix still proved too large to illustrate in totality. This was summarized for essential aspects and the rest of the TF-IDF matrix availed as a table in the appendix owing to its size.

2.3 Conceptual Framework

The conceptual framework below indicates our model-building and improvement process was iterative. The key dependent variable was the fake news classification which relies on the following independent variables: - Facebook posts' sources, bio-geo profiles of Facebook users, and TF-IDF.

Figure 3: Conceptual framework diagram



2.4 Operationalization of Variables

The variables mentioned and indicated in our Conceptual framework would be operationalized as follows in the below table: -

Figure 4: Variable operationalization

Variables	Sub-variables	Indicators(Symptoms)	Values(data)	Notes/Comments
Entity/Facebook Account Holder	Bio-profile	Age Gender Geo-location of posts concerning violence-prone attributes	Age group Sex Geographic locations	Fake news was highly consumed and generated by the old (40 to 64 years), yet the young (18 to 39) were the largest group consuming social media content.
	Geo-profile	Age Gender Geo-location of posts concerning violence-prone attributes	Age group Sex Geographic locations	
Text/Facebook Posts Source	None	Source of Facebook Post	Binary Values Fake, Genuine (1,0)	
Facebook Posts' Term Frequency – Inverse Document Frequency	None	Count of number of times specific terms appear Count of number of posts per Facebook profile or concerning crucial events within the specified periods	Score	
Presence on the Fact checker website	None	URLs flagged, in Fact, Checker websites. News or posts flagged on the Fact Checker website.	Binary Values Flagged, Not Flagged (1,0)	Used for pre-processing, if the content is flagged as Fake, it will not be genuine and should be filtered out in the data set for training and validation.
Fake news classification	None	Key terms denoting whether genuine or fake Code of words	Binary Values Fake, Genuine (1,0)	The binary values for this dependent variable could be more than binary for some of the unsupervised modeling techniques.
Title	None	The fundamental explanation of what subject matter is about	Unstructured text	
Subject	None	Matter of the text/ Facebook Post	Unstructured text	
Date	Year Month Day	Full date	Date value (Year, Month, Day)	

Variables	Sub-variables	Indicators(Symptoms)	Values(data)	Notes/Comments
Likes	None	Number of likes per post	Number of likes	
Category	Political Apolitical/Non-Political	Does subject matter have political connotations	Binary Political=1,Not Political=2	
Status	On Offline	The state of being online or offline	Binary Online=1, Offline=2	

CHAPTER THREE: METHODOLOGY

3.1 Introduction

We tackle the question of classification of fake news during political events, specifically general elections that occurred once every five years up to 2018 due to the runoff. We ensured to be conscious of the multiple dialects and hybrid mix of languages in expressing their sentiments in the Facebook posts. We started our research by seeking approval to request Facebook for access to large data volumes through its Graph API, distinctly from the election period events in Kenya that covered the period two months before the 2017 poll and two months after the 2018 poll. We then targeted posts from active users of social media accounts two months before and after the 2017 or 2018 repeat poll. Profile of the users included:- age of the user, education levels, socio-economic aspects, and geo-location, whether in urban, peri-urban, or rural regions of the country as categories. If posts or comments were triangulated to an individual in any of our violence-prone regions, we increased our focus on these regions of interest as mentioned in the posts or referenced: - Nakuru, Uasin Gishu, Kisumu, Mombasa, and Nairobi Counties.

Research instruments included specific API's, Tensorflow for computing power, and various python libraries and packages that ensure python powered AutoML processes could be run effectively: - Autosklearn and TPOT in Google Colab, Snorkel, and H20 were not effective after tests and checks with the platform providers. Depending on approvals, Facebook's graph API was used to fetch data targeting profiles of individuals highly likely to have been victims or perpetrators of fake news. This was then pre-processed and imported into the three platforms for our AutoML classification approaches. The best-performing model was selected and deployed using our deployment and management approaches after being evaluated using the K-Fold validation approach. We then compared the models and applied K-Fold validation. Natural language translation was vital for effective data collection and pre-processing. We anticipated some of the text would contain a mix of local dialects, including 'sheng' – Kiswahili-derived or hybrid language. All text from Facebook that contained mixed languages was manually translated using Facebook in post translation, then cross-checking this manually and assigning manual flags. Though a tedious process, building a translation engine for all possible languages spoken in Kenya would require the ground-up building of each language's dictionary, stop words, constructing context-specific corpus, generating lemmatization and TF-IDF that will work with multiple languages in pure or hybrid forms. After assessing this challenge, manual translation and annotation were chosen, and it took one month to complete before model building could be affected. Using factor analysis approaches assessed validity and reliability, ensuring we assessed all factors and settled only on the few necessary for classification. Cronbach's Alpha was used to achieve this requirement.

Depending on the data used and access policy, data was mined from Facebook's platform; Facebook Graph API was the primary tool, and where challenges arose, a manual crawl after key

search results were displayed using predefined vital terms. This was implemented when Facebook Graph API had challenges and ensured data was representative.

Data processing ensured we cleaned and prepped the Facebook posts using NLP principles of tokenization, removed stop words, and performed stemming before the corpora were created.

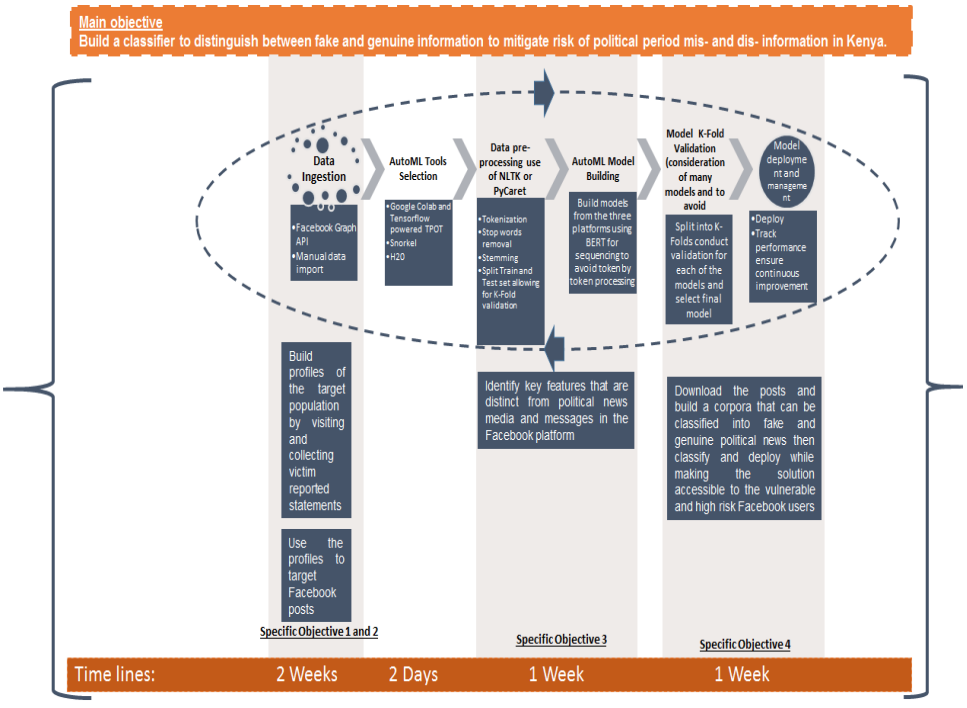
3.2 Research design

The study's main objective was to effectively build a fake and genuine political information classifier, ensuring it's unaffected by the local language, and continue to work on local languages.
The specific objectives entailed the following: -

1. Build credible personas and profiles of the target population.
2. Identify critical features that were distinct from political news media and messages on the Facebook platform
3. Downloaded the posts and built a corpora that classified them into fake and genuine political news.

These objectives were met by implementing the following steps as indicated in the diagram and abiding by the timelines, each of the specific objectives was mapped onto a specific section of our oval-shaped AutoML research design diagram, which indicated the machine learning lifecycle in building the classifier.

Figure 5: Research design illustration diagram



From the research design diagram, the main objective was achieved when we implemented the process in the entire diagram. The specific objectives 1 and 2 were met in 2 weeks when we implemented the processes in the data ingestion and acquisition phase. Tool selection, configuration, data import, and initial set-up, which was the second level in the ML lifecycle section of the research design diagram, was implemented within two days as a foundational process and did not need to align with any specific objectives through an essential step. Specific objective three was achieved in a week through feature engineering and model building. The final step, as entailed in the diagram, was achieved when we implemented specific objective 3, which entailed model deployment, use, and continuous management and improvement.

3.3 Target Population

There was a wide belief from other studies that specific personas and profiles of individuals were more accommodating to fake news than others. Studies have pursued this hypothesis with varying success. Global fake news targeted schemes for political gain had shown that mass targeting of victims was ineffective compared to strategically targeting a select few that impacted the overall vote. The study was targeted at Kenyans currently in the country. All Kenyans residing out of the country were not part of the study owing to the resource intensity in validating insights we got from them. In Kenya's case, we profiled a category of the most vulnerable population and a significant target for fake news.

Key aspects we looked at and disaggregated the population with included: - the age of the user, education levels, socio-economic vulnerabilities to and frequency of use of social media platforms, in this case, Facebook, populations residing in violence-prone urban, peri-urban, and rural regions of the country, the depth and degree of penetration of both mobile phone use and reliable internet connectivity (Guess et al., 2019).

Our population included Kenyan voting age population as of 2017, which stood at 19, 611, 423 registered voters according to IEBC voter registration records (IEBC, 2017), with 80% owning phones and more than 30% owning a smartphone with 50% owning feature phones (Kibuacha, 2021). The vulnerable age of Facebook users likely to consume fake news and spread it accounts for 12% in Kenya. These were the 40 to 64-year-olds; the youngest voting population accounts for 70% of these were 18 to 24-year-olds, according to (Tankovska, 2021). Our focus was to rack and mine posts targeting individuals fitting the vulnerable category who were over 25 years. The central unit of analysis was the Facebook posts and comments in posts disaggregated by political theme or topic in the message to draw out sentiment polarity and classify whether it was genuine or fake from Facebook original posts by individuals, media entities, re-shares, comments, likes and any form of reactions.

3.4 Sampling and Sampling Procedure

To guide the sampling of Facebook posts, geo-location information from users was used to sample and ensure representation. Ensured equal representation of the population in the

sampled Facebook posts and social network activity. The main driver of sampling was random purposive, stratified sampling that ensured representation.

From the finite population of 19,611,423, we used the finite population sample size calculator. We arrived at the below purposive sample distribution on the minimum number of Facebook users of voting age and the number of Facebook posts from each. If we did not have Facebook posts up to the requisite number indicated, we would oversample until the desired minimum was reached.

Equation 1: Sampling formula

$$\text{Finite population: } n' = \frac{n}{1 + \frac{z^2 \times \hat{p}(1-\hat{p})}{\varepsilon^2 N}}$$

Where;
z was the z score
ε was the margin of error
N was the population size
p̂ was the population proportion

Confidence level at 95%, Margin of Error at 5%, and Population proportional representation at a minimum of 50.

We got the below sample size and breakdown of sub-population representation.

Figure 6: Purposive sample results table

#	Metric	Figures	Unit	Comment
1	Total Voter Population	19,611,423	Registered Kenyan Voters	
2	30% of Voters who had smartphones	5,883,426	Estimate 30% with smartphones	
3	The sample size of individuals using finite population formula	385	Estimate sample size	
4	Nairobi County 11% of total registered voters	44	Nairobi estimate voters	Ensuring equal representation of Facebook
5	Mombasa County 3% of total registered voters	11	Mombasa estimate voters	

#	Metric	Figures	Unit	Comment
6	Kisumu County 3% of total registered voters	11	Kisumu estimate voters	posts from individuals coming from Mombasa, Nairobi, Kisumu, Uasin Gishu and Nakuru Counties.
7	Uasin Gishu County 2% of total registered voters	9	Uasin Gishu estimates voters	
8	Nakuru County 5% of total registered voters	19	Nakuru estimate voters	
9	Number of Facebook posts shared or original content from individual meeting profiles of registered voters above (count of 50, on the estimate for each meeting the criteria)	19,250		

3.5 Research Instrument

The critical research instruments were the following: -

1. Data collection was done using API calls to Facebook's Graph API
2. Python Software applied AutoML techniques on Google colab and used Tensorflow in a cloud computing environment.

3.6 Validity and Reliability of the instrument

Diagnostic tests were applied to the variables to check for reliability. Cronbach's Alpha statistic was used for this purpose, with the p-value being key and indicating relevance. Final model validation applied K-Fold validation across all models, and the best-performing model with reliable K-Fold scores was selected. Five folds for validation were used.

We were interested in the following model accuracy scores:

1. Accuracy - the number of correct predictions divided by the total number of predictions, expressed as a percentage.
2. Confusion Matrix - this showed a count of classification that was TP, FP, TN, FN
3. Precision - the ratio of true positives and total positives predicted.
4. Recall - the ratio of true positives to all the positives in the ground truth.
5. F1-score - the harmonic mean of the precision and recall statistics.
6. AU-ROC - Area under Receiver operating characteristics curve; this was the plot of the TPR and FPR on a linear curve.

3.7 Data collection procedure

The primary data collection was seamlessly done using Facebook's graph API. The Graph API process:

- Facebook login and access to the developer section of the platform at the following URL https://developers.facebook.com/.
- It was named, Created, and added the new APP from the Add-New-App menu item.
- Acquired the Facebook App ID and the necessary Facebook App Secret from the developer platform dashboard.
- Created and prepared the Access Token before using it in the python script, connected and ingested targeted data as needed.

3.8 Data Processing and analysis

The classification of the text followed a detailed process and sequential steps as outlined below:

Data ingestion

We kicked off the process by following these key steps: -

- Factor Analysis was achieved when Cronbach's alpha was applied

- Data Pre-processing was critical before model building. We started off by filtering out to ensure sample representation. Considerations were made for variables in Figure 4, which were key in model building. The variables, which include:- details of the Facebook account holder such as bio-profile, geo-profile like age, gender, location of posts with regard to violence, Facebook posts source, Facebook posts' term frequency–inverse document frequency, count of the number of times certain terms appear, count of the number of posts per Facebook profile or with regard to key events within the specified periods, were cleaned and used to filter out for the specific data required to build the model. The other preprocessing techniques include the following:-

    - Tokenization – This entailed breaking down a piece of text, in this case, the various posts on social media, specifically Facebook, from the users after conversion to English to small units called tokens.
    - Removed stop words – Stop words were key joining words like 'a, 'the', 'and'. These were removed to reduce the noise in the processing of the final tokens.
    - Stemmed words from the same word family – This ensured all derivative words were categorized as one type of word to avoid tokens with derivative words appearing multiple times.
    - Corpora created – Collection of corpus specific to our tasks of classifying and labelling fake and non-fake news.

  - AutoML Model building, we implemented classification by applying Autosklearn, Snorkel, and H20 were not effective and thus not utilized.

  - K-Fold Model Validated, the resampling into the K-Fold of our model, was split into a maximum of 5 folds as follows:
    - Shuffled the dataset randomly.
    - Split the dataset into k groups, in this case, five groups.
    - For each unique group, we took and treated the distinct group as a holdout or test set, and the remaining groups were then treated as training sets. After applying AutoML and fitting the model, we then evaluated its test set.

  - Model Selection. This entailed choosing the model. The most suitable model was selected using these machine learning approaches.
    - Confusion Matrix - this showed a count of classifications that were TP, FP, TN, FN
    - Precision - the ratio of true positives and total positives predicted.
    - Recall - the ratio of true positives to all the positives in the ground truth.
    - F1-score - the harmonic mean of the precision and recall statistics.

- AU-ROC - Area under Receiver operating characteristics curve; this was the plot of the TPR and FPR on a linear curve.

Various models were then tested before settling on an adaptive boosting neural network chosen for this binary classification task. The ML model was developed as follows:

1) In colab, various libraries key for the modelling need to be installed; these include:-
   i. requests
   ii. TPOT
   iii. json
   iv. sys
   v. io
   vi. google. colab
   vii. numpy
   viii. pandas
   ix. matplotlib
   x. nltk
   xi. re
   xii. sklearn
      a. sklearn. ensemble to access, then apply the AdaBoostClassifier
      b. sklearn. tree to access, then apply the DecisionTreeClassifier
   xiii. pickle
   xiv. seaborn
   xv. time.

These libraries were applied at the various model-building stages.
To scrape and acquire relevant data, I created an app on Facebook and linked REST API credentials – token and secret key with python for extraction. Then the scraped data was exported to my desktop in excel file format –xlsx for basic cleaning before importing back into Google colab.

2) Data ingestion.

We import our data set using pandas, io and google.colab library. The google.colab library opened a widget that enabled file interaction with our desktop directory to pull the file. The imported file was then prepared for preprocessing by removing unnecessary variables and columns and, most importantly, deleting blank or null cells.

3) Data preprocessing

Data preparation took most of our time and was the second compute-heavy task after modelling. This process began by checking variables and attributes meeting our requirements and selecting as the first step. For our binary classifier, the next step was to remove variables not required, including the variable “Likes” since it had little value in our binary classification from the factor analysis carried out. Advanced filtering was implemented to ensure sample representation was satisfied, and the variables named in Figure 4 for operationalization were also used in this step.

This included:- the use of the Fact Checker site to flag and filter adversely affected posts and/or URL's in those posts Media Bias/Fact Check, (2021). Then the columns containing the Facebook posts and the Label (that denotes Fake - 1, Genuine -0) were selected.
Then after the removal of punctuations in the sentences and thereafter, removal of stop words and rechecking through a word cloud and frequency count for any other stop words that our sklearn stop words dictionary might miss out. These stop words not included in the dictionary were then added to ensure we had no stop words in the final corpus. The words were then stemmed and lemmatized to ensure incomplete words and word families were removed. During the preprocessing, a secondary parse was performed to ensure stop words that were missed were added to the list of stop words, having checked the frequency of the words visualized in a word cloud to ensure all forms of stop words were completely removed.

4) Implementing the Term Frequency – Inverse Document Frequency

The TF – IDF was key in vectorising and made it possible to perform calculations on Facebook posts. The common terms and rare terms' importance in the document was measured and evaluated to ascertain importance. Keywords that were found in Fake (1) as opposed to Genuine (0) Facebook posts would then be evaluated in depth. It was expected that the TF – IDF value would increase proportionally as compared to how frequently a particular word in the document appears. This was then offset by the number of documents in the entire corpus that contained that specific word of interest.
This was achieved using the following approach, having pre-processed the data. Since TF-IDF was a function that checked the importance or lack of therein of a word across documents, and since we don't have documents or books by the conventional definition, we checked the importance of terms across Facebooks posts and treated each post as a document, then filtered out for posts that had on average 72 or more words to avoid our dense matrix created from TF-IDF from being very broad and not meaningful, not all posts were very large, having cleaned and filtered out Facebook posts containing terms that were less than the average count of all terms, our Facebook posts had an average of 72 Terms on the posts collected for the defined time frame 2017 to 2018.

The next step was to initialize the TF-IDF Vectorizer with key metrics to kick off the process. We then set our min_df = 5. This was a critical configuration set that handled terms appearing infrequently or less frequently. It meant we set our Vectorizer to ignore terms that appeared in less than five documents. It was a standard measure. We reviewed the results and then adjusted until we achieved TF-IDF that was meaningful and achieved the objective of discerning the importance of keywords or terms used in fake news. We then set our max_df = 0.95, and this meant that we would ignore terms that appeared in more than 95% of the documents. This was an optimization problem because when too low, the min_df, our Google colab notebook took too long to compute, and the dense matrix was not as meaningful since our rarity discerning metric for terms was too sparse. Having iterated severally on the metric, the appropriate set value was five documents. The max_df was set to ignore 95% of documents with the term in question being checked for rarity or how common it appeared. This then showed us very important documents.

The sublinear transformation, as denoted by "sublinear_tf = TRUE", was a function that performed sublinear transformation scaling on the Facebook posts. With due consideration and interpretation of the meaning of term frequency, it's unlikely when we observe x number of occurrences of a term in a document that the document truly carries x times the significance of a single occurrence. This has led to more research on its variations, going beyond merely counting the actual number of occurrences of a term or Facebook post. The most common modification to this was to use the logarithm of the term frequency instead, which assigned weights to the term frequency and provided a more accurate measure of term frequency. To this end, our TF-IDF calculation achieved varying max_df, min_df and sublinear_tf with respect to how the underlying data or the actual Facebook posts actually were represented. We could then ascertain any variations by tweaking and adjusting the parameters to see any distinct differences with more accuracy. The next step was to extract words or 'feature names' from our TF-IDF Vectorizer as implemented by the sklearn library in python in the Google colab IDE. The 'feature names' were the terms or Facebook post individual words in the matrix column headers that had been checked how many times they appeared in the Facebook posts. Recall we had filtered our posts with less than 72 or more actual words, having cleaned for stop words.

