Information Systems 96 (2021) 101662

Contents lists available at ScienceDirect

Information Systems

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

Topical affinity in short text microblogs

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

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

ARTICLE INFO

Article history:
Received 4 March 2020
Received in revised form 12 August 2020
Accepted 10 October 2020
Available online 24 October 2020
Recommended by Ioannis Katakis

Keywords:
Information retrieval
Short text mining
Taste profiling
Data mining
Neural networks
Social web

ABSTRACT

Knowledge-based applications like recommender systems in social networks are powered by complex network of social discussions and user connections. Short text microblog platforms like Twitter are powerful in this aspect due to their real-time content dissemination as well as having a complex mesh of user connections. For example, users on Twitter tend to consume certain content to a greater or less extent depending on their interests over time. Quantifying this degree of content consumption in certain topics is an arduous task. This is further compounded by the amount of digital information that such platforms generate at any given time. Formulation of personalized user profiles based on user interests over time and friendship network is thus a problem. Therefore, user profiling based on their interests is important for personalized third-party content recommendations on the platform. In this paper we address this problem by presenting our solution in a two-step process:- *(i) Firstly, we compute users' Degree of Interest (DoI) towards a certain topic based on the overall users' affinity towards that topic. (ii) Secondly, we affirm this DoI by correlating it to their friendship network.* Furthermore, we describe our model for DoI computation and follow-back recommendation system by learning a low-dimensional vector representation of users and their disseminated content. This representation is used to train models for prediction of correct cluster classifications. In our experiments, we use a Twitter dataset to validate our approach by computing degrees of interest for certain test users in three diverse and generic topics. Experimental results show the effectiveness of our approach in the extraction of intra-user interests and better accuracy in follow-back recommendations with diversities in the topics.

© 2020 Elsevier Ltd. All rights reserved.

1. Introduction

User interactions and proliferation of citizen journalism on microblogs such as Twitter and Facebook has led to the generation of enormous amounts of online content. The disseminated content on these platforms is also diverse e.g. text, videos, images or intra-user interactions via 'retweets', 'mentions' or 'likes'. Considering the nature of the content and dynamism in intra-user connections, and specifically on Twitter, it is difficult to infer users' interests. Ability to solve this issue is important in recommender systems research. Typically, recommender systems augment this personalization process by suggesting meaningful follower–followee relationships. On Twitter, user interests are majorly extrinsic based on their declared interests. However, dynamism in the disseminated content as well as changes in follower–followee relationships implies that user interests change over time. This poses the following questions : -

1. Is it possible to profile a user/group of users based on their disseminated microblog posts over time?
2. Does homophily in microblog platforms friendship networks influence formulation of online user profiles?

Furthermore, interests that microblog users have towards certain topics change over time e.g. hashtags. Regarding user's friendship networks, it is not correct to assume that a user and his/her followees share the same interests albeit to a certain threshold. It is typical for example for some microblog users to have lots of followers but not all of whom are followees. However, there is need for such users to create a network of influencers to propagate their content as well as receive relevant content from other actors with the same interests.

Therefore, we present a framework to compute the *Degree of Interest* that users may have towards a certain topic based on their disseminated content over time. As a case study, we consider interest in the *sports betting*, *Swahili related chatter* and *daily news chatter* as topics of interest in a generic Twitter dataset.

To the best of our knowledge, this work presents the first attempt at quantifying the degree of interest in a specific microblog topic by analyzing user's short, sparse and noisy content over time. More precisely, the paper made following contributions.

[*] Corresponding author.
E-mail addresses: herman.wandabwa@aut.ac.nz (H.M. Wandabwa), mnaeem@aut.ac.nz (M.A. Naeem), farhaan.mirza@aut.ac.nz (F. Mirza), russel.pears@aut.ac.nz (R. Pears).

https://doi.org/10.1016/j.is.2020.101662
0306-4379/© 2020 Elsevier Ltd. All rights reserved.

