

Detecting Polarizing Language in Twitter using Topic Models and ML Algorithms

Njagi Dennis Gitari^{1,2,*}, Zhang Zuping¹ and Wandabwa Herman³

¹*School of Information Science and Engineering, Central South University
Changsha, 410083, China*

²*Department of Information Technology, Jomo Kenyatta University of Science and
Technology (JKUAT), 62000, Kenya*

³*School of Information Science and Engineering, Xiamen University, Xiamen,
361005, China*

^{1,2}*gitaden2000@yahoo.com, ¹zpzhang@csu.edu.cn, ³wandabwa2004@gmail.com*

Abstract

The upsurge in the use of social media in public discourses has made it possible for social scientists to engage in emerging and interesting areas of research. Normally, public debates tend to assume polar positions along political, social or ideological lines. Generally, polarity in the language used is more of blaming the opposing group in such debates. In this paper, we investigated the detection of polarizing language in tweets in the event of a disaster. Our approach entails combining topic modeling and Machine Learning (ML) algorithms to generate topics that we consider to be polarized thereby classifying a given tweet as polar or not. Our latent Dirichlet allocation (LDA)-based model incorporates external resources in the form of a lexicon of blame-oriented words to induce the generation of polar topics. The Collapsed Gibbs sampling is used to infer new documents and to estimate the values of parameters employed in our model. We computed the log likelihood (LL) ratios using our model and two other state-of-the-art LDA-based models for evaluation. Furthermore, we compared polarized detection classification accuracy using the features extracted from polarized topics, bag of words (BOW) and part of speech (POS)-based features. Preliminary experiments returned higher overall accuracy results of 87.67% using topic-based features compared to BOW and POS-based features.

Keywords: LDA topic modeling, blame topics, ML Algorithms, multilingual sentiments

1. Introduction

The popularity of micro-blogging sites such as Twitter¹ presents an important opportunity for collaboration between ordinary people and the officialdom in sharing crucial information and opinions. Through Twitter, a government and her agencies can receive sustained and informed criticism and use it to gauge the public mood on particular events and decisions. Furthermore, the informal language used in Twitter coupled with its requirement for short posts acts as an inducement for users to react to emerging issues in real time. Yet, while blogs and social networking sites have been studied extensively in many geographical regions, these important data sources remain underutilized in African countries [1]. Furthermore, as a bilingual nation with Swahili and English recognized as national languages, online conversations in Kenya are often carried out using both languages, presenting unique challenges in processing these micro-blogging texts.

When disasters occur, such as the Mpeketoni attacks in Kenya between June 15 and June 17, 2014, public resentment and anger may be directed towards government agencies

¹ <http://twitter.com>

as well as other non-state actors. Often the public opinion is different from the official government position. As reported by Simon *et al* [1], support groups are more likely to receive positive sentiments than are the government agencies. Understanding the feelings of the public can help in reassuring them of their safety as well as help in mitigating the effect of such disasters. The use of Twitter in emergency response situations such as the Westgate Mall attack in Kenya [1] and the Japan earthquake [2] are some of the examples of collaborative information sharing where officialdom and the general public share information. Information relayed in disaster environments are quite understandably heavily laden with sentimental attitudes such as emotions, anger and helplessness. There is a tendency to allocate or pass blame in such situations either as to the cause of the disaster or to register displeasure to the kind of response undertaken to deal with the adversity.

The Merriam-Webster² dictionary defines the verb blame as to say or think that a person or thing is responsible for something bad that has happened. In a study of how linguistic cues influence liability, Fausey *et al* in [3] concludes that linguistic framing can shape how the events are construed even when they occurred similarly to a group of people. Thus, there is a sense in that the choice of words can deliberately lend certain expressions to be classified as blame or not. Intuitively, expression of blame falls within the ambit of sentiment analysis. Blame language is related to sentiment attitudes such as anger, stance, perspective and emotions. Past literature has elucidated on a number of related areas including stance taking, modeling opinion in meetings, perspectives, arguing subjectivity [4] and biased language [5]. However, to the best of our knowledge, no past work has looked at the use blame language within the realms of sentiment analysis.

Our research aims at leveraging on topic modeling techniques to present the dynamic nature of information in a disaster situation and to show the changing mood as more information unfolds as well as linking the positive and negative attitudes to the various actors in the situation. Topic models such as LDA [19] have been used widely for analyzing document collections in an unsupervised fashion. However, topic models such as LDA are only able to explain superficial and common aspects of the documents and are unsuitable for rare but important topics in the corpus. For our blame detection problem, we improve the LDA model through introducing a provision for seed information by the user. The seed words introduced aid the model in identifying the blame-related topics while at the same time maintaining the unsupervised nature of the model. Unlike supervised approaches, no training is required for either a part or the entire corpus.

