Knowledge-Based Systems 228 (2021) 107249

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Multi-interest semantic changes over time in short-text microblogs

Herman M. Wandabwa [a,*], M. Asif Naeem [b], Farhaan Mirza [a], Russel Pears [a]

[a] School of Engineering, Computer & Mathematical Sciences, Auckland University of Technology, New Zealand
[b] Department of Computer Science, National University of Computer & Emerging Sciences (NUCES), Islamabad, Pakistan

ARTICLE INFO

Article history:
Received 28 November 2020
Received in revised form 17 June 2021
Accepted 19 June 2021
Available online 7 July 2021

Keywords:
User profiling
Text mining
Neural Networks
Information retrieval
Short-text microblogs

ABSTRACT

Consumption of content in short-text microblogs is necessitated to a large extent by individual users and their friendship network interests. Based on the dynamism in the data throughput on such platforms, e.g., Twitter, prevailing conditions are bound to determine the nature of consumed or disseminated content. Therefore, semantic interests differ over time even for individual users. Detecting this semantic change over time is integral in mapping user profiles over a time period, especially in microblogs where only the extrinsic user profile identifiers provide metadata that seldom evolve. This is vital in serving relevant third-party content as well as in the computation of topical interest variations over time. In essence, current, and relevant topics of interest to a user on such a platform may not be representative of the same users' interests a few months later. In our quest to identify the most user-representative interests at any given time, each topical term was modelled as the inner product between word embeddings and a time-based embedding representation of assigned topics at varied time periods. The model was fitted onto tweets as time-series documents. To validate the model, changes in the extracted user-representative interests over time were semantically weighed against a mirrored, time-variant dataset. Interest weights across the time-variant datasets were computed and validated in five sub-topics for a period spanning two and a half years. Linearity in the relationships between the test and validation sets could be identified, more so in emerging topics. A Pearson correlation coefficient as high as 0.871 was achieved in interest change verification over the tested period.

1. Introduction

Currently, online content propagation has been proliferated by the surge in citizen journalism. This is partly attributed to the increase in the number of devices e.g. mobile phones, availability of internet as well as with the emergence of many social microblogging platforms like Twitter. Twitter as a short-text microblog has been instrumental in sharing near to real-time data in form of text, videos, hyperlinks and images. In essence, the number of disseminated tweets average about 500 billion per year, which roughly translates to approximately 6000 tweets per second.[1] On the platform, users are able to re-share the disseminated tweets, in form of "retweets", "comment" on and/or "like" the original tweet.

Generally, tweeters[2] consume content on the platform based on their prevailing interests at the time. For example, in times of political campaigns, many demographically relevant tweeters are likely to express interest in political content. However, this interest is likely to decay over time when the political season is over. The same scenario could be replicated in a sports season where support of teams fades as the season winds up. Modelling and extraction of such user dissemination patterns and representative interests is a challenging task especially for legacy systems. This is attributed to two factors *(i) The data volatility factoring its dissemination throughput (ii) Time-based variations in the nature of topics of interest.*

The ability to extract and present time-sensitive and accurate user representative profiles from such evolving short texts is important in recommender systems design research. The design goal of recommender systems on such platforms is to personalize both the third-party content, and the user identification process. This in turn presents the most relevant users as follower–followee suggestions, as well as delivery of more personalized third-party content for the users. This personalization process is based on the extrinsic interests from the disseminated content.

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding author.
E-mail address: herman.wandabwa@aut.ac.nz (H.M. Wandabwa).

[1] http://www.internetlivestats.com/twitter-statistics/.

[2] *user who posts tweets on the Twitter online messaging service.*

https://doi.org/10.1016/j.knosys.2021.107249
0950-7051/

The assumption in the design of the framework is that the semantics of disseminated content change over time for individual users. Therefore, we present a framework that is capable of discerning semantic user interests of short-text microblog users. This is challenging in short-text microblogs, as the level of expressivity is not always exhaustive due to character limitations per document/tweet. The text variable in a tweet's metadata is limited to 280 characters though realistically, the average tweet length is much shorter. This is in addition to factors relating to throughput and content decay and gain over time on such platforms.

In the quest to design such a framework, there was need to first extract topical representations of the data at specific periods. Since documents were short texts, vector representations worked well in discerning their semantic relevance. Each token in the pre-processed tweet was vectorized and modelled as the inner product between the vocabulary, and topic embeddings across specific periods. Topics of interest at each period were captured with the overall semantic representation being user interest weights across the dataset collection period. This way, semantic divergence across several topics of interest were accurately represented.

With the personalizing goal, there is need to develop a framework that is able to capture user-representative interests on such platforms, but with a factor of time. Such a framework has the ability to present semantic profiles to third-party content curators from an informative point of view on how user interests evolved with time. This can serve specific purposes related to the generation of profile information for users of interest in certain topics at given times. Such interest patterns can then be used in recommendations of time specific content, as well as related follower–followee networks. In the development of this framework, the following research questions are addressed:-

- Is it possible to extract user-representative interests in time-series based streaming short texts for profiling?
- Are the extracted patterns in the short-text sufficient in making time-dependent user/content recommendations/predictions for generalized short-text content disseminators?

This work is inspired by research in the computation of the degree of interest in selected topics as well as multi-interest user profiling in short-text microblogs [1,2]. In follow-back recommendations, user-representative interests recommendations were generated among short-text microblog users with shared interests. The social *theory of homophily* was applied in validating the semantic correlations among the users [3]. In multi-interest profiling, the quantification of user interests across topics was computed by generating a *responsibility matrix* across users depicting their interest levels across the topics. Algorithmically, Expectation Maximization (EM) and Gaussian Mixed Models (GMM) were applied over the vector representations to extract soft clusters [4]. Furthermore, the semantic distance to the clusters per user aggregated vector representations was assumed to be the interest level in the topical cluster.

The below points highlight the novelty in this research. In addition, we have emphasized how this work differs from other works at the end of Section 2.

1. The combination of word and topical embeddings as a time-variational distributional model differs from other works in the user profiling domain. Conventional modelling in related works in this domain, point to either word or topical representation models and not an amalgamation of the two.
2. In our approach, time sensitivity is definitive in the representation of the generated user-representative profiles. Inferencing topics over word embeddings where topical vectors are generated per mini-batch of documents, per timestamp, highlighting the dynamicity in identification of representative user interests in short texts. This time-sensitive approach differs from keyword, concept and hybrid profiling approaches that make use of external knowledge-bases in extraction of such interests. Vocabulary sparsity, volume, variety and velocity in short texts dissemination makes the above profiling approaches insufficient in this scenario.

In addition to the above, the following contributions were made in the paper: -

1. Formulation of semantically representative user profiling framework that considers the disseminated content as time-based entities.
2. Each word in the framework is modelled as a categorical distribution of word embeddings and a time-based representation of the word's assigned topic.
3. The model and data ingestion framework is tested on a generic set of tweets geolocated to an area over time. This ensured accuracy in validation of the end results using a true class dataset over certain periods in the collection. User-representative interests over a generic test set were computed by the methodology that qualitatively outperformed the other approaches across a range of measures.
4. Quantitatively, semantic weights between the control and test sets in five sub-topics across ten quarters were measured depicting their semantic correlations. Linearity in the correlations across the timestamps indicated the validity of our modelling approach.

The rest of the paper is organized as follows. Section 2 summarizes the background and related literature of our study. Our approach is described in Section 3. The experimental framework is presented in Section 4, with results discussed in Section 5. Discussion and application areas are elicited in Section 6.

2. Related works

Microblog user's content consumption patterns vary over time. This is mostly influenced by events at the specific timeframes, more so in short-text microblogs like Twitter. Designing a framework that is able to incorporate these semantic changes over time in user-representative profiling is pertinent. There are two approaches in the formulation of user profiles especially for use in recommender systems. The design either follows the *behavioural* [5] or *structural* [6] patterns in the modelling process. In structural modelling, extracted features normally stereotype groups of users. This means that the formulated user profiles are structurally representative of the user or groups of users [6] just like in the similarity-based algorithms [7,8]. On the other hand, behavioural modelling identifies observable parameters that reflect behaviour patterns in users. They can be probabilistic [5,9].

2.1. User profiles construction

Construction of a user's taste profile entails extracting the user's representative model which is largely dependent on the specific user's or group's interests. In short-text microblogs, users to some extent define their initial profile interests extrinsically when creating accounts. Normally, users are expected to select from a list of pre-determined interests at the time, in curation

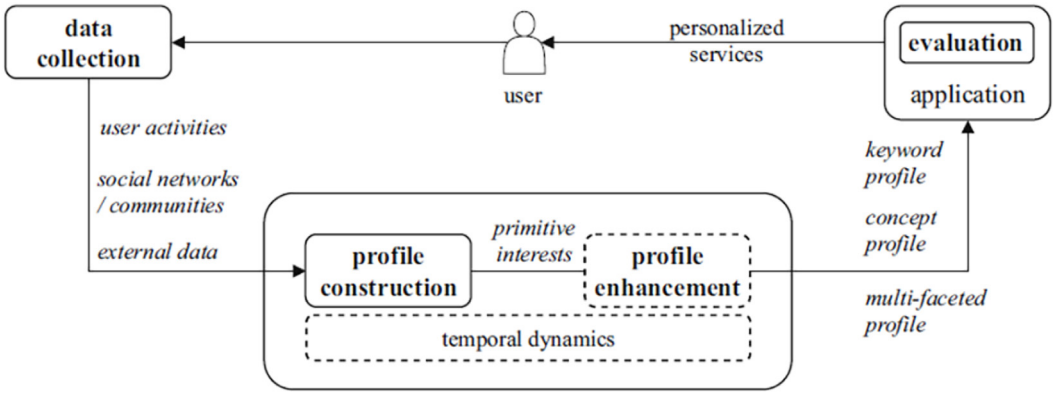

Fig. 1. User profiling process [10].

of third-party content, bound to be served to them over time. However, such interest(s) are supposed to be dynamic. In addition, the selected interest may not be of exclusive importance to users as they may for example never tweet anything semantically relevant to their declared interests. This explains why extraction of interests from the disseminated content was of significance in this work.

Therefore, a user model is a data structure that defines the users' characteristics [10]. This could be in terms of the friendship network or the data consumption patterns. Fig. 1 presents a sample framework in the user modelling and profiling process.

In Fig. 1, the profiling process involves (i) data collection in form of user activities, social networks or external data from ontologies and (ii) primitive interests extraction. The extraction is either done directly from the dataset or can be enhanced via external Knowledge Bases (KBs). The output is either a keyword-based or concept-based profile which can then be evaluated for use by recommender systems.

2.2. User profiles representation

Different approaches have been used in the representation of user taste profiles in microblogs. User taste profiles in Online Social Networks(OSNs) are broadly categorized into (i) keyword based (ii) concept based and (iii) multi-faceted based/hybrid profiles that were introduced by Gauch et al. [11].

2.2.1. Keyword-based profiles:

In the keywords based approach, *keywords* or *groups of keywords* are used in the representation of user interests. Weighting of each keyword is computed to infer the importance of the keyword in the profiling process. Techniques such as the Term Frequency Inverse Document Frequency (TFIDF) from information retrieval suffice in the generation of keyword based profiles [12].

Vectors of weighted keywords from tweets and user lists membership descriptions were used in formulating taste profiles [13,14]. Hashtags[3] in tweets have also been modelled as user interests thus were deemed representative taste profiles [15–17]. In addition, extracted keywords in tweets were also leveraged to represent taste profiles [18,19].

2.2.2. Concept profiles:

With limitations posed by keyword-based profiles, researchers proposed a *concept-based* approach to address the shortcomings. The advantage of using concepts from KBs e.g. Wikipedia is that they provided background knowledge for better extraction of concepts. Concepts have been used for varied purposes in user modelling. Their usage as KBs range from simple to complex taxonomies as presented by authors in [20–23].

Extraction of hierarchical semantic concepts from tweets for better profiling has also been experimented. Here, interests were presented as a hierarchical interest graph [24,25]. Tommaso et al. [26] proposed Wiki-mid that made use of Wikipedia in deriving better user interests in multilingual tweets. The authors used services such as Spotify to extract user preferences. Zheng et al. [27] came up with the Hierarchical Interest Overlapping Community (HIOC). In the approach, the authors computed user profiles relationships and further presented them as a personalized recommendation model.

2.2.3. Hybrid profiles:

Fusion of several aspects of interest about a target user in the modelling process results in a hybrid user profile. The assumption in this process is that the different aspects complement each other and improve on the extracted profile. Hybrid profiling in detection of fake news is one area that has been explored [28]. The authors evaluated the credibility of information sources on Twitter using node2vec where features from twitter followers/followees graph were extracted. In the approach, both user characteristics and their social graph were considered in profiling of fake information sources.