1. We develop a framework for formulating user profiles based on short and noisy texts.
2. We affirm the profiling process for the users by correlating it with user's friendship network.
3. We test the framework in three topics of interest, i.e. *sports betting*, *swahili related chatter* and *daily news chatter*. We also carried out an intensive experimental study in formulation of user representative profiles by analyzing their disseminated content over time.

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 and the generation and validation of the results we obtained is presented in Section 4 while results are shown in Section 5. Conclusions and future work are summarized in Section 6.

2. Background and related work

2.1. User preferences modeling/profiling

User interests are major contributors to the content that microblog users are prone to consume or disseminate in addition to them being integral in the design of recommender systems. In essence, recommender systems were developed to match users to resources they are most interested in. This could be in form of content or other users in cases of follow-back frameworks like Twitter. There are two main approaches in the formulation of user profiles in recommender systems i.e. *structural* [1,2] and *behavioral* [3] modeling.

Structural modeling approaches are primarily used to extract features from or stereotype a group of users [4]. In essence, the formulated user profile is structurally representative of the user(s) [4]. In collaborative filtering for example, a user profile can be modeled as a feature vector representing specific aspects of the user as entity descriptors e.g. geo-location in the metadata of a tweet. Similarity-based algorithms [5] can then be applied to measure similarity among users [6]. In the context of content-based recommender systems, a keyword-based vector represents the user profile. A similarity measure is then used to match the distance between the profile and other resources [7].

Behavioral modeling on the other hand identifies observable parameters that reflect user behavior. This can be to a certain level of certainty, thus are probabilistic e.g. Bayesian networks [3]. An extension of the Latent Dirichlet Allocation (LDA) probabilistic approach called Forum-LDA was formulated to extract useful interest topics in online forums by modeling the generative process of seed posts as well relevant and irrelevant responses to the seed posts using a Bernoulli distribution [8].

In light of recommendations in short text microblogs, Chen et al. proposed a tweet's recommendation system based on collaborative ranking to capture personal interests by integrating useful contextual information such as tweet topic level factors [9]. Authors in [10] proposed a methodology for User Interest Profile design by enriching tags generated by the user with friendship information via vector representations. Profiling malicious users in short text microblog users is also an area that has gained traction. Authors in [11] proposed a hybrid approach leveraging classifications and Petri net structure to profile malicious users on Twitter. On the other hand, researchers in [12] proposed a URL recommender system for Twitter users based on social voting and content sources. Generated topics and social interactions were discovered to be significant in presentation of recommendations.

Liang et al. addressed the user profiling problem on Twitter by formulating a dynamic user and word embedding model (DUWE) and a streaming keyword diversification model (SKDM) [13]. The embedding model tracked semantic representations of words and users in the same semantic space for similarity measurement. In addition, the authors propose a streaming keyword diversification model to characterize user profiles over time with top-K keyword. Our work differs from the above in follow-back recommendations which is a key objective in our study. As much as the initial modeling was through vector representations, the formulation of follow-backs as described in Section 5.1 is different. Our approach not only models individual users, but is also able to identify semantically relevant user relationships.

Enrichment of tweets in construction of semantic profiles using OpenCalais[1] ontology for detection of 39 different types of entities e.g. persons, events, products etc was suggested by Gao et al. [14]. Contrary to our work, the quality of the generated profiles is dependent on the quality of the external and abstract ontology. On the other hand, our modeling process is dependent on embeddings that are generated internally in the dataset. Therefore, our modeling process adapts better with decay in the short text microblogs content compared to the OpenCalais based approach.

User follow-back recommendations were also addressed by authors in [15,16] and [17].

2.2. Twitter in topical recommendation systems

Twitter has been instrumental in the dissemination of content in diverse domains. Users on the platform relate better with content that mirrors their timely interests based on the volatility of content on the framework. Hashtags recommendations is one such area in short text microblogs like Twitter. Figueiredo et al. [18] proposed a Topic Relevant Hashtag Identification (TORHID) for the retrieval and identification of hashtags with relevance to a certain topic on Twitter. The authors made use of a seed hashtag and further classifications to remove hashtags with less relevance. This resulted in more relevant and related hashtags that could be used to deepen the initial search.