Finally, in the process flow, we created a dense matrix. The density matrix was necessary owing to the very large dimension of the matrix we got from creating the TF-IDF. Visualizing this was a challenge after the vectorization step. To be able to visualize the entire TF-IDF, we created the density matrix, which allowed for the representation of a mixture of states using probability in a non-linear combination using an algorithm. The large matrix was thereby shrinking to a more meaningful and condensed to a much smaller representation of its original size (Cornel, 2015). The matrix, without transforming into a density function, took the shape of a (45246 rows, by 4000 columns) at the highest to (45246 rows, by 4000 columns) at the lowest. This kept on varying depending on our min_df, max_df and sublinear_tf parameters. The dense matrix was then converted to a list and finally a data frame, using the pandas library in Google colab python IDE. The size of the dense matrix remained very large, and limited visualization owing to the more than 1000 rows. We utilized the pandas library to export the results to excel and reviewed these. Notably, there were no differences in the TF-IDF between the data filtered for the labels of Fake (1) and Genuine (0) news. These metrics computed both for the category of data whose Facebook posts were genuine and the category of posts whose Facebook posts were fake, had no clear difference.

5) ML model building

Once the data pre-processing using the sklearn python library in Google colab was complete, the model was built and supported by the insights gained in vectorising using TF-IDF, which discerns any differences in the Facebook posts that were either fake or genuine. We also had critical insights into the descriptive statistics on the distribution of the data and the completeness and missingness of the data. The initial process in building our model made use of the sklearn library to implement the adaptive boost neural network algorithm known as AdaBoost (Mohri, 2021). This was an ensemble neural network model which improved the weak learners of the commonly used decision tree - a supervised machine learning approach.

There were many approaches for building ensemble ML Models (Dietterich, 2000) of the AdaBoost neural network type, which include: - bagging, boosting, and randomization. For this NLP problem, our focus would be the boosting approach, and we would focus on adaptive boosting mainly. This enabled the variation of the parameters that optimize various aspects of our model, as discussed in our implementation below. We implement our model as follows: -

Reviewed pre-processing results and ensured the data was completely clean as intended, and stopwords using the sklearn library of stopwords were updated with more stopwords evident in the data during data quality checks. The data set was then split between training, testing and validation to align with our proposal of 5 folds in our K-Fold validation. The training, test and validation data sets were very large. We would split these and optimize our compute engine in Google colab by moving from normal GPU to TPU in Google Tensorflow.

CHAPTER FOUR: ANALYSIS

4.1 Introduction

In this section, we delved into detailed analysis. This involved the key step of data comprehension by calculating descriptive statistics followed by the model building, which involved preprocessing, feature selection, model training, testing and validation, and lastly, tuning to improve deployment by the key model performance assessment metrics as illustrated. The model selected, which was actually an ensemble model called AdaBoost, or adaptive boost, was pitted against other models for binary classification of hate speech and compared using similar metrics. The distinct performance difference was then clearly outlaid to illustrate the superiority of the AdaBoost or Adaptive Boost model.

4. 2. Data demographics - descriptive analytics about the collected data

The data set we worked with was pre-processed and arranged into a data frame in python's IDE in Google, named Colab. It had the following key characteristics:-

It had 45,263 rows and 13 columns/variables in total. The data set contains posts from Facebook mined through the Facebook Graph API for the election period of 2017 and the repeat poll thereafter.

The dataset was from posts by Kenyans in the country, excluding diaspora posts as much as possible by location triangulation as indicated in the research design and the sampling framework. The posts consisted of sentiment from the 30% fraction of 19,611,423 registered voters as reported by IEBC prior to the elections.

The descriptive analytics was gained from analyzing the data frame in python of this dataset. We mainly apply natural language processing techniques and arrive at the following descriptive statistics.

Using pythons ".describe ()" and ".info ()" calls, we were able to get key meaningful statistics with regard to our type of data. It mostly contained text data and other variable types ranging from string to numeric.

Figure 7: Description of the data frame

status	
count	45263.00
mean	0.480768
std	0.499636
min	0.00
25%	0.00
50%	0.00
75%	1.00
max	1.00

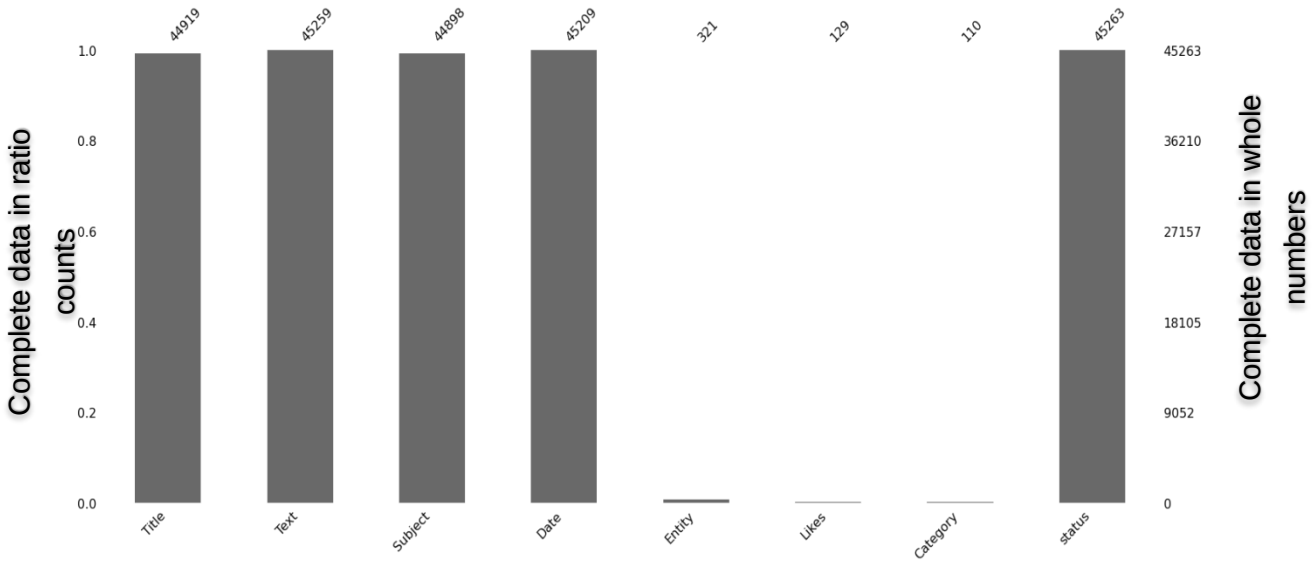
We then did a count of the non-null or non-empty/blank values in our text data frame downloaded and cleaned from Facebook. The result we got was from applying the python .info () command we got the following result.

Figure 8: Table showing counts of missing and non-missing variables and data types

# Column	Non-Null	Null Counts	Count	Dtype	Variable Type
0 Title	44919	344	Non-Null	object	String
1 Text	45259	4	Non-Null	object	String
2 Subject	44898	365	Non-Null	object	String
3 Date	45209	54	Non-Null	object	Date
4 Entity	321	44942	Non-Null	object	String
5 Likes	129	45134	Non-Null	object	Numeric
6 Category	110	45153	Non-Null	object	String
7 Status	45263	0	Non-Null	int64	String
8 Presence/Absence in Media Bias/Fact Check	1003	44260	Non-Null	object	String
9 Geo-location	44919	344	Non-Null	object	String
10 Age	40000	5263	Non-Null	object	Numeric
11 Gender	1337	43926	Non-Null	object	String
12 Count number of posts per Facebook Profile	67	45196	Non-Null	object	Numeric
13 Count of number of times certain terms appear	19732	25531	Non-Null	object	Numeric

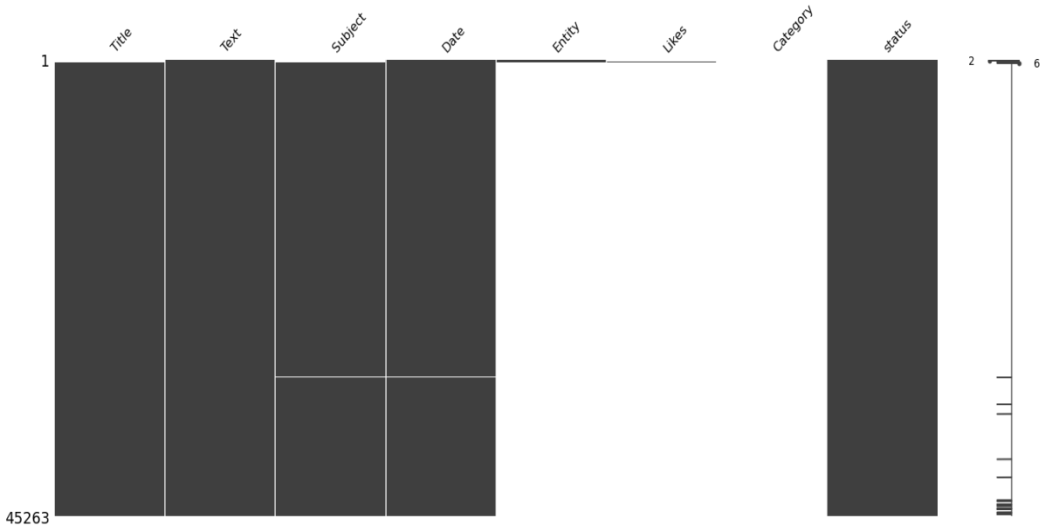
The column names of the data frame were indicated above, and counts of non-null and missing values are presented in the table; we could plot bar graphs showing data completeness in Figure 11 below, which indicated non-null value counts for the various variables as shown below:

Figure 9: Bar graph showing a sample of most and least complete data in the data frame



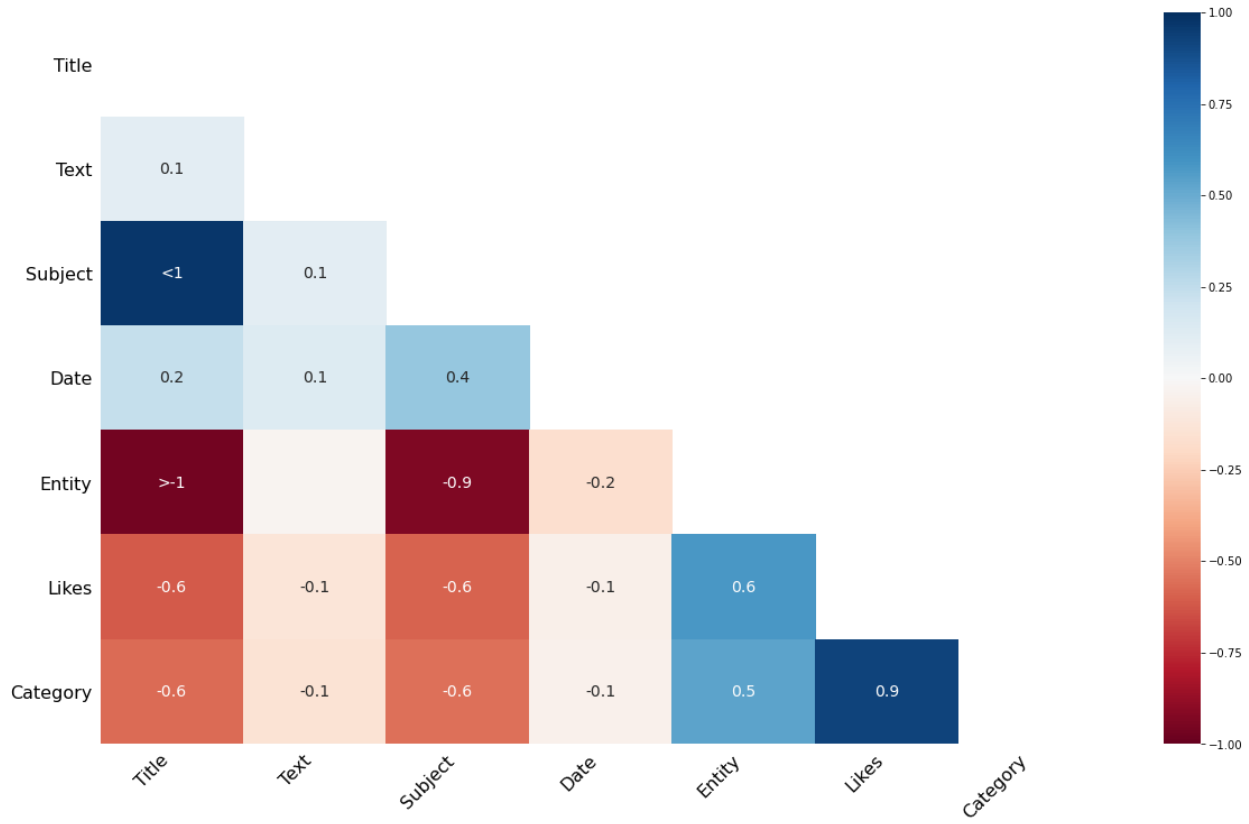
The graph above shows counts of the data. The bar graphs indicated data completeness. From the max total count of 45,263, the size of the bar below this count shows the data available, and it's represented as proportions on the left Y axis, with the maximum being 1. At the same time, they are represented as whole numbers on the right Y axis. The variables "Entity," "Likes," and "Category" were the least complete. They represented less than 50% of the whole dataset.

Figure 10: Matrix graph showing a sample least missing and most missing values



The graph above represents the completeness of data with more distinction on the actual proportion of available data from the three least represented variables, “Entity,” “Likes,” and “Category,” as compared to the rest of the variables.

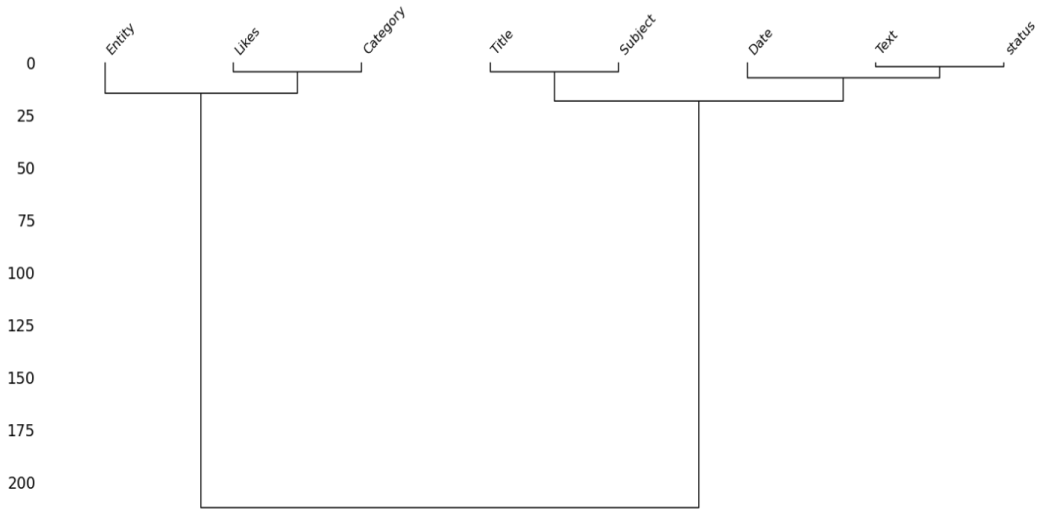
Figure 11: Heat map showing sample select correlation between nullities of some of the variables in the data frame



The heatmap above identified correlations of the nullity between each of the different columns. It identified the relationship in the presence of null values between all and each of the respective columns.

Values close to positive 1 indicated that the presence of null values in one column had a very strong correlation with the presence of null values in any of the other columns. The values that were close to negative 1 indicate the presence of missing or null values in one column was weakly or completely not correlated with the presence of null values in any of the other columns. This means that when we observe null values in one column, there were mostly always data values present in the other columns from our data frame. This was key in developing the model and in identifying features of interest.

Figure 12: Dendrogram showing hierarchical clustering of some of the select columns/variables that had strong correlations in nullity



The above dendrogram clustered the variables with correlated nullity; we observed that the more separated the columns in the tree were, then the less likely the null values could be correlated between the columns.

4.3 Objective one results

This objective entailed feature identification and finding distinct attributes that could be effectively utilized to develop our ML model for classifying fake news. I sought to investigate and identify attributes that could be used to classify fake news. The first attribute uncovered from our literature review that we would investigate further and test out was mentioned by (Conner-Simmons, 2018), (Media Bias/Fact Check, 2021), and (Bharadwaj et al., 2020). They elaborate more often than not with a high likelihood a Facebook post from a well-known fake news posting site would carry falsehoods.

The second key attribute, as mentioned (Palriwala, 2020), involved assessing the polarity of the sentiment in the post. We could assess this by finding the TF-IDF and checking which posts had terms with the highest TF-IDF from our hate speech, identifying key Labels or words identified. This was distinctly represented in the model to check if Facebook posts with terms carrying high TF-IDF were labeled fake depending on the polarity of the term as indicated by the TF-IDF values.

The last key attribute we investigated for classifying fake news was the profiles of account holders, as mentioned by (Rao et al., 2020, 95-100) and (Boshmaf et al., 2016). Accounts previously flagged for posting fake news had been known to post fake news repeatedly. It is expected that such flagged accounts would classify as fake and would help in the preprocessing of the data.

The attributes mentioned were key in labeling Fake (1) or Genuine (0) news covering political statements and collected during political periods. These terms included the following:-

Facebook post sources

The polarity of the sentiment in the Facebook post was a key attribute used to assess and classify Fake posts

Profiles of account holders.

The approach selected effectively utilized and applied all the attributes selected to increase accuracy and filter out data sets during preprocessing and while model building increased the effectiveness of our model as demonstrated.

Investigation of correlation relationships with our binary label Fake (1) or Genuine (0) for whether the news was genuine or fake.

In the process of identifying and investigating attributes that were used for classifying fake news, basic descriptive analytics, cross-tabulation, correlation, and factor analysis uncovered valuable insights. This was described in detail as follows:-

Binary Correlation Variables	Phi	Cramer's V	Pearson's R	Spearman Correlation	N of Valid Cases
Label (Genuine or Fake News – 0, 1) and Flagged in fact checker sites	0.000	0.000	0.000	0.000	44612
Labels (Genuine or Fake News – 0, 1) and Account Holder Profile Genuine	0.000	0.000	0.000	0.000	44612
Likes and Labels (Genuine or Fake News – 0, 1)	0.000	0.000			45242

The table above shows binary correlation statistics for the Label of interest showing fake or genuine news status and the variable that denoted whether the Facebook post was flagged in the fact checker sites we were checking to help in preprocessing. Since both were categorical variables, we used Phi and Cramer's V, which denoted a clear, strong correlation of -1. We observed that if a Facebook post had a hyperlink or links to a site that was flagged in the fact checker websites as mentioned by (Conner-Simmons, 2018), (Media Bias/Fact Check, 2021) and (Bharadwaj et al., 2020, pg2), we were using, then it was highly probable that the Facebook post was fake. Therefore this variable was statistically significant as denoted by a significance score of less than 0.5, at 0.0000.

The table above shows binary correlation statistics for the Label of interest showing fake or genuine news status and the variable that denotes whether the profile of the account holder of the Facebook post was genuine to help in preprocessing. Since both were categorical variables, we also used Phi and Cramer's V that denoted a clear, strong correlation of -1 between the profile of the account holder of a Facebook post and whether the label of a Facebook post was genuine or fake, as mentioned by (Rao et al., 2020, 95-100) and (Boshmaf et al., 2016) Therefore this variable was statistically significant as denoted by significance score of being less than 0.5, of 0.0000, and high binary correlation.

The table above shows binary correlation statistics for the Label of interest showing fake or genuine news status and the variable representing Likes of the Facebook posts. This information was critical in preprocessing. Since one was a categorical variable and the other a numeric variable, we used the Spearman correlation statistic that denotes a clearly very weak correlation of 0.05 between the number of Likes and whether the Facebook post was genuine or fake. We observed the statistical significance scores were below 0.05. Since our correlation relationship was weak, we concluded the number of likes was not a strong attribute or feature for model building.