The contributions of this study can be summarized as follows. First, we examine whether Twitter can be relied upon as a legitimate tool for expressing disaffection and blame in a meaningful and predictable fashion. Secondly, we use topic modeling to not only identify document's thematic areas, but also to establish its links with blame topics, dynamically, over time. Lastly, our research incorporates multi-lingual use in a micro-blogging environment, in particular the use of English and Swahili on tweets communication. The rest of the paper is organized as follows. In Section 2 we discuss related work. We then introduce our new LDA-based model and inference method in Section 3. We discuss the collection and processing of the dataset used in experiments in Section 4. In Section 5 we discuss the experiment results and conclude in Section 6.

2. Related Work

Our work relates to the use of Twitter in collaborative and disaster situations and also application of topic modeling techniques to blame-language detection. While, to the best of our knowledge, there is no specific work tackling blame detection problem, particularly in the social media, there is indeed a large body of works related to mining and analysis of opinions on Twitter. Intuitively, blame-language is subjective and therefore related to the

²<http://www.learnersdictionary.com/definition/blame>

broad area of sentiment analysis. In the past, Twitter has been used in modeling public mood and opinion [6]. Using a six-dimensional psychological instrument to represent mood features of tension, depression, anger, vigor, fatigue, and confusion, they show that tweets can be used to convey the mood of the author. To analyze public's emotional state over a sustained period they use a term-based Profile of Mood States (POMS) rating system.

User generated content (UGC) from Twitter has been used successfully in a number of web mining applications including election predictions [7], analysis of political sentiments [8] as well as detecting offensive messages [9]. While such successes are worthy, debate still ranges on the value of tweets as a dependable source of political expressions. The orthographic complexities and shortness of posts, poses serious challenges in identifying patterns for use by machine language algorithms as well as processing through NLP algorithms. It is on the basis of this that we investigate whether messages sent in Twitter can be of any value in modeling an abstract concept such as blame in user sentiments.

Twitter has been used widely in political deliberations. For sentiment analysis on political tweets, Bakliwal et al [10] employ a supervised learning approach to determine whether the sentiments towards a political party are positive, negative or neutral. The most successful model uses a combination of BOW, subjectivity lexicon and twitter specific features. In a study on the reflection of politics on Twitter, Tumasjan *et al* [7] found that the number of tweets or mentions is directly proportional to the probability of winning elections. A major concern in user generated content for political deliberations is the high number of negative sentiments in comparison to the positive ones [6]. The explanations for this scenario can be that people tend to be more reactive in case of disagreements and tend to express such frustrations more vehemently in matters that affect their lives like politics.

To model shifts in public opinion towards a candidate during the campaign period, Wang *et al* [8] created a real-time system for political tweets. Additionally, just like in our case, they addressed the use of vernacular, that is, a language different from the common language used by the majority (English in this case). To deal with languages other than English in texts, a number of translation options have been explored. Anta *et al* [11] use machine learning techniques and WEKA to process Spanish tweets. Within topic modeling, the multilingual problem has been handled through the use of multivariate normal distribution. However, the inference problem is made more difficult due to lack of conjugacy. Boyd-Graber and Rensick [12] therefore propose a tree-structure extension to the Dirichlet distribution by assuming vocabularies of all the languages can be represented using a common treelike semantic structure. Our key idea in dealing with the bilingual problem is the assumption that English and Swahili share the same semantic structure.

A number of linguistic approaches have been proposed to deal with various forms of subjective language. Some of the distinct aspects include stance or arguing subjectivity and biased language [5]. Common in blogs and other forms of UGC in social media, is the use of overt language that conveys support for certain positions. In polarized and controversial topics, metaphors and vocabularies are often used. For example, in a debate for or in support of government, a person may agree with, question or disagree with the government's official position. Recasens *et al* [5] identify two linguistic categories of biased language in the study of Wikipedia articles. Epistemological bias covers subtle forms of biases presented through linguistic cues such as hedges, factive verbs, and entailments. Framing bias is the more overt variety that uses language and metaphor to present decidedly subjective position.

Ahmed and Xing [13] extend the ideological-bias study to associate bias with certain topics. They use factored topic model to model ideological perspectives. Their model factors document collection alongside lexical variability influenced by an author's

inclination (ideological perspective), topical content and the interactions between the two. They present a multi-view topic model that represents a document as the interaction between topical and ideological dimensions. Their work is related to comparative text mining and sentiment analysis across different cultural setups. Under the cross-collection LDA (ccLDA) [14], a topic is associated with two sets of word distributions: one that is shared among all collections and one unique to the collection the document is drawn from.