A fusion of five personality traits and dynamic interests on Twitter have also been leveraged in the profiling of users via a graph-based representation model [29]. A dynamic representation of user interests and relaxation of the bag-of-words assumption was adopted in the modelling process. Identification of malicious profiles on Twitter is another area where a mixture of important features are used to identifying such users. Sahoo et al. [30] made use of Petri net structure to analyse user profiles and extract features to train classifiers for prediction of malicious and legitimate users. Research in the evolution of user interests and subsequent profiling over time in short-text microblogs has also been carried out albeit to a small scale. Jiang et al. [31] developed a framework for extraction of user interest changes over time on Twitter. The authors incorporated external knowledge sources to hierarchically generate more representative user interests. On the other hand, Zhu et al. [32] made use of a hierarchical tree structure factoring in interest decay over time in extracting user interests. Cami et al. [33] modelled user preferences using a Dirichlet Process Mixture Model where the authors considered user interactions to construct an evolving Bayesian non-parametric framework. The model performed well in the prediction of user preferences and behaviour. In addition, a temporal preference model for detection of changes in the user network structure based on node centrality change events on

[3] https://en.wikipedia.org/wiki/Hashtag.

Twitter and Jam social music datasets was also proposed [34]. Stai et al. [35] explored the temporal dynamics of topic-specific information spread in Twitter. The assumption was that each topic corresponded to a hashtag in the deduction of time-varying infection rates.

Regarding topical modelling in the user profiling realm, Wu et al. [36] incorporated an LDA model into density-based spatial clustering of applications with noise (DBSCAN). The model was used to quantitatively analyse the temporal evolution of topics. This was in relation to topics about natural disasters. In the same way, Serban et al. [37] proposed the SENTINEL for disease surveillance in tweets and news feeds. A Convolutional Neural Network (CNN) and a Long Short Term Memory Network (LSTM) to leverage unlabelled data for classification were proposed. The disadvantage of this approach is that it involved labelling tweets which is not feasible in a streaming temporal setup.

Liang et al. [38] made use of dynamic embeddings in profiling users on Twitter. The authors proposed a Dynamic User and Word Embedding Model (DUWE) and a Streaming Keyword Diversification Model (SKDM). DUWE in this approach tracks semantic representations of users and words over time. SKDM retrieves top-K relevant and diversified keywords in profiling user dynamic interests. Hybrid models combining Long Short Term Memory (LSTM) and Convolutional Neural Network (CNN) architectures to capture local and global features of sentences and documents have also been applied in contextualizing short and conventional text [39]. In addition, Ranjan et al. [40] studied similarities between brachygraphy and microtext in relation to social media language, more so with tweets. Similarly, Satapathy et al. [41] proposed PhonSenticNet, a concept-based lexicon that uses phonetic features to normalize out-of-vocabulary concepts to in-vocabulary ones. Classification accuracy improved in addition to the reduction in polarity detection time in the test set. Approaches deployed to measure topical sparsity across diverse set of datasets were also of interest [42,43]. However, the same approaches need to be modified to achieve the same objective as in this research.

As mentioned in Section 1, our approach differs from the above works and, especially with works by Liang et al. [38] in the utilization of word embeddings. The dynamicity in their modelling happens at a user level where a user and, representative word representations are embedded. The main idea in their approach was to map semantic similarities between users and words. Diversification of the words was then utilized to characterize user profiles over time. In our approach, dynamicity is at topical level per timestamp where each term is modelled as the inner product between word and topical embeddings, providing better generalization of the generated topics. The capture of topical embeddings at each period is probabilistic thus term weights over the vectorized set of topics are probability values. This approach is different from the other works in Sections 2.2.1, 2.2.2, 2.2.3 as keywords, augmentation from external sources and a mixture of both approaches respectively, were not utilized in the profiling process. The extracted time-based topical representations captured time sensitive interest's divergence.

3. Our approach

In modelling evolving user interests, the proposed framework encompasses several processes related to the generation of *topical interests, word embeddings* and a *time-variant distributional model*. The topical interests are distributed over word representations as time variational topics in the model. Time variations are user-defined as the dataset is time-series documents. Latent Dirichlet Allocation (LDA) is used to generate the topical information at each timestamp [44]. Summaries of the most optimal K topics are generated through a discrete probability distribution over terms. The topic model output is vectorized in the time-based distributional model as described in Section 3.3. This mitigated the vocabulary sparsity problem with LDA [45].

The core processes and their correlations in the proposed framework are discussed below: -

1. **Dataset Collection** — A generic set of tweets were collected via Twitter's search API[4] for pre-processing and further modelling.
2. **Short-Text Modelling** — Inputs in the modelling process are word tokens from the preprocessed tweets. Timestamp information is extracted from the tokenized tweet. Neural-network representations via *FastText* [46], *Word2Vec* [47] and *Glove* [48] were utilized in vectorizing the tokenized tweets. For comparative purposes, two baselines i.e. LDA [44] and Twitter-LDA [49] were also used in the computation of topical qualities as baselines. Other variants such as Latent Semantic Analysis (LSA) [50] and lda2vec [51] exist in the topic modelling domain. However, comparative inaccuracies between LSA and LDA for example, made LDA a preference. On the other hand, lda2vec uses Word2Vec in its vectorization. However, Word2Vec models words atomically which makes it insufficient in tweets that are often misspelled. There were variations in input parameters to these models as described in Section 4.1.
3. **Topics over Word Embeddings Inferencing** — For each term in the test set, a distribution model of topical representative words over word embeddings in the vectorized set was first computed. The output was a product of topic and word embeddings at each period. This way, better generalization of topics was achieved, ultimately allowing for smooth variations of topics over the specified period in the dataset.
4. **Multi-interest Semantic Changes** — At each timestamp, word weight probabilities were computed. The probabilities represented the word semantic weights over the vectorized set topics. Therefore, the assumption is that the generated topics at each timestamp are the representative topics of interest at a specific time. The interest changes are captured by weighting interest keywords in the topics across the specified time periods. Linearity in the semantic change indicated topical gain while the reverse represented topical decay.

As described above, topical diversity and capture of terms that are representative of those topics at different timestamps encompasses a few processes. Latent Dirichlet Allocation (LDA) is applied in computing the topical representations over vectors. Vector representations in the distributional model, helped capture topical semantics better. This mitigated LDA's vocabulary sparsity problem [45]. A brief description of the LDA modelling process follows in Section 3.1.

3.1. Latent Dirichlet allocation (LDA)

In the proposed approach, LDA is used to model topical interests in the distributional model at each period. It is an unsupervised technique in the discovery of knowledge in form of coherent topics in text where tweets in our instance are represented as a bag-of-words [44]. With LDA, a summary of pre-set topics is computed through a discrete probability distribution over words. A per-document distribution over the generated topics is then

[4] https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.

Table 1
Topic Diversity (TD), Topic Coherence (TC) and Topic Quality (TQ) values across five vector representation techniques and two topic modelling baselines. Criteria for the choice of these measures is here [52].

Model	Size	TD	TC	TQ
FastText Skip-Gram	100	**0.88**	0.7821	**0.6882**
Word2vec Skip-Gram	100	0.8	0.682	0.5456
FastText-CBOW	100	0.81	**0.847**	0.6861
Word2Vec-CBOW	200	0.86	0.73	0.6278
Glove	300	0.865	0.701	0.6064
LDA Baseline	10 topics	0.8	0.3833	0.3066
Twitter-LDA Baseline	10 topics	0.83	0.5545	0.4602

inferred where each tweet is a document. LDA considers k topics, each of which is a distribution over the vocabulary in the corpus. Furthermore, there is consideration of a topic proportion's vector θ_x for each tweet or document x in a corpus of X documents with V distinct terms. Let $b_{xn} \in 1, \ldots, V$ denote the nth word in the xth document. In the topic generation process, each term v is assigned to a topic k with a probability θ_{xk}. LDA weaknesses, more so in dealing with sparse data [45] are mitigated in this approach, by the use of topical vectors as opposed to the set of terms, better computing the semantic relevance in topical interests per mini-batch.

3.2. Vector representations:

Vector representations are integral in extraction of semantic knowledge from both short and conventional texts. In the selection of the best vector representation framework, a few neural language modelling approaches were adapted. Comparative experiments were then carried out to further select the neural language model of choice in the distributional model based on a few quality measures. The dataset was trained with *Word2Vec* [47], *Glove* [48] and *FastText* [46] based embeddings. A comparison with other conventional topical modelling algorithms was also carried out. The embedding algorithms were experimented as baselines with different but consistent dimensions. Results on their topic quality measurements as in Table 1 informed the embedding framework of choice for further modelling .

FastText algorithm modelling unlike other variants ignores word structures formulation. With FastText, each word token w is represented as a bag-of-character *n-grams* in the vector space. The word token itself is also included in the n-grams [1]. To incorporate most of the n-grams, $3 \leqslant n \leqslant 6$ are optimal [46]. With an n-gram dictionary of size Y and word w, $Y_w \subset \{1, \ldots, Y\}$. x_y is the vector representation of each n-gram b as in the below equation. A word is represented by the sum of the vector representations of its n-grams. Thus, the scoring function is formulated as $s(w, b) = \sum_{y \subseteq Y_w} x_y^{\top} {}_{v_b}$ where v is the word vector and b the context position of the word. Unlike *FastText*, *Word2Vec* and *Glove* modelling ignore word morphology i.e. the character sequence that forms a word, thus each term is treated as whole i.e. atomically. This makes vectorization of short and noisy text difficult in such modelling approaches.

3.3. Topics over word embeddings inferencing

As mentioned in Section 3.2, advantages of word vector representations and topic models were leveraged in this phase. This was pertinent in the distribution of topics as user interests over time variations, accommodating datasets that span over periods of time. Vector representations of terms and topics were modelled in the distributional model as follows: -

1. For each term v, the model considers an M-dimensional vector representation σ_v. At each timestamp t, the distributional model embedding is presented as $\alpha_k^t \in \mathbb{R}^M$ for each topic k in the semantic space of word tokens. This meant that each topic, with relevant probabilistic terms is denoted as a time-variational vector with documents separated in mini-batches [52]. Therefore, the probability of a word in document x (as in Section 3.1) to be modelled in given topic, is computed as the exponential inner product between the topic and word embeddings at corresponding timestamps (user defined) in the dataset. This probability function is represented in Eq. (1).

$$P(b_{xn} = v | z_{xn} = k, \alpha_k^{(t_x)}) \propto exp\{\sigma_v^{\top} \alpha_k^{(t_x)}\} \tag{1}$$

   z_{xn} is the topic assignment in LDA for each word n in the document as described in the probabilistic extraction of topics in Section 3.1.
2. Correlation between terms and their representative topics is higher when the term and topic embedding is also higher in the embedding space. Terms with such similarities just as in clustering, are bound to be grouped together. Markov Chain [53] enforces this smooth variation of topics over their embeddings. In addition, time priors are considered over the vector of topic proportions α_x. The model then captures variations in topics over the time series dataset. The prior over the vectors of topic proportions depends on a latent variable λ_{t_x} where t_x is the timestamp variation of document/tweet x. This probabilistic process is represented as $P(\theta_x | \lambda_{t_x}) = \mathcal{N}(\lambda_{t_x}, Q^2 I)$ where $P(\lambda_t | x_{t-1}) = \mathcal{N}(\lambda_{t-1}, \sigma^2 I)$. Q is a model parameter, while σ is a hyper parameter controlling topical smoothing of the Markov chains. $\mathcal{N}$ is the logistic normal distribution that transforms Gaussian variables (O, I) to the simplex.
   Overall, the combinatory modelling approach where terms are distributed over word and topic representations differs from conventional topic models. The difference is in the probabilistic distribution of topics over the vocabulary in the corpus just like in LDA.

The above modelling process is captured in Algorithm 1.

Algorithm 1: Embedded Interest Topic Modelling Process over Time.

Require: Tweet Tokens v, Corpus X, Time Stamp t_1, ..., t_x

1: **Initialization of model parameters** ▷ Dataset, model and model optimization arguments ($t = 1, ..., T, k = 1, ..., K$)
2: **for** $i = Iterationcounts$ **do**
3: Compute document(tokenized tweet) and topic embedding means
4: Compute the topics $\zeta_k^t = softmax(\sigma^{\top} \alpha^t{}_k)$
5: Extract tweets in mini-batches as per the set dataset argument
6: **for** each tweet x in each mini-batch **do**
7: Compute the topic proportions θ_x
8: **for** each word c in each tweet **do**
9: Compute $P(v_{xc} | \theta_x) = \sum_x \theta_{xk} \zeta_{k, v_{xc}}^{t_x}$
10: **end for**
11: **end for**
12: Update the dataset, model parameters and hyper-parameters
13: **end for**

In Algorithm 1, the model inputs are tweet tokens as time series documents. Modelling the tokens requires adjustment of model parameters, and hyper-parameters specifications (Line 1). Depending on the training iteration counts, terms and topical means are computed (Line 2,3 and 4). The word embeddings were

generated as described in Section 3.2. For each list of tweet tokens representative of a tweet in each mini-batch, topic proportions are computed via LDA from the topic embeddings and the closest terms assigned to the most representative topics (Lines 6 ... 9). The closer the topic and term means are, the higher the likelihood of being clustered together. Therefore, semantically close terms are assigned to similar topics as their representations are close in the embedding space. The process is repeated for all terms across the mini-batches until all tokens are modelled and all iterations completed. Model and data parameters as well as hyper-parameters can then be updated and the modelling re-run until convergence.

