information

MDPI

Article

Topic Modeling for Amharic User Generated Texts

Girma Neshir [1,2,*], Andreas Rauber [3,†] and Solomon Atnafu [4,†]

1 IT Doctoral Program, Addis Ababa University, Addis Ababa 28762, Ethiopia
2 Department of Software Engineering, Addis Ababa Science and Technology University, Addis Ababa 16417, Ethiopia
3 Institute of Information Systems Engineering, Technical University of Vienna, Favoritenstraße 9-11/194-01, A-1040 Vienna, Austria; rauber@ifs.tuwien.ac.at
4 Department of Computer Science, Addis Ababa University, Addis Ababa 1176, Ethiopia; solomon.atnafu@aau.edu.et
* Correspondence: girma1978@gmail.com or girma.neshir@aau.edu.et; Tel.: +251-913021313
† These authors contributed equally to this work.

Abstract: Topic Modeling is a statistical process, which derives the latent themes from extensive collections of text. Three approaches to topic modeling exist, namely, unsupervised, semi-supervised and supervised. In this work, we develop a supervised topic model for an Amharic corpus. We also investigate the effect of stemming on topic detection on Term Frequency Inverse Document Frequency (TF-IDF) features, Latent Dirichlet Allocation (LDA) features and a combination of these two feature sets using four supervised machine learning tools, that is, Support Vector Machine (SVM), Naive Bayesian (NB), Logistic Regression (LR), and Neural Nets (NN). We evaluate our approach using an Amharic corpus of 14,751 documents of ten topic categories. Both qualitative and quantitative analysis of results show that our proposed supervised topic detection outperforms with an accuracy of 88% by SVM using state-of-the-art-approach TF-IDF word features with the application of the Synthetic Minority Over-sampling Technique (SMOTE) and with no stemming operation. The results show that text features with stemming slightly improve the performance of the topic classifier over features with no stemming.

Keywords: machine learning; SMOTE; supervised topic detection; TF-IDF n-grams feature sets; topic modeling

check for updates

Citation: Neshir, G.; Rauber, A.; Atnafu, S. Topic Modeling for Amharic User Generated Texts. *Information* **2021**, *12*, 401. https://doi.org/10.3390/info12100401

Academic Editor: Arkaitz Zubiaga

Received: 7 August 2021
Accepted: 24 September 2021
Published: 29 September 2021

1. Introduction

With the rapid advancement of social media technologies, there is a vast accumulation of user generated content on different topics. Automated data analysis is urgently needed to get the most benefit out of it. One of the most useful ways to understand text is by analyzing its topic.

Because of the emergence of different online platforms—news posts, social media platforms and other sources—the contents are available in various forms—texts, audio, video, images and graphics. Among these contents, the volume of textual content takes up the larger proportion—80% of the existing content [1]. This content is not limited to the well-resourced languages but content for less-resourced languages, such as Amharic, is increasing quickly.

The availability of such content provides opportunities and challenges. Various text analytics applications can be developed including topic summarization [2], text classification [3], information extraction [4,5], sentiment analysis [6,7], lexicon building [8–10] and hate speech detection [11,12], just to name a few. This allows companies and governments to increase the quality of services, increase profits and decrease their costs. However, the analysis of texts has various challenges such as the curse of dimensionality, lack of quality of the data and, specifically, a lack of linguistic resources in less resourced languages like Amharic.

To reduce the dimensionality of text features, topic modeling is widely used. Textual contents are summarized and categorized by topics. Topic modeling is a way of recognizing and extracting different topics across a collection of documents. Topic detection is frequently applied before performing other applications. For example, sentiment analysis of the text is detected under a topic of interest, which helps with taking decisions for the daily activities of companies and government officials.

Topic modeling is strongly researched in resourceful languages like English. However, research on less-resourced languages, such as Amharic, cannot benefit much from these solutions. This is because there are few effective linguistic resources, preprocessing tools, part-of-speech taggers, named entity recognizers, and sufficient labeled datasets. Basic preprocessing operations include normalization, stop words detection, tokenization, lemmatization, punctuation removals, and so on.

This paper presents research that provides datasets which can be used for developing topic detection models for the Amharic language. It is further used to test the performance of topic detection for Amharic texts.

In the Amharic language, there have been few works carried out in topic modeling [13,14]. Both these studies are unsupervised topic modeling approaches. However, to develop effective topic detection, a supervised approach is preferred. A supervised approach to topic detection is more accurate at identifying the topics in a document and avoids the overlap of topics that cannot be resolved easily by clustering methods in unsupervised topic modeling. The lack of Amharic topic detection research, along with the lack of a topic detection dataset, has motivated us to contribute our share to fill this gap.

Prominent topic models include Latent Semantic Analysis (LSA), probabilistic Latent Semantic Analysis (pLSA), and Latent Dirichlet Allocation (LDA). LSA is an algebraic approach that discovers topics in a document relying on the document term TF-IDF weight matrices. However, this representation is less efficient as it requires large sets of documents and vocabulary. pLSA is a probabilistic variant of LSA, which uses a probabilistic method instead of the singular value decomposition of matrices.

The main idea of pLSA is to find a probabilistic model for the latent (hidden) topics that can generate the data we observe in the document-term matrix. Topic models assume that each document contains a mixture of topics and each topic is comprised of a set of terms. pLSA makes use of maximization expectation algorithms that try to find the most likely parameter estimate of the topic model, which depends on unobserved latent variables and which is flexible. Yet, it has the problem of assigning a probability for a new document. The other drawback is that the number of parameters increases as the number of documents increases. This causes the model to over-fit. To address these issues, LDA is applied as the most popular topic model.

In this research, we apply supervised approaches for the topic detection of user generated texts. This research deals with topic modeling for collections of Amharic user generated texts.

We address the following research questions: (1) Does LDA provide a suitable feature set for discriminating Amharic user generated texts into a specific topic category? (2) Do preprocessing operations, specifically stemmers, have a positive effect on the topic modeling of Amharic user generated text? (3) To what extent does supervised topic detection improve topic classification? (4) To what extent are the topic categories accurately predicted by the trained model?

The key contributions of this research are as follows:

- We provide annotated datasets of user generated content for supervised topic modeling in Amharic [15];
- We identify the most salient features (TF-IDF, LDA or combinations) to discriminate topics by machine learning models;
- We investigate the effect of stemming with TF-IDF word feature on identification of topics of Amharic texts;

- We test the effect of SMOTE with TF-IDF word grams features on the detection of the topics of Amharic user generated texts;
- We compare the performance of machine learning techniques on their identification of topics in Amharic user generated texts.

The paper is organized as follows. Section 2 presents related works on topic modeling. Section 3 describes the materials and methods, including the workflow of supervised topic modeling, and states the steps of the workflow. Section 4 presents the experimental set-up and the results, followed by a discussion of the results both quantitatively and qualitatively. An error analysis is also shown. The last section draws conclusions from the results of the research, followed by the answers to the research questions and future works.

2. Literature Review

The related work is organized into the following categories: (i) feature extraction, (ii) effects of stemming, (iii) topic classification with imbalanced datasets, and (iv) current approaches to topic modeling.

(i) Feature representation: Feature representation techniques of supervised topic modeling are surveyed. Studies revealed that bigram features are more important than unigram features for text classification [16].

In text processing for text categorization, the most widely used features include Bag of Words (BOW), Term Frequency-Inverse Document Frequency (TF-IDF), N-Gram (Char/word) language modeling features, topic modelling features and word embedding features. For instance, Gou et al. [17] proposed the term frequency-inverse topic frequency (TF-ITF) techniques to provide proper weight to the terms which discriminate the topics in a document for supervised topic modeling. In this research, we investigate the most discriminant features, which have more of a performance rise in supervised topic detection.

(ii) Effects of Stemming: Stemming has a prominent effect on the performance improvement of information retrieval systems. The effects of stemming from other natural language processing tasks, such as sentiment analysis and text classification, are also investigated in various languages.

For example, Alhaj et al. [18] studied the effects of stemming on Arabic document classification. The findings of this research revealed that using the ARLStem stemmer improved the performance of SVM for Arabic document classification with a micro-F1 value of 94.64% over the other Arabic stemmers, such as Information Science Research Institute (ISRI) and Tashaphyne. SVM has also achieved high performance gains compared to the effects of the above stemmer with other machine learning algorithms (such as K-NN and NB), using TF-IDF features.

Specifically, Duwairi et al. [19] investigated feature extraction, the effects of stemming and the characteristics of the different machine learning algorithms (i.e., SVM, NB and K-NN) in Arabic sentiment analysis. Using two datasets, such as movie reviews and political datasets, the results revealed that the use of a stemmer has shown improvement in performance on political datasets compared to movie review data. Besides, the NB classifier outperformed, with an accuracy of 97.2% on movie reviews. As K-NN depends on the closest k-reviews criterion, this classifier did not show a significant performance on both datasets.

For topic modeling applications, Schofield et al. [20] studied the effects of stemmers (Porter, Porter 1, Paice/Husk, Lovins, Krovetz stemmer, S-stemmer, and WordNet lemmatizer) on topic modeling. The findings of this research are that there is no meaningful improvement performance of topic modeling. It degraded topic stability because the effect of conflation seems to reduce the quality of performance in topic modeling.

In contrast, Swapna et al. [21] investigated the impact of stemmers on the performance of Telugu text classification. The findings showed an improvement of performance of the text K-NN classifier while using different stemming approaches (language dependent, language independent, and hybrid stemming).