An important aspect of our work is the incorporation of prior information into the unsupervised LDA process. Andrzejewski *et al* in [15] proposes the use Dirichlet Forest priors to incorporate Must Link and Cannot Link constraints into the topic models. A Must Link between a pair word type represents that the model should encourage both the words to have either high or low probability in any particular topic. A Cannot Link between a word pair indicates both the words should not have high probability in a single topic. This approach, however, requires supervision in terms of Must Link and Cannot Link information

3. Blame Detection Approach

We describe an LDA-based model for blame detection suitable for a social micro-blogging text-based dataset. Novel to our approach is the introduction of prior information that boosts the probability of the blame topics being drawn relative to the ordinary topics. Our approach is unique because at the point of initialization, the Gibbs sampler is made to “memorize” important words and ensure a certain portion of topics must be drawn based on the seed words. This ensures that at every iteration, we are assured that blame oriented topics will emerge. Besides generating blame topics, our model delivers a topic distribution that consistently models the relationship between the blame topics and the ordinary non-blame topics.

In the following, we describe the probabilistic approach that we undertake to generate the blame-detection model. We also describe the inference process, achieved through posterior inference of blame-specific parameters.

3.1. LDA-Based Topic Modeling

We extend the basic LDA model to deal with blame-specific characteristics of micro-blogging data. Like other variants of LDA, this model postulates a set of latent topics as variables with each topic corresponding to a multinomial distribution over a set of vocabulary of words. The purpose of our model, like with other LDA-based models, is to decompose a document into topics that are characterized by a multinomial distribution over words.

As a formal approach, assume a corpus collection of N documents denoted by $C = \{d_1, d_2, \dots, d_N\}$ with each document consisting of multi-set of words \bar{w} from a vocabulary σ of blame-oriented words and a set V of general topical and function words. Further consider the existence of separate collections C_1 and C_2 for blame-oriented and non-blame words respectively. The latent corpus distribution for the collection C_1 is given by a Dirichlet distribution ϕ_b while the Dirichlet distribution ϕ_a , defines the distribution for the collection C_2 . Our interest is to recover the set of topics ϕ that are simultaneously associated with the blame words and the topical words. Each topic is a multinomial distribution over the terms V and σ . We describe the generative story for the blame detection based on the plate diagram in Figure 1(right).

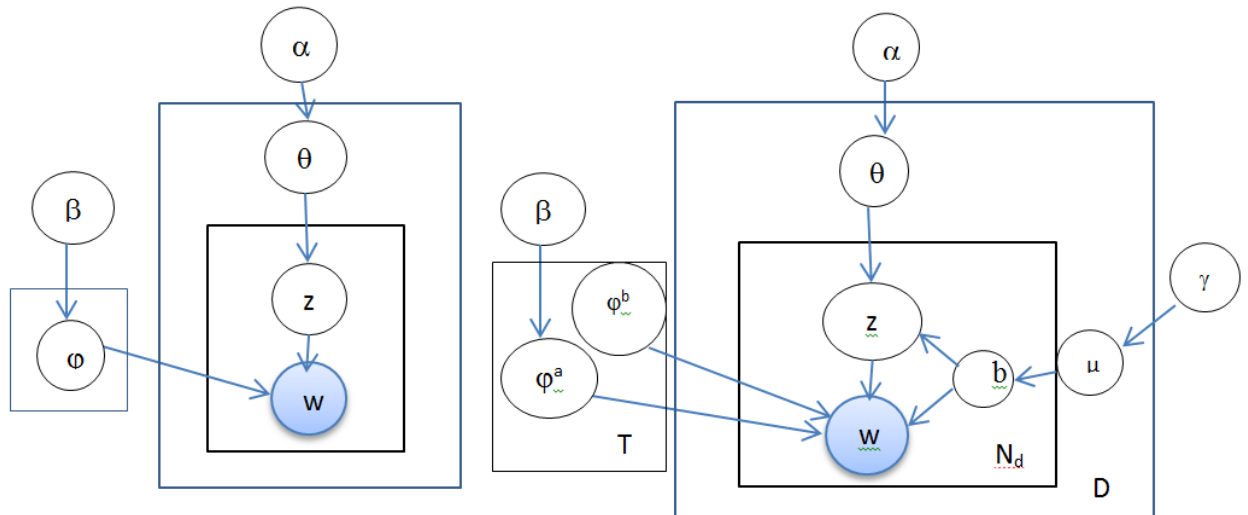


Figure 1. Left: LDA model (Griffiths).Right: proposed BLDA for Blame Detection. The Parameter B Is a Binomial Variable Representing Presence or Absence of Blame