4. Experimentation

This section validates the processes presented in Section 3. A few consecutive steps are followed in the modelling process in addition to the description of the time-variant dataset in the experimentation. The experimentation process aims to generate topics as interests that are sensitive to time spanning the dataset period. Therefore, the end result in this experimentation phase is a comparative evaluation of an agreement between the control set in Section 4.1.2 and the generated topics from the dataset over the test set period as described in Section 4.1.3.

4.1. Datasets

The goal of the experimental process was to compute topical interest evolution over a short-text dataset. This meant better extraction of user interest/preference changes over time in short texts. To simulate this process, a dataset of tweets that bore the following characteristics was collected:-

1. **Generic set of tweets** — Tweets were collected via Twitter's search API. The geolocated but generic set of tweets was collected from Kenya. Ideally, this presented a dataset that could be validated via a control set as described in Section 4.1.2. This does not mean that the framework is specific to just one location. As long as the dataset is of short nature and timestamp-based, this approach is applicable.
2. **Language independence** — About 90% of the tweets were disseminated in English. The rest were in Swahili and a mixture of the two. Ideally the modelling process especially with embeddings was language independent.
3. **Time Variance** — Extracted tweets spanned a period of five years segmented in quarters i.e. from *2015 Quarter One(Q1)* to *2020 Quarter two (Q2)*. Timestamp variation t_x for each tweet was incorporated.

4.1.1. Training set

The models were trained on a corpus of 828,789 generic, cleaned, and tokenised tweets with the collection geolocated to Kenya as mentioned in Section 4.1. The choice of Kenya was influenced by the availability of a control set. The same collection can be replicated across different geographical regions for extraction of larger datasets especially if the control set is not of essence, unlike in this scenario. The collection period was between *2015 Quarter One (Q1)* and *2020 Quarter Two (Q2)*. However, tweets in five quarters in this collection period were excluded from the final dataset as they had very few tweets. These were **Quarter 4 of 2015, Quarters 2 and 4 in 2016 and, Quarters 3 and 4 in 2017**. Less data with short very sparse vocabulary meant made it impossible to keep the data. Therefore, the resultant dataset spanned 17 quarters across 2015 to 2020. Each tweet in the collection included the tweet's metadata, e.g., geo-coordinates, hashtags, retweets, etc.

4.1.2. Control set

A *control set* was needed to validate the conversation in Section 4.1. This was to provide a semantic proof of topical evolvement over time [54,55]. The assumption is that the semantics in news items positively correlate with the disseminated tweets in specific demographics [56]. Furthermore, semantics in tweets and to a large extent evolves with changes in the news items.

Since the initial collection was geolocated in Kenya, it was prudent to collect news items in the same demographics as a control set. Therefore, 189,906 tweets from Twitter handles of two major media houses, i.e., *Nation Media*[5] and *The Star-Kenya*[6] were collected for validation. Retweets and tweets with just mentions of these Twitter handles were not incorporated in the collection. The aim was to collect tweets solely disseminated by the media houses as the assumption was that they semantically correlated with the initial generic collection.

4.1.3. Test set

A neutral but relevant set of test tweets was needed in the testing phase of the framework. The goal was to make sure that the model worked to the expected level on a neutral dataset. A collection of 161,675 generic tweets from the same geographical bounds as the training data was collected. The test set was not part of the training and control sets. In addition, each tweet entry also contained associated metadata such as hashtags, mentions etc. Retweets were filtered out of the dataset as they made up the bulk of the replicated tweets.

4.2. Parameters and model training

The proposed framework is segmented into two important parts. The vector representation part and the topic modelling one. The two had diverse but consistent parameters across the different algorithms as used in the experimentation process. The parameters are detailed as in the sections below: -

4.2.1. Word embeddings settings:

The neural language models were adopted in generating vector representations over the training set in Section 4.1.1. The idea was to trial out several modelling algorithms and pick the best performing one for further modelling. This meant subjecting the training data to *FastText*, *Word2Vec* and *Glove* algorithms for vectorization and comparisons in performance consistent with [1]. *FastText* is capable of extracting syntactic information from a textual corpus regardless of the language of expression and misspellings. It does not ignore the word morphology, which is a limitation in short texts. With *FastText*, vector representations are associated with each character n-gram. Words that are typically made up of characters are then represented as the sum of character vectors computed using a sliding window. This is one property that makes *FastText* work well with misspelled or shortened words such as in tweets.*Word2Vec* and *Glove* modelling works the same way with a slight variation in the term vectorization process. Words are modelled atomically and not at character level when compared to FastText. In the generation of embeddings, the *number of dimensions, learning rate, context window size, minimum count, window* and *epochs* were specified. They were as follows for *FastText* and *Word2Vec* models. The embedding settings were adapted from works by [2], where embeddings were applied in the extraction of multi-interests in short-text.

- *Size* of the dimensions. This ranged from 100 to 300 dimensions consistent with [47]

[5] https://twitter.com/dailynation.
[6] https://twitter.com/thestarkenya.

- *min_count* or minimum count of a word in the corpus. Any word with a co-occurrence count less than the specified parameter is not incorporated in the training set;
- *sg* parameter for training a skip-gram model if $sg = 1$, otherwise Continuous Bag of Words (CBOW).
- *word_ngrams* to enrich word vectors with sub-word (n-grams) information if specified as 1;
- *iter* or iterations. This specifies the number of epochs over the corpus.
- The *window* parameter specifies the maximum distance between the current and predicted word in a tweet;
- *Glove* model only provisions for the *epochs* and *learning rate(lr)* to be defined.

4.2.2. Topic modelling settings:

As stated in Section 3.3, the end result was a distributional model that encompasses modelling of topics over embeddings, with time as a factor. Variances of different priors $\lambda^2 = \sigma^2 = 0.005$ and $Q^2 = 0.5$, consistent with the baseline setup in [57] were adapted. The optimal number of topics was 10 after 50 passes across the models, including in LDA and Twitter-LDA baselines. This was based on the Elbow heuristic [58] measure. To compute this, several K values representing probable number of clusters were factored in the modelling process. For each value of K, *K-means++* was applied to calculate *heterogeneity*. Heterogeneity is a measure of compactness in the clustering process. Parameters such as *batch size* = 50, *dropout rate* = 0.1, *learning rate* = 0.005 were sufficient in this setup after several iterations of different batch values. A fully connected feed-forward inference network for topic proportions α_x with ReLU activation [59] was used. The training was run over 20 epochs.

4.2.3. Model generation

As mentioned in Section 4.1.1, the model was trained on 828,789 tweets. The goal of the modelling process was to make it possible for words in the training set to have contextual inference. This is important in the computation of inter-word/sentence distances. Ideally, words with close contextual similarity are likely to have a high co-occurrence in the training set. For example, *"Donald Trump"* is likely to have a high similarity score compared to *"Donald Biden"* in a US political dataset.

Pre-processing the tweets as input corpus to the model followed these steps:-

- Lower-cased all terms in each tweet.
- Filtered out numbers and encoding accented characters. Numbers were not of importance in this modelling process.
- Hyperlinks in the corpus were removed.
- Removal of hashtags in the dataset. They are user-generated words representative of a topic of interest and are normally prefixed by the hash (#) symbol.
- User mentions were also removed. They are usernames on the platform prefixed by the @ symbol.
- Words with a character count less than three were removed . Their contextual significance was not high as they were made of prepositions, etc., that were not in the stopword list.
- NLTK stopword list[7] was used to filter out words in the list out of the dataset.
- Tokenisation of tweets where individual terms in each tweet are split and appended in a list for further modelling.

[7] http://www.nltk.org/.

The output, which was a clean and tokenized set of terms in a list, was trained via the modelling algorithms mentioned in Section 4.2.1.

As shown in Algorithm 1 the tokenized list of terms topical proportions were modelled via LDA. This process was computed over each mini-batch, which was ideally on a quarterly basis. The closest terms in each mini-batch were then assigned to the closest topic. The semantic weight of a term in the LDA generational model relative to the model was extracted at each mini-batch. The same process was repeated across the modelling algorithms and their variants over all mini-batches. For consistency purposes, model generation hyper-parameters were all the same as in Sections 4.2.1 and 4.2.2.

5. Results

The framework's performance was measured quantitatively and qualitatively to ascertain that the results corroborated in both dimensions. Quantitatively, topical quality across the timestamps was the tested measure using the generated embeddings. Ideally, the best performing modelling algorithm was adapted in further qualitative evaluations. This related to intra-topical changes over time and correlations in topical interests, a key measure in the user profiling process in streaming texts.

5.1. Topical quality evaluation

Two metrics were used to compute the quality of generated topics over time. **Topic coherence (TC)** across the topics in each specified period [60] was one metric. Normally, topics were represented as sets of important words with semantic relevance towards a specific theme. Depending on the type of dataset, modelling algorithm and fine-tuned parameters, the output will either be topics that either make sense or not, as they need to be interpretable if they are to be judged by humans. Therefore, topic coherence measure distinguishes good topics from bad ones by computing the degree of similarity between high scoring words in the topics. Secondly, **topic diversity (TD)** was computed. Basically, this measures the percentage of unique words in the top 25 words of all topics [52]. A product of the two scores was an indicator of **topic quality (TQ)**. Results are captured in Table 1. Scores close to zero indicate redundancies in the generated topics, and thus the modelling algorithm is deemed inferior. Overall, *FastText* based models performed better than *Word2Vec* and *Glove* variants in the embedding based techniques. This is largely attributed to their character-level modelling capabilities making them ideal for datasets with misspelled or shortened words reminiscent of tweets. However, FastText-Skip Gram with 100 dimensions performed best across the two quality variables. LDA [44] as well as Twitter-LDA [49] baselines performed dismally in modelling of interpretable topics of interest. Therefore, *FastText-Skip Gram* with 100 dimensions was selected for further modelling of intra-topical changes over time as well as topical interests correlations.

5.2. Qualitative evaluation

Qualitatively, two approaches were used in discerning changes in interest levels for users over time : -

1. **Intra-topical changes over time** — This process involved tracking of interest changes over time. In essence, topical interests reflect the nature of the consumed content over a time period.
2. **Topical Interests Correlation** — The assumption is that topical interest weights in the test and control sets positively correlate [56]. Therefore, a correlation measurement score was computed to ascertain the correlations.

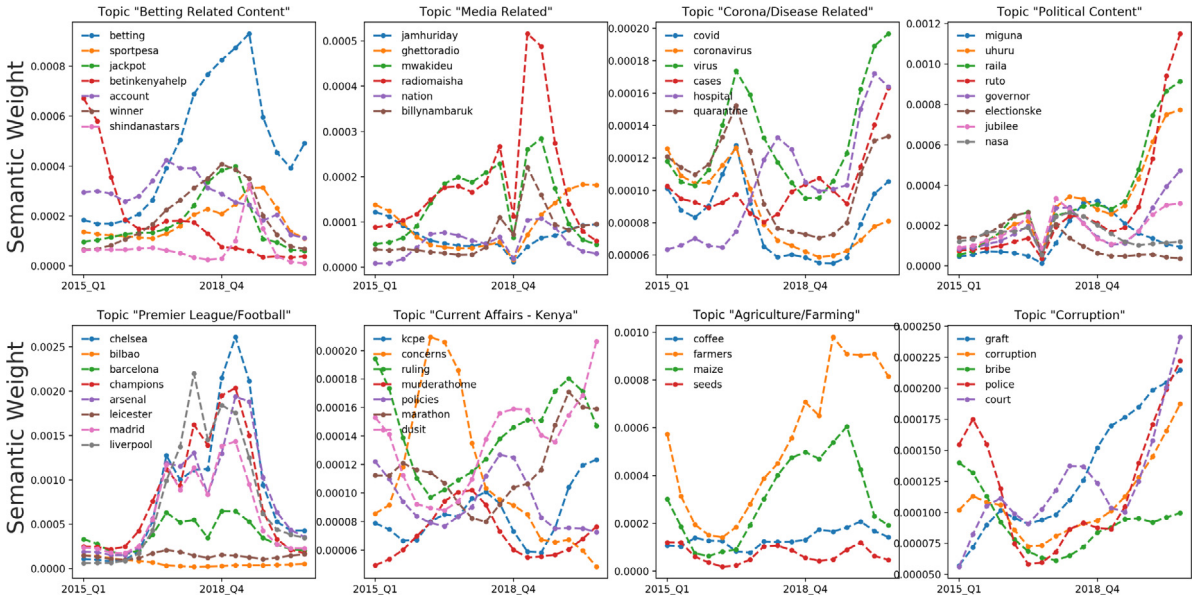

Fig. 2. Semantic Weights (Word probabilities) in eight topics across 17 quarters i.e. 2015–2020 in the test data. Probabilities shift with variations in time representing the overall interest levels across time. Interest in a word like "betting" rose exponentially from *2016 Q1* until about *2019 Q2* in the Betting related topic. Scaling varied per graph for better representation of individual semantic weights variations as the values differed largely across individual graphs.