(iii) Imbalanced Learning: In most domains, the distribution of classes in datasets is mostly imbalanced. That means that the proportion of the number of samples in target classes is not equal. The category that has a higher number of observations (samples) in the dataset is the majority class (i.e., over-represented category), whereas classes with the least number of samples are the minority class (i.e., under-represented category). If a machine learning model is trained with such datasets, it will be biased towards the majority class. That means the model may not work effectively for detecting samples of minority classes.

To reduce this problem, the Synthetic Minority Over-sampling TEchnique (SMOTE) method has shown performance gains of machine learning in different domains [22]. For instance, SMOTE is also used for non-textual imbalanced data viz. intrusion detection [23], applying SMOTE on three botnet datasets for malware and intrusion detection systems [24], credit card default detection [25], and tweet polarity detection [26]. Using three publicly available datasets, SMOTE has shown performance gains in tweet polarity classification [26]. Similarly, Neshir et al. [27] have shown that SMOTE is improving the performance of sentiment classification using four datasets.

(iv) Current approaches to topic modeling: Topic modeling has gained greater attention in text analytics communities. Topic modeling is widely used in different text mining tasks, including in recommender systems, lexicon generation [10], sentiment analysis, breaking news detection, text classification [3], information extraction [4], concept building, and so on. This section presents the related works on topic modeling and supervised topic detection in particular.

M. Naili et al. [28] studied topic identification of Arabic texts using LDA. They showed that applying a stemmer increases the performance of topic identification. The right choice of LDA hyper parameters alpha and beta has a high impact on the identification of topics. The result reveals that LDA with a light stemmer has a high impact in increasing the quality of LDA performance by reducing unnecessary repetition (LDA without stemmer) and the loss of meanings by stemmers that generate a root (e.g., LDA with Khoja Stemmer).

In addition, the values of alpha vary depending on the domain of the application of topic identification, whereas beta (which affects the identified topics) remains unchanged when LDA is applied to different applications.

Anoop et al. [29] proposed a topic modeling guided approach for efficient and scalable concept extraction and hierarchical learning from large text corpora.

In a similar work, Toubia et al. [30] described how regular LDA is adapted to guided LDA for topic modeling, relying on seeded words for shaping the distribution of topics in documents over regular LDA without seeds.

Li et al. [31] also developed seed-guided LDA for topic identification in the condition where there is a lack of labeled datasets. Their proposed method is a novel method in which the seed-guided model topics are explicitly guided by the seed words. The new model confirmed that it consistently outperforms state-of-the art text classifiers and non-seeded LDA. The model is not sensitive to tuning parameters, making it the right choice for real world topic modeling applications.

In [32], Jagarlamudi et al. proposed a topic modeling approach that is guided by seed words, that is, biasing topic distribution towards the selected seed words. This can also improve topic–document distributions by biasing documents to select the topics associated with the seed words. The authors reported that their proposed approach with the seeded model improved topic detection compared to models that use seed naïve information.

Kwon et al. (2019) [33] developed topic modeling and sentiment analysis of an airline using with over 14,000 online reviews collected from 27 airlines. First, significant topics are judged using frequency analysis, word cloud and topic modeling and, then, for each of these topics, sentiment analysis was carried out to measure the level of customer satisfaction of an airline. Six topics, such as in-flight meal, entertainment, seat class, seat comfort, staff service, Singapore airline are identified. These topics are likely to affect customer purchasing behavior. For the identified topics, the sentiment of the customer review is analyzed either positively (i.e., satisfied) or negatively (i.e., unsatisfied).

Gou et al. (2019) [17] constructed a supervised topic model relying on term frequency—inverse topic frequency (TF-ITF). The TF-ITF method is used to build the supervised topic model, by including the weight of each topic term to discriminate topics. This work tried to investigate both symmetric and asymmetric Dirichlet prior parameters and the result of the proposed supervised topic model with TF-ITF outperformed the SOTA supervised topic models.

Few works have been carried out on the topic modeling of Amharic texts. For instance, Kebede et al. [13] developed topic modeling using Latent Dirichlet Allocation (LDA) with and without word embeddings for Amharic short texts. The study showed performance gains of the topic clustering using LDA with word embeddings with an accuracy of 97% using a test set of six categories (such as "art", "health", "sports", "politics", "other" and "science and technology").

In another study, Yirdaw et al. [14] developed a topic-based summarization of Amharic documents using probabilistic latent semantic analysis (PLSA) by identifying keywords deciding the topic categories of a document. This research has two stages; first, keywords of documents are extracted and, second, sentence(s) which best contain those keywords are included in the summary.

The topic modeling related works are organized by approaches such as topic modeling (unsupervised) [13,14,28,33–36], seeded topic modeling (semi-supervised) [29–32], and supervised topic modeling [17,37–39].

To the best of our knowledge, the best features and the effects of stemming on supervised topic modeling have not been investigated. The existing studies on the topic detection of Amharic documents use only unsupervised topic modeling approaches [13,14].

Table 1 summarizes the approaches used, the findings and the limitations of some of the related work in different languages (i.e., English, Amharic, Arabic and Telugu). The above works differ from our proposed method in the following aspects: we use (i) supervised topic modeling rather than unsupervised, (ii) TF-IDF word feature, LDA feature and a combination of these two features for supervised topic detection, (iii) an investigation of the effects of stemming on the performance of supervised topic modeling, (iv) the application of the SMOTE strategy for balancing the datasets, and (v) applying it to large scale Amharic topic detection datasets.

In this research, we apply supervised approaches for Amharic topic identification and we also investigate the effect of text features such as LDA features, word TF-IDF features, and the stemmed texts on performing the topic model on Amharic corpora. Besides, we proposed SMOTE to balance the imbalanced datasets.

Table 1. Summary of Key Related Work of Supervised Topic Modeling.

Paper	Year	Approach	Findings	Metrics	Languages	Limitations
[18]	2019	Supervised algorithms (SVM, NB and KNN) + Stemmers	ARLStem stemmer has improved performance of SVM for Arabic document classification with micro-F1 value of 94.64% over the other Arabic stemmers.	Micro-F1 score 94.64%	Arabic	The datasets used is imbalanced. i.e., entertainment (474 samples) vs. Middle East News (1462 samples).
[19]	2014	Supervised Learning (SVM, NB and KNN) + Different Feature Strategies	The results show that stemming and light stemming combined with stopwords removal adversely affected the performance of the classification for the Movie dataset and slightly improved the classification for the Politics dataset	Accuracy 96.62%	Arabic	The built-in stemming algorithm in Rapidminer tool might have higher error rates which attributed to the less accuracy of experimental results,
[20]	2016	Topic Modeling	studied the effects of various stemmers of topic modeling and results has shown that stemming has not significant improvement of topic modeling.	-	English	Applying stemmers on keywords will reduce readability, to remedy this S stemmer or modified version of porter stemmer is recommended.
[21]	2019	KNN	Finding has shown an improvement of performance of text K-NN classifier while using different stemming approaches.	F1 score 82.89%,	Telugu	There is no procedures why KNN classifier is used over the other classifiers.
[40]	2015	Evaluation of topic modeling algorithms	Standard LDA has shown poor performance on short texts like in Twitter	-	English	Coherence of topics produced by the newly developed topic models are not judged by human experts
[28]	2017	Topic modeling of Arabic texts using LDA	Results shows that applying Arabic stemmers increase performance of topic identification Arabic	F1 Score 91.86%.	Arabic	Authors suggest to further study complete topic analysis based on topic models (LDA) and word embeddings.
[32]	2012	Seeded LDA	Guided LDA improved topic detection compared to model that use seed naive information	F1 score 81%	English	Allowing a seed word to be shared across multiple sets of seed words degrades the performance.
[33]	2019	Frequency based Topic detection and sentiment modeling	Topic based sentiment classification of Airline application review.	-	English	Since the limitations have not been fully resolved, future research may develop into a more feasible study if additional user information from the demographic site can be collected and utilized.
[17]	2019	supervised topic model	supervised topic model with TF-ITF outperformed the SOTA supervised topic models	Precision 53.76%	English	Better topic model will be obtained if using a variational autoencoder, which is a powerful technique for learning latent representations
[13]	2020	LDA + Word Embeddings	LDA with word embeddings with an accuracy of 97% using test set of six categories.	Accuracy of 97%	Amharic	Better feature enrichment extraction, preparation of datasets with more number of categories, and better design of LDA to cluster short texts are also recommended.
[14]	2012	PLSA + Keywords	Topic modeling using PLSA has shown encouraging result on Amharic topic summerization.	Preci-sion/Recall 51.38%	Amharic	does not work for multiple document, query focused, update based summarization and PLSA does not consider term weightings, lack of large scale datasets for evaluation

3. Materials and Methods

3.1. Supervised Topic Detection System

In this section, we investigate the framework for topic detection using three feature sets: TF-IDF ngram features, LDA features and the combination of these two feature sets, which are learned by supervised models to discriminate the topics of Amharic user generated texts. These documents are pre-processed (removal of punctuation marks, other non-

Amharic texts and stop words). Finally, the topic model feature is built and extracted from those pre-processed documents. The LDA is used as a feature vector for machine learning models. The performance of these models is evaluated using test sets both quantitatively and qualitatively.