As is the case with the LDA, the model posits that each word token (w) is jointly associated with a topic variable z , document's topic distribution (θ) and topic-word distribution variable (ϕ). The hyperparameters α and β are the prior Dirichlet distributions for the multinomial θ and ϕ . Based on the LDA, the generalized mixture model for BLDA can be represented using a joint distribution of the variables w, z, μ, θ and ϕ . Parameter μ is a document level variable accounting for the presence or absence of blame in a document. Because of the independence assumption under the Bayes, the joint distribution can be decomposed into a product of several factors as:

$$p(w, z, \theta, \phi, \mu | \alpha, \beta, \gamma) = p(\phi | \beta) \prod_{j=1}^D (p(\theta_j | \alpha) p(\mu_j | \gamma) \prod_{n=1}^N (p(w_{jn} | z_{jn}, \phi) p(z_{jn} | \theta_j))) \quad \dots(1)$$

An important part of the model is the probability of choosing a particular value of μ , given that the hyperparameters γ_0 and γ_1 are to represent the variable γ sampled from a Beta distribution. Thus we derive $P(\mu | \gamma_0, \gamma_1)$ as:

$$p(\mu | \gamma_0, \gamma_1) = c \mu^{\gamma_0 - 1} (1 - \mu)^{\gamma_1 - 1} \quad \dots(2)$$

Where C is a normalizing constant ensuring that $P(\mu | \gamma)$ sums to 1 over all the values of μ . Indeed, without the constant C the prior probability can be presented as:

$$p(\mu | \gamma_0, \gamma_1) \propto \mu^{\gamma_0 - 1} (1 - \mu)^{\gamma_1 - 1} \quad \dots(3)$$

Further, the variable μ influences the state of word token $w_{d,n}$ and the topic index $z_{d,n}$ through the word-level variable b . This is important because though a document may be potentially blame-oriented there are still many words that are either topical or functional words in it.

More formally we present the topic generation using the following algorithm:

Algorithm: Blame-Topic Generation

Input: text collection D , seed lexicon s

Output: topic index z

Begin

```
For all topics  $t \in [1, T]$  do  
  //randomly boost with seed lexicon  
  Choose blame topic  $\varphi_k^b \sim Dir(\beta_b)$   
  Choose Non-blame topic  $\varphi_k^a \sim Dir(\beta_a)$   
End for  
For each document  $d \in [1, D]$   
  Select mixture proportion  $\theta_d \sim Dir(\alpha)$   
  Sample blame proportion  $\mu_t \sim Beta(1,1)$   
    For each word  $w_i \in W$  in topic  $z$   
      //Select topic proportion  
      Sample a blame label  $b \sim beta(1, 1)$   
      If  $b_i=1$   
        Sample from blame topic  $z_{d,n} \sim mult(\varphi_k^b)$   
      Else if  $b_i=0$   
        Sample from non-blame topic  $z_{d,n} \sim mult(\varphi_k^a)$   
      End if  
    End for
```

End for


As described by the network diagram in Figure 1, the algorithm above is an extension to the basic LDA. Topic proportion parameters φ_k^a and φ_k^b determine the distribution of blame and non-blame topics respectively. For every document in a collection, topic proportion is governed by a Dirichlet distribution through parameters θ_d and μ_t for the general mixture proportion and blame proportion respectively.

4. Data Collection

Our dataset is a collection of raw tweets following criminal attacks in Mpeketoni, Kenya, where more than 60 people were killed. While, Somali-based Al-Shabaab militant group claimed responsibility, the Kenyan President affirmed that the attacks were organized by local politicians with a network to gangs. The Kenyan Opposition blamed it on the state's inability to protect its people and their property. As a result of this conflicting information, differing viewpoints emerged on Twitter in support of either the government position or the side taken by the Opposition.

We collected raw tweets using the ScraperWikiAPI³, a web-based platform for collaboratively building programs to extract and analyze public (online) data, for about 5 days from 18th to 23rd June, 2014. In our case, Scraper Wiki incorporated use of the Twitter Search, REST and Stream APIs where historical tweets (approximately 2 weeks old) could still be retrieved as well as the ability to stream tweets in real-time. This therefore made collection of tweets that were streaming since 15th June when the attack was first reported. We only collected tweets with the Hash Tag (#Mpeketoniattack)⁴, the official Twitter handle for any message directed at informing the world in any way about the attack. We stored the tweets both in Excel as well as JSON formats for later

³<https://scraperwiki.com>

⁴  <https://twitter.com/hashtag/mpeketoniattack>

processing using NLP tools. Our corpus consists of the textual message alongside other metadata such as timestamp, ‘Twitterer’s’ ID and ‘retweet’ count.

Out of the 87503 tweets, only 42418 were distinct, the rest were retweets with a significant 35828 consisting of retweets more than 10 times. The number of retweets was significantly higher than earlier studies in [7] that showed the percentage of retweets at around 40%. Almost 50% had been retweeted at least 5 times. We felt that two main reasons could possibly explain the unusually high level of retweets: 1) the tweet originated from an influential public personality; 2) content of the tweet largely represents the views of the sender [16].