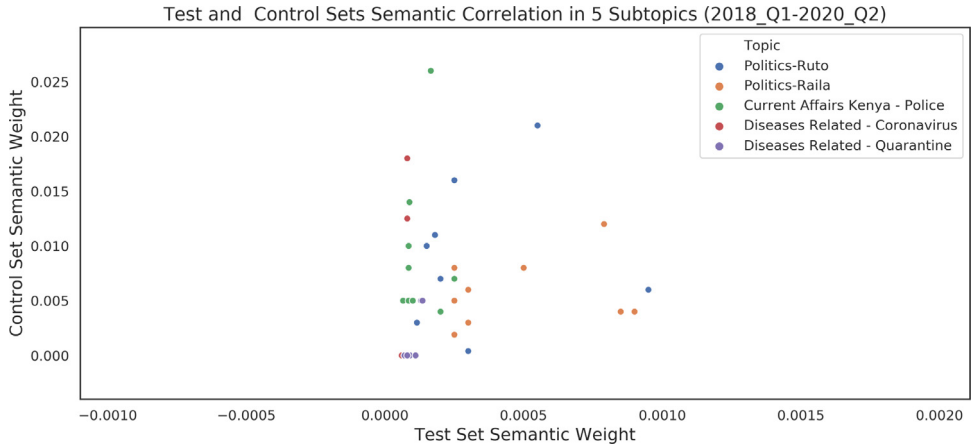

Fig. 3. Semantic weights (Word probabilities) as the interest score in five subtopics across 5 topics between *2018 Q1* and *2020 Q2*. Each data point is an interest score for each of the subtopics in the test and control sets per quarter. Individual values per quarter can be referenced in Table A.2.

5.2.1. Intra-topical changes over time

A representation of the topical interests over time is depicted in Fig. 2 from *2015 Q1* to *2020 Q2*, for a period of 17 quarters as described in Section 4.1.1. The interest changes are the probabilities of specific topical interest words across the timestamps in the test set. Basically, the probabilities are simply weights/influence a word has in a topic and across the time periods. The same evaluation measure is adapted in the control set. From Fig. 2, it is evident that overall interest in for example *"Football"* by audience in Kenya reduced in 2020. This is characterized by the lack of any footballing activity at this time when most European countries banned all forms of sporting activity,[8] a favourite activity for the Kenyan population. On the contrary, sub-topics with terms related to *"virus","coronavirus"* and *"quarantine"* gained attention at the start of 2020. The weights of these terms are probabilistic relative to the topic across the dataset timestamps. This period was characterized by lots of chatter about the virus across the world, Kenya included. In addition, this chatter time coincided with the COVID-19 mitigation measures by the government of Kenya.[9]

5.2.2. Topical interests correlation

As much as eyeballing gives an idea of the patterns in the dataset, it is not a scientific measure to draw any plausible statistical conclusions. To ascertain the topical interest change was a true representation especially over time, a control set was made use of as described in Section 4.1.2. This dataset was a collection of generic news items geolocated to Kenya just like the test dataset. The assumption is that news items are largely a reflection of what most tweeters disseminate [56]. Therefore, agreement in the topical relevance in both the control and test sets is a possibility.

Figs. A.4–A.13 show the topical distributions in the control set. In this set, the topical changes span a period of ten quarters. The correlations between the control (news items dataset) and the

[8] https://bbc.in/2Onmvi0.

[9] https://bit.ly/3ew5InE.

test set was in the semantic weights meaning the topical interests across the timestamps varied. To ascertain the correlation between the control and test set weights, a correlation measure was computed at each timestamp between the two sets. For consistency, the computation was done for the last 10 quarters, a period the control set was available. Table A.2 has records of these semantic weights in the two sets across the quarters. For demonstration purposes, five sub-topics were selected from the control and test sets as in depicted in Table A.2. The choice was driven by the presence of the sub-topics across the dataset collection period.

Fig. 3 is the resultant scatter plot of the subtopic weights in the test and control sets from *2018 Q1* to *2020 Q2*. From the figure, the values between the two sets largely correlated more so, in sub-topics such as "Diseases Related - Coronavirus" and "Diseases Related - Quarantine" as they are in the same semantic space in the plot. Their weights in both sets were close to 0 in the initial 8 quarters. Since the weights are probabilistic relative to each topic, the assumption is that weight will still be slightly greater than 0. This is attributed to the fact that the term at least exists, and has influence on the topic thought to a lesser extent. This explains why their topical influence was very low prior to *2020*. Their semantic weights increase in the last two quarters, i.e., *quarters one* and *two* of *2020* the period in which *COVID-19* was declared an epidemic in Kenya. The same pattern in the subtopics is replicated in the test set as shown in Fig. 2. On the other hand, *"Politics-Ruto"*, *"Politics-Raila"* and *"Current Affairs - Kenya"* depicted sinusoidal patterns over the 10 quarters in the two sets. This correlated to the nature of political content shared in the country over the time frame as shown in Fig. 2.

Furthermore, the correlation in the test and control sets was evaluated by measuring the Pearson Correlation Coefficient (PCC) between the two sets at each timestamp [61] across the 10 quarters. Ideally, there is a linearity assumption in the relationships between the weights. This is in the sense that an increase or decrease in the semantic topical weight of the test set should have the same effect as in the control set, depicting uniformity in interest over time. The coefficient is computed as the covariance of the two variables per time stamp divided by the product of the standard deviation in the control and test weight sets as $PCC = \mathrm{Cov}(a, b)/\sigma(a)*\sigma(b)$. Here, PCC is the Pearson's correlation coefficient. σ is the standard deviation that is applied to variables a and b. The correlation is expressed with a value between -1 and 1, where -1 depicts a negative correlation while 1 indicates positive correlation.

The results in Table A.2 depict positive correlation between the semantic weights for both the control and test sets across the subtopics. It is worth noting that the most positive correlations were noted in emerging topics such as *"Diseases Related - Coronavirus"* and *"Diseases Related - Quarantine"*. The same positivity in the correlations is noted in the *political* and *current affairs* subtopics, albeit to a lesser extent. This is attributed to the sinusoidal pattern in the interest weights across the evaluation timestamps.

6. Discussion and application areas

6.1. Discussion

The goal of our work has been to identify evolving topical interests in short texts and eventually build representative user profiles. Prior research in evolution of interests in streaming platforms encompassed several factors. Usage of external data, annotation, and a mixture of features in the augmentation of user interests in the profiling process has been studied [31,32,35]. Word embeddings representations have also been utilized in the generation of dynamic user profiles [38]. Inspired by these insights, dynamicity in our approach was at topical level, where the inner product between word and topical embeddings representing each topical interest at different timestamps was derived.

Our proposed representation model allows us to investigate relationships between topics and word embeddings at each timestamp. To do so, each term is distributed over topical embeddings where each topic is denoted as a time-variational vector. Therefore, each word is probabilistically assigned to a topic. The word-topic correlation is higher in the semantic space, when the semantic relevance between the two is also high. With the consideration of time priors over the topical embeddings, variations of topical interests are captured over time.

To evaluate the modelling framework, both quantitative and qualitative measures were implemented. Quantitatively, topical qualities across the embeddings and baselines were computed. A product of topical diversity and coherence was the topical quality measure. FastText Skip-Gram variant gave the best results in terms of the topical quality thus was the word embedding algorithm of choice for further evaluations. Qualitatively, a measure of intra-topical changes over time and topical interests' correlation was evaluated. Topical terms with their respective probabilistic weights across the timestamps were generated with results depicted in Figs. A.4–A.13. Ideally, topical words that generated attention had higher weights and this varied across the timestamps.

Further, there was need to use a control set in depicting the correlation of interests with the test. The results are presented in Table A.2 as well as in Fig. 3. The idea was to correlate the semantic weights in the two sets. This way, it was possible to extract variations in the topical interests over the timestamps. The Pearson Correlation Coefficient between the values in the two sets validated the semantic changes. From the results, emerging topical words like "coronavirus" depicted greater correlations in the two sets. However, co-occurring terms across the timestamps e.g., political interests fluctuated over time, typical occurrence even in conventional news.

6.2. Application areas

A few practical implications and application areas of our results are in the following spheres particularly relevant to third-party content providers as well as short-text data dissemination platforms:-

- Dissemination of semantically relevant time-variational content to users on the short-text platforms.
- Accuracy in the forecasts of the type/nature of content users are bound to consume in certain demographics and the likelihood of their future consumption by modelling their current consumption patterns.
- Content engagement patterns over time as content related to certain topics may be of more interest at specific times.
- Identification of the most relevant users to serve content by third-party content disseminators. This is relevant in cold-start scenarios too.

7. Conclusion and future work

Microblogging platforms like Twitter help present intrinsic and extrinsic user profiles to third party content providers. This to a large extent is based on the nature of content that users and their friendship networks consume over time. A framework that factors variational timestamps on topical embeddings was proposed. Several embeddings techniques and baselines were modelled and tested for topical quality. FastText-based embeddings

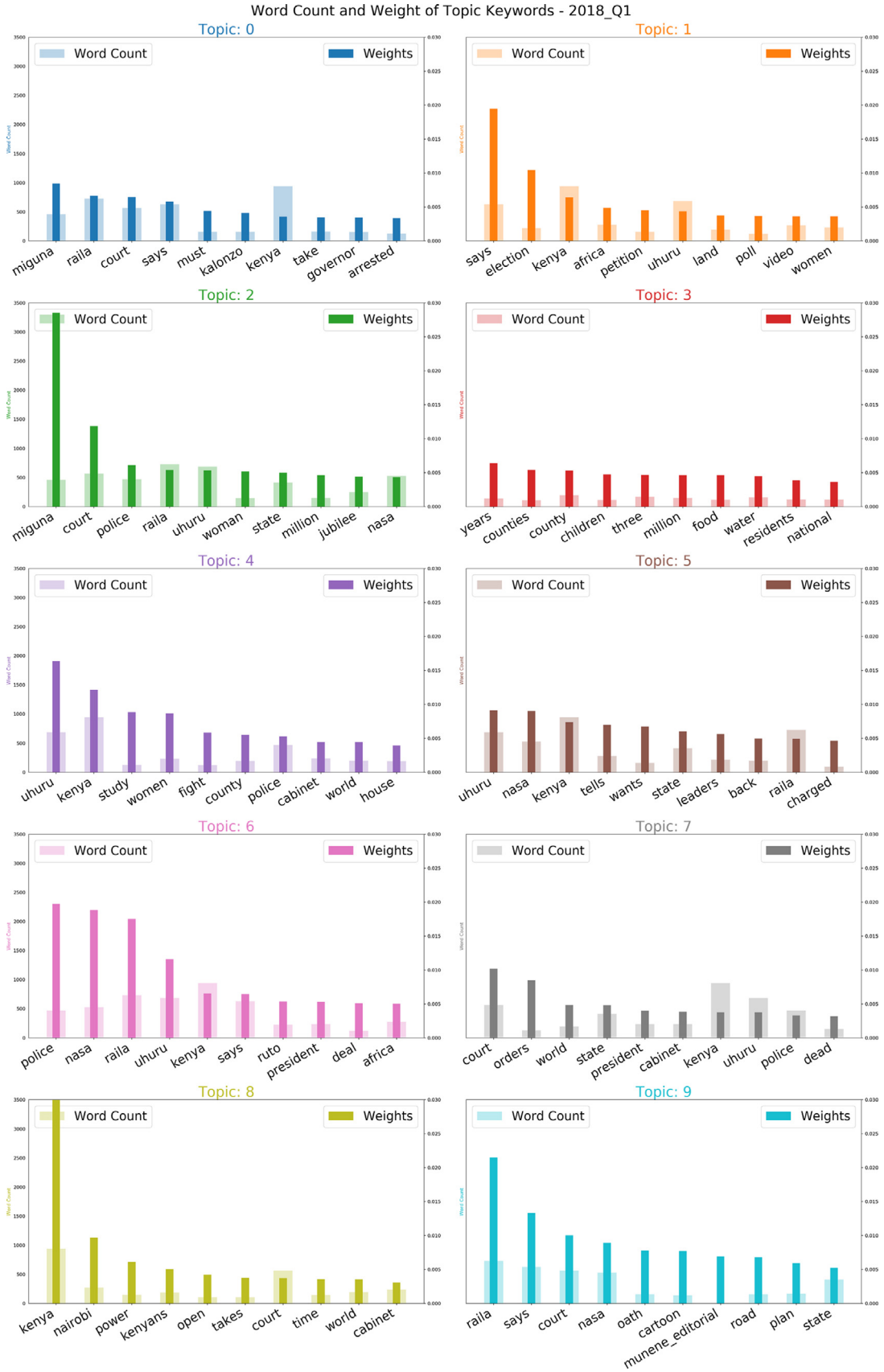

Fig. A.4. Sample word count versus term weights in each topic in Q1 of 2018.