After assessing the relationship using binary correlation, the next step was to assess factor analysis and gain some insight into reliability using Cronbach's Alpha. Factor analysis was then used to determine how reliable our measures ensure the selection of factors that were most promising in effectively classifying the labels.

The next step was assessing the reliability of the selected variables selected. The Cronbach's alpha when we performed it on standardized items was 0.448, which indicated the weakness or inability of all 4 items to measure reliability cumulatively. However, on checking whether Cronbach's alpha improved on the deletion of each of the items/variables one by one. It was evident that Likes when removed, we had a reliable increase and improvement of the reliability tests in our factor analysis. Cronbach's alpha improved from 0.448 to 0.596, which was still rejected, though improved, but below the threshold of 0.7.

4.4. Objective two results

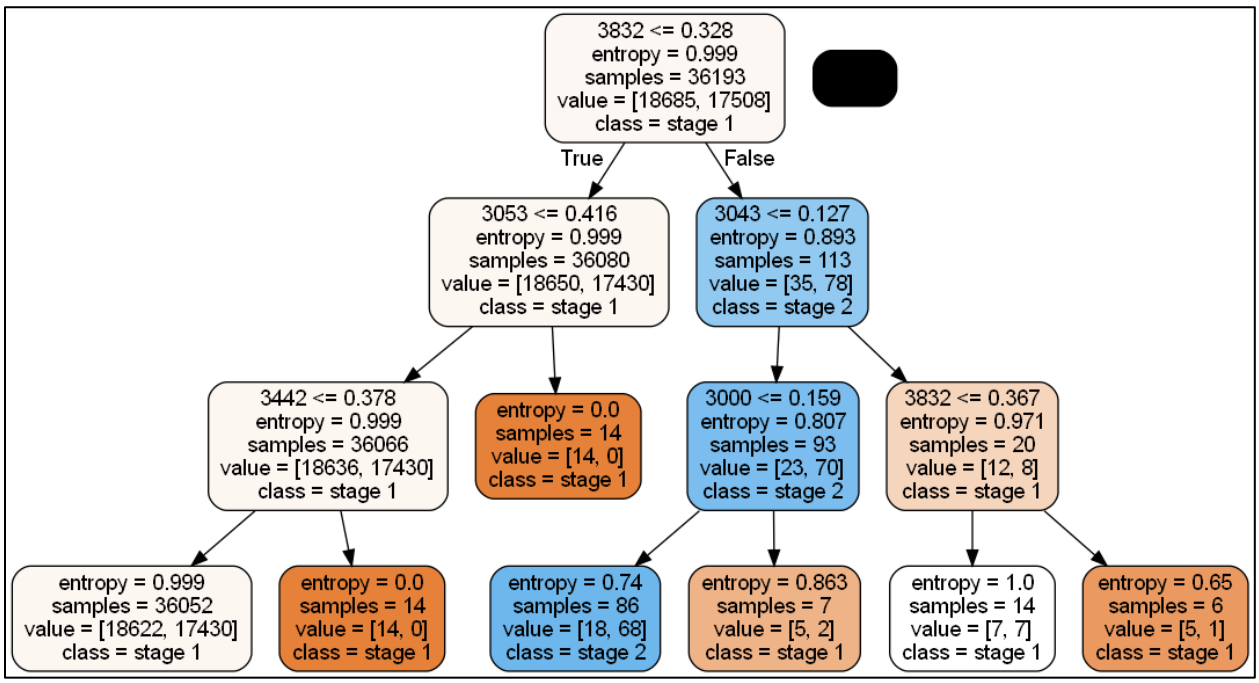
To develop a machine learning model that uses the identified attributes to classify fake news in Kenya. The ML model selected was built on neural networks that applied adaptive boosting of decision trees to improve the performance of the weak learners and enhance overall performance in our model.

The resulting tree-based adaptive boosting model is run by defining the base estimator as a decision tree and pruning, with the max depth set at 3, as shown below.

Figure 13: Defining AdaBoost parameters

```
# We'll use 100 weak learners to build a strong learner
classifier = AdaBoostClassifier(base_estimator=DecisionTreeClassifier(criterion='entropy', random_state=7, max_depth = 3,
                                                                       min_samples_leaf=5,max_leaf_nodes=7),n_estimators=100)
classifier.fit(x_train,y_train)
```

Figure 14: The combined decision tree and AdaBoost model developed



The originally developed model had a very large unpruned max_depth. After effecting pruning and confining the max_depth to 3, above is a visualization of the developed decision tree.

From our model above, we make use of entropy to define the decisions made in our model as opposed to gini. The rules, therefore, are as follows:

1. If the task at the root node indicated by Index 3832 <= 0.328 and the task at index 3053 <=0.416, then in the decision tree, we are currently in stage 1. There are thirty-six thousand and sixty-six cases (36066) covered by this rule. These were observed to be in stage 1.
2. If the task at the root node indicated by Index 3832 <= 0.328, the task at index 3053 <=0.416, and the task at index 3442 <= 0.378, then in the decision tree, we are currently in stage 1. There are thirty-six thousand and fifty-two cases (36052) covered by this rule. These were observed to be in stage 1.
3. If the task at the root node indicated by Index 3832 <= 0.328, the task at index 3053 <=0.416, and the task at index 3442 <= 0.378, then in the decision tree, we are currently in stage 1. There are fourteen cases (14) covered by this rule. These were observed to be in stage 1.
4. If the task at the root node indicated by Index 3832 <= 0.328 and the task at index 3053 <=0.416, then in the decision tree, we are currently in stage 1. There are fourteen cases (14) covered by this rule. These were observed to be in stage 1.
5. If the task at the root node indicated by Index 3832 <= 0.328, the task at index 3043 <=0.127, and the task at index 3000 <=0.159, then in the decision tree,

we are currently in stage 2. There are eighty-six cases (86) covered by this rule. These were observed to be in stage 2.

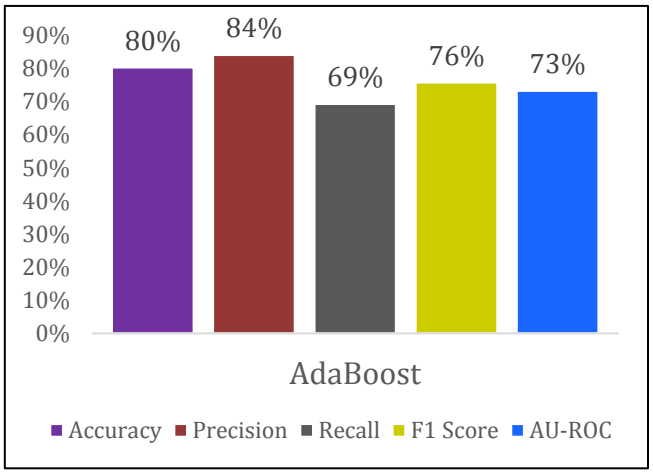
6. If the task at the root node indicated by Index 3832 <= 0.328, the task at index 3043 <=0.127, and the task at index 3000 <=0.159, then in the decision tree, we are currently in stage 1. There are seven cases (7) covered by this rule. These were observed to be in stage 1.
7. If the task at the root node indicated by Index 3832 <= 0.328, the task at index 3043 <=0.127, and the task at index 3832 <=0.367, then in the decision tree, we are currently in stage 1. There are fourteen cases (14) covered by this rule. These were observed to be in stage 1.
8. If the task at the root node indicated by Index 3832 <= 0.328, the task at index 3043 <=0.127, and the task at index 3832 <=0.367, then in the decision tree, we are currently in stage 1. There are six cases (6) covered by this rule. These were observed to be in stage 1.

4.5. Objective three results

1. To evaluate the accuracy and assess the ease of deployment of the model.

The AdaBoost model chosen yielded promising results because the data cleaning and processing made a combined use of various techniques. This ensured the data ran through the model met some necessary key pre-requisites before being selected for model building, testing, and validation with actual data.

Figure 15: Showing AdaBoost Model Using Decision Tree as Base Estimator



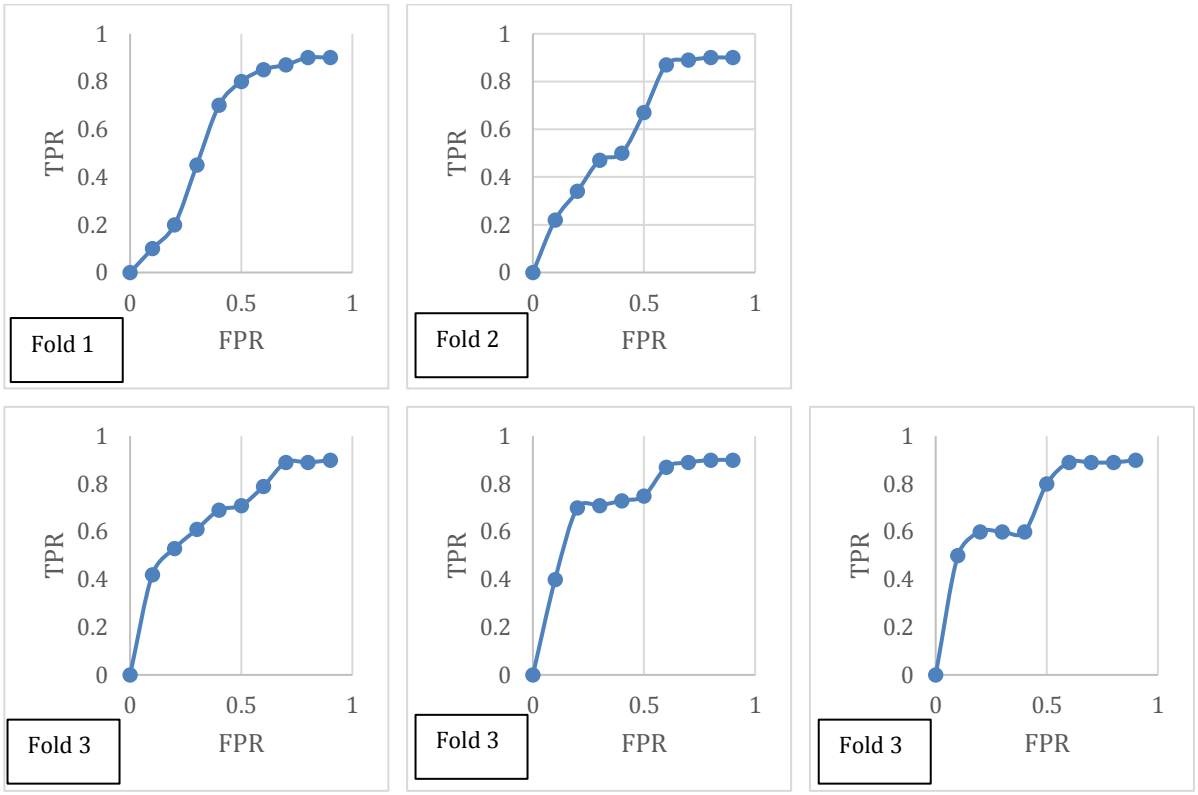
The figure alongside shows how our selected model performed with regard to the measured accuracy, precision, recall, F1 score, and AU-ROC values after implanting the K-Fold validation approach.

Figure 16: Showing the chosen AdaBoost Model confusion matrix for each of the five times it was performed in our K-Fold/5-Fold approach

Key											
0	1	Fold 1		Fold 2		Fold 3		Fold 4		Fold 5	
T N	F P	83 9	16 2	97 7	14 6	87 9	82	100 1	56	79 3	79
F N	T P	31 2	49 7	17 4	51 3	27 2	57 7	150	60 3	35 8	58 0
0	1										

The figure above shows how our model performed on a confusion matrix. The TN values were higher than the FN values, as well as the TP for each FP. This would, therefore, clearly lead to a high TPR and low FPR, which is critical in determining AUC-ROC.

Figure 17: Showing TPR and FPR and AUC-ROC on the test data



The model was then deployed to check and test TPR and FPR on the 5 folds of data sets. It performed as shown above to give the above FPR and TPR values and respective AUC-ROC values.

Figure 18: Showing model accuracy, with respect to other model performance metrics, precision, recall, F1 score, and AUC values for each of the models

Neural Network Model	Accuracy	Precision	Recall	F1 Score	AU-ROC
Model 1 - AdaBoost Model (Decision Tree as the Base Estimator with 100 weak learners)	80%	84%	69%	76%	73%
Model 2 - Decision Tree	54%	44%	32%	37%	52%
Model 3 - LSTM	51%	56%	39%	46%	50%
Model 4 - Naïve Bayes	43%	39%	15%	22%	49%
Model 5 - Random Forest	41%	80%	35%	49%	43%
Model 6 - SVM	54%	48%	41%	44%	47%

The figure above shows a summary of the results for the other models as compared to the selected model. The AdaBoost model using a decision tree as the base estimator outperforms the LSTM, Naïve Bayes, Random Forest, SVM, and traditional decision tree models.

4.6 Discussion of results- Compare and contrast your results with results obtained by other studies

The results, as compared to other studies, had key similarities and differences in the entire process of data collection to model building and final model tuning.
Specific to the attributes that could be used to classify fake news, our investigations bore the following key findings and observations with regard to the first objective:-

- **Decision Trees**

  University of Minnesota CSE, 2021, as compared to our AdaBoost model with the decision tree as the base estimator, were very different in their approaches. This model takes a mathematical approach to calculating the decision tree without necessarily presenting the final model outlook as we did in our approach. The two completely contrast with each other because in using our model, we identified the attributes of interest and started off by testing relationships and correlations of these attributes to the ident direction of the relationship. University of Minnesota CSE, 2021, did not apply this initial step in attribute identification on the same attributes we identified. The key approach here is separating records or data points using the inherent difference in characteristics. There was no attempt to identify attributes close to the ones we used for our approach.

- **Random Forests**

  Islam et al., 2019 presented a Random Forest model that is more effective when the data, in this case, text data of political period posts on Facebook, had very high or large dimensions and there was a need to reduce the white noise or non-intuitive aspect of the text data in the model. We limited our approach to the

following key attributes – 'Title,' 'Text_Resized,' 'Embedded URLs,' 'Actual_URLs,' 'Subject,' 'Date,' 'Month,' 'Month words,' 'Entity,' 'Likes,' 'Flagged_in_fact_checker_sites,' 'Account_Holder_Profile_Genuine' and 'Category. Although the Random Forest Model, as shown by Islam et al., 2019 could take more variables than the ones mentioned in my model. They used completely different attributes, with "Facebook posts" and 'URLs' being the only similar attributes.

- **Simple Vector Machine (SVM)**

  Bedi, 2018 applied binary classification of a very large dataset just as we did. The only key difference is that our model had a much larger volume of data, and our classification was informed by 12 attributes at the outset, which helped in increasing our model accuracy and deployment. Bedi, 2018 made use of an SVM which is known to efficiently perform a non-linear classification process while effectively applying a kernel shrinking. This ensures it consistently maps inputs into very high-dimensional feature space. However, the attributes only involved the test in being classified and the label. There was no consideration for multiple languages. These differences ultimately limited the performance of their model, as shown in Figure 17.

- **K-Means Clustering**

  This was an unsupervised machine learning model, where Foley, 2019 made use of unlabeled data, and the model would later categorize the data set after preprocessing and standardization. It proved incomparable or unusable in this case since our data was already labeled. The data set we were working with had binary labels of being either Fake or Genuine. The use of this model would not prove useful or relevant. Therefore our modeling discarded this model completely since unsupervised learning would not yield the kind of output we sought from the outset.

  We next assessed the key similarities and differences after we developed various machine learning models that used the identified attributes to classify fake news in Kenya. These are discussed in detail below:-

- **Decision Trees**

  In its use of Decision Trees, the University of Minnesota CSE, 2021 is able to perform labeling just as we were able to in our AdaBoost model. The key difference is that they go into the mathematical approach of creating the decision tree without coding or programming language and technology use. The AdaBoost model proposed was developed with the decision tree as a base estimator of the ensemble model built. The classification model we advance goes further in ensuring attributes that will further enhance the model building. For example, filtering out data sets already flagged in fake news tracker sites are removed in the preprocessing step, as well as Facebook posts that contained URLs known to link to Fake news posts.

- **Random Forests**

  Islam et al., 2019 built a SARF – Semantics Aware Random Forest that effectively extracts the attributes and features used by the Decision Trees to generate the labeling predictions and thereafter selects a subset of the relevant predictions in the predicted classes. However, this model performed very well on 30 known real-world text datasets in a test environment. When used on our data set, it fell short in terms of accuracy and all the other performance metrics used for comparing with our AdaBoost model. These performance statistics will be further advanced in our next section on the actual performance of all the models. Both models can effectively handle high-dimensional datasets, with the only limitation being computing power. However, since the Sentiment Aware Random Forest by Islam et al., 2019 is a traditional model based on several Decision Trees as compared to the AdaBoost model, which uses a decision Tree as a base estimator and goes further to optimize the weak learners from the initial classification function, we would, later on, have a higher accuracy from the AdaBoost model as compared to the Random Forest because of this key difference.

- **Simple Vector Machine (SVM)**

  Bedi, 2018 proposed the use of an SVM for binary classification that is linear, doesn't take into account higher than binary dimensions, and made use of our approach in pre-processing and the development of TF-IDF. Their model, however, misses updating the dictionary used for removing stop words and doesn't update the dictionary for lemmatization as well. These are key since some of the words in our Facebook posts are mixed dialects or non-English words. The final prediction function for our model is non-linear, while the SVM model is linear. Because of these very different approaches, we end up with two very different levels of accuracy and overall performance for the models, as we would later on see.

We evaluated the accuracy and deployment of each model as follows below:-

The Decision Trees proposed by the University of Minnesota CSE, 2021 had foundationally more similarities that made up our chosen model but were not enhanced to handle multi-dimensions to the degree that the chosen AdaBoost model could work due to how the model was optimized, and the parameters were tuned. Hence the – accuracy, precision, recall, F1 Score, and AU-ROC values were all below our chosen model with a huge margin of the highest gap at 37% and the lest gap at 21%, the AUC-ROC and Recall values, respectively. The overall model accuracy had the AdaBoost outperforming the Decision Tree at 80% and 54%, respectively.

The Random Forest, proposed by Islam et al., 2019, due to its ability to handle high dimensionality and work with several trees before aggregating its final model to a set of high-performing trees, enabled it to perform as close to the AdaBoost model since its contained an aspect near similar for the AdaBoost in optimizing weak learners and using the decision tree as

the base estimator. There was only a 4% gap in precision value between the Random Forest at 80% and the chosen AdaBoost model at 84%. However, the accuracy for the latter at 80% was 40% higher than for the former.

The Bedi, 2018 SVM was built founded on linear modeling with no consideration for a high count of attributes of very large data sets. The pre-processing was nearly similar but without due consideration for language richness or the need to enhance the stop words and lemmatization dictionaries to counter multiple languages. This, therefore, led to an (SVM) with performance metrics below the AdaBoost model. The Accuracy, Precision, Recall, F1 Score, and AU-ROC values were at 54%, 48%, 41%, 44%, and 47%, respectively.

Comparatively, the chosen AdaBoost model only came close in performance metrics to the Random Forest, as proposed by Islam et al., 2019. The other models came inferior in performance. In terms of use, though all models can be manually used, the AdaBoost model can be ported or applied by anyone by simply importing the "AdaBoost.Pickle" file generated in python as indicated below:

Figure 19: AdaBoost.Pickle file used for exporting and transferring the built model for use.

```
def analyseText(text):

    cls = pickle.load(open("adaboost.pickle","rb"))
    vct = pickle.load(open("vectorizer.pickle","rb"))

    # First we need to clean the text given
    text = cleanText(text)

    # Then we need to vectorize the text
    text = vct.transform([text])

    # And let's predict results using vector
    pred = cls.predict(text)

    decision = "neutral_indeterminate"

    if pred[0] == 0:
        decision = "Fake news"

    elif pred[0] == 1:
        decision = "Genuine news"

    return decision
```

This ensures the model can be used by anyone and retrained for further enhancing performance if the metrics indicate a drop in performance.

4.7. Summary of results

The models were assessed on various performance metrics, including - accuracy, precision, recall, F1 score, and AUC-ROC values. Accuracy, which evaluates the number of correct predictions divided by the total number of predictions, is expressed as a percentage. Precision which denoted the ratio of true positives and total positives predicted. The recall is a statistic that measures the ratio of true positives to all the positives in the ground truth. F1-score measured the harmonic mean of the precision and recall statistics in the respective models. Lastly, the AUC-ROC, which means Area under the Receiver operating characteristics curve, was the plot of the TPR and FPR on a linear curve. Overall the AdaBoost model had higher accuracy and performed better on the test and validation data, using the K-Fold validation approach as

compared to the SVM, Random Forest, Naïve Bayes, LSTM, and Decision Tree ML Models. We observe that AdaBoost outperformed all the other models on the same dataset, with only random forest coming close in the precision metric at 80% and 84% for random forest and AdaBoost, respectively. This was expected because AdaBoost improved the weak learners from the Decision Tree model as its original algorithm.