The proposed framework shown in Figure 1 includes different modules, such as input texts collection, preprocessing texts, supervised topic classifier approach and prediction (or evaluation).

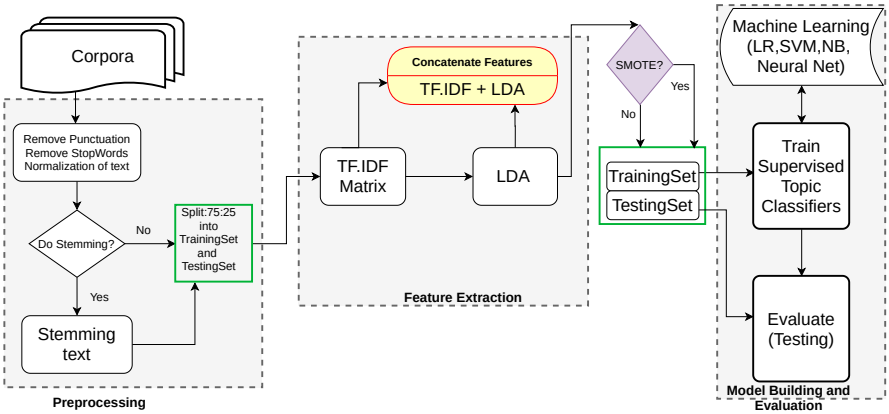

Figure 1. Workflow of the Topic Detection with Supervised Approach.

I. Input Texts: This corpus is collected from the archive of Ethiopian Reporter Amharic text collections. The corpora comprise 14,751 documents, which are used to train and test our proposed approach. The Ethiopian reporter media service publishes weekly Amharic articles from 2011 to 2017. These documents are collected and are categorized into different themes or topics such as law, arts, social, culture, local news, business, opinion, entertainment, politics, sport, nature and so on, as summarized in Table 2.

These themes are used to guide the proposed supervised machine learning algorithm to predict the theme of Amharic user generated texts in social media such as Facebook. The corpus is used to train and test the supervised topic classifiers. The complete summary of the dataset is presented in Table 2.

Table 2. Topic Modeling Amharic Corpora Description (2011–2017).

Category	#Docs	Proportion	Avg. Length	Description of the Target
News	5953	39.5%	254	Local and international news
Politics	1611	10.6%	593.5	Election and other political issues
Business	2502	16.5%	449.4	Financial, economical and other businesses
Sports	1031	6.8%	520.8	States kinds of sports
Culture	265	1.9%	504.8	States cultural events and values
Social	1635	10.8%	457.6	Social activities in community
Laws	371	2.5%	1246.5	States legal activities
Arts	749	4.9%	299.9	States arts, music, film and entertainment
Nature	176	1.2%	198.6	Related to natural resources and life
Opinion	458	3.1%	1197	Comments on current issues

As we can see in Table 2, the annotated topic categories are not balanced. The news class is the majority class as it contains 39.5% of the corpus, whereas most classes viz. culture (1.9%), laws (2.5%), nature (1.2%), and opinion (3.1%), are classes with the least number of samples in them. These classes are the minority (under-represented samples) in the dataset. So, the corpus needs to be balanced by generating synthetic samples of the minority class.

In this research, we proposed seeing the Synthetic Minority Oversampling TEchnique (SMOTE) as a balancing strategy to see its effect on the performance of the topic classifiers [41]. In addition, the average length of the samples in those categories ranges from 198.6 (nature class)

to 593.5 (politics class). The size of the samples in all categories is long enough to extract more representative features to discriminate the samples by the machine learning model.

Figure 2 evinces the distribution of samples into topic classes and it shows that the corpus is an unbalanced dataset. News is the majority class, whereas nature, laws, and culture are the minority classes. Thus, we proposed applying the SMOTE strategy and report its effect on the performance of the topic classifiers.

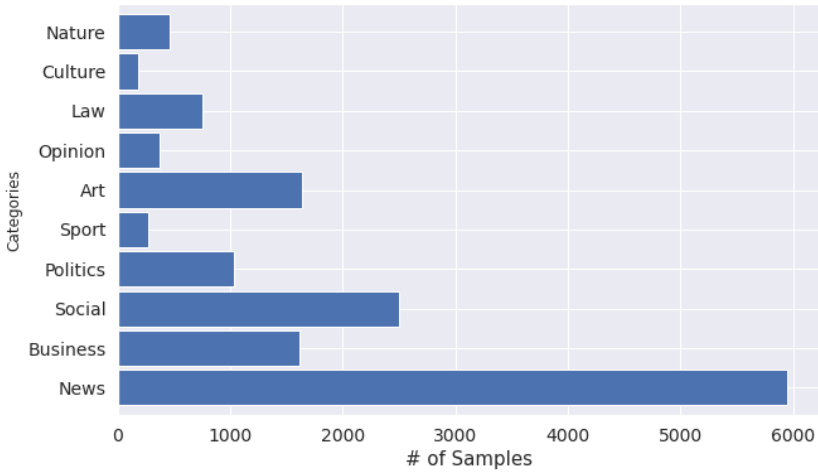

Figure 2. Distributions of Amharic user generated documents in each category.

II. Pre-process Texts: As preprocessing plays a significant role in text mining, we applied tokenization, removal of all numbers, removal of all punctuation marks and non-Amharic letters, stop-words removal, and normalized all various letters of the same sound by a common letter. To see the effect of stemming on the performance of topic detection, we used input texts with or without applying stemming. In this setup, a stemmer [42] was used.

III. Topic Modeling: Different topic modeling algorithms exist, such as Latent Semantic Indexing (LSI), Non-negative Matrix Factorization (NMF) and Latent Dirichlet Allocation (LDA). LDA is the most popular, generative probabilistic approach, which predicts/discovers topics of a document relying on probability distributions of words and documents. Thus, LDA is proposed as a feature for supervised machine learning for topic classification with the intention that it can reduce dimensionality and capture salient global features that determine topics. This means that we used LDA features for building topic detecting models to test whether it is more discriminant topic categories.

For a document $d \epsilon$, collections D and for each word $w \epsilon W$ in that document, the probability of that word w belonging to each topic $z_i \epsilon Z$ is computed [43]. Mathematically,

$$p(z|w,d) = \frac{n_{w,z} \cdot m_{w,d,z}}{N_{W,z}}\cdot, \tag{1}$$

where $p(z|w,d)$ is the probability of each topic z given word w and the document d containing it, $n_{w,z}$ is the count of w in z, $m_{w,d,z}$ is the count w in d in z and $N_{W,z}$ is the total number of words in z.

LDA makes use of Bayesian prior estimation techniques, which are specified as follows. Suppose a collection of Amharic documents of size N is represented by document-term matrix X. Then, LDA decomposes the document–term matrix into two lower dimensional matrices, such as document–topic matrix, $M1$ and topic–word matrix, $M2$ with dimensions (N, K) and (K, M), respectively, where K is the number of topics and M is the vocabulary size (i.e., unique words in the collection). The primary aim of LDA is to improve the document–topic and topic–word distributions matrices relying on sampling techniques.

The inputs to LDA are the document collections, number of topics, vocabulary size, and number of documents. The output of LDA is the topic probability distribution of each

word in the document. For each word in each document in the collection, it readjusts topic–word probability assignment.

A new topic is assigned to a word relying on the probability distribution p, which is the product of $p1$ and $p2$, where for topic k, two probability distributions are computed $p1$ = p(topic k |document d) = proportion of words in doc d that are assigned to topic k and $p2$ = p(word w| topic k) = the proportion of assignments to topic t over all docs that come from this word w. Thus, the probability of topic k for the generated word w, p(topic k| word w) = $p1 \cdot p2$.

After several iterations when a steady state is achieved, the document–topic distributions and topic–word distributions are fairly good.

Each of the above topic models have the following two issues in common: (i) each of the topic models take inputs such as number of topics and document-term matrices and (ii) each of the topic models yields two matrices, that is, word–topic matrix and topic–document matrix.

IV. Word ngram TF-IDF feature: TF-IDF is a statistical technique in information retrieval to transform text documents to vectors that are suitable for search algorithms. The term-frequency (TF) is the total count of words per document, whereas the inverse document frequency (IDF) is used to tell how rare a word is. When the words are unique, they are important features to differentiate documents in the collection.

TF-IDF score of a word w in document d is given by:

$$TF - IDF(w,d) \ = \ TF(w,d) \cdot log(\frac{N}{DF(w)}) \cdot, \tag{2}$$

where N is the total number of documents in the collection, $TF(w,d)$ is the number of occurrence of word w in document d, and $DF(w,d)$ is the number of documents containing word w.

The uni-gram word TF-IDF features are not only the feature input to the topic models, such as LDA, but also input to supervised topic classifiers such as SVM, NB, LR, and Neural Nets. So, for this work, we investigate the text features: the TF-IDF ngram features, LDA as a feature and a combination of the two features (i.e., TF-IDF and LDA feature sets) for building supervised topic models.

V. Supervised Topic Classifiers: the most popular supervised machine learning algorithms include SVM, NB, LR, and Neural Nets, which are briefly stated as follows: (i) LR is originated from statistics where this method is used for training binary categorical classes rather than continuous variables. This algorithm relates the independent variable x to dependent variable y, which has binary categorical values [44]. If a linear regression function is $y = c + m \cdot x$, then its logistic counterpart becomes $\frac{1}{(1+e^{-(c+m \cdot x)})}$, the same as:

$$\frac{e^{(c+m \cdot x)}}{(1 + e^{(c+m \cdot x)})} \cdot, \tag{3}$$

where c is a constant and m is the slope. In the case of logistic regression, Equation (3) maps the values of x to the values of y, which range from 0 to 1.