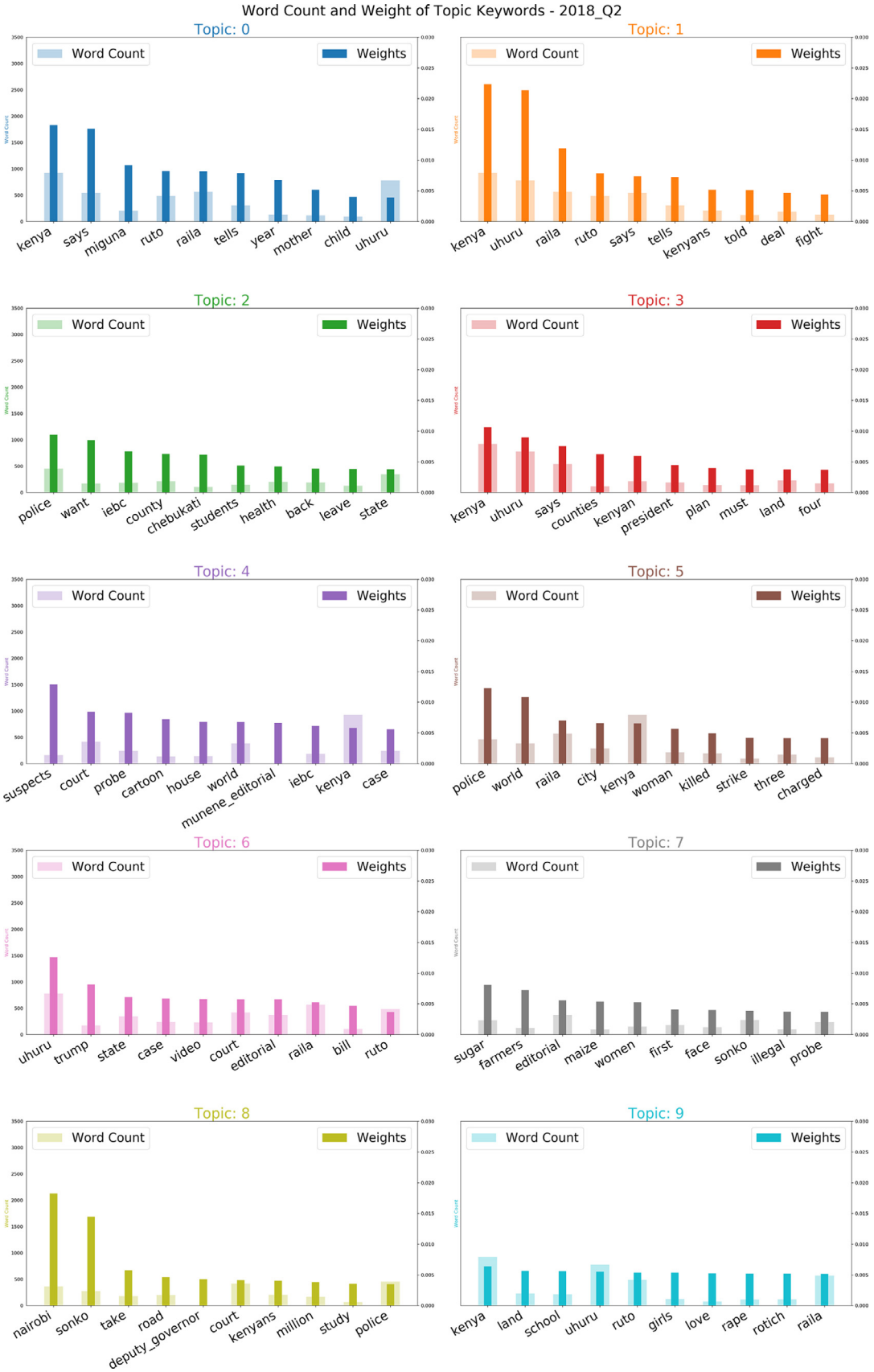

Fig. A.5. Sample word count versus term weights in each topic in Q2 of 2018.

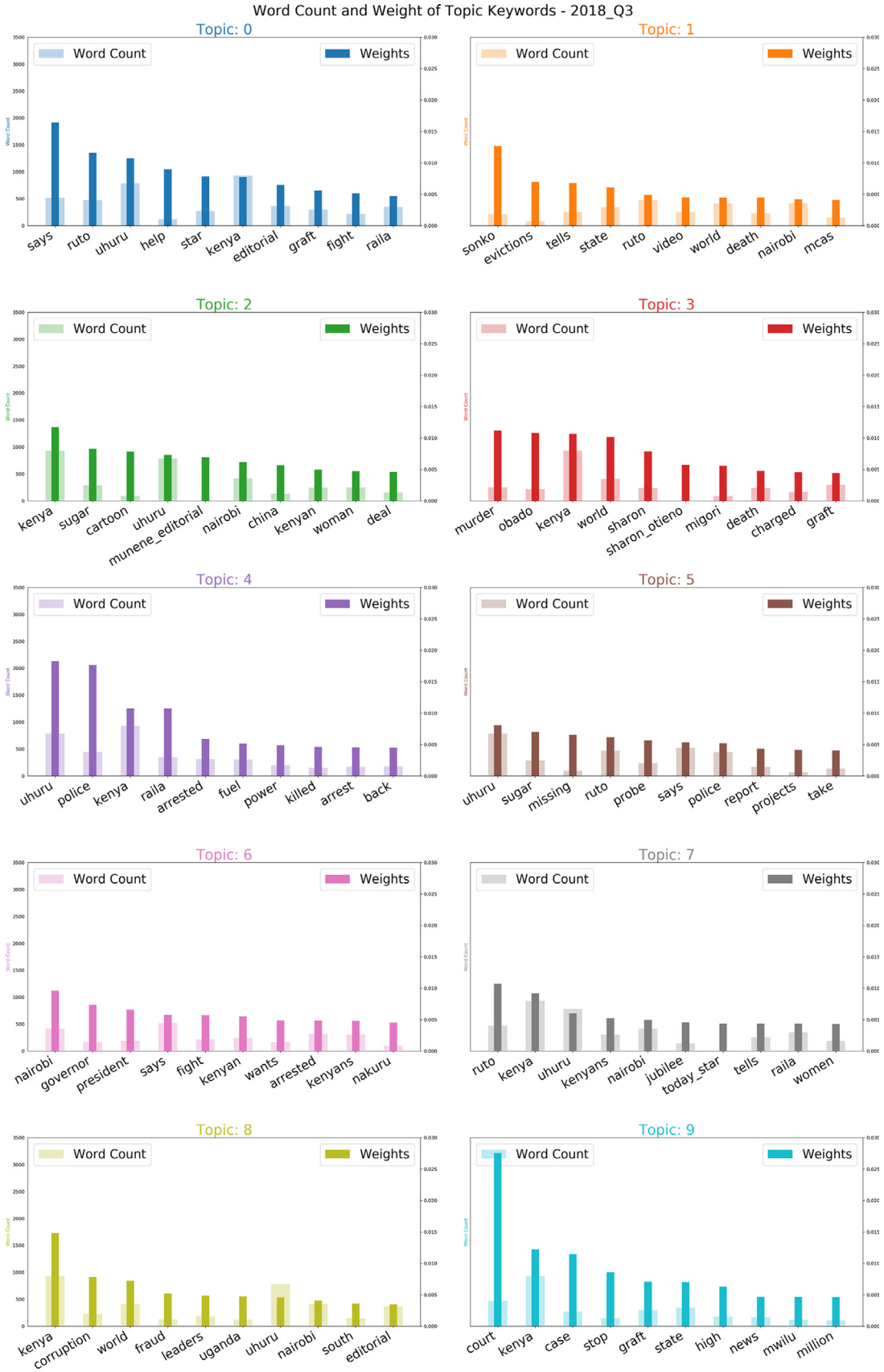

Fig. A.6. Sample word count versus term weights in each topic in Q3 of 2018.

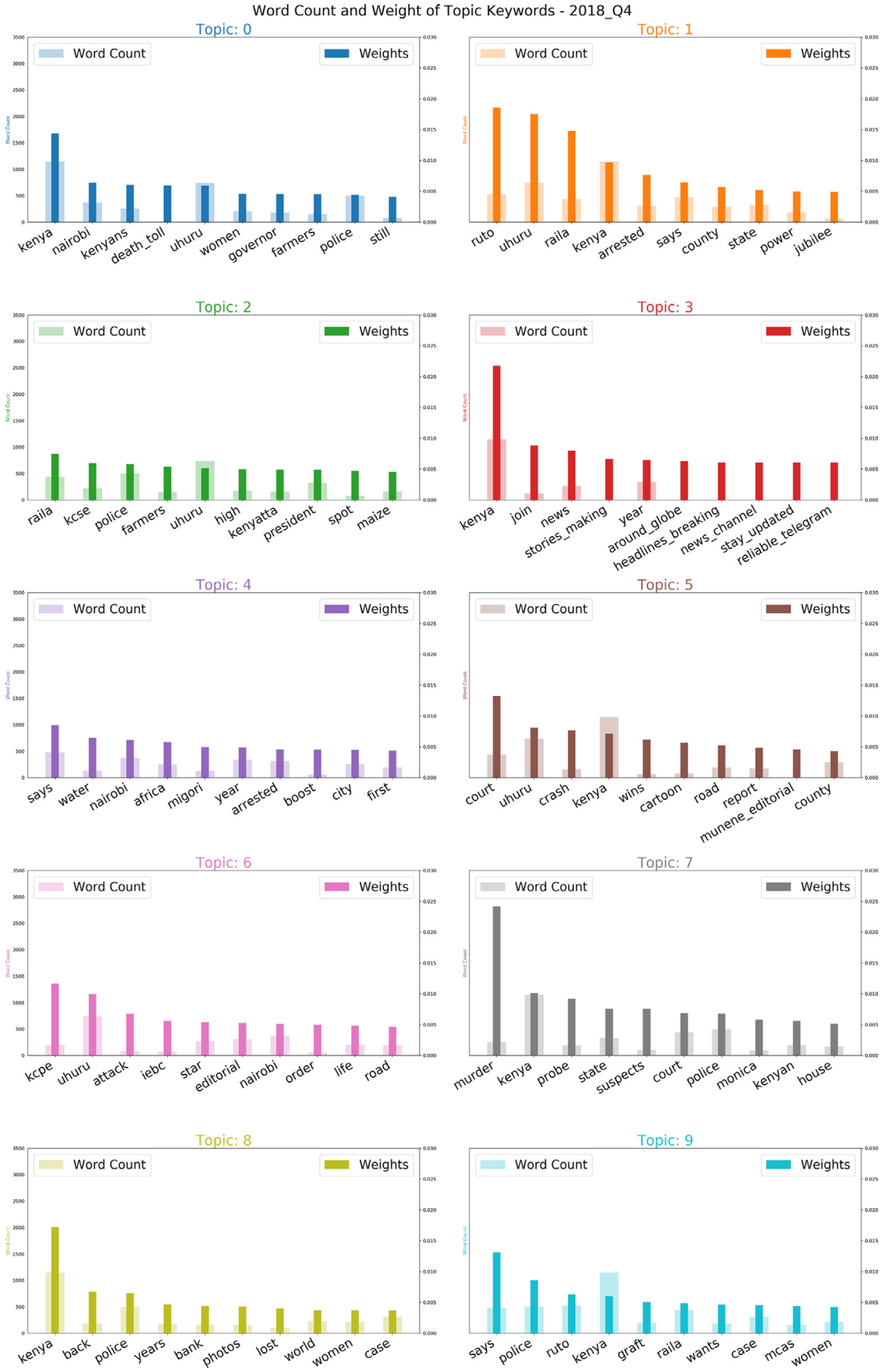

Fig. A.7. Sample word count versus term weights in each topic in Q4 of 2018.