Out of the 42418 tweets and based on the timestamp feature to ensure representative samples, we selected 3000 tweets for manual annotation. The main purpose of the annotated tweets is to model the use of blame language in twitter. After a general analysis of our twitter corpus, we identified three primary targets of the tweets as: Government, opposition and the terrorists group Al-Shabaab. Included in the government group are the president, cabinet secretary in charge of security and security forces in general. In the opposition, we include the political party leaders at the national as well as the regional level where the violence occurred. We asked annotators to label each tweet with both a target and a blame tag. The target tags are “gov”, “opp”, “Asb” for government, opposition and Al-Shabaab respectively, and “oth” if the target is neither of the three. For the blame tag a “bl” label blames the target while a “def” label defends the target. Table 1 shows the distribution of blame among the various targets.

Table 1. Shows the Distribution of Blame Topics among Government, Opposition, Al-Shabaab and Others

| Gov | Opp | Asb | Oth |
|-----|-----|-----|-----|
| 79% | 9% | 9% | 3% |

4.1. Preprocessing

A notable feature among Kenya micro- blogging community is the use of a combination of both English and Swahili words even in the same sentence. Thus, the first step in the preprocessing stage is to translate all the Swahili words into English. Swahili, like English language, largely, follows a subject–verb–object (SVO) sentence typology. Therefore, word tokens drawn from either of the languages are replaceable in a sentence without loss of meaning. Though a popular variant of Swahili language called Sheng is popular in Kenya, we are only concerned with grammatically correct Swahili words.

In example (1), using Google Translate, a tweet written in both Swahili and English is converted to the English only word tokens, while maintaining its original meaning.

Example (1): “Dear mr president fire ole lenku *ama ni ule ole wetu*”
 Translates to: “Dear Mr President fire ole lenku *or is then our woes*”

We perform further preprocessing steps, including removing all the duplicate instances in our dataset appearing in the form of retweets(indicated by “rt”).From word vocabulary, we also eliminate all URLs , hashtags, user mentions, punctuations and other symbols frequently found on twitter such as @ &# \$ * + , and %. To deal with orthographic nuances, all words with a character repeated three or more times were reduced to a single character. Further, we use a dictionary of slangs⁵ to convert some of the more common acronyms in our dataset. We then performed tokenization using O’Oconur’s TweetMotif tokenizer [18].

⁵<http://www.noslang.com/dictionary/>

5. Experimental Results

In this section, we provide the details about the results of our experiments and also perform evaluation of our modeling technique. Our experiments focus on discovering blame topics in tweets as well as modeling the dynamic trends as events unfolds. We begin by exploring the predictive capability of our model by comparing it with two other state-of-the-art LDA-based models. The topic features obtained from our model are then used for classification. Additionally, we also compare the results accuracy of the proposed blame topic features with other different kinds of features.

5.1. Evaluation using Log Likelihood

We evaluate the performance of our LDA-based model against related other models before using the results to perform the classification task. We rely on a traditional approach of estimating the likelihood of the unseen held-out documents given some training documents. We divide out twitter corpus into 200 documents as the training set and 40 documents as the held-out set. To estimate the values of the parameters in the held-out document, we use the 'fold-in' approach, by taking a new document and adding it in to the corpus in turn. We only run the Gibbs sampling on the words in the new document while keeping the topic assignment of the old documents the same.

Using the log likelihood drawn from our model, the LDA, and the Cross-Collection LDA (ccLDA) models, we compare predictive capabilities of the three models. To calculate the likelihood of the held-out set, we use the formula: $L(w) = \log p(w|\Phi, \alpha) = \sum_d \log p(w_d|\Phi, \alpha)$ of a set of unseen documents w_d given the topics Φ and the hyperparameter α for topic-distribution θ_d of documents d . A higher likelihood indicates a better generalization performance over the held-out document. Figure 2 below shows the comparison, of the likelihood values in different sets of topics, among our model approach and the other two models.

The LL results suggest that our model outperforms the other two in the prediction of blame-oriented topics. The prior information supplied through the seed lexicon makes the model more attuned to generalize on blame-specific topics. The figure also reveals that in all the models, the LL is highest within the first 10 topics decreasing as more topics are discovered. This can be explained by the relatively limited lexicon of blame-dedicated words making it difficult to influence the generation of thematically coherent topics.

5.2. Feature Engineering with BLDA

For feature extraction, we follow the approach proposed in [9]. However, instead of treating each tweet as an independent document, we amalgamate a series of around 20 tweets posted within a contiguous time frame using the timestamp feature for a total of 2122 unique documents. Organizing them in this way allows us to capture trending words and themes and associate them to events as they occurred. Moreover, as noted in [14], topic detection methods based on a word or n-grams co-occurrences, or any other type of statistical inference, suffer when documents are short.