(ii) NB is a probabilistic approach that relies on Baye's rule, where the input features are assumed to be independently determines the output variable. Even though this approach worked well in many problems, this independence assumption is rare in reality. The other strength is that it can learn incrementally, scalably and update its probability distribution [44–46]. Thus, this research uses multinominal Naive Bayes, where it models the same probability. It is given by:

$$P(y|x) = \frac{P(x|y) \cdot P(y)}{P(x)} \cdot, \tag{4}$$

where $P(y|x)$ is posterior, $P(x|y)$ is likelihood probability, $P(y)$ is class prior probability and $P(x)$ is predictor prior (marginal probability). For maximized $P(y|x)$,

$$P(y|x) = P(x_1|y) \cdot P(x_2|y) \cdot P(x_3|y) \cdot ... \cdot P(x_n|y) \cdot P(y) \cdot . \quad (5)$$

The maximized probability is found by discarding the divisor probability as it is the same in both classes, that is, positive and negative.

(iii) SVM is the most popular method, which tries to divide each input feature using a hyper-plane by maximizing the distance between data points of positive and negative outputs. Besides maximizing margin distance, it minimizes errors using objective function. It considers a possible hyper-plane, which divides the outputs but optimizes and selects the best hyper-plane with a maximum distance from data points of either positive or negative class.

For linearly separable data, the hyper-plane is:

$$c + m_1x_1 + m_2x_2 + .. + m_nx_n = 0 \cdot . \quad (6)$$

This is the decision boundary, which is used to classify samples (or data points). It is re-written as:

$$W^TX = 0 \cdot \quad (7)$$

Using this equation, one can find a class of input sample x whether it is either above or below the hyper-plane. All samples less than the negative hyper-plane ($W^TX = -1$) are classified as negative class, whereas all samples greater than positive hyper-plane ($W^TX = +1$) are classified as positive class. Samples on either negative hyper-plane ($W^TX = -1$) or positive hyper-plane ($W^TX = -1$) are said to be support vectors. The distance between positive and negative hyper-planes is the margin.

The objective function of SVM is to maximize the margin. Terms of SVM such as margins, support vectors, decision boundary and hyper-planes for binary classification tasks are shown. SVM makes use of quadratic programming for optimization of margin distance of data points from hyper-plane. If data is not lineally separable, SVM kernels (e.g., polynomial, radial) maps the data to a higher dimensional space, where the data might became linearly separable.

This is called kernel tricks, which are done depending on the characteristics of the data points. Transformation of data is done in two steps: (a) finds optimal tuning parameters and (b) training SVM using such optimal parameters.

(iv) Neural Network (NN): Neural Network (NN) is one of the best machine learning techniques, inspired by the functioning of neurons in the human brain. NN is comprised of neurons as a basic unit. NN has mostly three layers (input layers, hidden layer, output layer). Each of the layers are composed of two or more neurons. The input layer accepts the input feature values (e.g., word frequency) and each neuron is associated with weights to compute the activation function and then the result is propagated to the next layer neuron and then to the output layer.

The process is repeated by adjusting weights until the error is minimized at a certain value. NN has a variety of architectures depending upon the complexity of the problem. NN is applied in various domains. For example, NN can be employed for document level sentiment classification. However, it requires much time for training the models [6].

3.2. Evaluation Metrics

To measure the performance of a classification system, evaluation is measured using metrics, including accuracy, precision, recall, F-score, and confusion matrix.

Confusion matrix is an C by C matrix used for the performance of a classification algorithm, given that C is the number of classes. The purpose of the confusion matrix is to compare the actual target values with predicted values by the classifier model. This matrix reports the summary of what the model predicted correctly and what it did not. We can

see in more detail by using a two by two confusion matrix, which is a binary classification task. So, for binary classification, the 2×2 confusion matrix is represented in Table 3.

Table 3. The confusion matrix for evaluating a binary classification system, where TP = True Positive, FP = False Positive, FN = False Negative and TN = False Negative.

		Actual Values		
		Positive	Negative	Total
Predicted Values	Positive	TP	FP	$TP + FP$
	Negative	FN	TN	$FN + TN$
	Total	$TP + FN$	$FP + TN$	$TP + FP + FN + TN$

In Table 3, the column represents the actual values, that is, Positive and Negative, whereas the row represents the predicted values of the target class. Now, the most important terms, such as TP, TN, FP and FN are:

- (i) True Positive (TP) is the number of samples in the positive class which are correctly detected by the model;
- (ii) True Negative (TN) is the number of samples in the negative class which are correctly detected by the model;
- (iii) False Positive (FP) (also called Type I Error) is the number of samples in the positive classes that are wrongly detected by the model;
- (iv) False Negative (FN) (also called Type II Error) is the number of observations in the negative classes that are wrongly detected by the model.

The model's performance evaluation metrics are presented below:

- (i) Accuracy is the proportion of observations that are correctly predicted by the model, that is,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{8}$$

- (ii) Precision (P) is a measurement of the correctly predicted observations that are actually turned out to be positive class, that is,

$$Precision(P) = \frac{TP}{TP + FP}. \tag{9}$$

- (iii) Recall (R) is a measurement of the proportion of actual positive observations, that are correctly predicted by the model, that is,

$$Recall(P) = \frac{TP}{TP + FN}. \tag{10}$$

- (iv) F-score: is a measurement metrics which returns a balanced score of both recall and precision, that is,

$$F - score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}. \tag{11}$$

4. Results and Discussions

4.1. Experimental Settings

In this section, we present the above corpus of size 14,751 is partitioned into a 75:25 ratio (i.e., a training set of size 11,063 samples and a testing set of size 3688 samples), respectively. We have implemented these algorithms relying on the Scikit learn python libraries [47].

As described above, the corpora has ten class labels, such as business, law, culture, politics, nature, arts, sports, news, opinion and social. The class labels are used to evaluate the quality of topics or themes generated by the trained supervised machine learning

model. To select the best feature sets (LDA, TF-IDF and combination of these features), three experiments have been carried out (with no stemming, with stemming and with application of SMOTE).

And the TF-IDF feature set (uni-gram, minimum document frequency = 5, maximum document frequency = 0.8, with use of IDF = True), LDA feature set (with several components = 10, learning decay = 0.7, maximum iteration = 10) and combinations of these feature sets are also compared to see which feature set has a more positive impact on topic classification performance.

The experimental settings of the hyper-parameters of machine learning algorithms for carrying out experiments are specified as: LR(C = 1, penalty = l2, tol = 0.0001), SVM (C = 1, kernel set to rbf), NB (alpha = 1), and Neural Net has been set with hidden layer size = 8,8,8, activation function set to relu, optimization solver set to adam, and number of epochs set to 500. The summary of the parameters settings are depicted in Table 4.

Table 4. Hyper-parameters values of the machine learning and the TF-IDF Vectorization.

Algorithm	Hyper-Parameter	Type	Default Value	Selected Value
TF-IDFVectorizer	analyzer	discr	word	word
	max_df	cont	1	None
	max_features	discr	None	None
	ngram_range	disc	(1,1)	(1,1)
LR	C	cont	1	1
	alpha	cont	None	None
	average	discr	None	None
	penalty	disc	l2	l2
	power_t	cont	None	None
	tol	cont	0.0001	0.0001
NB	alpha	con	1	1
	fit prior	cat	TRUE	TRUE
SVM	C	con	1	1
	coef0	con	0	0
	degree	discr	3	3
	gamma	con	scale	scale
	kernel	disct	rbf	linear
	tol	con	0.001	0.001
Neural Nets	hidden_layer_sizes	discr	-	(8,8,8)
	activation	discr	-	'relu'
	solver	discr	-	'adam'
	max_iter	discr	-	500

Using the above mentioned parameter set ups three experiments have been conducted (Exp I, Exp II and Exp III). Experiment I (Exp I) is undertaken with LDA, TF-IDF word uni-gram, and a combination of these two features, whereas experiment II (Exp II) uses the same feature sets as experiment I, but experiment II applies stemming on the feature sets. In experiment III, the datasets with the same feature sets as experiment I and II; however, in this case, it applies SMOTE data augmentation strategies for adding more augmented data samples to increase the size of the minority class.

4.2. Results

This research conducted 48 runs (i.e., three experiments × four machine learning algorithms × three feature sets). In each experiment, the above machine learning is trained and tested. The evaluations of the performance of the topic classifier models on the testing set are generated in-terms of the performance metrics such as accuracy, precision, recall and average f-score as shown in Table 5.

Table 5. Comparison of Performance of Supervised Topic Classifier using TF-IDF word, with (no) Stemming and with (no) SMOTE relying on Amharic text datasets. WoS = TF-IDF Word uni-gram + NoStem, WS = TF-IDF word uni-gram + Stemming, and WSM = TF-IDF word uni-gram + SMOTE, Exp = Experiment. The values in bold shows the highest performance scores of the topic classifier.