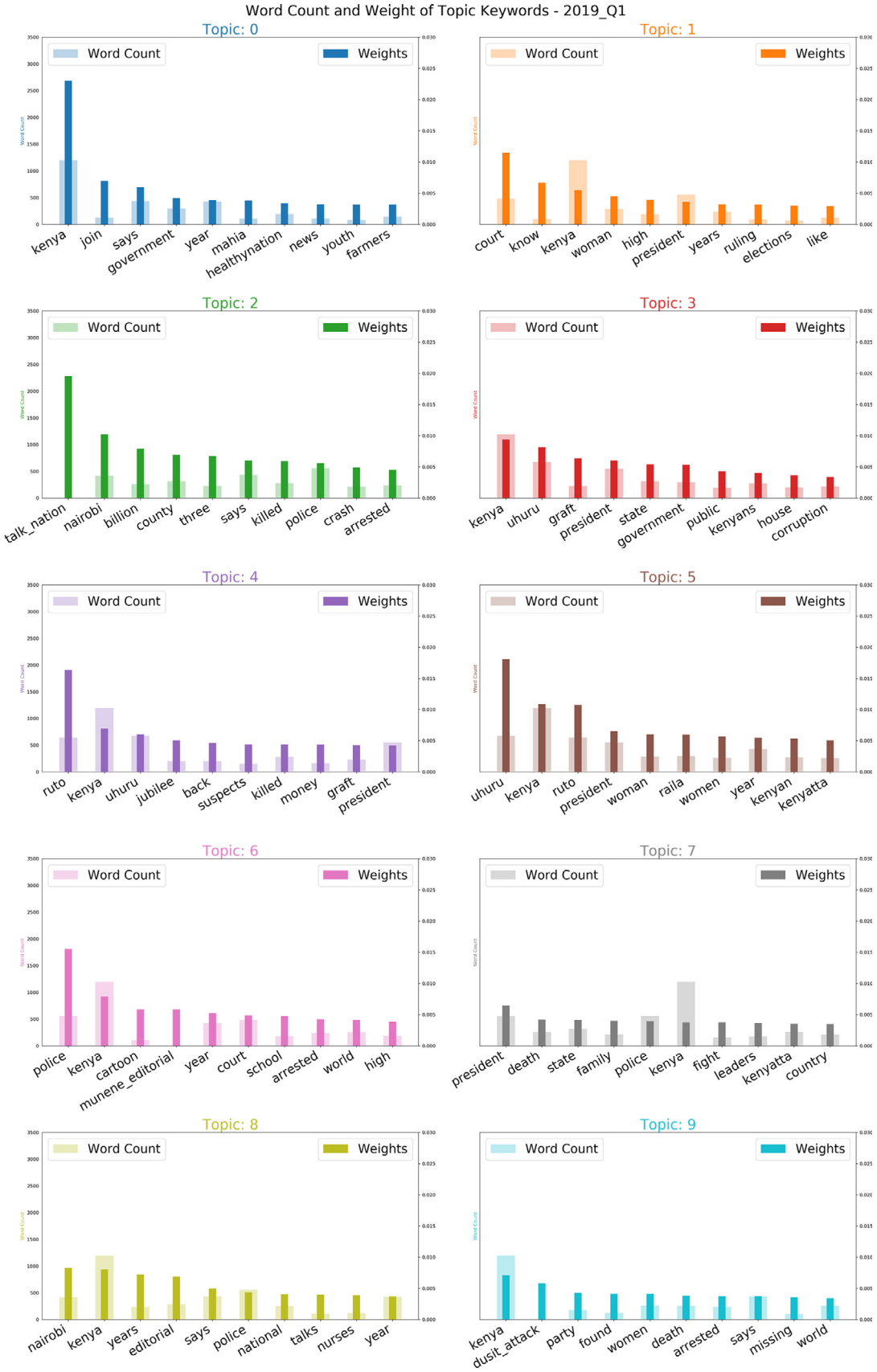

Fig. A.8. Sample word count versus term weights in each topic in Q1 of 2019.

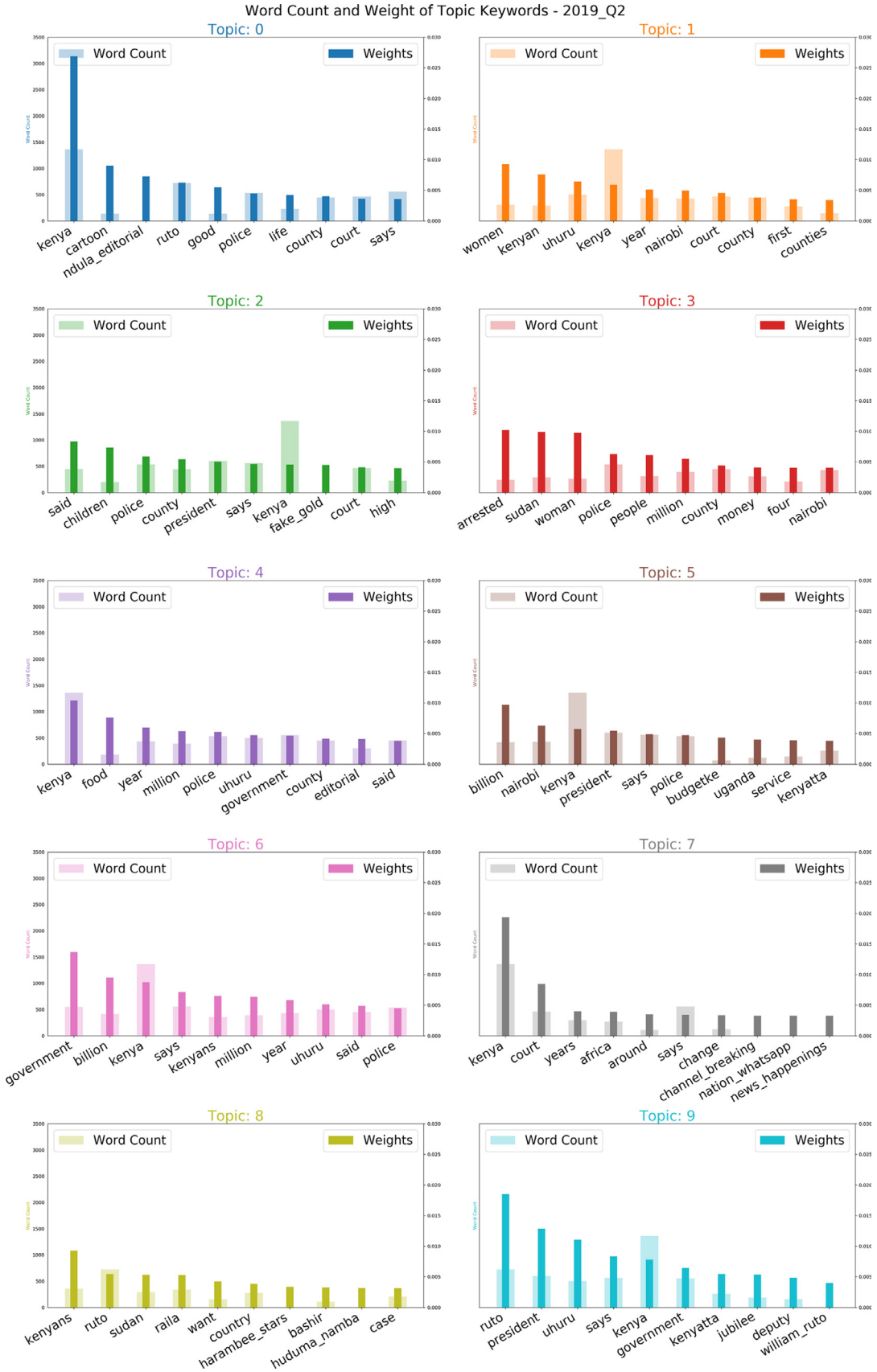

Fig. A.9. Sample word count versus term weights in each topic in Q2 of 2019.

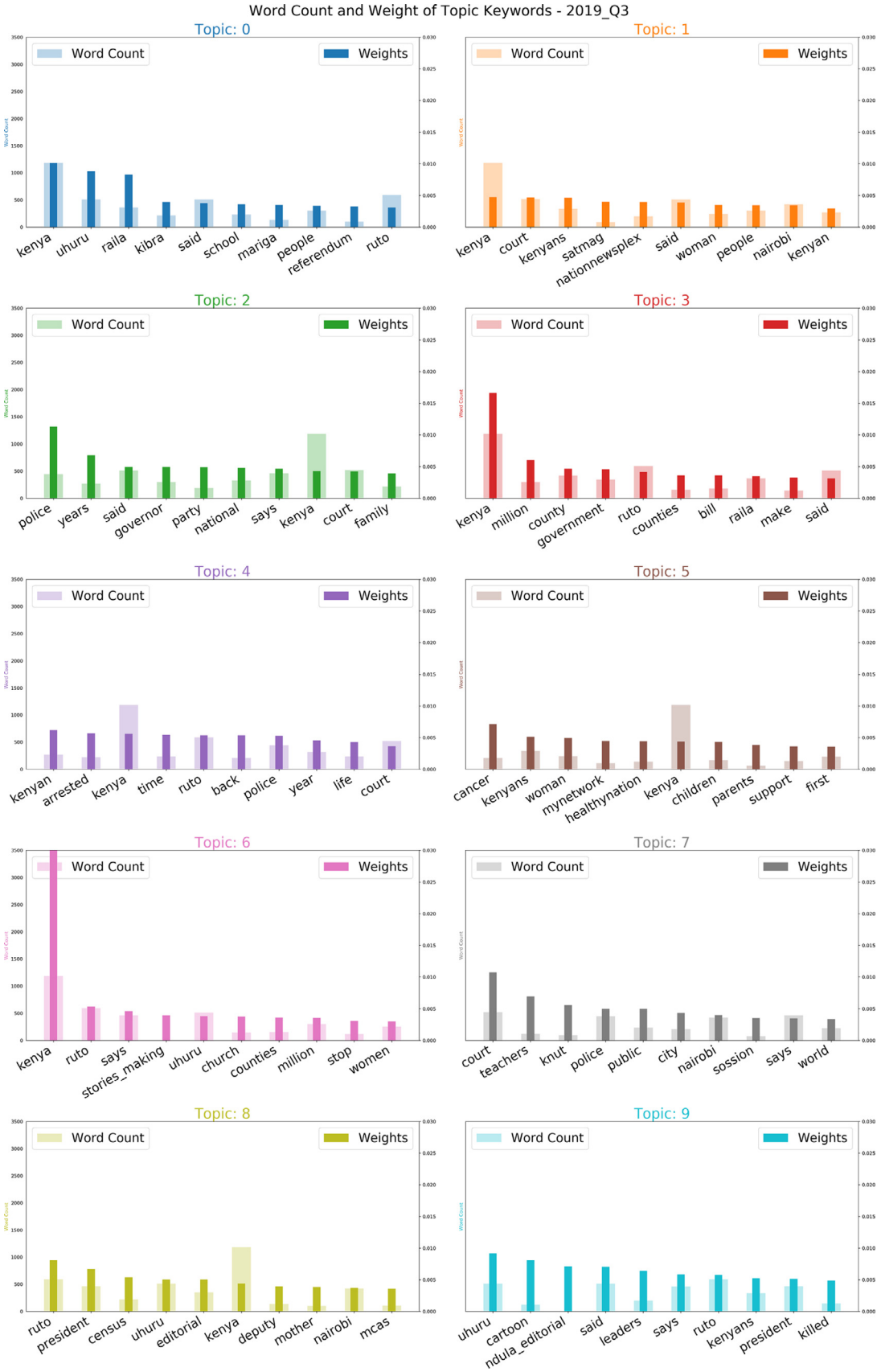

Fig. A.10. Sample word count versus term weights in each topic in Q3 of 2019.