With regard to accuracy, the AdaBoost model had the highest accuracy at 80%, while the least accuracy was attained by the Random Forest model, which barely attained an accuracy of 41%. Naïve Bayes was just 2% above the performance of the Random Forest. The decision tree and SVM models all had an accuracy of 54% each. With regard to Precision, only the AdaBoost and Random Forest models managed to attain values of 80%. The Naïve Bayes model attained a precision value of 39%, with the Decision Tree and SVM managing 44% and 48%, respectively. With regard to Recall values, all the models didn't attain high values for Recall. With AdaBoost at 69% and the lowest score of 15% attained by Naïve Bayes., F1 scores were below 50% other than for the AdaBoost model, which attained a score of 76%. The AU-ROC curve values were at 73% for the chosen model, with the rest of the tested models attaining AU-ROC values of 50% and less.

CHAPTER FIVE: CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

The attempt we made to develop an ML model for hate speech classification in the Kenyan contexts specific for political messaging was successful as tested and validated using five folds for testing and comparing among various models, namely:- SVM, Random Forest, Naïve Bayes, LSTM and Decision Tree ML Model.

This chapter provides a detailed conclusion of the findings we got after carefully assessing the results. The results summary in chapter 4 highlights this in broader detail. The contributions we have made to similar studies done in the past are also highlighted. A careful assessment of all relevant studies mentioned in our literature review and how the study conducted has contributed to the advancement in knowledge in this field is highlighted and broadly discussed. We hope the work done will be critical in informing further advancements in research in this area. We also made an observation on the opportunities for future research in this area. In this chapter, we will also highlight policy actions that can be effected by all knowledge generated from this study.

5.2 Conclusions

While this study focused on ensemble model building and how this was utilized to build a model for classifying fake and genuine news in the Kenya political period environment. Particularly Adaptive boosting model, specifically the AdaBoost model, had the best performance across all measurement metrics. Though other adaptive boosting models exist, our model was mainly chosen, having iterated and optimized various parameters using AutoML frameworks in google colab. The model is also usable since we tested using five folds for validation instead of just splitting the data set into training and test sets. 82 adaptive boost model pipelines were generated and ranked from the initial parameter optimization process, and the selected model was far superior to the others. The ensemble building process was comparative among 81 other ensembles, clearly providing superior performance compared to other models. Hate speech classification and model building greatly benefit from ensemble models as opposed to traditional models.

We can also conclude that not all ensemble models are superior. As exhibited, different aspects of the ensemble models make the model usable and efficient. The critical aspect of the tradeoff between computing power and time is vital. The model built was tested among various activation functions while adjusting multiple parameters to develop many pipelines within the adaptive boosting neural network approach. The parameter optimization deeply assessed various parameter effects on the model from the pipeline profiler map and table generated that assessed the effect of each of the following: - class balancing, random forest presence, Mlp presence effect, different trees and pruning, Sgd, whether the model being passive aggressive was effective, feature type effect and gradient boosting as some the critical adjustments in the pipelines that improved or worsened the value of the AUC – ROC as shown in figures 19, 20 and 21.

5.3 Contributions of the study

The model was run on primary data for validation and testing in a 5-fold validation process. This shows how the model performs with actual live data. Our model is built and can be visualized and transferred, hence complimenting transfer learning. The study used neural networks to enhance a commonly used algorithm called decision trees. We were able to identify weak learners and boost their learning rate. This was achieved by the use of AutoML approaches to test the AdaBoost model in up to 82 pipelines. Though approaches from previous studies, none involved the use of AdaBoost for binary labeling. We evaluated this model from other previous models and optimized parameters through the 82 pipelines, going above and beyond what the other models had achieved. The essential contribution is in the specific choice of parameter values to adjust the thresholds for these adjustments and the ultimate performance accuracy realized by these adjustments.

The study had two main contributions:

1. We moved from manual ML model building to automated machine learning and parameter optimization of hybrid or ensemble models. From a list of models consisting of deep learning neural networks and hybrid combinations by Riedal et al. 2017; Largent, 2017; Chopra and Jain, 2017; Thorne et al., 2017; Akshay et al. 2017; Aymanns et al., 2017; Karadzhov et al., 2017; Ruchansky et al., 2017; University of Minnesota CSE, 2021; Islam et al., 2019; Bedi, 2018 and Foley, 2019. The ability to manually test among various activation functions, which include:- Linear Function, Sigmoid/Logistic Function, Tanh Tangent Hyperbolic Activation Function, Tanh derivative RELU, Leaky ReLU Function, Parametric ReLU Function, Exponential Linear Units (ELUs) Function, Softmax, Swish, Gaussian Error Linear Unit (GELU) and Scaled Exponential Linear Unit (SELU). Is not only a detailed process but error-prone if done manually and time-consuming in value adjustment. AutoML and pipeline profiling reduce the time taken and manually driven parameter optimization, where from the 82 pipelines generated, we could select one with the best performance.
2. Within the ensemble models composed of adaptive boosting, we can show that classification functions using this form of boosting are far superior and efficient in the results compared to the other models.
3. The model we also built, as compared to other Adaptive boosting models used to classify fake news, especially models by Largent 2017 and Thorne et al. 2017, are uniquely different in performance and approach in building from the proposed approach where we use AutoML and the manual approach used by, Largent 2017 and Thorne et al. 2017. While we can test automate the testing across essential activation functions and parameter value adjustments, review this in the pipeline profiler and ultimately select the superior model. Largent 2017 and Thorne et al. 2017, use a manual selection of activation functions and parameter values.

5.4 Recommendations for Future Research

It is implementing fake news classification, not on binary classifiers but on the multi-classification of labels. The fake or genuine binary labeling does not consider multiple classes aspect with the existence of both mis- and dis- information. The reality is that there exists a category of facebook posts that contain a mixture of genuine and fake aspects in reporting or that have URLs already flagged in fake news tracking platforms. While the latter, as a pre-check before classification, improves model performance, the former creates a grey area on whether a portion or section of a post is genuine or fake. Binary labels will not improve classification as compared to multi-class labels. This is a possible research area. Multi-class classification is an opportunity for further research that can be further developed.

The adaptive boosting model performed better across all 5 folds of validation data. The model created a test and tweaked parameters in calling the model fitting functions. Other parameters are tuning and optimizing techniques that were not utilized due to resource constraints. For enterprise-grade solutions that can be scaled, there is an opportunity to test resource-intensive approaches like the use of diverse activation functions that the google computes environment was inhibited from implementing owing to huge costs and robustly compare results.

The model was built on text data type in a natural language setting involving posts in Kenya, regardless of the language used. The language was later translated before being used. There is an opportunity to implement fake or genuine news classification in multi-media content containing a mix of images, emojis, emoticons, and motion pictures. The reality is that fake and genuine news is never purely an expression containing only text data but is a mix of multiple data types that the model-building process needs to be robust enough to deal with, such as in the event of audio and video data formats.

References

Abdulrauf, A. A. (2016). Cognitive engagement and online political participation on Facebook and Twitter among youths in Nigeria and Malaysia (Doctoral thesis). Universiti Utara Malaysia, Changlun.

StatCounter GlobalStats. (2021, June 1). *Social Media Stats Kenya*. gs.statcounter.com. Retrieved July 15, 2021, from https://gs.statcounter.com/social-media-stats/all/kenya

Newman, N., Fletcher, R., Schulz, A., Andı, S., & Nielsen, R. K. (2020). *Reuters Institute Digital News Report 2020*. Reuters. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf

Madowo, L. (2019, May 24). Was Facebook undermining democracy in Africa? *BBC News*. https://www.bbc.com/news/world-africa-48349671

Facebook. (2020, November 19). *Here's how we're using AI to help detect misinformation*. Facebook AI Blogs. Retrieved July 15, 2021, from https://ai.facebook.com/blog/heres-how-were-using-ai-to-help-detect-misinformation/

The Guardian. (2016, 11 1). Facebook's failure: did fake news and polarized politics got Trump elected? *Facebook's failure: did fake news and polarized politics got Trump elected?* https://www.theguardian.com/technology/2016/nov/10/facebook-fake-news-election-conspiracy-theories

Madrigal, A. C. (2017, October 1). What Facebook Did to American Democracy and why it was so hard to see it coming? *The Atlantic*. https://www.theatlantic.com/technology/archive/2017/10/what-facebook-did/542502/

Kertysova, Katarina. (2018). Artificial Intelligence and Disinformation. Security and Human Rights. 29. 55-81. 10.1163/18750230-02901005.

Watson, A. (2021, May 28). *Fake news worldwide - statistics & facts*. Fake news worldwide - statistics & facts. Retrieved July 15, 2021, from https://www.statista.com/topics/6341/fake-news-worldwide/

Conner-Simmons, A. (2018, October 4). Detecting fake news at its source. Machine learning system aims to determine if an information outlet was accurate or biased. https://news.mit.edu/2018/mit-csail-machine-learning-system-detects-fake-news-from-source-1004

Palriwala, R. (2020, July 11). *Fake News Detection Using TFIDF Vectorizer and Passive Aggressive Classifier*. Fake News Detection Using TFIDF Vectorizer and Passive Aggressive Classifier. https://medium.com/analytics-vidhya/fake-news-detector-cbc47b085d4

Bharadwaj, A., Ashar, B., Barbhaya, P., Bhatia, R., & Shaikh, P. Z. (2020, June). Source Based Fake News Classification using Machine Learning. *Source Based Fake News Classification using Machine Learning*, *9*(6), 7.

Rao, D.K. S., .Sreeram, D.G., & RAJU, D. B.D. (2020, April 21). DETECTING FAKE ACCOUNT ON SOCIAL MEDIA USING MACHINE LEARNING ALGORITHMS. *International Journal of Control and Automation*, *13*(1), 95-100.

Y. Boshmaf, D. Logothetis, G. Siganos, J. Lería, J. Lorenzo, M. Ripeanu, K. Beznosov, H. Halawa, "Íntegro: Leveraging victim prediction for robust fake account detection in large scale osns", Computers & Security, vol. 61, pp. 142-168, 2016.

Allcott, H., Gentzkow, M., & Yu, C. (2018). Panel A: Facebook Engagement. In *Trends in the Diffusion of Misinformation on Social Media* (1st ed., Vol. 1, p. 8). New York University, Microsoft Research, and NBER. https://web.stanford.edu/~gentzkow/research/fake-news-trends.pdf

Poynter. (2021, July 15). *A guide to anti-misinformation actions around the world*. A guide to anti-misinformation actions around the world. Retrieved July 15, 2021, from https://www.poynter.org/ifcn/anti-misinformation-actions/

Guess, A., Nagler, J., & Tucker, J. (2019, Jan 09). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Less than you think: Prevalence and predictors of fake news dissemination on Facebook*, *5*(1), 9. 10.1126/sciadv.aau4586

The Wire. (2017, August 05). *Facebook Offers Tool to Combat Fake News in Kenyan Elections*. The Wire. https://thewire.in/external-affairs/facebook-offers-tool-combat-fake-news-kenyan-elections

Tankovska, H. (2021, May 5). *Distribution of Facebook users in Kenya as of April 2021, by age group*. Distribution of Facebook users in Kenya as of April 2021, by age group. https://www.statista.com/statistics/1029198/facebook-user-share-in-kenya-by-age/

Ahmed, H., Traore, I., & Saad, S. (2017, October). Detection of online fake news using n-gram analysis and machine learning techniques. In *International conference on intelligent, secure, and dependable systems in distributed and cloud environments* (pp. 127-138). Springer, Cham.

Setiawan, R., Ponnam, V. S., Sengan, S., Anam, M., Subbiah, C., Phasinam, K., & Ponnusamy, S. (2021). Certain Investigation of Fake News Detection from Facebook and Twitter Using Artificial Intelligence Approach. *Wireless Personal Communications*, 1-26.

Poddar, K., & Umadevi, K. S. (2019, March). Comparison of various machine learning models for accurate detection of fake news. In *2019 Innovations in Power and Advanced Computing Technologies (i-PACT)* (Vol. 1, pp. 1-5). IEEE.

BBC. (2019, October 31). Facebook moderation firm Cognizant quits. *Facebook moderation firm Cognizant quits*. https://www.bbc.com/news/technology-50247540

Hakak, S., Alazab, M., Khan, S., Gadekallu, T. R., Maddikunta, P. K. R., & Khan, W. Z. (2021). An ensemble machine learning approach through effective feature extraction to classify fake news. Future Generation Computer Systems, 117, 47-58.

Moore, J. H. (2021, January 1). *The Tree-Based Pipeline Optimization Tool (TPOT)*. AutoML. http://automl.info/tpot/

Ratner, A. (2020, July 14). *Snorkel AI: Putting Data First in ML Development*. Snorkel AI: Putting Data First in ML Development. https://snorkel.ai/platform/#how-it-works

Pandey, P. (2019, October 16). *A Deep Dive into H2O's AutoML*. A Deep Dive into H2O's AutoML. https://www.h2o.ai/blog/a-deep-dive-into-h2os-automl/

IEBC. (2017, January 1). *Statistics of Voters*. Statistics of Voters. https://www.iebc.or.ke/registration/?stats

Kibuacha, F. (2021, January 13). *Mobile Penetration and Growth in Kenya*. Mobile Penetration and Growth in Kenya. https://www.geopoll.com/blog/mobile-penetration-kenya/

University of Minnesota. (2021). Classification: Basic Concepts, Decision Trees, and Model Evaluation. University of Minnesota. https://www-users.cse.umn.edu/~kumar001/dmbook/ch4.pdf

Islam, M. Z., Liu, J., Li, J., Liu, L., & Kang, W. (2019, November 7). A Semantics Aware Random Forest for Text Classification [A Semantics Aware Random Forest for Text Classification]. http://184pc128.csie.ntnu.edu.tw/presentation/20-02-03/A%20Semantics%20Aware%20Random%20Forest%20for%20Text%20Classification.pdf.

Chengsheng, Tu & Huacheng, Liu & Bing, Xu. (2017). AdaBoost typical Algorithm and its application research. MATEC Web of Conferences. 139. 00222. 10.1051/matecconf/201713900222.

Bedi, G. (2018, November 9). A guide to Text Classification (NLP) using SVM and Naive Bayes with Python. A guide to Text Classification (NLP) using SVM and Naive Bayes with Python. https://medium.com/@bedigunjit/simple-guide-to-text-classification-nlp-using-svm-and-naive-bayes-with-python-421db3a72d34

Foley, D. (2019, February 8). K-Means Clustering. https://towardsdatascience.com/k-means-clustering-8e1e64c1561c

Ahmad, I., Yousaf, M., Yousaf, S., & Ahmad, M. O. (2020). Fake news detection using machine learning ensemble methods. Complexity, 2020.

Cornell University (n.d.) (2015). https://www.classe.cornell.edu/~dlr/teaching/p6574/lectures/lecture1.pdf

Mohri, M. (2021). Foundations of Machine Learning Boosting. Foundations of Machine Learning Boosting, 1(1), 41. 2021

Thomas G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Machine Learning*, 40(2):139-158, 2000.

Hansen, T. J. (2021, October 1). Sklearn.ensemble. AdaBoostClassifier — scikit-learn 1.0.2 documentation. Scikit-learn. Retrieved April 8, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html

Aarshay. (2016). Complete Machine Learning Guide to Parameter Tuning in Gradient Boosting (GBM) in Python. *Complete Machine Learning Guide to Parameter Tuning in Gradient Boosting (GBM) in Python*, *1*(1), 5. https://www.analyticsvidhya.com/blog/2016/02/complete-guide-parameter-tuning-gradient-boosting-gbm-python/

Moore, A. W. (2015). *Cross-validation for detecting and preventing overfitting*. https://www.cs.cmu.edu/~./awm/tutorials/overfit10.pdf

Benjamin Riedel, Isabelle Augenstein, Georgios P. Spithourakis, and Sebastian Riedel. 2017a. A simple but tough-to-beat baseline for the fake news challenge stance detection task. CoRR abs/1707.03264. http://arxiv.org/abs/1707.03264.

Largent, W. (2017, June 20). Talos Targets Disinformation with Fake News Challenge Victory. Cisco Talos Blog. https://blog.talosintelligence.com/talos-fake-news-challenge/

Delenn Chin Kevin Chen Akshay Agrawal. ???? Cosine siamese models for stance detection. Technical report, Stanford University, year = 2017.

Aymanns, C., Foerster, J.N., & Georg, C. (2022). Fake News in Social Networks. Political Methods: Computational eJournal.

Karadzhov, Georgi & Nakov, Preslav & Màrquez, Lluís & Barrón-Cedeño, Alberto & Koychev, Ivan. (2017). Fully Automated Fact Checking Using External Sources. 344-353. 10.26615/978-954-452-049-6_046.

Tiwari, S. (2018). Activation functions in Neural Networks - GeeksforGeeks. GeeksforGeeks. https://www.geeksforgeeks.org/activation-functions-neural-networks/

Baheti, P. (2022). 12 Types of Neural Networks Activation Functions: How to Choose? Www.v7labs.com. https://www.v7labs.com/blog/neural-networks-activation-functions

Ong'ong'a, Oloo. (2021). Countering the New Media Podia: Youth and "Fake news" in Kenya. 9. 1-23. 10.34293/sijash.v8i4.3033.