Model	Metric	Exp I: WoS			Exp II: WS			Exp III: WSM				
		LDA	TF-IDF	Combined	LDA	TF-IDF	Combined	LDA	TF-IDF	Combined	Train Time	Test Time
LR	Accuracy	0.41	0.78	0.78	0.40	0.81	0.81	0.49	0.87	0.84	184.11 s	0.23 s
	Precision	0.25	0.88	0.88	0.11	0.86	0.86	0.53	0.88	0.86		
	Recall	0.13	0.67	0.67	0.10	0.75	0.75	0.49	0.87	0.84		
	F1	0.11	0.74	0.74	0.06	0.80	0.80	0.46	0.87	0.84		
NB	Accuracy	0.40	0.54	0.50	0.40	0.56	0.55	0.45	0.85	0.84	0.93 s	0.21 s
	Precision	0.40	0.53	0.44	0.04	0.59	0.52	0.52	0.85	0.84		
	Recall	0.10	0.25	0.20	0.10	0.28	0.25	0.45	0.85	0.84		
	F1	0.06	0.25	0.20	0.06	0.31	0.27	0.43	0.84	0.84		
SVMLin	Accuracy	0.41	0.83	0.83	0.41	0.84	0.84	0.49	**0.88**	0.87	17.8 s	0.21s
	Precision	0.24	0.87	0.87	0.09	0.86	0.86	0.52	**0.89**	0.88		
	Recall	0.12	0.81	0.81	0.12	0.82	0.82	0.49	**0.88**	0.87		
	F1	0.09	0.83	0.83	0.09	0.84	0.84	0.45	**0.88**	0.87		
NeuralNet	Accuracy	0.51	0.76	0.69	0.53	0.78	0.72	0.53	0.75	0.77	471.8 s	0.23 s
	Precision	0.33	0.78	0.72	0.40	0.79	0.70	0.61	0.82	0.81		
	Recall	0.28	0.72	0.66	0.29	0.71	0.67	0.53	0.75	0.77		
	F1	0.29	0.73	0.67	0.31	0.75	0.68	0.51	0.76	0.77		

In the subsequent subsections, the performance of each of the models is evaluated both quantitatively and qualitatively.

4.3. Quantitative Evaluation

This subsection presents the discussion of the results of the experiments in the following perspectives: (i) effects of feature enrichment, (ii) effects of stemming, (iii) effects of SMOTE, (iv) comparing performance classifiers across class categories and (v) comparing performance of classifiers.

(i) Effects of feature enrichment: As shown in Table 5, the TF-IDF word uni-gram feature has shown better performance of topic classification not only over the other features, such as LDA and combination of LDA and TF-IDF. That means, TF-IDF word uni-gram feature is the best discriminant of topics of Amharic user generated texts across all classifiers shown in blue bars in Figure 3.

Even though the performance of the neural network model is the least of the other classifiers, the combination of LDA and TF-IDF (bars in orange) has shown a better performance over the individual features in this model. The LDA features set is the least discriminant feature for topic detectors (the bars in blue).

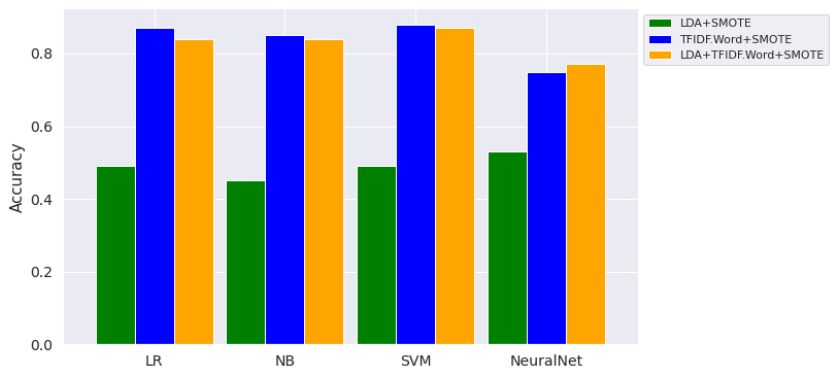

Figure 3. Effects of Features on Performance of Topic Classifiers.

(ii) Effects of stemming: Table 5 is revealing that applying stemming on the text is slightly improving the performance of the topic classification machine learning algorithms

over the features without stemming. Specifically, topic classifiers with TF-IDF with Stemming have shown tiny increment of performance among those classifiers without stemming.

That means, as is depicted in Figure 4, the TF-IDF with stemming (bars in blue) is slightly better than the TF-IDF feature without stemming (bars in green) in all classifiers. This is because the stemmer might be a heavy stemmer, which might remove semantic features of the texts, that used to discriminate topic category of a text. So, a lightweight stemmer needs to be used to preserve this information and achieve significant performance improvement of the topic classifiers.

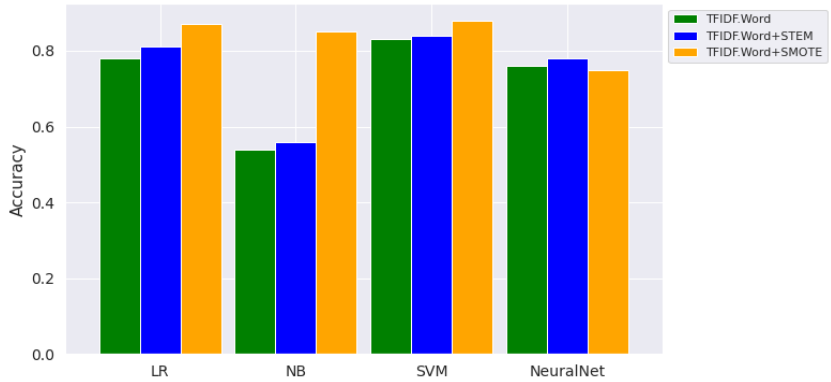

Figure 4. Effects of Stemming and SMOTE on Performance of Topic Classifiers.

(iii) Effects of SMOTE: As depicted in Table 5, the application of SMOTE has shown performance gains on topic classification of Amharic texts. In particular, TF-IDF + SMOTE (bars in orange) is the best feature, which shows the high performance gains of topic classification of most classifiers (LR, NB and SVM) as visualized in Figure 4.

However, with neural nets, the application of SMOTE has shown poor performance of topic detection as compared with all other classifiers. One of the reasons might be the lack of sufficient samples for generating augmented samples for increasing the size of minority classes. This in turn adds up samples which are not representative of the actual samples. The neural net needs large scale labeled data to increase its performance.

(iv) Comparing Performance of Classifiers: Table 5 shows that SVM outperforms in all the experimental settings, with an accuracy of 88%, precision of 89%, recall of 88% and F1 score of 88% on TF-IDF feature set with the application of SMOTE.

The result with this setup is closer to the combination of feature sets (i.e., TF-IDF + LDA) with applying SMOTE, whereas the lowest performance is reported by NB with an accuracy of 40% using LDA feature sets in both stemming and without stemming operations. One of the reasons for this is that SVM is less sensitive to noise than NB.

With the application of SMOTE on the three feature sets, the neural net (accuracy 75%, precision 82%, recall 75%, and F1 score 76%) has shown the worst performance of all other classifiers as shown in Figure 5.

Regarding training and prediction time, the performance of classifiers is also compared, as is shown in Figure 6. The neural network takes a lot of time (i.e., 471.8 s) in training the model whereas the NB classifier is trained the quickest (0.93 s). Regarding prediction, the time elapsed to predict the testing samples is almost negligible (i.e., 0.23 s) and the same time is spent by all classifiers. This shows that deploying either of the models in real-time application has no prediction time difference.

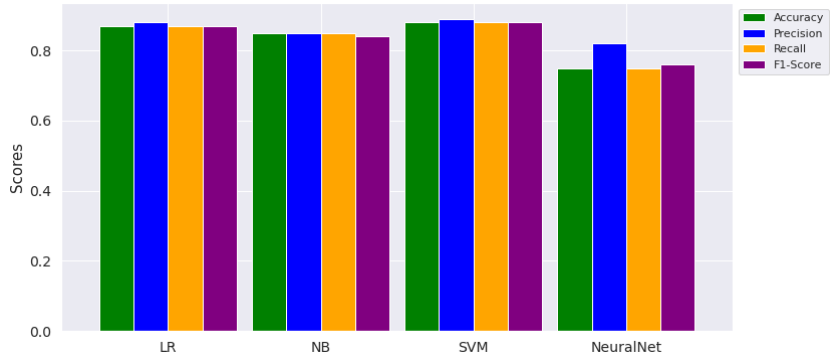

Figure 5. Comparing Performance of Topic Classifiers.

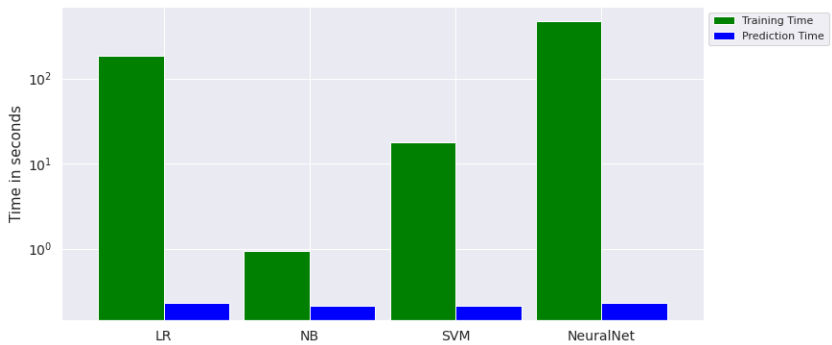

Figure 6. Comparing Performance of Topic classifiers in terms of Training and Prediction time.