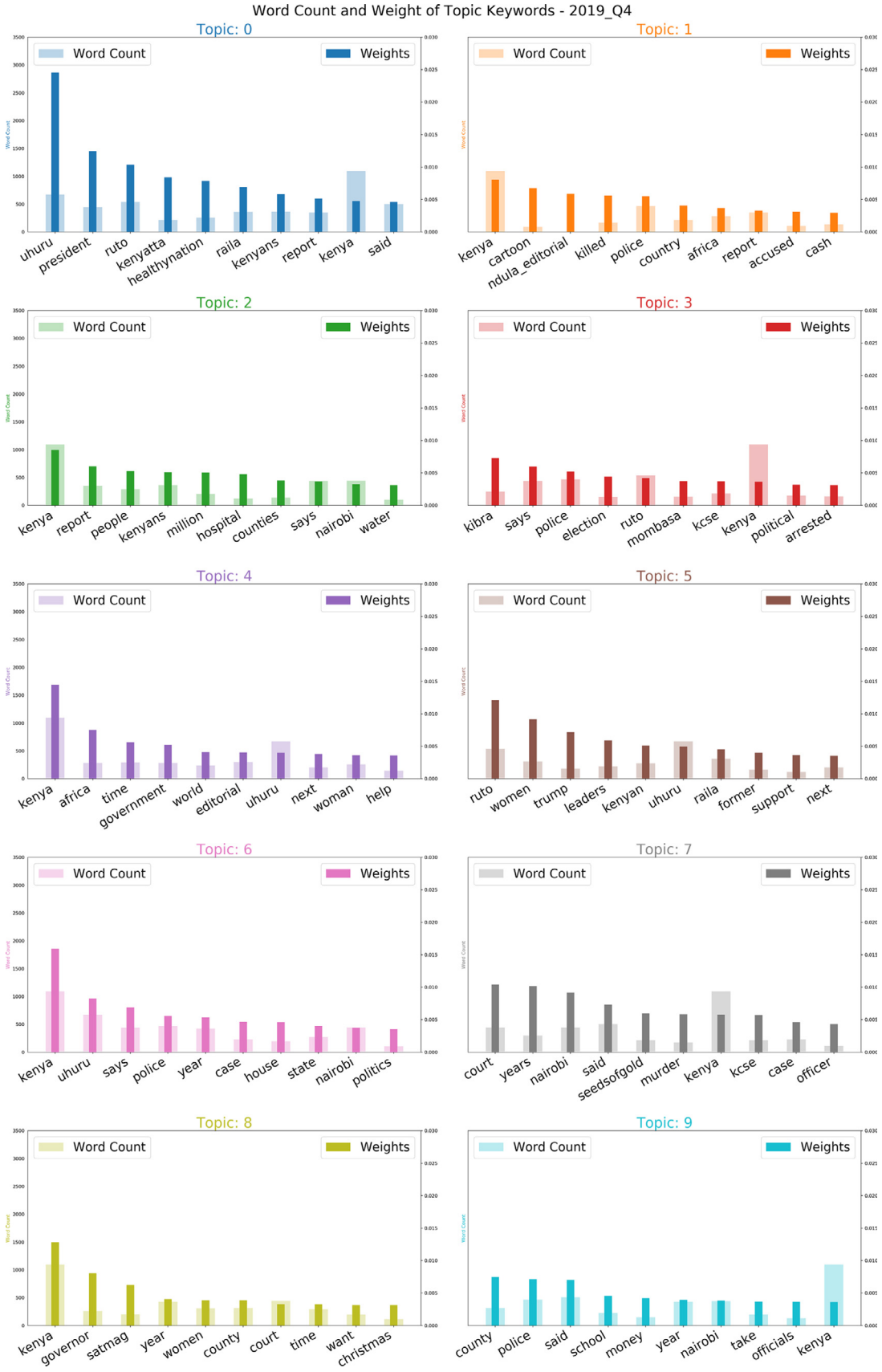

Fig. A.11. Sample word count versus term weights in each topic in Q4 of 2019.

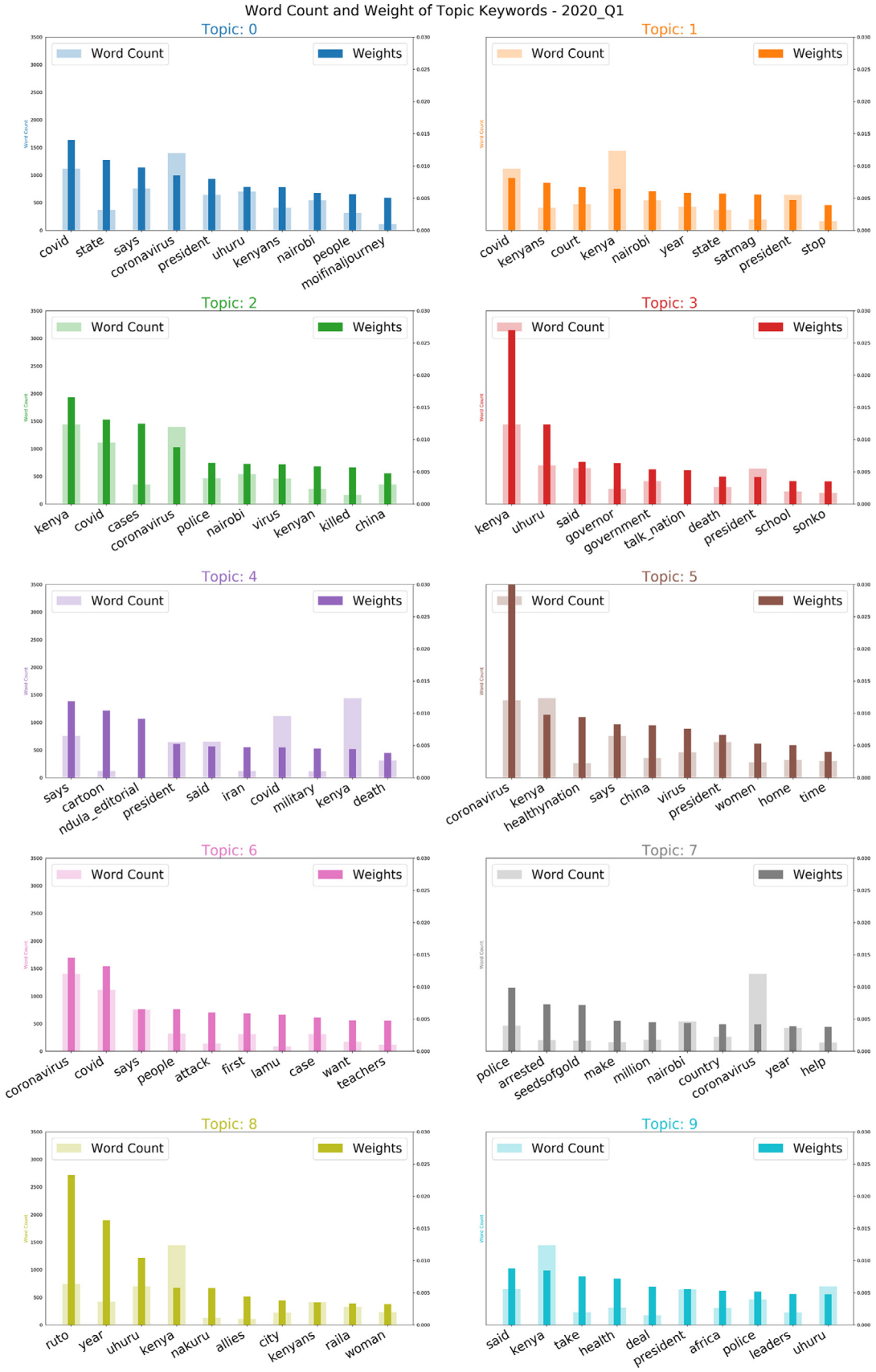

Fig. A.12. Sample word count versus term weights in each topic in Q1 of 2020.

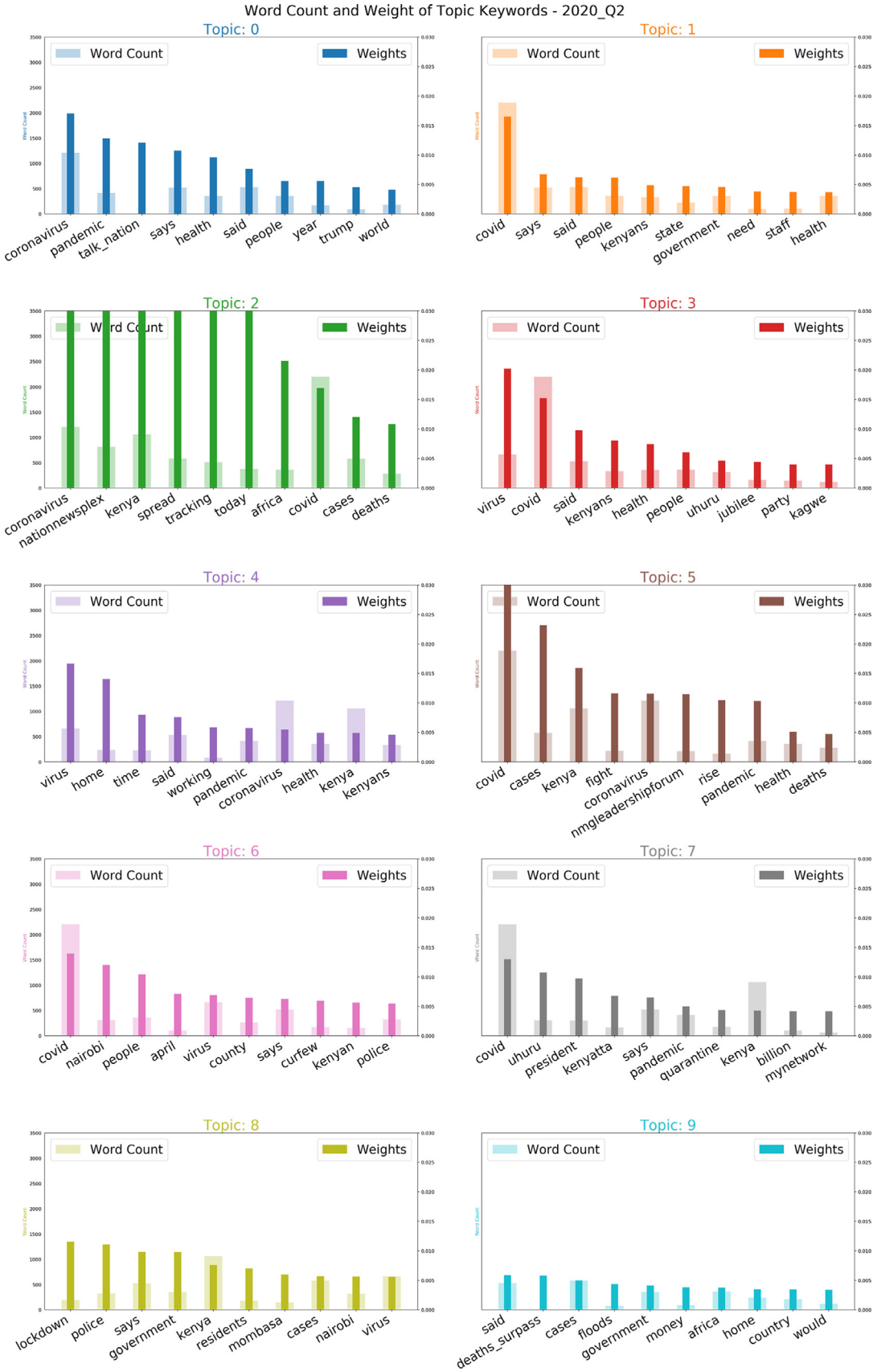

Fig. A.13. Sample word count versus term weights in each topic in Q2 of 2020.

Table A.2
Sample sub-topics with corresponding semantic weights in test and control sets over a period of 10 quarters. Pearson Correlation Coefficient (PCC) between test and control sets for each subtopic per time stamp was computed depicting the semantic changes as validation.

	Sub-topic	Dataset	2018_Q1	2018_Q2	2018_Q3	2018_Q4	2019_Q1	2019_Q2	2019_Q3	2019_Q4	2020_Q1	2020_Q2	PCC
1	Diseases Related - Quarantine	*Test Set*	0.00009	0.000075	0.00007	0.00007	0.00007	0.00007	0.00008	0.00011	0.00013	0.000135	**0.871**
		Control Set	0	0	0	0	0	0	0	0	0.005	0.005	
2	Diseases Related - Coronavirus	*Test Set*	0.00008	0.00007	0.000065	0.00006	0.00006	0.00006	0.00006	0.00007	0.00008	0.00008	**0.635**
		Control Set	0	0	0	0	0	0	0	0	0.0125	0.018	
3	Current Affairs Kenya - Police	*Test Set*	0.000065	0.000085	0.000088	0.000085	0.000085	0.0001	0.000135	0.000165	0.0002	0.00025	**0.090**
		Control Set	0.005	0.01	0.014	0.005	0.008	0.005	0.005	0.026	0.004	0.007	
4	Politics-Raila	*Test Set*	0.00025	0.00025	0.00025	0.0003	0.00025	0.0003	0.0005	0.00079	0.00085	0.0009	**0.239**
		Control Set	0.0019	0.008	0.005	0.006	0.005	0.003	0.008	0.012	0.004	0.004	
5	Politics-Ruto	*Test Set*	0.0002	0.00025	0.00025	0.0002	0.00015	0.00018	0.0003	0.00055	0.00095	0.000115	**0.162**
		Control Set	0.005	0.016	0.005	0.007	0.010	0.011	0.0004	0.021	0.006	0.003	

were selected for further computation based on their success in generating quality topics across the timestamps. Technically, the dataset had a factor of time and topics representing interests modelled over the timestamps. The test set was subdivided in quarters from 2015 Quarter one (Q1) to 2020 Quarter two (Q2) and subjected to the framework for modelling. The resultant output was the semantic weight of subtopics within broader topics over 17 quarters in the test set. To validate the results, a semantically relevant set(control set) was collected and modelled. Semantic weights corresponding to the last 10 quarters were extracted. A correlation measure between the control and test set weights was then computed based on the topical linearity assumption. Results indicated positivity in the correlations justifying the validity of the results in the proposed framework.