GeoPoll and Portland, (2017)..Fake News Of In Kenya https://portland-communications.com/pdf/The-Reality-of-Fake-News-in-Kenya.pdf

Appendix 1: Research Schedule

Figure 20: Illustration of the research schedule

Task Name	Duration	Start	Finish
Data Ingestion	5 days	Mon 8/2/21	Fri 8/6/21
•Facebook Graph API	5 days	Mon 8/2/21	Fri 8/6/21
•Manual data import	5 days	Mon 8/2/21	Fri 8/6/21
•AutoML Tools Selection	1 day	Mon 8/9/21	Mon 8/9/21
•Google Colab and Tensorflow powered TPOT	1 day	Mon 8/9/21	Mon 8/9/21
•Snorkel	1 day	Mon 8/9/21	Mon 8/9/21
•H20	1 day	Mon 8/9/21	Mon 8/9/21
•Data pre-processing used of NLTK or PyCaret	2 days	Mon 8/9/21	Tue 8/10/21
•Tokenization	1 day	Mon 8/9/21	Mon 8/9/21
•Stop words removal	1 day	Mon 8/9/21	Mon 8/9/21
•Stemming	1 day	Mon 8/9/21	Mon 8/9/21
•Split Train and Test set allowing for K-Fold (5-Folds) validation	2 days	Mon 8/9/21	Tue 8/10/21
•AutoML Model Building	1 day	Wed 8/11/21	Wed 8/11/21
•Build models from the three platforms using BERT for sequencing to avoid token by token processing	1 day	Wed 8/11/21	Wed 8/11/21
•Model K-Fold Validation (consideration of many models and to avoid overfitting)	1 day	Thu 8/12/21	Thu 8/12/21
•Split into K-Folds (5-folds) conduct validation for each of the models and select final model	1 day	Thu 8/12/21	Thu 8/12/21
•Model deployment and management	3 days	Fri 8/13/21	Tue 8/17/21
•Deploy	3 days	Fri 8/13/21	Tue 8/17/21
•Track performance ensure continuous improvement	2 days	Fri 8/13/21	Mon 8/16/21

Appendix 2: Resources and Budget

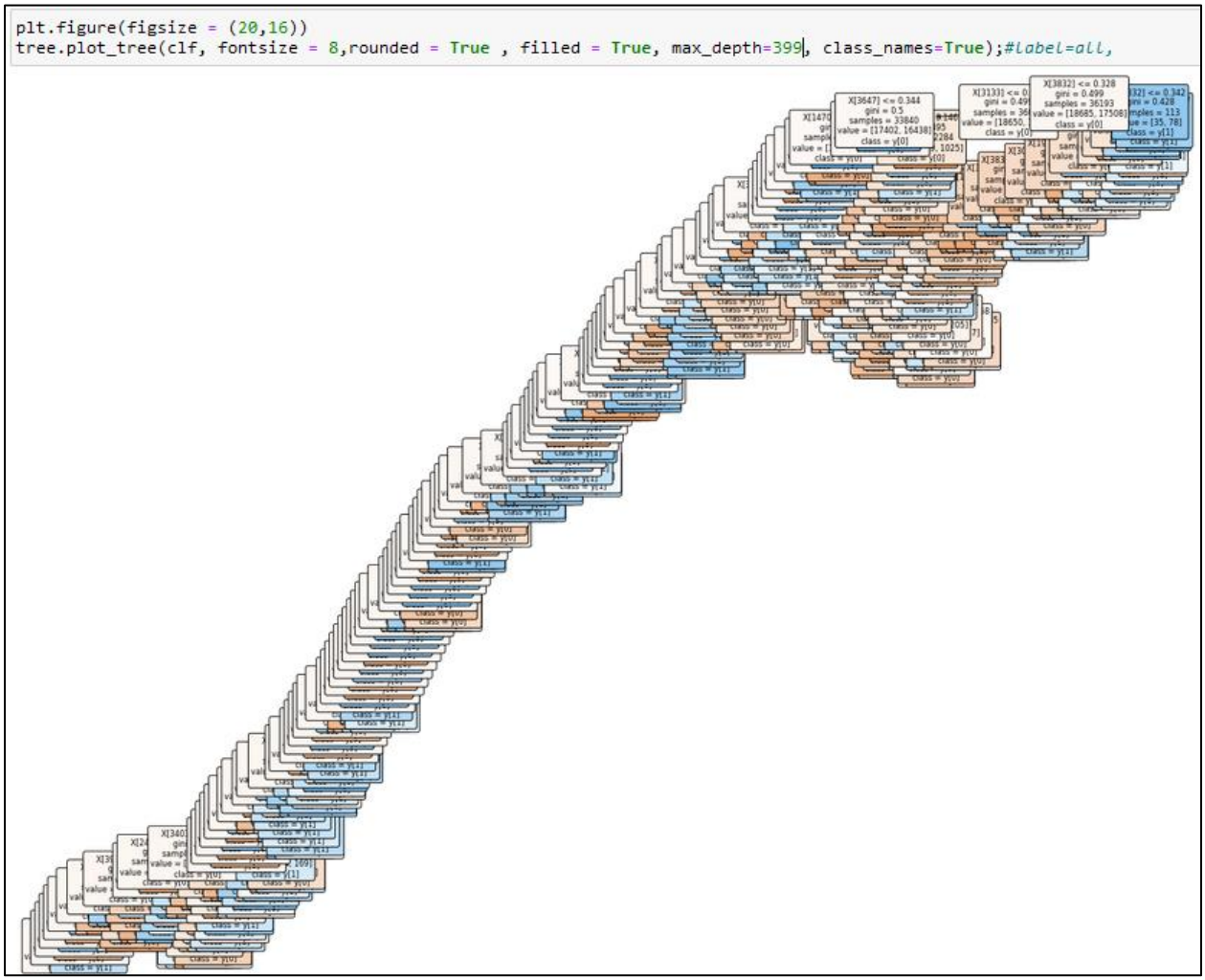
The Table below indicated all the required resources and expected cost.

Figure 21: Budget breakdown

#	Item/Budget Line	Unit	Per Unit Cost	Number of Units	Total Cost
1	Adequate Internet bandwidth	20 MBPS	5,999	4	23,996
2	Ethical approvals	1	10,000	1	10,000
3	Facebook Academic Research Approvals	1	N/A	1	0
4	Transportation	1	500	16	8000
5	Communication	1 Month	1000	4	4000
6	Software costs (covers google cloud platform storage, tensorflow processing unit for processing)	1	10,000	1	10,000
7	Hardware costs (laptop)	1	90,000	1	90,000
	Total				145,996

Appendix 3: The full model with max_depth of 399

Figure 22: Decision tree model visual with complete max_depth at 399



Appendix 4: TF-IDF breakdown

The detailed results from our TF-IDF were too significant to illustrate as a table. We illustrate the columnwise aggregates of TF-IDF in the below table.

Figure 23: Showing TF-IDF column-wise sums

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
daily	373.364	Kenya	422.597	nation	381.411	citizen	611.507
century	166.411	house	144.775	president	347.691	Uhuru	440.458
magufuli	131.468	Friday	132.046	presidential	145.546	said	263.577
berlin	127.153	former	125.036	Monday	121.755	Raila	244.297
Beijing	119.454	democratic	117.092	ruto	115.345	st	172.838
china	98.425	kenyas	83.455	new	111.128	wire	171.676
Nairobi	96.135	government	74.175	one	110.872	says	166.850
Mombasa	89.909	jubilee	72.457	people	107.311	republic	166.127
candidate	84.282	first	70.034	members	88.607	Wednesday	152.038
campaign	83.791	german	69.174	party	84.738	Thursday	150.422
court	75.450	election	67.297	minister	82.842	Tuesday	143.073
bill	75.293	Germany	62.455	media	78.431	video	134.538
administration	73.529	group	61.114	news	71.541	state	114.397
day	58.083	foreign	56.509	political	71.295	would	112.131
called	57.834	governor	54.964	representatives	64.126	white	71.659
Nakuru	57.020	general	54.500	prime	59.646	Sunday	70.007
Barack	56.193	judge	52.188	national	59.625	th	68.677
citizen	56.138	DCI	52.181	paul	58.249	us	67.429
another	54.724	director	47.069	north	57.663	two	67.049
could	52.022	democrat	45.468	last	57.204	secretary	66.664
chancellor	47.717	Iran	42.358	president-elect	57.131	senate	66.561
Cairo	46.543	James	42.006	may	56.479	week	59.687
back	46.391	Islamic	41.491	like	56.013	union	57.204
black	44.856	even	39.237	police	52.587	Saturday	55.606
Bernie	44.800	killed	38.762	leader	52.011	senator	54.286
conservative	43.039	department	38.529	made	51.585	story	53.871
Somalia	42.388	Uganda	36.612	nominee	50.504	speaker	53.439
committee	41.964	democrats	36.444	man	50.335	security	52.974
Chinese	41.918	executive	36.412	military	49.292	Syrian	52.019

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
defense	41.915	john	34.982	Merkel	47.573	time	50.945
chief	41.857	know	34.756	lawmakers	45.388	Ryan	50.062
city	40.341	going	34.489	many	43.868	told	49.103
attorney	39.845	host	34.056	law	42.144	turkey	48.762
attack	39.261	intelligence	33.967	parliament	40.397	years	47.487
asked	36.403	got	33.564	leaders	39.370	year	45.898
air	33.710	got	33.535	officials	39.258	Tanzania	45.141
decision	33.106	Iraq	33.178	make	37.655	world	45.119
accused	32.798	Iraqi	32.806	night	36.844	drc	44.384
conservatives	31.880	highlights	32.262	must	35.015	top	44.267
according	31.804	forces	32.042	next	34.451	Texas	43.748
days	31.253	justice	31.405	left	34.299	Washington	41.591
corrects	30.940	jeff	31.403	paragraph	33.067	sanders	41.334
announced	30.246	go	30.850	least	32.708	support	41.297
Comey	29.791	fox	30.378	million	32.665	vote	41.228
apparently	29.545	great	29.637	Muslim	32.375	show	39.675
congress	29.300	interview	29.513	plan	31.948	tax	39.311
British	29.274	end	28.133	press	31.652	twitter	38.243
deal	29.217	force	27.537	rally	31.590	supreme	37.964
around	27.856	ever	27.162	major	31.564	Syria	37.101
co	26.187	four	27.003	released	31.005	women	36.810
debate	25.957	GOP	26.225	order	30.630	Turkish	36.526
decided	25.841	every	26.173	Lebanon	30.388	urged	36.136
Britain	25.643	Illinois	25.994	Nairobi	30.367	social	36.080
coalition	25.589	head	25.921	morning	30.300	set	34.815
Caracas	25.585	fake	25.547	public	29.113	woman	34.476
Cleveland	25.540	Florida	25.278	office	29.084	supporters	33.579
anyone	25.165	elections	25.079	policy	28.209	senior	33.068
Bangkok	24.861	investigation	24.140	plans	28.030	took	32.942
arrested	24.761	following	23.874	meeting	27.701	take	32.829
army	24.744	help	23.544	nuclear	27.370	right	32.558
business	24.719	health	23.447	report	26.365	three	32.403
Brasilia	24.510	found	23.072	months	26.184	since	32.285
already	24.475	international	22.940	liberal	25.299	war	32.229
agreed	24.288	hold	22.728	legislation	25.297	south	31.642
ban	23.992	enough	22.333	official	25.252	way	31.327
big	23.575	hurricane	22.209	put	24.708	team	31.090
best	23.298	held	21.938	much	24.668	trying	29.687
close	22.902	getting	21.718	member	24.324	talks	29.597

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
adviser	22.283	gave	21.665	making	24.269	watch	29.211
agency	22.094	free	21.622	really	23.788	say	29.096
authorities	21.972	Kurdish	21.519	old	23.629	wants	28.766
commission	21.851	known	21.321	release	23.112	want	28.505
Athens	21.465	groups	21.320	latest	22.950	still	28.171
budget	21.451	good	21.173	mike	22.934	speech	27.125
ago	21.297	federal	20.919	presidency	22.702	see	26.839
Chris	21.294	Jim	20.775	lot	22.643	today	26.453
approved	20.866	five	20.735	opposition	22.539	voters	26.026
continues	20.842	Museveni	20.704	nations	22.474	sessions	25.796
aires	20.781	home	20.633	recent	22.229	went	25.495
Buenos	20.781	including	20.380	Michael	22.041	school	25.148
change	20.610	face	20.344	part	21.933	think	25.070
clear	20.603	high	19.999	healthcare	21.757	work	24.974
came	20.512	effort	19.960	month	21.730	united	24.902
become	20.308	efforts	19.864	mayor	21.682	ruto	24.304
criticized	20.141	Jinping	19.763	point	21.415	vice	24.137
come	20.086	fired	19.750	racist	21.394	ruling	24.058
border	19.995	expected	19.232	move	21.224	Venezuela	23.862
control	19.891	George	18.698	likely	21.048	rights	23.345
climate	19.757	keep	18.670	ordered	21.024	ted	22.712
Barcelona	19.528	everyone	18.444	parties	20.962	year old	22.507
billionaire	19.357	history	18.429	lead	20.836	republicans	22.442
allies	19.342	far	18.354	Raila	20.830	shot	22.332
activist	19.173	Egyptian	18.159	need	20.735	whether	22.184
crisis	19.087	financial	18.144	legal	20.693	signed	22.100
across	19.028	issued	18.103	power	20.652	seen	21.731
California	18.926	illegal	17.914	never	20.626	warned	21.222
call	18.648	finally	17.751	northern	20.097	visit	20.729
considering	18.508	gun	17.560	person	20.092	Saudi	20.400
claimed	18.367	fact	17.468	race	20.047	several	20.385
appeals	18.324	facebook	17.047	Mattis	20.010	taking	20.316
criminal	18.131	human	17.000	Pelosi	19.816	Tayyip	20.300
Boston	18.106	expressed	16.995	nothing	19.777	wall	20.281
Austin	17.833	Egypt	16.966	mainstream	19.276	stop	20.151
ben	17.768	young	16.881	long	19.203	used	20.137
claims	17.715	family	16.876	name	19.101	running	19.686
companies	17.370	healthcare	16.837	launched	19.014	times	19.472
attacks	17.319	knows	16.830	real	18.848	warning	19.472

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
college	17.259	full	16.682	meet	18.839	well	19.426
ambassador	17.184	frontrunner	16.619	recently	18.726	university	19.123
calling	17.175	final	16.458	Ohio	18.581	sent	19.038
Brexit	16.992	hit	16.362	Patrick	18.574	supporter	19.010
Carolina	16.946	fire	16.313	near	18.519	something	18.934
away	16.909	Korean	16.193	leading	18.514	taken	18.577
children	16.894	immigration	16.168	reported	18.228	saad	18.535
believe	16.756	join	16.091	nancy	18.097	trade	18.251
congressman	16.749	hard	15.730	proposed	18.090	travel	18.117
defended	16.648	Irma	15.666	Lebanese	17.868	second	17.914
comes	16.489	given	15.275	live	17.734	Robert	17.877
biggest	16.175	joe	15.243	protesters	17.704	working	17.583
Arabia	16.087	elected	15.206	lost	17.623	Virginia	17.532
CIA	16.058	economic	15.145	met	17.478	senators	17.518
Christian	16.003	inauguration	15.144	ministry	17.408	states	17.509
broke	15.770	jersey	14.970	open	17.214	used	17.366
bipartisan	15.734	evidence	14.890	late	17.026	tv	17.247
civil	15.730	key	14.844	officer	16.952	talk	17.245
candidates	15.541	issue	14.808	money	16.948	thousands	17.165
communist	15.446	email	14.469	parliamentional	16.918	six	17.142
almost	15.379	fight	14.322	mark	16.822	seeking	16.951
caught	15.201	hearing	13.945	NASA	16.682	Thailand	16.886
David	15.200	environmental	13.942	oil	16.588	voted	16.684
chairman	15.136	language	13.940	planned	16.584	seems	16.674
bad	15.133	female	13.767	representative	16.395	students	16.642
Bangladesh	15.109	independence	13.755	ready	16.381	suspected	16.533
Copenhagen	15.040	filed	13.723	Michel	16.317	step	16.498
backed	15.025	episode	13.692	Muslims	16.255	sexual	16.436
Brazilian	15.010	give	13.603	rejected	16.211	thinks	16.411
attacked	14.969	guy	13.457	panel	16.189	threat	16.090
anything	14.927	fo	13.452	Putin	15.967	response	16.082
action	14.908	Flynn	13.384	lives	15.691	sanctions	15.990
declared	14.797	find	13.317	nato	15.667	Vladimir	15.941
began	14.729	deputy	13.261	lawmaker	15.487	weeks	15.926
aboard	14.647	fighting	13.233	let	15.374	soon	15.715
death	14.547	global	13.074	little	15.212	words	15.569
Catalonia	14.504	heard	13.059	Michigan	15.193	win	15.558

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
bogota	14.478	early	13.026	pa	15.168	sign	15.510
congressional	14.430	king	12.811	Puerto	15.159	victory	15.235
brazil	14.382	idea	12.804	nearly	15.153	version	15.185
cut	14.375	hope	12.749	needs	15.125	violence	15.131
always	14.369	earlier	12.718	na	15.079	start	15.130
charged	14.284	important	12.671	reporter	15.068	voter	14.979
continue	14.279	hate	12.666	men	14.947	tried	14.822
Amman	14.194	front	12.646	number	14.913	spoke	14.819
agreement	14.087	eight	12.634	problem	14.794	weekend	14.804
calls	14.020	energy	12.612	pressure	14.750	using	14.638
among	13.808	Hollywood	12.577	look	14.693	shows	14.593
bush	13.807	facing	12.547	matter	14.601	rule	14.518
billion	13.640	finance	12.313	refugees	14.580	special	14.461
absolutely	13.593	food	12.268	past	14.496	strong	14.412
citizens	13.524	Iowa	12.221	life	14.475	thing	14.357
community	13.448	documents	12.151	love	14.464	se	14.327
continued	13.332	job	12.143	park	14.427	role	14.282
case	13.326	kelly	12.094	November	14.367	town	14.108
address	13.238	despite	12.022	march	14.342	wa	14.039
allow	13.221	detained	11.878	proposal	14.339	staff	14.038
Colombia	13.170	handed	11.756	push	14.326	scandal	13.978
block	13.136	everything	11.730	missile	14.104	tom	13.969
cabinet	13.126	eastern	11.576	letter	14.082	run	13.917
ask	13.125	journalist	11.573	main	14.051	student	13.894
charges	13.046	jr	11.546	leftist	14.029	thought	13.852
car	13.020	establishment	11.543	photo	13.999	tweet	13.738
defend	13.015	funding	11.443	received	13.881	turned	13.705
yesterday	13.013	french	11.414	pay	13.837	system	13.269
activists	12.936	Kenyatta	11.380	lying	13.804	Rico	13.175
corp	12.893	largest	11.327	place	13.785	wife	13.146
bazar	12.867	ha	11.247	primary	13.639	rohingya	13.132
coming	12.813	flag	11.245	lawyer	13.547	service	13.042
breaking	12.749	jobs	11.157	nominate	13.516	secret	13.023
better	12.696	homeland	11.137	raised	13.480	west	12.980
conference	12.654	hundreds	11.055	reports	13.460	tell	12.978
crowd	12.574	dismissed	11.037	message	13.301	saying	12.964
care	12.467	Kenya backed	10.823	others	13.295	shooting	12.945
Budapest	12.454	industry	10.749	militants	13.240	truth	12.930
daughter	12.444	embassy	10.674	looking	13.165	whose	12.855

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
claim	12.428	independent	10.556	presi	13.109	sure	12.794
Benghazi	12.350	investigating	10.434	percent	12.955	trip	12.786
battle	12.226	father	10.412	lawsuit	12.886	stand	12.786
council	12.191	intends	10.337	peace	12.880	small	12.776
com	12.150	hopes	10.322	nov	12.840	ruled	12.776
central	12.077	fighters	10.316	professor	12.797	terrorist	12.634
controversial	12.069	hear	10.314	offered	12.746	resigned	12.481
bomb	12.066	form	10.310	pretty	12.558	William	12.445
act	12.063	emails	10.296	Libya	12.540	unveiled	12.439
alleged	11.583	Iranian	10.257	rep	12.521	wanted	12.390
dec	11.525	failed	10.234	reform	12.480	western	12.315
attention	11.447	east	10.219	reached	12.479	try	12.197
current	11.341	girl	10.216	might	12.445	rules	12.129
blocked	11.158	denied	10.193	protest	12.434	ties	12.102
Bucharest	11.148	Greece	10.076	mass	12.346	watching	12.068
capital	10.993	goes	10.059	majority	12.343	Tillerson	12.009
corruption	10.986	endorsed	10.057	pence	12.253	un	11.942
concerned	10.956	gets	10.036	question	12.223	sean	11.929
Baltimore	10.946	details	9.994	pass	12.126	return	11.767
armed	10.894	education	9.960	manager	12.120	rex	11.723
acting	10.889	dollars	9.900	region	11.922	telling	11.668
actually	10.872	instead	9.893	makes	11.885	third	11.624
appears	10.852	Kansas	9.792	position	11.832	wrote	11.583
crime	10.838	died	9.715	poll	11.807	takes	11.341
appeared	10.819	hours	9.690	opened	11.794	serious	11.300
Catalan	10.751	desperate	9.669	powerful	11.772	speak	11.274
company	10.747	Israel	9.640	passed	11.770	spd	11.270
comment	10.685	district	9.637	note	11.737	shut	11.201
believes	10.566	documentary	9.626	refused	11.620	swearing	11.080
ass	10.534	exposed	9.564	possible	11.618	scheduled	11.076
build	10.522	handle	9.561	Myanmar	11.616	voting	11.070
actor	10.508	future	9.555	outside	11.600	veteran	11.055
anchor	10.504	June	9.488	regional	11.599	someone	11.052
attempt	10.503	joint	9.421	politicians	11.598	transition	10.927
Christmas	10.480	fellow	9.279	probably	11.542	spent	10.861
blame	10.466	Georgia	9.254	list	11.521	street	10.846
behind	10.445	joined	9.210	mexico	11.425	statement	10.825
arab	10.398	discuss	9.194	middle	11.375	victim	10.805
also	10.364	Elizabeth	9.189	repeal	11.371	showed	10.750