Our idea is to identify the blame topics and then identify the sentiments and sentiment targets which are associated with such topics. Intuitively, topics that exonerate an entity from blame are likely to carry more positive sentiments while those that heap blame have negative sentiments. To learn a model that can simultaneously identify potential blame topics as well as blame sentiments, we place informative word priors over the word distribution in order to incorporate knowledge from external sources. In particular, we include SentiWordNet[20], a lexical resource which assigns a triplet of numerical scores for positivity (PosScore), negativity (NegScore) and objectivity as $(1 - (\text{PosScore} + \text{NegScore}))$ to each of the vocabulary terms in our corpus. Words with a

negative orientation have a greater chance to be generated than positive or neutral terms. This prior information is supplied at the initialization stage of the Gibbs sampling.

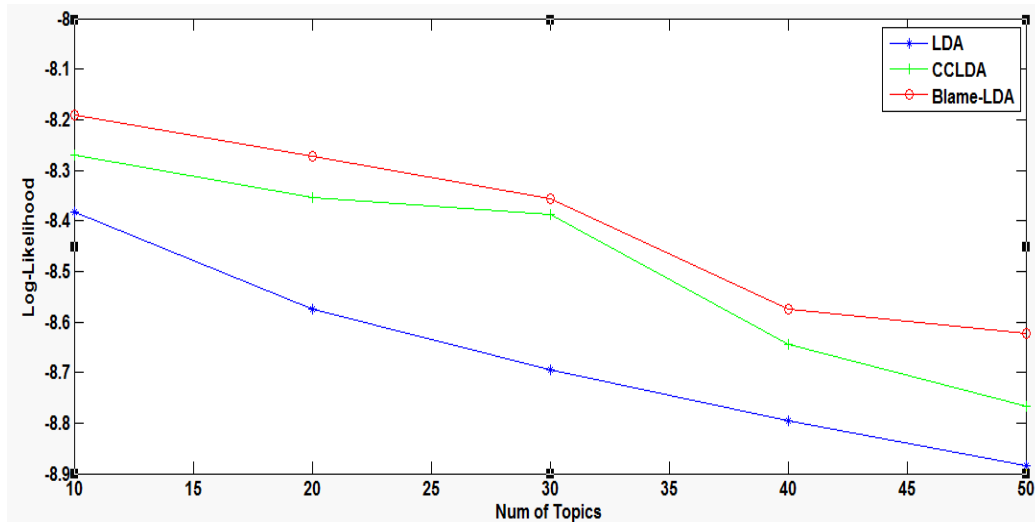


Figure 2. Comparison of the Log-Likelihood with the 3 Models

We adopt MALLET LDA model implementation⁶, however, we tweak the model to enable prior parameterization via the Gibbs sampling. For the input k for the expected number of topics, we experiment with different values. We also explore with different number of iterations, since under LDA each series of iteration potentially randomizes a different set of words. For lexicon building using the LDA features, we use 50 topics and top 20 words for each topic. Table 1 below is an example of topics generated from our corpus.

Table 2. Examples of Blame Topics Associated With the Government and the Opposition

| | Government | | Opposition | |
|-----------|--------------|----------------|-------------|----------------|
| blame | shame | attacks | lenku | Shabab |
| shabab | surveillance | Raila | raila | Twitter |
| game | confirms | responsibility | politics | Account |
| stop | negligence | breaking | red | Responsibility |
| politics | allegations | jubilee | crossed | Suspects |
| survivors | delay | Failed | blame | Cord |
| analysis | reaction | Resign | politicians | Operating |
| leaders | nis | following | political | Omondi |
| issues | Kenyatta | Nis | line | Arrested |
| people | ole | Cord | security | Claim |

5.3. Classification using LibLinear SVM

In the past research, support vector machine (SVM) has been hailed as one of the most effective classifier in text categorization applications. LibLinear SVM is especially suited for twitter data due to high number of instances (tweets) and word features as well as well

⁶<http://mallet.cs.umass.edu>.

as sparsity of the data [21]. We classify as a two class problem of blame and not blame using the SVM binary linear classifier. Given the instance-label pairs (x_i, y_i) where x_i is a tweet instance and label $y_i \in \{+1, -1\}$ with +1 representing the presence of blame and -1 a non-blame class. Using the equation: $\min_w \frac{1}{2} w^T w + c \sum_{i=1}^l \mathcal{E}(w; x_i, y_i)$, we solve the unconstrained optimization problem with the objective function $\mathcal{E}(w; x_i, y_i)$ where $C > 0$ is a penalty parameter for an incorrect classification and w is a support vector.

In our case we optimize the utility function: $\max(1 - y_i w^T x_i, 0)$. We compare the results of LibLinear SVM with two other ML algorithms, the Naïve Bayes and multinomial Naïve Bayes.