The paper provided an enhanced modelling framework with time as a factor. The framework is a advancement in the design of short-text third party recommender systems with volatility in interest gain and decay. In this work, it was possible to deduce the interest changes in topics of interest by tracking the dissemination patterns over time. Such a process has the potential to improve on the current state-of-the-art in time variational topic modelling and interests identification processes.

In future, an inclination towards usage of larger language representation model such as BERT to enhance the interest extraction process in short texts is a possibility. Incremental learning via transformers to enhance interest extraction and eventual user profiling is another area worth exploration especially for language independent models.

CRediT authorship contribution statement

Herman M. Wandabwa: Conceptualization, Methodology, Data curation, Formal analysis. **M. Asif Naeem:** Resources, Writing - review & editing, Supervision. **Farhaan Mirza:** Writing - review & editing, Supervision, Validation. **Russel Pears:** Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix. Topical term distributions

The graphs Figs. A.4–A.13 represent sample quarterly word counts versus term weights in topical interests across 2018, 2019 and 2020. Word counts and related weights in the identified topics are pertinent in most topical modelling algorithms. On the other hand, Table A.2, shows a side by side comparison of semantic weights in the test and control sets over ten quarters.

References

[1] H. Wandabwa, M.A. Naeem, F. Mirza, R. Pears, Follow-back recommendations for sports bettors: A Twitter-based approach, in: Proceedings of the 53rd Hawaii International Conference on System Sciences, 2020, pp. 2569–2578.

[2] H. Wandabwa, M.A. Naeem, F. Mirza, R. Pears, A. Nguyen, Multi-interest user profiling in short text microblogs, in: International Conference on Design Science Research in Information Systems and Technology, Springer, 2020, pp. 154–168.

[3] Y. Halberstam, B. Knight, Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter, J. Publ. Econom. 143 (2016) 73–88.

[4] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. Ser. B Stat. Methodol. 39 (1) (1977) 1–22.

[5] H. Yin, B. Cui, L. Chen, Z. Hu, X. Zhou, Dynamic user modeling in social media systems, ACM Trans. Inform. Syst. (TOIS) 33 (3) (2015) 1–44.

[6] P. Yang, J. Liu, J. Qi, Y. Yang, X. Wang, Z. Lv, Comparison and modelling of country-level microblog user and activity in cyber-physical-social systems using weibo and Twitter data, ACM Trans. Intell. Syst. Technol. (TIST) 10 (6) (2019) 1–24.

[7] Z. Ghaemi, M. Farnaghi, A varied density-based clustering approach for event detection from heterogeneous twitter data, ISPRS Int. J. Geo-Inf. 8 (2) (2019) 82.

[8] C.C. Aggarwal, Content-based recommender systems, in: Recommender Systems, Springer, 2016, pp. 139–166.

[9] C. Chen, J. Ren, Forum latent Dirichlet allocation for user interest discovery, Knowl.-Based Syst. 126 (2017) 1–7.

[10] G. Piao, J.G. Breslin, Inferring user interests in microblogging social networks: a survey, User Model. User-Adapted Interact. 28 (3) (2018) 277–329.

[11] S. Gauch, M. Speretta, A. Chandramouli, A. Micarelli, User profiles for personalized information access, in: The Adaptive Web, Springer, 2007, pp. 54–89.

[12] A. Alsaeedi, A survey of term weighting schemes for text classification, Int. J. Data Mining Modell. Manage. 12 (2) (2020) 237–254.

[13] P. Bhattacharya, M.B. Zafar, N. Ganguly, S. Ghosh, K.P. Gummadi, Inferring user interests in the twitter social network, in: Proceedings of the 8th ACM Conference on Recommender Systems, 2014, pp. 357–360.

[14] I. Paul, A. Khattar, P. Kumaraguru, M. Gupta, S. Chopra, Elites tweet? Characterizing the Twitter verified user network, in: 2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW), IEEE, 2019, pp. 278–285.

[15] J.R. Chowdhury, C. Caragea, D. Caragea, On identifying hashtags in disaster Twitter data, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34(01), 2020, pp. 498–506.

[16] S. Xu, A. Zhou, Hashtag homophily in twitter network: Examining a controversial cause-related marketing campaign, Comput. Hum. Behav. 102 (2020) 87–96.

[17] Y. Wei, Z. Cheng, X. Yu, Z. Zhao, L. Zhu, L. Nie, Personalized hashtag recommendation for micro-videos, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 1446–1454.

[18] R. Cui, G. Agrawal, R. Ramnath, Tweets can tell: activity recognition using hybrid gated recurrent neural networks, Soc. Netw. Anal. Min. 10 (1) (2020) 1–15.

[19] X. Zheng, A. Sun, Collecting event-related tweets from twitter stream, J. Assoc. Inform. Sci. Technol. 70 (2) (2019) 176–186.

[20] J. Kang, H. Lee, Modeling user interest in social media using news media and wikipedia, Inf. Syst. 65 (2017) 52–64.

[21] P. Dooley, B. Božić, Towards linked data for wikidata revisions and Twitter trending hashtags, in: Proceedings of the 21st International Conference on Information Integration and Web-Based Applications & Services, 2019, pp. 166–175.

[22] G. Piao, J.G. Breslin, Exploring dynamics and semantics of user interests for user modeling on Twitter for link recommendations, in: Proceedings of the 12th International Conference on Semantic Systems, 2016, pp. 81–88.
[23] C. Nishioka, A. Scherp, Profiling vs. time vs. content: What does matter for top-k publication recommendation based on Twitter profiles?, in: 2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL), IEEE, 2016, pp. 171–180.
[24] D. Yu, D. Xu, D. Wang, Z. Ni, Hierarchical topic modeling of Twitter data for online analytical processing, IEEE Access 7 (2019) 12373–12385.
[25] Y. yeon Sung, S.B. Kim, Topical keyphrase extraction with hierarchical semantic networks, Decis. Support Syst. 128 (2020) 113163.
[26] G. Di Tommaso, S. Faralli, G. Stilo, P. Velardi, Wiki-MID: a very large multi-domain interests dataset of Twitter users with mappings to wikipedia, in: International Semantic Web Conference, Springer, 2018, pp. 36–52.
[27] J. Zheng, S. Wang, D. Li, B. Zhang, Personalized recommendation based on hierarchical interest overlapping community, Inform. Sci. 479 (2019) 55–75.
[28] T. Hamdi, H. Slimi, I. Bounhas, Y. Slimani, A hybrid approach for fake news detection in Twitter based on user features and graph embedding, in: International Conference on Distributed Computing and Internet Technology, Springer, 2020, pp. 266–280.
[29] S. Dhelim, N. Aung, H. Ning, Mining user interest based on personality-aware hybrid filtering in social networks, Knowl.-Based Syst. 206 (2020) 106227.
[30] S.R. Sahoo, B.B. Gupta, Hybrid approach for detection of malicious profiles in twitter, Comput. Electr. Eng. 76 (2019) 65–81.
[31] B. Jiang, Y. Sha, Modeling temporal dynamics of user interests in online social networks, Procedia Comput. Sci. 51 (2015) 503–512.
[32] Z. Zhu, Y. Zhou, X. Deng, X. Wang, A graph-oriented model for hierarchical user interest in precision social marketing, Electron. Commer. Res. Appl. 35 (2019) 100845.
[33] B.R. Cami, H. Hassanpour, H. Mashayekhi, User preferences modeling using dirichlet process mixture model for a content-based recommender system, Knowl.-Based Syst. 163 (2019) 644–655.
[34] F.S. Pereira, J. Gama, S. de Amo, G.M. Oliveira, On analyzing user preference dynamics with temporal social networks, Mach. Learn. 107 (11) (2018) 1745–1773.
[35] E. Stai, E. Milaiou, V. Karyotis, S. Papavassiliou, Temporal dynamics of information diffusion in Twitter: Modeling and experimentation, IEEE Trans. Comput. Soc. Syst. 5 (1) (2018) 256–264, http://dx.doi.org/10.1109/TCSS.2017.2784184.
[36] W. Wu, J. Li, Z. He, X. Ye, J. Zhang, X. Cao, H. Qu, Tracking spatio-temporal variation of geo-tagged topics with social media in China: A case study of 2016 hefei rainstorm, Int. J. Disas. Risk Reduc. 50 (2020) 101737.
[37] O. Şerban, N. Thapen, B. Maginnis, C. Hankin, V. Foot, Real-time processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification, Inf. Process. Manage. 56 (3) (2019) 1166–1184.
[38] S. Liang, X. Zhang, Z. Ren, E. Kanoulas, Dynamic embeddings for user profiling in twitter, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 1764–1773.
[39] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, J. Gao, Deep learning–based text classification: A comprehensive review, ACM Comput. Surv. 54 (3) (2021) 1–40.
[40] R. Satapathy, E. Cambria, A. Nanetti, A. Hussain, A review of shorthand systems: from brachygraphy to microtext and beyond, Cogn. Comput. (2020) 1–15.
[41] R. Satapathy, A. Singh, E. Cambria, Phonsenticnet: A cognitive approach to microtext normalization for concept-level sentiment analysis, in: International Conference on Computational Data and Social Networks, Springer, 2019, pp. 177–188.
[42] T. Lin, Z. Hu, X. Guo, Sparsemax and relaxed Wasserstein for topic sparsity, in: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, 2019, pp. 141–149.
[43] S. Burkhardt, S. Kramer, Decoupling sparsity and smoothness in the Dirichlet variational autoencoder topic model, J. Mach. Learn. Res. 20 (131) (2019) 1–27.
[44] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (Jan) (2003) 993–1022.
[45] M.S. Tajbakhsh, J. Bagherzadeh, Semantic knowledge LDA with topic vector for recommending hashtags: Twitter use case, Intell. Data Anal. 23 (3) (2019) 609–622.
[46] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Trans. Assoc. Comput. Linguist. 5 (2017) 135–146.
[47] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.
[48] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2014, pp. 1532–1543.
[49] W.X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, X. Li, Comparing twitter and traditional media using topic models, in: European Conference on Information Retrieval, Springer, 2011, pp. 338–349.
[50] T.K. Landauer, P.W. Foltz, D. Laham, An introduction to latent semantic analysis, Discourse Process. 25 (2–3) (1998) 259–284.
[51] C.E. Moody, Mixing dirichlet topic models and word embeddings to make lda2vec, 2016, arXiv preprint arXiv:1605.02019.
[52] A.B. Dieng, F.J. Ruiz, D.M. Blei, Topic modeling in embedding spaces, Trans. Assoc. Comput. Linguist. 8 (2020) 439–453.
[53] J. Besag, An introduction to Markov chain Monte Carlo methods, in: Mathematical Foundations of Speech and Language Processing, Springer, 2004, pp. 247–270.
[54] N. Hassan, F. Arslan, C. Li, M. Tremayne, Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 1803–1812.
[55] A. Hannak, D. Margolin, B. Keegan, I. Weber, Get back! you don't know me like that: The social mediation of fact checking interventions in twitter conversations, in: Proceedings of the International AAAI Conference on Web and Social Media, vol. 8(1), 2014.
[56] G. Brena, M. Brambilla, S. Ceri, M. Di Giovanni, F. Pierri, G. Ramponi, News sharing user behaviour on twitter: A comprehensive data collection of news articles and social interactions, in: Proceedings of the International AAAI Conference on Web and Social Media, vol. 13, 2019, pp. 592–597.
[57] D.M. Blei, J.D. Lafferty, Dynamic topic models, in: Proceedings of the 23rd International Conference on Machine Learning, 2006, pp. 113–120.
[58] P. Bholowalia, A. Kumar, EBK-means: A clustering technique based on elbow method and k-means in WSN, Int. J. Comput. Appl. 105 (9) (2014).
[59] Y. Li, Y. Yuan, Convergence analysis of two-layer neural networks with relu activation, in: Advances in Neural Information Processing Systems, 2017, pp. 597–607.
[60] D. Korenčić, S. Ristov, J. Šnajder, Document-based topic coherence measures for news media text, Expert Syst. Appl. 114 (2018) 357–373.
[61] J. Benesty, J. Chen, Y. Huang, I. Cohen, Pearson correlation coefficient, in: Noise Reduction in Speech Processing, Springer, 2009, pp. 1–4.