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
asks	10.286	fraud	9.189	radio	11.360	southern	10.748
center	10.260	described	9.173	Nigeria	11.351	Spain	10.745
challenge	10.159	fiscal	9.136	pres	11.332	willing	10.717
delivered	10.153	large	9.132	movement	11.314	son	10.662
confirmed	10.079	labor	9.116	reason	11.229	sick	10.582
announcement	10.044	Detroit	9.113	officers	11.180	stunning	10.569
condemned	10.033	field	8.999	named	11.173	troops	10.512
bank	10.030	Haider	8.926	promised	11.042	rival	10.500
argentine	10.007	greek	8.921	monic	11.022	sought	10.443
agencies	9.973	Emmanuel	8.917	meetings	11.002	side	10.442
ca	9.894	film	8.916	politics	10.983	spokesman	10.441
conspiracy	9.838	events	8.903	local	10.829	violent	10.397
crooked	9.791	jail	8.887	overhaul	10.825	strike	10.378
closer	9.726	kind	8.747	red	10.747	services	10.298
aide	9.725	emergency	8.702	protection	10.647	tr	10.260
christie	9.682	enforcement	8.662	mo	10.605	responsibility	10.252
Arizona	9.680	eric	8.654	pick	10.566	Spanish	10.215
able	9.582	friends	8.626	moment	10.446	wh	10.197
allowed	9.577	growing	8.624	prominent	10.406	seven	10.188
asking	9.555	injured	8.609	London	10.403	workers	10.135
base	9.549	Hungary	8.593	phone	10.286	whole	10.082
concern	9.545	Jeanine	8.578	presumptive	10.250	sheriff	9.963
deep	9.459	keeps	8.563	offensive	10.167	series	9.946
appointed	9.437	gone	8.554	low	10.090	revealed	9.924
crazy	9.427	huge	8.550	lies	10.079	started	9.922
consider	9.416	done	8.543	reality	10.058	television	9.912
couple	9.398	jones	8.524	prosecutor	10.044	Taiwan	9.868
carter	9.323	half	8.498	prosecutors	10.037	surprise	9.846
communications	9.321	dropped	8.457	pope	10.008	star	9.782
chair	9.303	forward	8.445	led	9.942	within	9.762
Argentina	9.286	Jerusalem	8.429	repeatedly	9.924	wit	9.741
Carson	9.284	issues	8.367	protect	9.838	tha	9.732
based	9.275	Israeli	8.263	martin	9.827	warren	9.724
choice	9.271	forced	8.239	record	9.809	sentenced	9.696
became	9.245	Gabriel	8.237	Michelle	9.747	worst	9.605
along	9.229	divided	8.187	mother	9.742	Venezuelan	9.551
bid	9.161	kicked	8.176	less	9.741	turn	9.548
comments	9.143	faced	8.157	remarks	9.735	soldiers	9.546
cyber	9.082	la	8.098	quite	9.718	struck	9.535

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
convention	9.067	Kellyann	8.069	leave	9.702	sh	9.534
ahead	9.002	helped	8.060	Maher	9.390	speaking	9.508
cities	8.990	epic	8.058	Pennsylvania	9.388	strikes	9.434
captured	8.981	greater	8.022	posted	9.285	term	9.421
add	8.949	Hampshire	7.966	relationship	9.256	threatening	9.377
clearly	8.936	fund	7.917	macron	9.252	stage	9.371
comedian	8.883	event	7.910	personal	9.189	ways	9.272
anti-trump	8.836	February	7.898	Reilly	9.135	sought	9.222
Aden	8.824	investor	7.896	millions	9.127	super	9.188
backlash	8.764	internet	7.872	legislative	9.048	tucker	9.170
chuck	8.759	far-right	7.861	network	9.033	tweeted	9.146
Conway	8.755	gas	7.849	line	9.016	returned	9.119
course	8.751	floor	7.814	nomination	8.994	threw	9.114
author	8.720	Dennis	7.814	mind	8.967	wi	9.113
begin	8.643	land	7.800	liberals	8.960	retired	9.088
bring	8.622	introduced	7.792	Melania	8.916	temporary	9.077
church	8.569	information	7.791	popular	8.810	temer	9.038
august	8.564	January	7.780	price	8.740	votes	9.013
approval	8.561	holding	7.776	planning	8.729	terrorists	8.987
CEO	8.558	destroy	7.751	loves	8.721	tired	8.938
book	8.537	especially	7.745	program	8.672	sa	8.902
boy	8.531	highly	7.729	Maine	8.633	threatened	8.901
actions	8.492	formally	7.685	quickly	8.602	tonight	8.794
Carlson	8.487	dozens	7.656	often	8.557	vowed	8.754
charleston	8.464	investigators	7.645	regarding	8.550	saw	8.748
caused	8.463	god	7.644	questions	8.546	wow	8.741
aimed	8.460	hoped	7.626	Paris	8.533	target	8.718
commerce	8.450	entire	7.560	paying	8.528	terror	8.670
class	8.402	editorial	7.525	owner	8.509	ridiculous	8.659
aware	8.378	feel	7.512	Palestinian	8.505	source	8.642
colorado	8.369	freedom	7.495	Palin	8.481	transgender	8.565
aid	8.289	hack	7.464	navy	8.472	uh	8.559
building	8.227	discussed	7.453	pledged	8.457	sued	8.540
accept	8.186	journalists	7.410	read	8.455	view	8.537
brought	8.152	husband	7.394	progress	8.451	spending	8.476
Belfast	8.147	expect	7.365	project	8.294	resignation	8.450
africa	8.134	diplomatic	7.344	organization	8.291	tells	8.439
agree	8.099	historic	7.325	polls	8.265	share	8.408
created	8.095	guilty	7.300	means	8.265	website	8.374

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
dead	8.051	giving	7.276	preside	8.264	request	8.348
believed	8.051	fro	7.274	Odinga	8.237	rhetoric	8.338
answer	8.047	hired	7.256	relations	8.198	treasury	8.278
constitution	8.025	due	7.252	praised	8.126	sworn	8.276
credit	8.022	dispute	7.248	President	8.095	winning	8.224
con	8.017	inside	7.245	private	8.046	wo	8.165
claiming	8.007	immigrant	7.242	las	8.007	spicer	8.116
ally	7.928	happy	7.229	potential	7.935	seem	8.086
arrest	7.806	democracy	7.226	massive	7.910	steve	8.028
brilliant	7.806	economy	7.218	politician	7.867	toward	7.968
Carles	7.774	kids	7.207	protesting	7.866	ross	7.960
airport	7.718	lady	7.183	rebels	7.862	welcomed	7.943
allowing	7.712	jordan	7.170	militias	7.816	weapons	7.937
completely	7.709	jailed	7.155	mi	7.764	rightwing	7.925
becoming	7.654	imagine	7.143	published	7.744	without	7.819
changes	7.611	fans	7.103	oklahoma	7.666	romania	7.803
born	7.600	drug	7.061	play	7.658	tim	7.787
ar	7.595	fattah	7.053	october	7.639	seriously	7.787
amateur	7.593	kurdistan	6.999	neil	7.630	socalled	7.785
demanding	7.483	dr	6.966	moore	7.605	stories	7.782
dangerous	7.482	impose	6.960	pushing	7.603	wounded	7.767
boom	7.467	different	6.955	parents	7.573	review	7.746
attacking	7.465	fled	6.952	lower	7.572	yes	7.727
anthony	7.444	employee	6.912	nominated	7.561	true	7.697
concerns	7.436	ireland	6.897	oct	7.556	trial	7.693
alternative	7.424	gives	6.874	lie	7.503	wonder	7.641
belgian	7.383	figure	6.850	protests	7.459	scott	7.599
bit	7.347	fresh	6.825	reporters	7.450	stepped	7.572
defeat	7.339	friend	6.811	light	7.320	sarah	7.533
consortium	7.335	ken	6.811	militia	7.313	shocking	7.530
chance	7.314	july	6.803	murder	7.295	wrong	7.528
accepted	7.299	fully	6.783	refugee	7.295	socialist	7.521
appeal	7.251	ho	6.782	provide	7.231	sweeping	7.489
allegations	7.236	employees	6.764	preparing	7.228	stupid	7.470
breitbart	7.198	id	6.750	rebel	7.222	uhur	7.449
blasted	7.192	hands	6.749	played	7.205	talking	7.449
counsel	7.150	ivanka	6.709	refiled	7.165	sept	7.431
biden	7.106	failing	6.677	puigdemont	7.155	uhu	7.418
april	7.075	firm	6.677	page	7.129	send	7.409

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
carrying	7.063	embarrassing	6.669	radical	7.121	thai	7.395
debt	7.054	knew	6.667	nbc	7.106	usual	7.345
announce	7.054	hussein	6.651	points	7.100	roy	7.329
aides	7.028	ite	6.646	longer	7.090	wilbur	7.328
declined	7.021	else	6.618	removed	7.070	round	7.325
committed	7.009	dutch	6.606	replace	7.064	together	7.316
child	7.007	grand	6.597	living	7.038	shi	7.314
angry	6.996	exchange	6.596	presiden	7.028	wikileaks	7.311
clarify	6.990	happen	6.561	marched	7.011	short	7.305
bruce	6.984	killing	6.556	league	6.962	ta	7.285
appearance	6.929	firing	6.519	mission	6.952	respect	7.257
alliance	6.870	foundation	6.518	religious	6.914	starting	7.242
crimes	6.869	hand	6.515	parts	6.905	safe	7.242
charlottesville	6.858	fear	6.512	needed	6.886	shown	7.237
assault	6.843	expects	6.497	pro	6.885	straight	7.172
bureau	6.826	follow	6.496	prison	6.872	towards	7.169
bi	6.817	increase	6.473	ousted	6.864	turns	7.160
asylum	6.811	interior	6.439	militant	6.860	rick	7.139
cuomo	6.796	game	6.433	pathetic	6.844	worth	7.124
canada	6.763	judicial	6.398	rant	6.823	wait	7.095
decide	6.760	editor	6.365	po	6.813	seeing	7.087
counterpart	6.745	hospital	6.362	likes	6.809	test	7.087
corporate	6.688	disgusting	6.324	maduro	6.804	republic	7.072
certainly	6.667	designed	6.322	nicolas	6.804	suspended	7.057
belgrade	6.661	judiciary	6.317	newly	6.789	upon	7.056
arrived	6.655	disturbing	6.309	please	6.788	san	7.039
banking	6.603	internal	6.307	literally	6.774	trouble	6.991
charlotte	6.598	exactly	6.293	louisiana	6.757	teacher	6.986
additional	6.589	gay	6.235	prepared	6.754	result	6.947
cause	6.588	intel	6.235	promises	6.727	threats	6.932
criticism	6.570	invited	6.206	outrage	6.722	walking	6.928
convicted	6.562	isis	6.199	rather	6.715	total	6.921
avoid	6.557	google	6.188	longtime	6.693	wake	6.889
amid	6.548	holds	6.162	lawyers	6.684	risk	6.851
corrupt	6.528	ke	6.150	raqqa	6.662	tennessee	6.807
art	6.518	kushner	6.128	pastor	6.640	water	6.789
account	6.475	jared	6.128	moved	6.637	society	6.755
cops	6.457	hacking	6.124	readers	6.629	republicancontrolled	6.749
corrected	6.445	insurance	6.118	losing	6.613	results	6.748

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
ap	6.395	interference	6.099	mccain	6.546	rt	6.741
charles	6.386	denounced	6.073	meltdown	6.517	suicide	6.736
bedminster	6.374	disastrous	6.072	pointed	6.514	stay	6.702
awesome	6.345	extremely	6.072	poor	6.504	setting	6.695
area	6.333	investigations	6.061	megyn	6.503	views	6.669
closed	6.317	donors	6.052	referendum	6.498	testify	6.664
definitely	6.300	gov	6.029	reporting	6.457	though	6.659
caucus	6.284	germans	6.026	measure	6.456	thanks	6.643
bully	6.273	helping	5.986	mitt	6.423	rubio	6.615
ba	6.262	favor	5.984	pic	6.417	terrible	6.604
delay	6.252	headline	5.954	racism	6.415	tensions	6.572
brian	6.248	johnson	5.944	measures	6.402	september	6.532
cnn	6.206	hateful	5.929	pentagon	6.401	resign	6.527
conflict	6.184	elite	5.923	promise	6.399	taxes	6.485
allegedly	6.141	door	5.916	remove	6.365	weekly	6.462
appearing	6.132	easily	5.906	reelection	6.363	showing	6.443
advocates	6.130	development	5.902	lynch	6.335	winner	6.435
carl	6.124	irish	5.892	nationalist	6.333	room	6.402
atlanta	6.114	doubt	5.870	mexican	6.331	richard	6.394
brings	6.101	familiar	5.865	massachusetts	6.328	tough	6.352
blow	6.099	heart	5.840	operations	6.321	upset	6.338
contenders	6.023	diplomats	5.814	marco	6.272	wisconsin	6.301
actress	6.013	fa	5.801	playing	6.269	standing	6.300
areas	5.984	governments	5.798	later	6.264	word	6.287
cuts	5.983	fell	5.782	officially	6.239	watchdog	6.260
amendment	5.974	islamist	5.730	period	6.226	strategist	6.250
choose	5.973	emerged	5.728	mueller	6.221	soros	6.249
complete	5.972	feed	5.718	ran	6.219	throwing	6.244
appear	5.961	keny	5.672	lee	6.217	sen	6.221
ballot	5.922	detention	5.659	pushed	6.201	ukraine	6.209
attend	5.908	homes	5.656	oregon	6.163	worker	6.187
considered	5.881	japan	5.641	nfl	6.098	strategy	6.168
correct	5.854	fit	5.639	quit	6.086	reserve	6.114
consumer	5.851	interest	5.623	probe	6.040	sta	6.114
cbs	5.845	fi	5.615	manafort	6.029	returns	6.099
ch	5.835	gorsuch	5.603	offer	6.010	speaks	6.047
curated	5.824	drops	5.594	reach	5.968	summoned	6.024
check	5.793	hopeful	5.585	mostly	5.965	spy	6.020

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
beijingtaipei	5.791	hysterical	5.571	rauner	5.944	responsible	6.015
answers	5.778	immediately	5.569	mean	5.930	safety	5.994
accusations	5.766	island	5.567	lose	5.919	schools	5.986
attended	5.755	ensure	5.543	launch	5.918	ron	5.981
belgium	5.753	ju	5.531	le	5.901	rise	5.967
bundy	5.747	hour	5.525	leaked	5.892	romney	5.959
confidence	5.738	hacked	5.520	records	5.858	supremacist	5.957
antifa	5.733	image	5.511	ne	5.852	statements	5.917
brown	5.708	hits	5.508	pulled	5.810	zone	5.912
amazing	5.695	jeanclaude	5.499	mosque	5.807	temporarily	5.909
decades	5.690	hilarious	5.488	ongoing	5.802	warrant	5.874
beyond	5.684	funny	5.480	picture	5.761	visa	5.853
advisor	5.683	keeping	5.472	ni	5.689	unusual	5.851
arkansas	5.668	founder	5.471	raid	5.668	seat	5.817
bob	5.663	happened	5.469	numbers	5.650	sigmar	5.809
chaos	5.662	devos	5.457	negotiator	5.649	unit	5.804
danish	5.656	humiliated	5.444	miss	5.648	sit	5.790
demand	5.648	everywhere	5.432	los	5.637	wolfgang	5.778
bizarre	5.641	goldman	5.430	leadership	5.598	worried	5.751
certain	5.631	entered	5.408	prevent	5.581	sexually	5.740
ballistic	5.626	kim	5.382	opposed	5.574	serve	5.739
common	5.625	fourth	5.376	ove	5.568	undercover	5.731
aliens	5.614	hotel	5.364	ministers	5.527	targeted	5.729
break	5.608	guns	5.323	related	5.476	shocked	5.706
damage	5.595	jeb	5.322	maryland	5.462	va	5.706
begun	5.584	insane	5.321	leaving	5.461	victims	5.693
attorneys	5.575	dozen	5.309	peninsula	5.457	spend	5.686
coal	5.567	investigative	5.304	multiple	5.446	secure	5.685
cast	5.557	easy	5.302	presented	5.429	resumes	5.669
basically	5.547	hell	5.286	reference	5.426	sri	5.664
access	5.534	hall	5.284	minnesota	5.424	streets	5.658
cl	5.522	influence	5.282	recount	5.420	viral	5.622
demanded	5.522	fix	5.278	previously	5.416	worse	5.619
corey	5.505	electoral	5.265	orders	5.388	scalia	5.615
australia	5.499	facts	5.258	legacy	5.374	tweets	5.606
arms	5.475	el	5.256	listen	5.374	station	5.603
awkward	5.446	extended	5.236	outrageous	5.353	usa	5.592
active	5.425	green	5.214	raising	5.315	tries	5.580
atlantic	5.411	largely	5.206	powers	5.310	space	5.573

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
cooperation	5.404	kill	5.203	lo	5.300	surprised	5.568
airlines	5.397	dirty	5.199	learned	5.294	vegas	5.555
aoun	5.388	laid	5.177	nazi	5.269	suspect	5.544
canadian	5.371	desire	5.132	proven	5.243	supporting	5.541
absolute	5.362	julian	5.124	rare	5.226	rips	5.540
colombo	5.320	disaster	5.121	remains	5.217	riot	5.533
brave	5.319	jets	5.090	negotiators	5.208	train	5.475
backing	5.303	favorite	5.077	progressive	5.187	wins	5.386
deliver	5.302	fill	5.064	loretta	5.184	reputation	5.377
antonin	5.293	easier	5.063	promote	5.182	summer	5.375
changed	5.265	focused	5.053	proof	5.178	schaeuble	5.357
audience	5.265	drop	5.053	problems	5.177	terrorism	5.349
bu	5.243	diplomat	5.050	owned	5.165	trust	5.342
assembly	5.243	example	5.049	process	5.145	via	5.291
december	5.226	gold	5.044	positions	5.121	veterans	5.291
bravo	5.212	experts	5.028	moving	5.116	summit	5.275
analysis	5.208	greatest	5.020	onto	5.112	segment	5.252
banks	5.200	hot	5.014	polling	5.099	wang	5.222
congresswoman	5.193	interesting	5.010	previous	5.096	thugs	5.207
bus	5.187	halt	5.007	nine	5.069	scaramucci	5.200
computer	5.183	directly	5.005	maria	5.069	ru	5.199
colin	5.178	kenyan	5.003	perhaps	5.058	vi	5.198
catholic	5.142	kennedy	4.981	marginalise	5.052	slammed	5.196
coast	5.138	highest	4.967	leak	5.047	tens	5.193
beginning	5.126	girls	4.966	medical	5.047	tusk	5.193
data	5.102	intended	4.964	rating	5.043	rigged	5.182
agenda	5.093	doug	4.964	mentioned	5.041	sam	5.177
astana	5.085	involved	4.955	nd	5.024	sitting	5.172
clueless	5.084	golf	4.950	possibly	5.020	separate	5.165
create	5.083	giant	4.948	players	5.000	wearing	5.163
apple	5.078	exclusive	4.939	performance	4.971	totally	5.151
burns	5.072	divisive	4.938	organizations	4.943	thurs	5.150
columbia	5.064	horrific	4.906	overseas	4.925	teen	5.148
carried	5.052	envoy	4.901	oh	4.921	submarine	5.123
contrived	5.052	fed	4.880	mental	4.917	writing	5.108
civilians	5.022	impact	4.874	lewis	4.902	unhinged	5.090
currently	5.018	haley	4.868	protester	4.895	signs	5.086
attempts	5.004	gon	4.854	nice	4.881	trail	5.082
conduct	5.000	housing	4.833	policies	4.876	truly	5.074

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
chosen	4.993	investment	4.833	pe	4.874	wave	5.070
ceremony	4.991	immigrants	4.797	propaganda	4.872	sat	5.056
agents	4.982	explosion	4.797	quick	4.870	sachs	5.052
alternate	4.961	draft	4.791	markets	4.843	upcoming	5.044
bar	4.961	directed	4.784	presidents	4.842	stein	5.036
collusion	4.941	juncker	4.763	rape	4.840	ten	5.022
bee	4.940	greg	4.762	prove	4.836	roles	5.013
businesses	4.937	judges	4.758	pull	4.835	rich	5.003
beating	4.935	drive	4.750	proving	4.827	woods	5.000
capitol	4.926	direct	4.728	missouri	4.822	waters	4.989
clapper	4.918	endorsement	4.726	netanyahu	4.811	visited	4.983
busy	4.897	dismiss	4.725	missing	4.775	su	4.974
bunch	4.886	institute	4.720	nevada	4.773	worked	4.936
chelsea	4.880	graphic	4.700	questioned	4.773	theory	4.925
although	4.869	japanese	4.694	lied	4.764	serial	4.898
cape	4.850	justin	4.687	migration	4.762	tour	4.891
dealt	4.845	jan	4.661	pressed	4.754	treatment	4.876
date	4.841	headquarters	4.651	liar	4.751	writer	4.868
candidacy	4.829	highprofile	4.635	reportedly	4.743	sounds	4.862
arguments	4.812	fan	4.635	moderate	4.739	sex	4.860
cold	4.812	eyes	4.619	migrants	4.730	simply	4.856
confirm	4.786	enemy	4.617	opposes	4.728	sacked	4.838
businessman	4.734	forming	4.606	province	4.716	update	4.816
betsy	4.723	endorse	4.598	predicted	4.707	successful	4.809
dan	4.716	kurds	4.593	market	4.705	shares	4.765
band	4.704	kasich	4.579	rallies	4.704	trey	4.754
dakota	4.703	jon	4.574	parry	4.690	stuart	4.753
deir	4.702	hide	4.573	paid	4.682	urging	4.749
defending	4.699	desperately	4.544	laws	4.674	tehran	4.731
austria	4.690	failure	4.538	mauricio	4.672	table	4.726
dealing	4.689	hypocrite	4.532	opinion	4.667	talked	4.711
accusing	4.684	hosts	4.525	magazine	4.658	testimony	4.692
commander	4.678	denmark	4.521	owners	4.640	sue	4.682
clark	4.675	explain	4.518	matt	4.636	suffering	4.677
barcelonamadrid	4.671	item	4.511	opening	4.635	surrounding	4.670
confirmation	4.670	drew	4.503	marine	4.623	requested	4.665
body	4.667	ended	4.499	port	4.604	sea	4.652
cost	4.659	hungarian	4.448	profile	4.561	residents	4.635