An improved Twitter-LDA model with the assumption that topics and background words differ per user was proposed by Sasaki et al. [19]. The model was further improved considering the time sequence and capability of online inference based on Topic Tracking Model (TTM) [20], an LDA based probabilistic consumer behavior model for tracking user interests. Our work differs from this work in the modeling process bearing in mind in two ways. Contextual similarity of tweets in Twitter-LDA is an issue as each tweet cannot be assumed to be representative of an independent topic even with the variation of topics and background word per user. Secondly, we formulated a robust homophily measure for follow-back recommendations as in Section 5.1

A hashtag recommender system based on semantic embeddings representation of tweets is another approach for content recommendations on short text microblogs [21]. Authors used a pre-trained Google news based embeddings to represent tweets by averaging per tweet vectors. The features were then clustered using density-based spatial clustering of applications with noise to eventually extract the most similar tweets and recommendation of top-K suitable hashtags in relation to the extracted cluster centroids. As much as the model performed well, a correlation between the language in Google news as well as word-level embeddings are practically not very relevant to the language and word structures in tweets.

On the other hand, Liu et al. [22] proposed a dynamic graph-based embedding model for recommendations of relevant temporal texts and users. The authors modeled a heterogeneous

[1] https://rapidapi.com/.

user–item that evolved as content changed. The model captured temporal relationships and related texts by embedding their representation in a low dimensional space.

Access to Twitter related content in specific topical classifications contributes to a better understanding of tweeter's social patterns and content. Our goal in this research is with regard to improvement of application methods in the evaluation of the degree of interest that users may have towards a topic. Once computed, this metric can be used as input in modeling the tweeter's interest in sports betting.

Our work differs from the works described above with respect to the state-of-the-art given the lack of studies in influence based recommender systems on Twitter's network. In addition to the contributions in Section 1, we contributed to knowledge in user modeling and recommendations as follows: -

1. **Topical affinity in short texts** - Affinity for the topics of interest is what ultimately determines the clustering of users in the same embedding space. In this respect, our modeling approach is able to identify the interests better more so in short and often misspelled words as evidenced in Section 3.1. This presents quality clusters whose centroids are pivotal in computation of affinities from test users in Table 2. This computation process is different from the works in user interests and preferences modeling as detailed in Section 2.1
2. **Follow-back recommendations** - The process of determining the most representative users to follow-back based on a users distance to the centroids of interest distinguishes our work from research in the topical recommendations domain in Section 2.2. This is complemented by the application of the *Theory of homophily* as in Section 4.3.1. We were able to accurately identify contextually representative users proven in Section 5.3 by real tweeters.

We made use of low vector representations in form of sentence embeddings to comprehend the underlying semantic structure of tweets in order to profile tweeters more accurately. In addition, a responsibility matrix in the computation of the interest level in certain topical content. Neural network representations in this instance worked well with misspelled/shortened words, a common occurrence in tweets as well as making intelligent guesses for out of the vocabulary words that had some form of character level consistency with the terms in the vocabulary.

3. Our approach

We present the core processes for the framework to automatically compute the degree of interest in microblog topics of user interest. The framework encompasses computation processes related to *short text modeling* as well as a *follow-back recommendations*. Comprehension of short and informal texts reminiscent of microblog posts presents a challenge for modeling algorithms [23,24]. This is despite their success in modeling of conventional texts [25]. We adopted a neural network approach in modeling text, more so at character level due to the nature of the short texts. Their success in modeling such texts has been demonstrated [26–28]. Generally, most microblog posts encompass shortened and mispelled words in addition to the informality in the language of expression. Therefore, the below processes were influenced by the nature of the dataset and research problem.