(v) Comparing performance classifiers across class categories: As seen in Figure 7, SVM has performed the best in the identification of topic categories such as opinion, laws, nature and sport, whereas SVM has performed the least (precision 65%, recall 87% and F1 score 75%) in identifying Amharic texts in the news category. This is due to the fact that the contents of the news category are highly mixed up with other categories.

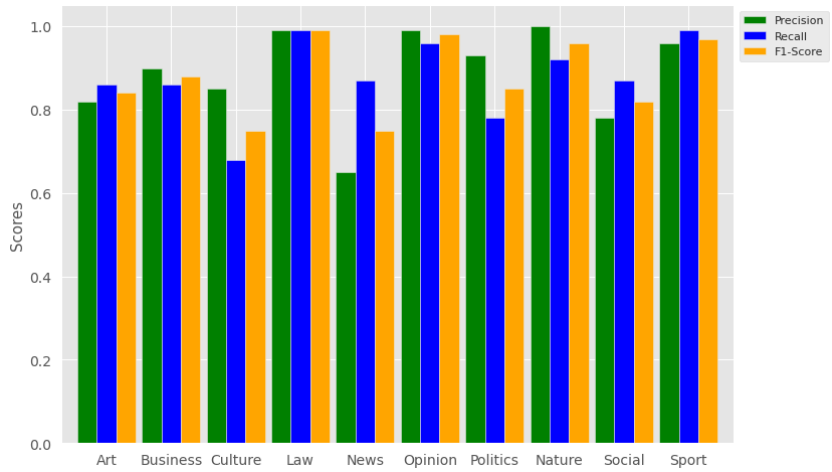

Figure 7. Comparing Performance of SVM with TF-IDF + SMOTE across Classes.

4.4. Qualitative Evaluation

This section presents a qualitative analysis of topics generated by the supervised topic classifier in terms of the topmost word features generated under each topic category.

(i) Topics Generated by Supervised Classifier: the supervised topic classifier generates the top salient words, which uniquely distinguish each topic or theme. The topics and corresponding words are presented in Table 6.

As depicted in Table 6, the top ten topic terms under each target topic are generated by the best performing classifier, that is, SVM. SVM generates the top word features learned from the corpora to distinguish the target topics using the scikit learn library built-in function of the SVM algorithm [47].

Table 6. Sample Topic-Terms Generated by SVM Topic Classifier. Asterisks indicates meaningful topic words.

Topic	Topic Words	Analysis of the Topic Words
ART	**የሙዚቃ* ሙዚየም* አርቲስት* ሙዚቃ* ፊልም* ቴአትር* ጥበብ* ዐውደ* መጽሐፍ* ኮንሰርት***/ music* museum* artist* music* film* theater* art* exhibit* book* Concert*/	These words in Asterisks are about music, museum, Artists, film, theater, arts, books. These are all talking the topic arts
BUSINESS	**ንግድና* ጠቅሰዋል ስለመሆኑ ስለመሆኑም ተጠቅሷል እንዲህ ባሻገር ያሉት አስታውቋል ተደርጎ** /business*, mentioned, about, being, about, being mentioned, so beyond, existing, announced/	Only one word is talking about the topic business.
CULTURE	**ምኒልክ* አገላለጽ ብሔረሰብ* ዳግማዊ* የቅርስ* ምዕት ዘመን* ባህል* በዓል* ባህላዊ***/Menelik * statement* national *heritage* century* culture* festival* cultural* /	These words in Asterisks is referring to the same theme or topic what is culture and history.
LAW	**አድራሻቸው ሕጉ* ይሁን ይቻላል ጸሐፊውን* ግዴታ* ጽሑፍ* ሕግጋት* ወዘተ መርሆች*** address, the law* possible, author* duty* article* laws* etc., principles*/	The words with Asterisks are under law/or justice domain.
NEWS	**አንበርብር ባለፈው በማለት አስረድተዋል* መሆኑን ምንጮች* ገልጸዋል* አንበርብር የገለፁት* አስረድተዋል***/ Anberber, explained* in the past, that, sources*, explained*, Anberber, explained*/	The words in Asterisks are usually said by the journalists in media.
OPINIONS	**ታዲያ እንገልጻለን* አድራሻቸው በኢሜይል አመለካከት* የጸሐፊውን ከአዘጋጁ ጽሑፉ የሚያንፀባርቅ* ይመስለኛል***/ So, explain*, address, email, opinion*, author, author's article, reflecting*, think*/	The words in Asterisks reflects opinions.
POLITICS	**አክሷል ጠቅላይ* ነበር ሪፖርተር የፖለቲካ* ያስረዳሉ አህመድ* ዓብይ* አድማ* የኢህአደግ*** / remarked, PM*, was, reporter, political* explains, Ahmed* Abiy*, strike*, EPRDF*/	The words in Asterisks are highly connected to politics domain
NATURE	**ሽንቁጥ* ወንዶቹ* አጥቢዎች* ወፎች* ማንይንገረው ዛፍ* ጂደብሊው እንስሳ* ባለአከርካሪዎች* ዝርያዎች***/ The squirrel*, the males*, the mammals*, the birds*, unspoken tree*, JW animal* vertebrae*, species*/	The words in Asterisks are belongs to nature and life domain
SOCIAL	**ስለዚህም ጉባኤ* የጤና* ገብረማርያም ትምህርት* በታደሰ ጤና* እንደሚሉት መልኩ በሽታ***/ therefore, conference*, health*, G. mariam, education*, renewed, health*, the form of, disease*/	The words in Asterisks are related to social activities such as meetings, education and health domain
SPORT	**ውድድር* የስፖርት* ጨዋታ* ዋንጫ* ኦሊምፒክ* እግር* ኩጫ* ስፖርት* አትሌቶች* ኳስ***/ competition*, sports*, game*, cup*, olympics*, foot*, running*, sports*, athletes*, football*/	All words are terms connected to sports domain

Most of the word features (as shown by Asterisks) are highly meaningful for distinguishing the corresponding topics. SPORT, CULTURE and ART topics are almost 100% coherent topics, as all the top ten topic words are meaningful and highly associated with the corresponding topics.

We can also describe the generated topics and the corresponding topic terms quantitatively, using an approach similar to that in [48,49].

$$A_{topics} = \frac{n\_topics}{N\_topics}\cdot, \tag{12}$$

where A_{topics} = the accuracy of topics, n_{topics} = number of relevant topics discovered and N_{topics} is the total number of predefined topics. Based on this, Equation (12), all the ten topics are relevant, that is, 10/10, which is 100% accurate.

$$A_{terms} = \frac{n_{terms}}{N_{terms}}\cdot, \tag{13}$$

where A_{terms} is the ratio of relevant terms which belong to the topic per total number of terms generated, n_{terms} = number of relevant terms in all topics and N_{terms} = total number of terms in all topics.

The accuracy of the total number of relevant terms generated by the model is given by the total number of relevant terms divided by the total number of terms generated by the model. So, based on this, Accuracy (Terms) = 63%. This shows that 63% of the generated topic word features are relevant and meaningful to distinguish the corresponding topics.

4.5. Error Analysis

The errors are detected from the results using the confusion matrix by the best classifier, that is, SVM. Using the test set samples, the confusion matrix is shown as Figure 8.

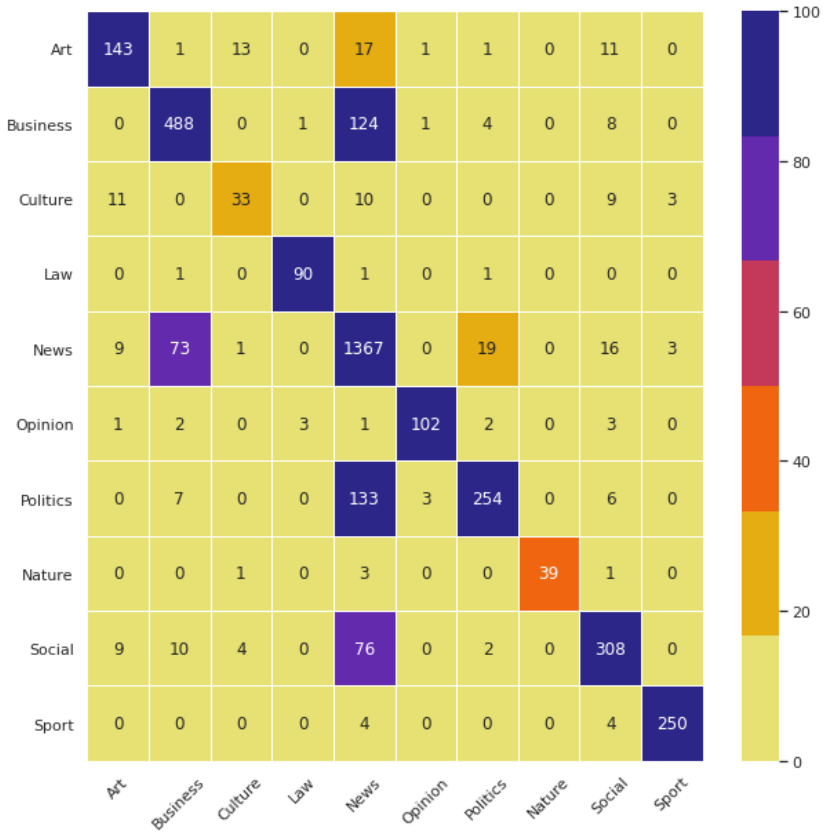

Figure 8. The confusion matrix result by SVM topic detection using TF-IDF with no application of SMOTE.