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
apply	4.650	identified	4.428	pac	4.556	situation	4.620
canceled	4.642	inaugural	4.421	newspaper	4.551	taxpayer	4.597
colo	4.624	imposed	4.418	patriot	4.548	road	4.583
ash	4.623	disgraced	4.414	parliamentary	4.547	shout	4.582
bigoted	4.615	guest	4.410	lawless	4.539	upheld	4.574
board	4.608	idiot	4.399	operation	4.538	whining	4.563
camp	4.606	funded	4.389	negotiations	4.537	thug	4.561
comp	4.603	faces	4.385	moves	4.524	responded	4.558
brand	4.603	dp	4.372	reduce	4.522	stephen	4.552
conversation	4.601	furious	4.367	partner	4.506	vicious	4.532
blew	4.599	di	4.366	pirro	4.496	sides	4.525
becomes	4.594	expanded	4.355	qualified	4.493	village	4.524
cam	4.587	destroyed	4.350	renewed	4.483	rouge	4.524
counting	4.551	ignorance	4.350	movie	4.470	rod	4.521
confident	4.549	fair	4.341	lewandowski	4.460	wear	4.512
begins	4.534	drills	4.333	phoenix	4.452	russians	4.498
baba	4.534	industrial	4.330	protecting	4.435	viewers	4.491
benjamin	4.528	gowdy	4.330	refuse	4.435	unity	4.487
comedy	4.526	illegals	4.294	positive	4.433	visas	4.485
baton	4.524	football	4.286	pl	4.431	sharply	4.466
coverage	4.521	hackers	4.278	meant	4.427	site	4.461
brain	4.507	ferguson	4.264	raped	4.417	suffered	4.457
amount	4.487	ideas	4.264	minority	4.415	temper	4.456
celebrate	4.483	dialogue	4.264	loss	4.408	throughout	4.454
blood	4.480	devastating	4.252	picked	4.400	waiting	4.449
convoy	4.449	feud	4.247	philippines	4.380	schumer	4.448
banning	4.439	domestic	4.230	levels	4.377	scene	4.445
continuing	4.437	joy	4.226	quote	4.351	vetoed	4.436
brennan	4.412	difficult	4.215	maralago	4.328	rioters	4.429
confederate	4.379	driver	4.199	plane	4.327	saudiled	4.419
bashar	4.377	elect	4.184	macri	4.314	vehicle	4.400
brag	4.371	interviewed	4.175	presidentele	4.312	science	4.386
behavior	4.354	discussion	4.173	ra	4.296	unless	4.381
bigotry	4.354	heavily	4.164	openly	4.293	spelling	4.374
charge	4.351	dictator	4.163	pol	4.286	silence	4.353
da	4.347	diversity	4.163	lgbt	4.268	throw	4.341
bombs	4.332	jay	4.149	proved	4.267	suggested	4.332
damascus	4.325	exist	4.147	level	4.265	seized	4.315
critics	4.325	kentucky	4.108	nikki	4.249	twice	4.307

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
ability	4.322	gathered	4.102	reaction	4.245	supports	4.304
assange	4.318	jake	4.098	obsessed	4.241	unlikely	4.300
blamed	4.318	guard	4.096	phony	4.234	uk	4.297
cincinnati	4.315	doors	4.092	libyan	4.229	reveals	4.296
alert	4.308	dhabi	4.089	moscow	4.224	smart	4.289
campus	4.306	finds	4.084	possibility	4.220	rising	4.276
contains	4.281	donations	4.078	male	4.198	suspend	4.271
albany	4.281	intense	4.073	milwaukee	4.185	spying	4.249
campaigning	4.276	gove	4.070	package	4.185	terms	4.249
bloomberg	4.269	improve	4.057	li	4.164	seekers	4.244
christians	4.255	eye	4.050	platform	4.157	strategic	4.243
berkeley	4.255	heat	4.046	managed	4.142	seemed	4.240
abuse	4.241	fdp	4.045	pair	4.140	shawn	4.239
connecticut	4.241	disappointed	4.036	placed	4.132	watters	4.214
carry	4.221	hates	4.032	protected	4.120	supported	4.205
classic	4.221	however	4.031	ov	4.118	resume	4.196
borders	4.210	democr	4.030	portland	4.114	scandals	4.176
blacks	4.201	increasingly	4.023	represent	4.098	string	4.174
bannon	4.168	kevin	4.020	refuge	4.076	singer	4.172
apart	4.137	ease	4.019	lol	4.072	suspension	4.167
clean	4.136	italian	4.013	poised	4.070	rnc	4.160
birthday	4.129	dossier	4.007	observers	4.065	riots	4.158
advanced	4.128	earthquake	4.003	proves	4.062	turning	4.146
construction	4.122	flint	3.995	publicly	4.055	smith	4.145
demands	4.121	exercise	3.975	peaceful	4.054	teachers	4.141
calif	4.120	farc	3.968	ratings	4.027	trillion	4.141
complex	4.119	dog	3.955	provided	4.024	springs	4.131
career	4.115	difference	3.945	zero	4.020	subject	4.125
br	4.114	heavy	3.942	oversight	3.998	tu	4.116
comm	4.110	honor	3.931	reacted	3.976	voice	4.115
banned	4.107	discovered	3.931	pop	3.967	tantrum	4.101
criminals	4.104	jesse	3.917	remember	3.958	tapper	4.098
baby	4.098	famous	3.915	petition	3.950	somalia	4.098
category	4.096	draw	3.913	offend	3.943	sense	4.095
broken	4.083	heads	3.899	mor	3.943	speculation	4.086
dawn	4.067	fun	3.867	receive	3.942	signing	4.078
associated	4.058	jimmy	3.864	native	3.939	visiting	4.067
coulter	4.056	income	3.856	mocked	3.910	understand	4.033
bombshell	4.056	function	3.852	loyal	3.906	session	4.032

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
barring	4.052	js	3.852	prolife	3.906	tape	4.029
costs	4.048	fail	3.848	misconduct	3.896	technology	3.999
dean	4.035	jackson	3.839	probusiness	3.884	resolve	3.983
aircraft	4.034	hurt	3.838	peo	3.884	standards	3.981
association	4.030	either	3.837	raise	3.864	sur	3.967
cha	4.028	ed	3.835	mr	3.858	territory	3.966
baghdaderbil	4.017	exploratory	3.830	modern	3.839	storm	3.966
crew	4.016	determine	3.823	paragraphs	3.838	res	3.966
arm	3.999	kfnx	3.820	poland	3.838	ugly	3.963
camera	3.996	encouraged	3.799	names	3.837	sell	3.950
affordable	3.973	houston	3.799	lashing	3.835	wolf	3.946
bel	3.965	fb	3.789	lately	3.833	sentence	3.926
club	3.949	focus	3.772	notice	3.829	todd	3.925
advice	3.947	disputed	3.770	launching	3.825	threeway	3.923
convince	3.938	devin	3.768	reminds	3.822	rob	3.919
app	3.937	goal	3.758	mom	3.810	slams	3.904
added	3.937	fighter	3.749	rein	3.805	stood	3.889
bullied	3.928	jury	3.747	partners	3.796	runner	3.889
clarke	3.916	jewish	3.745	pakistan	3.795	widely	3.882
closest	3.905	jong	3.743	nat	3.793	soninlaw	3.873
christopher	3.905	happening	3.741	learn	3.777	restrictions	3.863
delegation	3.890	horrible	3.741	mouth	3.776	shocker	3.863
che	3.879	expert	3.716	nunes	3.768	var	3.852
appealed	3.876	islands	3.714	oppose	3.767	susan	3.851
agriculture	3.856	journal	3.713	piece	3.745	soldier	3.845
declaration	3.856	fla	3.711	online	3.744	theories	3.837
culture	3.848	griffin	3.703	per	3.740	uses	3.831
asia	3.835	families	3.690	promising	3.736	search	3.831
authority	3.833	determined	3.688	recognize	3.728	stronghold	3.829
apologized	3.832	hotels	3.682	maxine	3.725	serving	3.828
defeated	3.817	individuals	3.681	proposals	3.704	resolution	3.827
audio	3.813	idaho	3.678	regulation	3.700	southeast	3.827
analyst	3.796	joke	3.669	pi	3.695	representing	3.815
bond	3.771	doctor	3.650	mayors	3.691	stance	3.810
awarded	3.759	follows	3.649	oval	3.680	sixth	3.803
customer	3.759	gover	3.631	laying	3.676	tower	3.803
causing	3.756	huma	3.622	orlando	3.673	sheila	3.791
commentator	3.746	document	3.613	nra	3.670	restrict	3.775
bills	3.742	gingrich	3.612	putting	3.662	wish	3.773

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
clown	3.735	deployment	3.606	nationwide	3.652	romanian	3.768
airbase	3.732	introduce	3.603	mate	3.652	written	3.758
conducted	3.718	interested	3.602	prior	3.651	type	3.720
deals	3.717	hat	3.598	range	3.649	signaled	3.719
critical	3.716	greens	3.598	loser	3.649	supposed	3.704
aside	3.710	hosted	3.597	memo	3.637	required	3.685
busted	3.695	dream	3.592	moon	3.625	solution	3.685
constitutional	3.687	incredible	3.578	newt	3.612	strongly	3.680
books	3.675	incredibly	3.577	perfect	3.601	te	3.677
congratulated	3.661	dick	3.575	mccarthy	3.589	terry	3.674
delegates	3.654	fears	3.571	places	3.588	welcome	3.671
blocking	3.643	display	3.570	peshmerga	3.583	scale	3.667
commissioner	3.641	kalonzo	3.565	mic	3.583	sc	3.660
ahmed	3.637	islam	3.561	natural	3.568	rush	3.645
cash	3.630	gathering	3.561	reid	3.559	sector	3.640
benefits	3.623	insult	3.558	nominees	3.554	thur	3.639
convinced	3.616	focusing	3.557	netherlands	3.550	sdf	3.624
commentary	3.605	fool	3.557	register	3.549	russiagate	3.620
capitalism	3.604	factory	3.553	refusal	3.537	wing	3.616
brzezinski	3.598	england	3.547	path	3.526	stores	3.616
clashed	3.596	file	3.520	miami	3.519	securit	3.605
ages	3.596	emanuel	3.513	perry	3.515	toll	3.604
baseball	3.591	hannity	3.503	lock	3.509	spied	3.603
buckle	3.586	kirkuk	3.497	refuses	3.504	rowe	3.595
delusional	3.583	demonstrators	3.485	necessary	3.504	trudeau	3.589
creating	3.582	institutional	3.478	queen	3.502	violated	3.582
advisers	3.569	hoping	3.475	numerous	3.501	tomi	3.560
amend	3.558	funds	3.472	pleaded	3.496	secured	3.550
connection	3.535	initiative	3.467	offering	3.490	significant	3.540
clock	3.523	illegally	3.467	meets	3.482	requiring	3.531
code	3.519	dnc	3.464	mnuchin	3.476	require	3.522
deadly	3.513	effectively	3.462	leg	3.474	speed	3.519
abandon	3.508	deplorable	3.459	peter	3.449	republi	3.514
brooklyn	3.496	fjs	3.452	links	3.444	responding	3.510
contender	3.495	insurgents	3.448	parent	3.440	revised	3.509
brigitte	3.495	duty	3.432	pharmaceutical	3.423	unidentified	3.506
cutting	3.488	ground	3.430	religion	3.419	stability	3.504
creepy	3.479	failures	3.426	net	3.418	sinai	3.485

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
cooper	3.478	fifth	3.403	regime	3.411	standoff	3.481
alive	3.477	kenyaled	3.397	remain	3.400	wednesda	3.478
campaig	3.476	founded	3.392	rand	3.387	videos	3.474
crowds	3.473	ethnic	3.391	negotiating	3.387	suit	3.465
announces	3.472	korybko	3.388	lobbyists	3.383	stelter	3.464
awful	3.471	geneva	3.383	lodged	3.382	taiwanese	3.462
confronted	3.467	expose	3.373	opponents	3.376	thinking	3.454
comparison	3.464	exit	3.369	outlets	3.370	track	3.445
commit	3.435	describing	3.367	lavish	3.362	sanctuary	3.434
addressed	3.433	expand	3.350	parenthood	3.360	sweden	3.431
cons	3.428	ivory	3.335	passing	3.358	screams	3.430
assaulted	3.425	executives	3.330	marked	3.350	size	3.426
arrogant	3.420	fir	3.325	remind	3.346	warrants	3.425
commercial	3.395	forms	3.321	prince	3.346	rouhani	3.421
accord	3.393	extra	3.319	mea	3.342	vs	3.421
count	3.388	fought	3.310	partisan	3.342	ultimate	3.402
deeply	3.375	jesus	3.302	mcconnell	3.340	uni	3.401
breaks	3.364	hysteria	3.298	premier	3.335	screwed	3.393
alarm	3.348	investors	3.285	prayer	3.327	rest	3.387
deeper	3.337	finding	3.281	reg	3.319	sold	3.386
article	3.336	exposing	3.280	lemon	3.315	training	3.382
african	3.325	ethiopia	3.271	receiving	3.313	weak	3.381
afraid	3.319	invitation	3.265	memorial	3.311	securities	3.379
cars	3.319	jerry	3.249	legend	3.309	serbian	3.372
brownback	3.305	flashback	3.245	praise	3.300	restore	3.370
crossing	3.299	kept	3.245	mount	3.298	unacceptable	3.362
commitment	3.296	granted	3.245	poverty	3.296	spokesperson	3.347
acknowledged	3.289	insisted	3.237	occupy	3.292	teenage	3.344
consumed	3.285	edward	3.236	minimum	3.284	vietnamese	3.342
cleared	3.280	jus	3.235	montana	3.284	tuesd	3.341
belt	3.280	experience	3.232	passenger	3.283	thank	3.341
card	3.276	jaein	3.230	minorities	3.281	testified	3.340
administratio	3.274	disney	3.228	publication	3.256	sun	3.331
buying	3.272	gift	3.227	potus	3.255	trum	3.321
daniel	3.271	ignored	3.225	pissed	3.242	republic	3.318
antiamerican	3.260	finished	3.222	personality	3.226	worry	3.315
bishkek	3.259	destroys	3.218	monster	3.223	task	3.312
brothers	3.251	except	3.214	mohamed	3.220	whe	3.294

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
apology	3.250	holiday	3.205	notorious	3.218	somehow	3.288
champion	3.235	environment	3.203	pressing	3.218	resumed	3.286
bringing	3.232	farmers	3.196	min	3.218	unprecedented	3.281
criticizing	3.230	infamous	3.191	outgoing	3.216	shoot	3.266
celebrated	3.230	jill	3.185	prospect	3.210	speeches	3.259
cover	3.227	erupted	3.183	mick	3.199	woke	3.251
admitted	3.217	division	3.176	regards	3.197	thu	3.247
afternoon	3.215	info	3.173	leads	3.194	thursd	3.243
behalf	3.210	guevara	3.158	mooch	3.188	tweeting	3.242
censorship	3.207	ir	3.154	migrant	3.184	thin	3.233
concerning	3.207	fine	3.151	negative	3.173	victories	3.232
boat	3.195	getelements bytagname	3.149	protested	3.172	scarborough	3.231
barry	3.195	getelementb yid	3.149	postponed	3.167	roll	3.200
columnist	3.193	discussing	3.147	mon	3.163	veritas	3.198
bragging	3.176	extend	3.142	podesta	3.161	stern	3.195
bl	3.173	icon	3.140	ref	3.160	thursda	3.194
crack	3.172	klein	3.140	professional	3.153	stone	3.188
closing	3.162	harassment	3.134	liberties	3.152	research	3.187
crude	3.161	hunt	3.127	portion	3.148	sink	3.178
celebrities	3.154	eliminate	3.126	pundit	3.140	unreal	3.175
allen	3.153	gore	3.125	pu	3.133	targeting	3.162
buy	3.152	incoming	3.121	lepage	3.133	vision	3.154
collapse	3.150	influential	3.107	regulator	3.133	scam	3.150
controversy	3.148	kerry	3.107	pictures	3.132	visits	3.149
approve	3.147	expanding	3.105	pledge	3.131	tal	3.148
delaying	3.136	embattled	3.103	options	3.124	strongest	3.147
charlie	3.133	era	3.102	philadelphia	3.121	steven	3.140
barely	3.129	investigate	3.102	legislators	3.115	values	3.137
ammon	3.121	golden	3.097	proud	3.114	tw	3.135
controls	3.119	hypocrisy	3.094	nigerian	3.111	rushed	3.135
consistent	3.111	lahren	3.087	reagan	3.101	tea	3.135
ci	3.101	joining	3.085	rahm	3.101	shift	3.132
alassad	3.094	gary	3.084	quiet	3.099	study	3.132
aggressive	3.089	interests	3.080	rank	3.097	success	3.129
bentley	3.088	estate	3.077	privilege	3.082	witch	3.127
clashes	3.088	individual	3.076	navarro	3.081	wore	3.118
bloody	3.081	forme	3.074	mc	3.074	violating	3.107
bla	3.074	flags	3.072	missiles	3.071	silent	3.103

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
annual	3.072	involving	3.070	nationals	3.052	stands	3.091
creation	3.065	lankan	3.067	letting	3.046	statue	3.085
crackdown	3.064	jason	3.065	neighbors	3.042	unfortunately	3.084
ball	3.063	huckabee	3.063	model	3.035	walk	3.076
danger	3.062	featuring	3.063	meddling	3.035	sending	3.068
briefing	3.054	exploring	3.060	legislature	3.034	skin	3.067
celebrity	3.054	expressing	3.055	libertarian	3.032	rolling	3.066
dark	3.052	handful	3.048	qatar	3.032	seattle	3.063
asian	3.050	lanka	3.048	repeated	3.025	sane	3.061
brutally	3.049	kaine	3.046	orange	2.993	sources	3.047
chemical	3.048	excon	3.044	pension	2.989	schulz	3.039
cli	3.037	donnell	3.029	particularly	2.985	screaming	3.039
aim	3.031	gig	3.025	rate	2.982	season	3.031
bloc	3.027	imam	3.023	maintained	2.975	shutting	3.027
assassination	3.027	kick	3.021	reiterated	2.970	taxpayers	3.018
checkpoint	3.020	fall	3.019	newest	2.956	suggestion	3.016
burn	3.017	ibrahim	3.019	physically	2.953	revolutionary	3.013
boeing	3.016	guesses	3.012	ou	2.945	runs	3.003
baldwin	3.007	gotten	3.009	membership	2.944	slovakia	2.989
argument	3.007	foot	3.006	lay	2.930	voiced	2.985
compared	3.003	emerging	3.000	marriage	2.923	sound	2.975
brother	2.994	indeed	2.998	lets	2.917	scorches	2.974
assaulting	2.990	lack	2.992	participate	2.917	ship	2.974
burka	2.987	examining	2.973	proposes	2.914	thrown	2.970
bratislava	2.980	initial	2.963	potentially	2.912	rock	2.957
core	2.976	fallen	2.957	maintain	2.911	text	2.945
clintons	2.976	doubts	2.950	lea	2.909	wallace	2.943
ads	2.975	honest	2.949	neurosurgeon	2.905	scary	2.941
barnier	2.965	holocaust	2.948	overnight	2.903	thanked	2.932
deadline	2.950	fleeing	2.947	original	2.900	stunned	2.926
besides	2.943	ignore	2.945	leftists	2.889	wells	2.923
box	2.931	existence	2.941	ramping	2.882	vermont	2.908
corps	2.927	flight	2.938	reflect	2.881	spin	2.907
claire	2.925	generally	2.936	marijuana	2.881	van	2.906
ayatollah	2.918	dying	2.928	pleased	2.879	swamp	2.899
boss	2.913	jose	2.909	registered	2.877	violate	2.895
beijingseoul	2.909	heading	2.897	obviously	2.876	socialism	2.880
adults	2.904	gulf	2.894	par	2.875	setback	2.878