Our main task is to identify the most appropriate features for blame detection. We begin by using the 3000 labeled examples, described in section 4 above, to create model classifiers with each of the three ML algorithms. Then, we apply the model to learn the unlabeled examples. With the now fully annotated tweets, we extracted different set of features to be used as feature vectors. We experimented with different varieties of unigram word features including all bag-of-words (BOW) features, POS-tagged noun and adjective-based features. For the BOW features we used a vector of term weight values, tf-idf, of a term occurring in the tweet. The Stanford tagger was used to learn the POS for each word in a tweet. To test the efficacy of using LDA-based features, we build model classifiers using topic-based lexicon created from section 4.1 above. Because of the imbalanced dataset that favors the blame class, we leverage on the SMOTE feature of WEKA to create a feature vector that creates a balance between the blame and non-blame class.

We further experimented with a subset of features using the information gain (IG) algorithm and the ranker function. Table 3 shows accuracy results for 200 most discriminating features under different classifiers with a 10-fold cross-validation. A graphical representation of Table 3 is presented in Figure 4.

Table 3. Blame Classification Accuracy Results with a Different Set of Features and ML Algorithms. LDA-Topic Features Return the Highest Accuracy

| | Naive Bayes | LibLinear SVM | Naïve Bayes Multinomial |
|----------------|--------------|---------------|----------------------------|
| BOW | 77.75 | 80.80 | 85.97 |
| POS-Nouns | 84.08 | 85.17 | 85.98 |
| POS-adjectives | 74.25 | 76.83 | 78.92 |
| BLDA-Topic | 87.23 | 89.67 | 86.12 |

Interestingly, noun-tagged features were more effective in blame detection than POS-tagged adjectives. We report a significant association of typical blame words with names of personalities. This indicates that blame language often includes the subject of blame. Among the top mentioned words, is Mr. Lenku the cabinet secretary in charge of security and the terror group ‘Shabbab’. This suggests that blame targets are important features in determining blame. We observe that features induced using statistical topic modeling is more discriminating in blame classification than all the other features. The semantic association between named entities such as names of individuals and institutions and sentiment features induced through word prior contributed to the positive results.

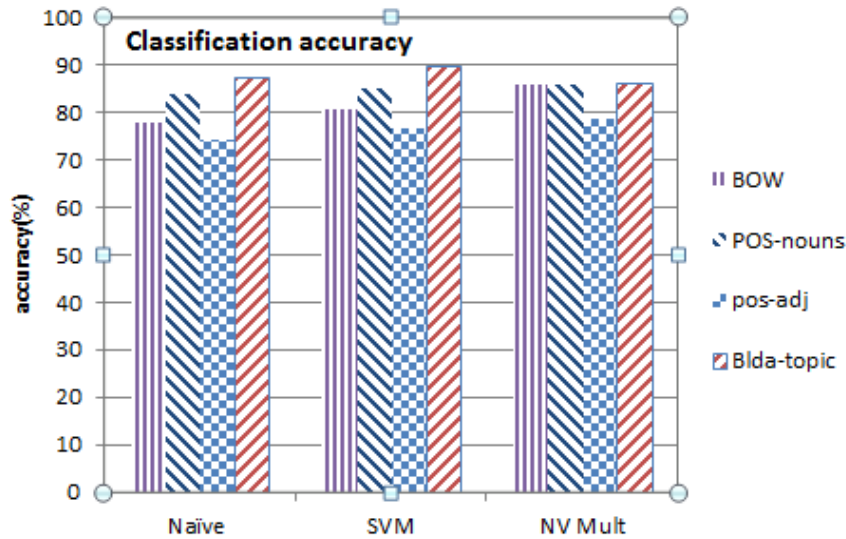


Figure 4. Comparison of the Accuracy of Classification with Different Kind of Features

6. Conclusion

We modified a topic modeling approach based on LDA and applied the same in detection of blame in polar opinions in short texts in form of Tweets. Our model was able to discriminate blame words from non-blame ones as well as track the blame usage trends over time. In evaluating the performance of our approach, we compared the predictive capability of our model with two other LDA-based approaches using log likelihood ratio. Furthermore, on the classification task, topical features derived through the model had a higher accuracy compared with features induced through machine learning algorithms.

However, a major drawback of our method reminiscent of all other topic-based approaches is reduced log likelihood as the number of topics increase. This can be explained by the abstractness of the blame-detection problem making it increasingly difficult to supply enough words to build correlations around the concept of blame. Furthermore, there are inherent limitations in using tweets for topic modeling tasks due to shortness of the messages and other orthographic nuances.

A number of potential extensions to our work can be considered for future work. Though the model has been developed specifically for blame detection, we believe a similar approach can be applied to other domains with similar dichotomous characteristics as well as on different sets of data and not just short texts such as tweets. Blog data as well as Facebook posts may be the dataset source in future. Furthermore, prior information can be applied not only at the word level but also at the topic level to improve topic-document distribution. This way, a multiple classification with 3 or more labels can be considered. This study can also be extended to detect blame in languages other than English using deep learning techniques especially at character level.