Discussion: As depicted in the non-normalized result of the confusion matrix, three topics such as business, politics and social topics, are highly confused with news topics. From the above three topics, 124, 133 and 76 samples, respectively, are wrongly predicted as news topics. One of the main reasons is that the nature of news content can be in any domain. In addition, 11 samples from the culture category are wrongly classified as the art domain, as art and culture domains are highly correlated.

In Table 7, partial contents of three wrongly predicted samples are presented. The two samples in the first and the second rows are from political and business domain, respectively, which are wrongly predicted as news domain. The reason for this error in prediction by the model is that the nature of news content is not a unique domain, that is, it can hold the contents of any domain as far as it is news. Because of this, the model treats those samples as news. In the third row, the sample from the art domain is wrongly predicted to be a culture topic. As mentioned earlier in the confusion matrix, art topics are highly associated with culture topics.

Table 7. Sample of Incorrectly Predicted Texts by SVM Topic Classifier.

Sample Text	Target Class	Predicted Class
በአስረኛው ዙር የኮንዶሚኒየም ቤቶች እጣ እድል ከደረሳቸው የአድስ አበባ ነዋሪዎች መካከል ፈጠነ ዋቅጅራ ይገኝበታል.../Among the residents of Addis Ababa who were lucky enough to win the tenth round of condominiums was Fetene Waqjira./	Politics	News
የቢራ የጅምላ ዋጋ ተመን ምርቱን ከሚመረትበት ከተማ ውጭ እንዳይወጣ እያደረገው....ከማከፋፈያ ዋጋቸው በታች እንድሸጡ የሚያስታውቁ መሆኑም ችግር እንደፈጠረም ታውቋል .../It is also known that the wholesale price of beer is keeping the product out of the city where it is produced ..../	Business	News
ከአክሱም ከተማ በስተሰሜን ምስራቅ አቅጣጫ ኪሎ ሜትር ርቀት ጥንታዊቷ ይሀ ከተማ ትገኛለች መነሻችን ከነበረው አክሱም አድዋና አድግራት የሚወስደውን ዋና ጎዳና ጎን ትተን ኪሎ ሜትር ከተጓዝን .../This ancient city is located a few miles northeast of Axum ..../	Art	Culture

5. Conclusions

In this section, we present the conclusions drawn from the experiments carried out in the proposed supervised topic detection approach. We have seen the effect of the stemmer on the performance of topic classifiers. According to the experimental results, stemming slightly improves the performance of topic classifiers.

So, light weight stemming is recommended to use for topic modeling. In addition, we compare the performance of supervised topic classifiers with different feature sets, such as TF-IDF feature sets, LDA feature sets and a combination of these two feature sets. From the experimental results on four classifiers (i.e., SVM, NB, LR, and Neural Net), the TF-IDF feature set has better discriminating power than the other feature sets. Of all the four topic classifiers, SVM performed best for topic classification of Amharic user generated texts.

We also investigate the performance of a supervised topic classifier both quantitatively and qualitatively. With the supervised topic classifier, we obtain relevant topics as they are pre-defined in the datasets. The topic words generated under each topic have also been shown to be highly relevant topic terms of the corresponding topics. However, there is a wrong prediction of samples from business, politics and social into the news topic as the contents of news can be of any domain as far as it is an event or a recent issue for journalists.

The formulated research questions are briefly answered as follows.

(1) Does LDA provide suitable feature set to discriminate Amharic user generated texts into a specific topic category? The answer to this research question is shown in Table 5 that LDA feature have the least importance in recognizing the topic of Amharic texts.

(2) Do preprocessing operations, specifically stemmers, have a positive effect on topic modeling of Amharic user generated text? As is reported in Table 5, applying a stemmer to Amharic texts has a slightly improved performance of the topic classifiers compared with classifiers using features without the application of stemming. As stemming worked well in most languages [18,19,21], it could have worked well for Amharic topic classification. One of the reasons for having a poor performance of topic modeling classifiers might be the errors in the stemming algorithm itself. The other reason is that the stemmer might be a heavy stemmer, which might remove the semantic features of the input texts. In contrast, the findings can also be accepted as there is a study revealing that most of the English stemmers have also shown a negative performance of topic modeling in [20].

(3) To what extent does the supervised topic detection approach improve topic classification? The answer to this research question is presented in Table 5, that there is a large performance difference between the best topic detector (SVM with accuracy of 88%) and the worst topic detector (i.e., NB with accuracy of 40%), which is over a 48% improvement.

(4) To what extent are the topic categories accurately predicted by the trained model? The answer to this research question is shown and discussed through a qualitative evaluation of the topics learned by the models relying on the top important topic terms in the samples. The correctness of the topic terms generated by the trained model is validated and manually confirmed that it is really under that topic. This is reported in detail in the qualitative evaluation section, Section 4.4. Figure 7 also reports the performance of the topic classifiers across topic class categories. That is, the classifier performed poorly in the news category (i.e., precision 65%, recall 87%, F1 score 75%), whereas it performed the best in the law category (i.e., precision 99%, recall 99%, F1 score 99%).

We recommend other approaches to topic modeling such as semi-supervised, guided topic modeling, and the short text topic modeling approach, which needs to be examined to categorize Amharic texts into topics in social media user generated content.

We also plan to develop topic-based sentiment classification for Amharic user generated texts. This will be more beneficial to government officials and industries for supporting their decision making. Such a system can provide customer feed back/opinion summaries under a topic.

We achieved a slight performance increase of topic classification by applying stemming. However, the increment might not be significant. So, this requires further investigation to see the effect of stemming, and might require the development of a lightweight stemmer to preserve the semantic information of the input texts.

We recommend increasing the size of the datasets for supervised topic modeling in order to improve the performance of the topic detection model. The researchers also suggest adding more topic categories to the documents collection.

Author Contributions: Conceptualization, G.N. and A.R.; methodology, G.N.; software, G.N.; validation, A.R., G.N. and S.A.; formal analysis, G.N.; investigation, G.N.; resources, G.N.; data curation, S.A.; writing—original draft preparation, G.N.; writing—review and editing, A.R.; visualization, G.N.; supervision, A.R.; project administration, G.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: It is available in zenodo data repository at doi: 10.5281/zenodo.5504175 ref [15] and the code is available at github repository: https://github.com/girmaneshir/GNTopicAmharicTexts, accessed on 20 September 2021.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

SVM	Support Vector Machine
RF	Random Forest
NB	Naive Bayesian Classifier
LR	Logistic Regression
TF-IDF	Term Frequency Inverse Document Frequency
LDA	Latent Dirichlet Allocation
LSI	Latent Semantic Indexing
NN	Neural Network
KNN	K-Nearest Neighbor Classifier
pLSA	probabilistic Latent Semantic Analysis
SMOTE	Synthetic Minority Oversampling TEchnique