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
cameroon	2.903	disgraceful	2.887	patriots	2.874	wirethe	2.876
crossed	2.892	hole	2.881	music	2.870	steps	2.867
boycott	2.889	inspired	2.878	pervert	2.867	weiner	2.860
angel	2.886	endless	2.873	mus	2.866	schwarzenegger	2.857
condition	2.884	enterprise	2.873	outspoken	2.848	sad	2.856
convincing	2.877	editors	2.871	rachel	2.843	restaurant	2.855
blames	2.873	ethics	2.860	manhattan	2.826	universe	2.851
bigot	2.865	flights	2.859	pat	2.826	whine	2.850
blatant	2.860	iii	2.853	prote	2.825	warns	2.849
debates	2.859	fiorina	2.852	preliminary	2.813	sally	2.848
alongside	2.858	iranianbacked	2.852	puts	2.811	sec	2.844
arnold	2.857	knocked	2.849	obamaera	2.799	uber	2.840
carly	2.852	developing	2.843	nyc	2.798	toxic	2.838
castro	2.852	importance	2.840	poli	2.791	whistleblower	2.837
declaring	2.850	ht	2.837	looked	2.787	ultimately	2.836
chances	2.846	inner	2.832	oth	2.787	watched	2.828
aviation	2.842	firs	2.831	laura	2.782	software	2.825
antimuslim	2.842	hitler	2.830	none	2.779	reviews	2.824
auto	2.827	emirates	2.827	probing	2.778	returning	2.818
brace	2.826	fringe	2.820	regulations	2.776	wednes	2.816
confused	2.817	investigated	2.819	mulvaney	2.770	stopped	2.812
bristol	2.809	keefe	2.815	lunch	2.760	triggered	2.811
bodies	2.808	illustration	2.809	ob	2.758	republicanled	2.809
bolton	2.808	hedge	2.801	remained	2.755	shooter	2.803
airline	2.802	deportation	2.794	narrative	2.753	rose	2.802
administra	2.799	fel	2.787	pattern	2.752	vehicles	2.800
dem	2.792	ending	2.778	malta	2.751	waste	2.792
apparent	2.791	kathy	2.775	operatives	2.749	single	2.791
bo	2.791	incompetence	2.775	loans	2.743	vietnam	2.787
assistance	2.789	ku	2.772	raids	2.742	settled	2.782
daughters	2.778	guests	2.770	mary	2.737	requests	2.774
basketball	2.778	fame	2.768	laundering	2.722	sudden	2.765
cancel	2.777	implement	2.766	lobbyist	2.721	seal	2.760
approximately	2.757	drawing	2.759	nightmare	2.720	restraint	2.759
compromise	2.748	habit	2.758	mississippi	2.720	sort	2.758
assistant	2.743	grow	2.756	punish	2.720	tanks	2.757
concert	2.742	eln	2.750	marxist	2.719	spokeswoman	2.752

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
clever	2.737	filmmaker	2.750	medicine	2.715	vacation	2.746
contest	2.732	fast	2.750	rarely	2.711	shared	2.744
chain	2.723	farm	2.746	remark	2.710	wonderful	2.743
ambitious	2.722	formal	2.745	opponent	2.709	wednesd	2.739
advisory	2.722	dishonest	2.742	organized	2.707	surfaced	2.737
amnesty	2.718	helicopter	2.736	orban	2.704	supposedly	2.728
curious	2.717	governmen	2.735	locked	2.698	whenever	2.724
bamako	2.714	enter	2.735	popularity	2.694	stronger	2.720
anonymous	2.713	exports	2.731	pompeo	2.694	treated	2.719
australian	2.713	kor	2.729	offi	2.689	split	2.712
delayed	2.710	dems	2.726	obamac	2.682	viktor	2.704
bil	2.708	guess	2.723	planet	2.679	roads	2.695
babies	2.706	demonstrated	2.716	logic	2.668	struggling	2.690
combat	2.706	feels	2.716	perfectly	2.659	shootings	2.689
adopted	2.698	jet	2.712	nigel	2.654	scalise	2.688
classified	2.697	jo	2.705	releases	2.649	served	2.680
admit	2.694	donor	2.703	reject	2.648	welfare	2.678
del	2.690	eat	2.700	reforms	2.647	taxcut	2.669
accident	2.687	lame	2.696	principle	2.646	respects	2.669
attempted	2.686	denying	2.694	pruitt	2.643	steinmeier	2.665
cambridge	2.668	harder	2.693	nature	2.641	senato	2.663
brandon	2.658	immediate	2.690	match	2.641	theme	2.663
coffers	2.658	increasing	2.688	limits	2.638	sovereignty	2.660
burning	2.658	donated	2.686	pretending	2.633	unlike	2.657
declare	2.657	junta	2.680	precious	2.631	sp	2.657
communities	2.649	filled	2.679	leaks	2.627	uncontrollable	2.656
consequences	2.648	gates	2.671	reads	2.626	touch	2.656
charter	2.643	detail	2.668	mitch	2.624	websites	2.653
delivers	2.643	effect	2.668	mahmoud	2.621	warming	2.652
arrive	2.642	guards	2.668	mouthpiece	2.621	wiretapping	2.648
cool	2.641	humanity	2.667	population	2.617	sho	2.647
beijingdaily	2.641	jonathan	2.663	prepare	2.611	ret	2.644
age	2.635	garland	2.663	missed	2.609	republicanbacked	2.637
consumers	2.625	koch	2.662	outlet	2.608	seth	2.636
contributor	2.624	followers	2.658	medal	2.599	switzerland	2.629
billy	2.623	imprisoned	2.657	popcorn	2.599	snyder	2.621
attending	2.621	grab	2.654	presidenti	2.594	unfit	2.618
affair	2.621	farage	2.654	mohammed	2.590	scientists	2.618

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
allied	2.619	horror	2.651	lines	2.583	ri	2.600
delta	2.611	democra	2.650	lockheed	2.577	sudan	2.596
decade	2.608	demo	2.648	paper	2.576	wedn	2.594
boko	2.606	intensified	2.648	mond	2.576	snap	2.593
colombian	2.606	highway	2.644	oba	2.568	shinawatra	2.590
bikers	2.604	insults	2.641	normal	2.568	staged	2.586
bridgewater	2.594	deserve	2.637	payments	2.562	ronald	2.584
antonio	2.594	jamie	2.632	presidentia	2.562	spot	2.582
buzzfeed	2.592	heated	2.618	pact	2.548	valley	2.577
arlington	2.588	ford	2.616	material	2.543	suppo	2.576
barzani	2.588	figh	2.605	peterson	2.538	tuesda	2.574
affairs	2.577	explode	2.602	petty	2.537	timing	2.571
declassified	2.574	denial	2.597	percentage	2.535	shootout	2.568
bars	2.573	dump	2.595	paramilitary	2.534	works	2.565
beyonc	2.563	fundraising	2.592	peopl	2.531	swept	2.561
cited	2.553	ensued	2.588	luther	2.530	stick	2.559
agreements	2.552	experienced	2.588	misogynist	2.530	silicon	2.559
cruise	2.552	int	2.586	nobody	2.529	rigging	2.558
content	2.543	doj	2.575	latino	2.529	veto	2.554
corker	2.543	grave	2.575	racists	2.525	secretarygeneral	2.553
bigger	2.538	hitting	2.574	parliamentman	2.522	responses	2.553
contact	2.537	karl	2.570	refiles	2.520	swedish	2.550
arabian	2.528	discredit	2.569	politically	2.520	volunteer	2.539
camerota	2.527	invest	2.565	perform	2.517	rid	2.534
completed	2.526	fuel	2.564	operating	2.516	rice	2.533
customers	2.524	economist	2.564	maddow	2.513	status	2.532
challenger	2.522	die	2.563	nixon	2.511	rocked	2.525
cat	2.521	hardline	2.562	nati	2.509	stealing	2.523
boil	2.520	houthi	2.557	pyongyang	2.506	suing	2.515
cable	2.517	feb	2.551	manufacturing	2.505	veiled	2.513
beautiful	2.517	landslide	2.550	rallied	2.498	revelation	2.512
anger	2.515	flies	2.550	raping	2.495	spec	2.512
brighton	2.513	katie	2.550	lists	2.487	unbelievable	2.510
caribbean	2.512	historically	2.546	muslimmajority	2.475	sale	2.495
artist	2.510	everybody	2.546	replacement	2.474	withdrew	2.494
beach	2.508	financier	2.545	poorly	2.474	shifted	2.486
bombers	2.502	differences	2.543	observed	2.473	roughly	2.473
approach	2.500	hungry	2.543	reasons	2.461	vetting	2.471

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
acts	2.496	equipment	2.543	mevlut	2.458	stephanopoulos	2.470
anymore	2.496	falsely	2.537	modi	2.458	various	2.468
command	2.495	engage	2.532	particular	2.449	reshuffle	2.468
bible	2.491	instructed	2.532	puppet	2.447	restoring	2.464
brief	2.489	earth	2.530	relief	2.436	rosie	2.454
catalans	2.481	flew	2.528	regular	2.435	selfies	2.453
alphabet	2.479	kurdishled	2.524	migh	2.433	shafik	2.446
allegation	2.479	indicted	2.519	opportunities	2.432	sites	2.439
backs	2.475	dividing	2.509	nominating	2.428	struggled	2.438
coleader	2.473	inte	2.508	observatory	2.425	retailers	2.417
administrati	2.472	eyebrows	2.507	promoted	2.424	traveled	2.417
bauchi	2.469	ice	2.502	pm	2.423	tragedy	2.413
activity	2.467	fundraiser	2.500	mariano	2.421	suspicious	2.403
agent	2.466	insider	2.500	opportunity	2.411	serbia	2.402
bonds	2.465	evil	2.498	looks	2.407	utah	2.401
alone	2.458	deny	2.497	momentum	2.407	wise	2.400
cards	2.454	favored	2.493	progovernment	2.400	vans	2.398
advocacy	2.448	incorrect	2.492	papal	2.396	supremacists	2.395
bathrooms	2.443	faux	2.490	preparations	2.396	style	2.376
customs	2.442	dire	2.489	practice	2.393	telephone	2.376
custody	2.441	incident	2.486	reconsider	2.384	worldwide	2.374
brotherhood	2.440	heckler	2.484	products	2.380	tend	2.370
commander inchief	2.436	embarrassment	2.482	mont	2.378	spread	2.370
critic	2.429	giuliani	2.482	participated	2.377	sees	2.367
cooperate	2.428	interim	2.463	pundits	2.371	secur	2.367
bern	2.420	embarrassed	2.462	magnitude	2.369	sparked	2.366
chanting	2.420	execu	2.459	mika	2.368	stud	2.362
campaigned	2.417	gr	2.456	mccabe	2.360	spoken	2.362
columbus	2.416	elec	2.450	patton	2.358	rescue	2.361
cohn	2.416	janet	2.441	ralph	2.357	respond	2.360
clip	2.407	freedoms	2.437	removal	2.355	verdict	2.355
alleging	2.403	integration	2.432	property	2.354	tone	2.351
bombing	2.401	frustrated	2.431	partial	2.352	stepping	2.351
birmingham	2.401	dissident	2.415	milo	2.349	socialized	2.348
bold	2.395	fly	2.408	prefer	2.345	toilet	2.344
challenges	2.391	dershowitz	2.404	letters	2.341	resolved	2.342
benefit	2.389	extending	2.396	mohammad	2.340	ripped	2.342
anthem	2.382	fantastic	2.394	raises	2.339	song	2.341

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
clash	2.371	flake	2.393	promotion	2.337	whites	2.340
aired	2.370	jumped	2.391	lou	2.335	revive	2.339
crushed	2.369	electio	2.391	maybe	2.333	scrap	2.335
catherine	2.368	forthcoming	2.387	northwest	2.332	walmart	2.333
batch	2.367	detailed	2.387	questioning	2.330	swing	2.332
beware	2.367	increased	2.386	lefty	2.330	sports	2.331
bet	2.365	follo	2.357	listening	2.327	successfully	2.322
actual	2.365	firestorm	2.354	minis	2.317	whatever	2.320
conducting	2.362	gross	2.354	pers	2.316	shy	2.317
cozy	2.354	diego	2.351	referring	2.316	wil	2.316
courage	2.349	effective	2.340	legally	2.316	stressed	2.310
admiral	2.344	downright	2.338	nazis	2.311	witnessed	2.307
adding	2.344	infighting	2.336	lit	2.311	tougher	2.301
constantly	2.343	du	2.333	omar	2.309	whi	2.298
cameras	2.340	headed	2.332	mudavadi	2.304	stabbed	2.296
conan	2.337	endorsements	2.327	mocks	2.300	syr	2.294
ambush	2.337	gabbard	2.322	nationalists	2.294	semiautonomous	2.292
coffee	2.333	firmly	2.317	moments	2.293	unilateral	2.291
advance	2.327	irancontra	2.316	lobby	2.292	terrori	2.286
cashstrapped	2.327	laptop	2.315	paulo	2.284	stuck	2.285
deflect	2.327	introducing	2.311	ministe	2.283	screening	2.285
associates	2.323	hacker	2.310	lobbying	2.282	secu	2.284
alle	2.321	gi	2.309	naval	2.275	si	2.276
cleric	2.318	falling	2.306	racial	2.272	tiny	2.276
appropriate	2.316	harm	2.303	predecessor	2.268	traveling	2.269
agai	2.310	inappropriate	2.299	nig	2.262	rescind	2.269
cheering	2.305	entering	2.296	postpone	2.262	southeastern	2.268
barbara	2.301	estimates	2.289	eac	2.262	stat	2.265
catholics	2.301	kidnapped	2.287	opposing	2.261	tent	2.261
chiefs	2.297	khan	2.286	learning	2.258	sanction	2.255
barron	2.294	grandparents	2.286	literal	2.258	store	2.253
contempt	2.293	electi	2.281	narcissist	2.257	utter	2.252
conclude	2.291	fundamentally	2.281	qaeda	2.255	tel	2.250
buried	2.289	disclosure	2.276	outraged	2.250	sai	2.248
album	2.288	inability	2.272	networks	2.246	smoke	2.248
communism	2.287	hooper	2.271	michele	2.245	teenagers	2.247
ashraf	2.275	dragged	2.268	offices	2.245	spotted	2.245

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
blue	2.275	gains	2.260	registering	2.243	unveil	2.241
advised	2.274	ele	2.260	regions	2.237	soil	2.240
candid	2.272	fined	2.259	nsa	2.235	wildlife	2.240
complaint	2.272	junior	2.257	mich	2.233	saturd	2.234
brusselsberlin	2.269	highlevel	2.253	murdered	2.233	surely	2.229
cycle	2.267	haram	2.251	painting	2.230	standard	2.222
anarchists	2.266	grasp	2.247	phase	2.230	tears	2.221
consulate	2.265	knife	2.243	remembered	2.226	sharpton	2.220
civilian	2.256	kyrgyzstan	2.239	massacre	2.223	republ	2.220
circuit	2.250	explained	2.238	quietly	2.222	tru	2.217
chat	2.250	electronic	2.229	pri	2.222	uphold	2.217
conyers	2.249	kenyans	2.229	northwestern	2.220	stable	2.213
affiliated	2.246	examine	2.229	ord	2.219	stewart	2.209
cdata	2.242	extreme	2.225	napolitano	2.216	secrets	2.203
acceptance	2.241	investigati	2.222	manuel	2.215	wed	2.198
column	2.241	demolished	2.219	recommendation	2.215	widespread	2.197
baker	2.237	exposes	2.216	panic	2.214	warplanes	2.197
concept	2.236	hosting	2.212	mtv	2.214	reveal	2.196
aiming	2.230	employment	2.211	relatively	2.212	tshirt	2.192
conv	2.229	evening	2.207	regulate	2.208	uc	2.185
challenging	2.225	intervention	2.204	pretend	2.207	satellite	2.177
anticorruption	2.223	driving	2.199	pride	2.204	teenager	2.177
defence	2.221	integrity	2.197	master	2.204	wealth	2.176
accountability	2.217	ga	2.197	mexicans	2.202	resistance	2.174
afford	2.217	divorce	2.196	nort	2.197	row	2.173
antics	2.217	forum	2.194	picking	2.196	square	2.171
assuming	2.217	images	2.193	loyalty	2.194	transform	2.171
accounts	2.216	frightening	2.191	narrow	2.193	wher	2.170
careful	2.215	expelled	2.188	rabid	2.191	wedne	2.170
ceasefire	2.209	fargo	2.173	remaining	2.188	wee	2.169
carla	2.206	harsh	2.172	rampant	2.184	virus	2.166
crashed	2.205	ignorant	2.168	nationally	2.183	server	2.163
competitive	2.203	idlib	2.166	producer	2.182	vows	2.162
commenting	2.203	identify	2.161	prediction	2.181	trends	2.160
assured	2.196	extremism	2.158	plea	2.181	severity	2.160
bat	2.181	govern	2.155	map	2.179	sunni	2.158
award	2.180	halted	2.154	op	2.172	str	2.158

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
cake	2.177	idiots	2.153	quarter	2.168	tech	2.158
dallas	2.175	demonstration	2.150	pending	2.167	solid	2.157
animal	2.173	evacuation	2.143	rap	2.165	rosenstein	2.156
bor	2.173	downtown	2.139	purchase	2.159	shinzo	2.153
ambassadors	2.170	entry	2.138	rage	2.159	wage	2.149
coordinated	2.165	findings	2.135	priest	2.155	targets	2.146
crucial	2.162	hostile	2.134	lastditch	2.154	unknown	2.144
beyonce	2.160	estimated	2.131	noticed	2.150	somali	2.144
colleagues	2.160	kindness	2.126	mode	2.148	sout	2.143
breached	2.160	juan	2.124	random	2.147	strange	2.135
chan	2.158	gang	2.124	liz	2.143	urban	2.126
defector	2.156	escape	2.123	ok	2.142	spring	2.123
associate	2.154	involvement	2.119	nex	2.142	threatens	2.122
aug	2.154	fareed	2.114	miners	2.134	tribal	2.118
beat	2.140	infrastructure	2.112	reckless	2.131	slam	2.109
bashing	2.137	guys	2.108	polish	2.131	snowflake	2.107
contract	2.136	depicting	2.103	medicaid	2.113	ticket	2.104
bias	2.135	donation	2.100	recorded	2.106	sy	2.099
apologize	2.133	gottlieb	2.098	onc	2.105	resident	2.088
cj	2.123	feature	2.096	pocket	2.105	romanians	2.087
authorized	2.121	humiliation	2.096	nails	2.103	statistics	2.085
accuse	2.120	headlines	2.094	matthews	2.089	solidarity	2.079
dam	2.116	judging	2.079	patrol	2.082	televised	2.073
bitter	2.113	device	2.074	petry	2.080	sayyed	2.073
alien	2.111	driven	2.073	repo	2.080	stoltenberg	2.071
blast	2.109	gunmen	2.072	mounting	2.074	usually	2.066
contain	2.108	derail	2.072	nasrallah	2.073	warriors	2.064
boost	2.108	feeling	2.071	narendra	2.067	surgery	2.060
addicting	2.107	jens	2.071	rajoy	2.066	sons	2.059
addressing	2.104	intent	2.062	muhammadu	2.061	settlement	2.057
clooney	2.090	lab	2.056	panther	2.060	rip	2.051
buhari	2.061	fresno	2.053	progressives	2.057	sexist	2.047
accidentally	2.049	footage	2.043	removing	2.053	wer	2.027
classes	2.034	frequency	2.041	monda	2.045	solve	2.026
commun	2.031	exists	2.039	rapper	2.036	walls	2.016
cancer	2.026	kid	2.034	packed	2.036	sharia	1.992
cliven	2.012	kazakhstan	2.027	priorities	2.030	stuff	1.974
blaming	1.988	devastated	2.024	oliver	2.008	tulsi	1.951

Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary	Terms	Sum of TF-IDF Columnwise Summary
containing	1.957	freeport	1.979	mistake	1.995	walked	1.924
blasts	1.874	dis	1.969	posing	1.942	varney	1.904
blasio	1.753	holder	1.957	rancher	1.938	untold	1.876