References

- [1] T. Simon, A. Goldberg, L. Aharonson-Daniel, D. Leykin and B. Adini, "Twitter in the Cross Fire—The Use of Social Media in the Westgate Mall Terror Attack in Kenya", *PLoS ONE* 9(8): e104136. doi:10.1371/journal.pone.0104136, (2014).
- [2] T. Sakaki, M. Okazaki and Y. Matsuo, "Earthquake shakes Twitter users: realtime event detection by social sensors", *International World Wide Web Conference*, (2010).
- [3] C.M. Fausey and L. Boroditsky, "Subtle linguistic cues influence perceived blame and financial liability". *Psychonomic Bulletin & Review*, vol. 17, no. 5, (2010), pp. 644-650.
- [4] W. Lin, T. Wilson, J. Wiebe and A. Hauptmann, "Which side are you on?: identifying perspectives at the document and sentence levels", In *CoNLL, ACL*, (2006), pp 109-116.
- [5] M. Recasens, C. Danescu-Niculescu-Mizil and D. Jurafsky, "Linguistic Models for Analyzing and Detecting Biased Language", *Proceedings of ACL*, (2013).
- [6] P. Alexander and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", In *Proceedings of LREC*, (2010).
- [7] A. Tumasjan, T.O. Sprenger, P.G. Sandner and I.M. Welp, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment", In *Proceedings of the International Conference on Weblogs and Social Media*, (2010).
- [8] H. Wang, D. Can, A. Kazemzadeh, F. Bar and S. Narayanan, "A system for realtime Twitter sentiment analysis of 2012 U.S. presidential election cycle", In *ACL (System Demonstrations)*, (2012).
- [9] G. Xiang, B. Fan, L. Wang, J. I. Hong and C. P. Rosé, "Detecting offensive tweets via topical feature discovery over a large scale twitter corpus", *CIKM*, (2012), pp. 1980-1984.
- [10] A. Bakliwal, J. Foster, J. V. Puil, R. O'Brien, L. Tounsi and M. Hughes, "Sentiment Analysis of Political Tweets: Towards an Accurate Classifier", In *Proceedings of the Workshop on Language in Social Media (LASM 2013)*, Association for Computational Linguistics, (2013), pp. 49–58.
- [11] A.F. Anta, L.N. Chiroque, P. Morere and A. Santos, "Sentiment Analysis and Topic Detection of Spanish Tweets: A Comparative Study of NLP Techniques", (2013), pp 45-52.
- [12] J. Boyd-Graber, P. Resnik, "Holistic Sentiment Analysis Across Languages: Multilingual Supervised Latent Dirichlet Allocation", *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, MIT, Massachusetts, USA, (2010), pp.45–55.
- [13] A. Ahmed, E.P. Xing, "Staying Informed: Supervised and Semi-Supervised Multi-view Topical Analysis of Ideological Perspective", *Conference on Empirical Methods in Natural Language Processing*, (2010).
- [14] M. J. Paul and R. Girju, "Cross-Cultural Analysis of Blogs and Forums with Mixed-Collection Topic Models", *Conference on Empirical Methods in Natural Language Processing*, (2009).
- [15] D. Andrzejewski, X. Zhu and M. Craven, "Incorporating domain knowledge into topic modeling via dirichlet forest priors", In *Proceedings of the 26th Conference on Machine Learning*, New York, NY, USA. ACM, (2009), pp. 25–32.
- [16] M. Cha, H. Haddadi, F. Benevenuto and K. P. Gummadi, "Measuring User Influence in Twitter: The Million Follower Fallacy", In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, (2010), pp. 10-17.
- [17] L.M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Göker, I. Kompatsiaris and A. Jaimes, "Sensing trending topics in Twitter", *IEEE Transactions on Multimedia*, (2013), pp 1268-1282.
- [18] B. O'Connor, M. Krieger and D. Ahn, "TweetMotif: Exploratory search and topic summarization for Twitter", In *Proc. of ICWSM (demotrack)*, (2010).
- [19] D. Blei, A. Ng and M.I. Jordan, "Latent Dirichlet allocation", *Journal of Machine Learning Research*, (2003).
- [20] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang and C.-J. Lin, "LIBLINEAR: A library for large linear classification *Journal of Machine Learning Research*", (2008), pp. 1871-1874.
- [21] D. A. Shamma, L. Kennedy and E. F. Churchill, "Peaks and persistence: Modeling the Shape of Microblog Conversations", in *CSCW: ACM Conference on Computer Supported Cooperative Work*. New York, NY, USA: ACM, (2011), pp. 355–358.