References

1. Shanmugam, R. *Practical Text Analytics: Maximizing the Value of Text Data*; Anandarajan, M., Hill, C., Nolan, T., Eds.; Springer Press: Cham, Switzerland; Taylor Francis: London, UK, 2019; ISBN 978-3-319-956663-3.
2. Allahyari, M.; Pouriyeh, S.; Assefi, M.; Safaei, S.; Trippe, E.D.; Gutierrez, J.B.; Kochut, K. Text Summarization Techniques: A Brief Survey. *arXiv* **2017**, arXiv:1707.02268. Available online: https://arxiv.org/pdf/1707.02268.pdf (accessed on 9 September 2021).
3. Kowsari, K.; Jafari, M.K.; Heidarysafa, M.; Mendu, S.; Barnes, L.E.; Brown, D.E. Text Classification Algorithms: A Survey. *Information* **2019**, *10*, 150, doi:10.3390/info10040150. [CrossRef]
4. Shaukat, K.; Shaukat, U. Comment extraction using declarative crowdsourcing (CoEx Deco). In Proceedings of the 2016 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube), Quetta, Pakistan, 11–12 April 2016; pp. 74–78.
5. Claro, D.B.; Souza, M.; Castellã Xavier, C.; Oliveira, L. Multilingual Open Information Extraction: Challenges and Opportunities. *Information* **2019**, *10*, 228. [CrossRef]
6. Medhat, W.; Hassan, A.; Korashy, H. Sentiment Analysis Algorithms and Applications: A Survey. *Ain Shams Eng. J.* **2014**, *5*, 1093–1113. [CrossRef]
7. Alemneh, G.N.; Rauber, A.; Atnafu, S. Negation handling for Amharic sentiment classification. In Proceedings of the Fourth Widening Natural Language Processing Workshop, Seattle, WA, USA, 5 July 2020. [CrossRef]
8. Augustyniak, Ł.; Szymański, P.; Kajdanowicz, T.; Tuligłowicz, W. Comprehensive Study on Lexicon-based Ensemble Classification Sentiment Analysis. *Entropy* **2016**, *18*, 4. [CrossRef]
9. Alemneh, G.N.; Rauber, A.; Atnafu, S. Dictionary Based Amharic Sentiment Lexicon Generation. In Proceedings of the International Conference on Information and Communication Technology for Development for Africa, Bahir Dar, Ethiopia, 28–30 May 2019; pp. 311–326. [CrossRef]
10. Shaukat, K.; Hameed, I.A.; Luo, S.; Javed, I.; Iqbal, F.; Faisal, A.; Masood, R.; Usman, A.; Shaukat, U.; Hassan, R.; et al. Domain Specific Lexicon Generation through Sentiment Analysis. *iJET* **2020**, *15*, 9. [CrossRef]
11. Tesfaye, S.G.; Kakeba, K. *Automated Amharic Hate Speech Posts and Comments Detection Model Using Recurrent Neural Network*; Research Square: Durham, NC, USA, 2020. [CrossRef]
12. Vashistha, N.; Zubiaga, A. Online Multilingual Hate Speech Detection: Experimenting with Hindi and English Social Media. *Information* **2021**, *12*, 5. [CrossRef]
13. Deboch, K. Short Amharic Text Clustering Using Topic Modeling. Master's Thesis, Jimma University, Jimma, Ethiopia, 2020.
14. Yirdaw, E.; Ejigu, D. Topic-based Amharic Text Summarization with Probabilistic Latent Semantic Analysis. In Proceedings of the International Conference on Management of Emergent Digital EcoSystems, Bangkok, Thailand, 26–29 October 2010; pp. 8–15.
15. Neshir, G. Corpus for Amharic Topic Classification. 2021. Available online: https://zenodo.org/record/5504175#.YU3KV30RVPY (accessed on 9 September 2021). [CrossRef]
16. Hofmann, M.; Chisholm, A. *Text Mining and Visualization: Case Studies Using Open-Source Tools*; CRC Press: Boca Raton, FL, USA, 2016.
17. Gou, Z.; Huo, Z.; Liu, Y.; Yang, Y. A Method for Constructing Supervised Topic Model based on Term Frequency-Inverse Topic Frequency. *Symmetry* **2019**, *11*, 1486. [CrossRef]
18. Alhaj, Y.; Xiang, J.; Zhao, D.; Al-Qaness, M.; Abd Elaziz, M.; Dahou, A. A Study of the Effects of Stemming Strategies on Arabic Document Classification. *IEEE Access* **2019**, **7**, 32664–32671. [CrossRef]
19. Duwairi, R.; El-Orfali, M. A Study of The Effects of Preprocessing Strategies on Sentiment Analysis for Arabic Text. *J. Inf. Sci.* **2014**, *40*, 501–513. [CrossRef]
20. Schofield, A.; Mimno, D. Comparing Apples to Apple: The Effects of Stemmers on Topic Models. *Trans. Assoc. Comput.* **2016**, *4*, 287–300. [CrossRef]
21. Swapna, N.; Subhashini, P.; Rani, B. Impact of Stemming on Telugu Text Classification. *Int. J. Recent Technol.* **2019**, *8*, 2767–2769.
22. Padurariu, C.; Breaban, M. Dealing with Data Imbalance in Text Classification. *Procedia Comput. Sci.* **2019**, *159*, 736–745. [CrossRef]
23. Yan, B.; Han, G.; Sun, M.; Ye, S. A Novel Region Adaptive SMOTE Algorithm for Intrusion Detection on Imbalanced Problem. In Proceedings of the 2017 3rd IEEE International Conference On Computer And Communications (ICCC), Chengdu, China, 13–16 December 2017; pp. 1281–1286. [CrossRef]
24. Gonzalez-Cuautle, D.; Hernandez-Suarez, A.; Sanchez-Perez, G.; Toscano-Medina, L.; Portillo-Portillo, J.; Olivares-Mercado, J.; Perez-Meana, H.; Sandoval-Orozco, A. Synthetic Minority Oversampling Technique for Optimizing Classification Tasks in Botnet and Intrusion-Detection-System Datasets. *Appl. Sci.* **2020**, *10*, 794. [CrossRef]
25. Alam, T.; Shaukat, K.; Hameed, I.; Luo, S.; Sarwar, M.; Shabbir, S.; Li, J.; Khushi, M. An Investigation of Credit Card Default Prediction in The Imbalanced Datasets. *IEEE Access* **2020**, *8*, 201173–201198. [CrossRef]
26. Ah-Pine, J.; Soriano-Morales, E. A Study of Synthetic Oversampling for Twitter Imbalanced Sentiment Analysis. In Proceedings of the Workshop on Interactions Between Data Mining And Natural Language Processing (DMNLP 2016), Skopje, Macedonia, 22 September 2017.
27. Neshir, G.; Rauber, A.; Atnafu, S. Meta-Learner for Amharic Sentiment Classification. *Appl. Sci.* **2021**, *11*, 8489. [CrossRef]
28. Naili, M.; Chaibi, A.; Ghézala, H. *Arabic Topic Identification Based on Empirical Studies of Topic Models*; Revue Africaine De La Recherche En Informatique Et Mathématiques Appliquées (ARIMA): Trier, Germany, 2017; Volume 27.

29. Anoop, V.; Asharaf, S.; Deepak, P. Unsupervised Concept Hierarchy Learning: A Topic Modeling Guided Approach. *Procedia Comput. Sci.* **2016**, *89*, 386–394. [CrossRef]
30. Toubia, O.; Iyengar, G.; Bunnell, R.; Lemaire, A. Extracting Features of Entertainment Products: A Guided Latent Dirichlet Allocation Approach Informed by The Psychology of Media Consumption. *J. Mark. Res.* **2019**, *56*, 18–36. [CrossRef]
31. Li, C.; Xing, J.; Sun, A.; Ma, Z. Effective Document Labeling with very few Seed Words: A Topic Model Approach. In Proceedings of the 25th Association of Computing Machinery (ACM) International on Conference on Information and Knowledge Management, Indianapolis, IN, USA, 24–28 October 2016; pp. 85–94. [CrossRef]
32. Jagarlamudi, J.; Daumé, H., III; Udupa, R. Incorporating Lexical Priors into Topic Models. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, 23–27 April 2012; pp. 204–213.
33. Kwon, H.; Ban, H.; Jun, J.; Kim, H. Topic Modeling and Sentiment Analysis of Online review for Airlines. *Information* **2021**, *12*, 78. [CrossRef]
34. Tong, Z.; Zhang, H. A Text Mining Research-based on LDA Topic Modelling. In Proceedings of the International Conference on Computer Science, Engineering and Information Technology, Vienna, Austria, 21–22 May 2016; pp. 201–210.
35. Liu, L.; Tang, L.; Dong, W.; Yao, S.; Zhou, W. An Overview of Topic Modeling and its Current Applications in Bioinformatics. *Springerplus* **2016**, *5*, 1–22. [CrossRef] [PubMed]
36. Foulds, J.; Smyth, P. Robust Evaluation of Topic Models. In Proceedings of the Neural Information Processing System (NIPS), Stateline, NV, USA, 5–10 December 2013.
37. Korshunova, I.; Xiong, H.; Fedoryszak, M.; Theis, L. Discriminative topic modeling with logistic LDA. In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019; Volume 32, pp. 1–11.
38. Ramage, D.; Hall, D.; Nallapati, R.; Manning, C. Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–7 August 2009; pp. 248–256.
39. Inkpen, D.; Razavi, A. Topic Classification using Latent Dirichlet Allocation at Multiple Levels. *Int. J. Linguist. Comput. Appl.* **2014**, *5*, 43–55.
40. Jónsson, E.; Stolee, J. An Evaluation of Topic Modeling Techniques for Twitter. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), Beijing, China, 26–31 July 2015; pp. 489–494.
41. Chawla, N.; Bowyer, K.; Hall, L.; Kegelmeyer, W. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
42. Alemayehu, N.; Willett, P. Stemming of Amharic Words for Information Retrieval. *Lit. Linguist. Comput.* **2002**, *17*, 1–17. [CrossRef]
43. Likhitha, S.; Harish, B.; Kumar, H. A Detailed Survey on Topic Modeling for Document and Short Text Data. *Int. J. Comput. Appl.* **2019**, *178*, 1–9. [CrossRef]
44. Brownlee, J. Master Machine Learning Algorithms: Discover How They Work and Implement Them from Scratch. Available online: https://bbooks.info/b/w/5a7f34e12f2f40dc87fbfda06a584ef681bc5300/master-machine-learning-algorithms-discover-how-they-work-and-implement-them-from-scratch.pdf (accessed on 3 June 2021).
45. Llombart, O. Using Machine Learning Techniques for Sentiment Analysis. Available online: https://ddd.uab.cat/pub/tfg/2017/tfg_70824/machine-learning-techniques.pdf (accessed on 6 May 2021).
46. Ho, R. Big Data Machine Learning: Patterns for Predictive Analytics. DZone Refcardz. Available online: https://www.bizreport.com/whitepapers/big_data_machine_learning_patterns.html (accessed on 5 June 2021).
47. Scikit-Learn Machine Learning in Python. Available online: https://scikit-learn.org/stable/ (accessed on 2 June 2019).
48. Yang, T.; Torget, A.; Mihalcea, R. Topic Modeling on Historical newspapers. In Proceedings of the 5th Association for Computational Linguistics (ACL)-Human Language Technologies (HLT) Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Portland, OR, USA, 24 June 2011; pp. 96–104.
49. Resnik, P.; Armstrong, W.; Claudino, L.; Nguyen, T.; Nguyen, V.; Boyd-Graber, J. Beyond LDA: Exploring Supervised Topic Modeling for Depression-related Language in Twitter. In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Colorado, CO, USA, 5 June 2015; pp. 99–107.