1. **Short text modeling** - A corpus of tweets is used to train a model using *FastText*[2] algorithm. The model generates low vector representations of words in the corpus [29]. In *FastText*, the model learns embeddings at character-level thus relevant to the nature of our dataset [30]. This modeling procedure is of choice after a performance comparison with other state-of-the-art methodologies as shown in Table 2.
2. **Extraction of centroids and clustering** - A clustering approach is used to group similar tweets by averaging vectorized intra-word similarities. In this case, cluster centroids are used to infer clusters for new tweets depending on the semantic distance between the model's and test tweets vectors.
3. **Computation of users Degree of Interest (DoI)** - Sample tweets are vectorized via the trained model and our proposed method is applied to measure tweeter's level of interest in the topic based on the disseminated tweets over time.
4. **Association of tweeter's DoIs with their friendship networks** - A correlation measure of a selected list of tweeters and their friendship network is computed. This follows the *homophily* theory where users with similar interests tend to relate, an important process in modeling follow-back recommender system.

3.1. Short text modeling

The adapted modeling methodology was primarily based on the neural language model, *FastText* [30]. The model works by extracting syntactic information in a textual corpus independent of the language of expression, vocabulary size and misspelling reminiscent of tweets. *FastText* creates a model with a vector representation of a word(s) based on the context within which the word is commonly used.

To validate our modeling approach, we trained the same dataset with *Word2Vec* [31] and *Glove* embeddings [32] baselines with different but consistent dimensions across the frameworks. *FastText* unlike other word embedding algorithms, does not ignore the morphology of words. This is a limitation especially in languages with a large vocabulary as well as ones with rare words. With *FastText* modeling, a vector representation is associated with each character n-gram where words are modeled and represented as the sum of character vectors in those words. This makes it the most appropriate algorithm in dealing with mispelled and out of the vocabulary words. It also uses a sliding window of characters in computing word vectors.

Vector representations of words make it possible to infer their contextual usage. This is important in computation of inter-word/sentence distances. Words that are contextually similar are usually used together in expressions like *"good morning"* will have a higher similarity score i.e. close to 1 compared to dissimilar ones. Therefore, a cosine value close to 0 depicts very low term/sentence contextual similarity. We followed the below process to model tweets :-

- **Pre-processing** - For better analysis of unstructured short texts, preprocessing is necessary. For instance, prepositions and punctuation in documents do not provide any meaning, more so contextually; therefore, they were removed. To clean up the tweets which form the model training corpus, we followed the below steps:-
  - Lower cased all words for uniformity.
  - Removed all numbers and encoding accented characters.
  - Removed all hyperlinks in the corpus.

[2] https://fasttext.cc/.

  - Removed all hashtags as they were not of interest in this instance. Hashtags are just words inserted manually by tweeters in a tweet and are prefixed by the hash (#) symbol. Their function is to help identify topically similar tweets.
  - Removal of user mentions. Syntactically, they are presented just like hashtags except that they are prefixed by the @ symbol.
  - Removal of words whose length was less than three characters. Their contextual relevance was not significant in successive experiments.
  - Removal of stopwords. We used the NLTK stopword list.[3]
  - Tweets tokenization, where individual terms in each tweet are split and appended in a list for modeling.

- **Model Training** — A cleaned and tokenized corpus of tweets was the input in the model training pipeline across the state-of-the-art frameworks mentioned above. The models were trained to learn conceptual knowledge of the dataset by mapping each word to a continuous vector space from its distributional properties observed in the corpus.
  The following parameters are to be specified in order to train *FastText* and *Word2Vec* models:

  - *size* or the number of dimensions;
  - *min_count* or minimum count of a word in the corpus for it to be factored in the training;
  - *sg* for training a skip-gram model if $sg = 1$, otherwise Continuous Bag of Words (CBOW).
  - The *window* parameter specified the maximum distance between the current and predicted word in a tweet;
  - *word_ngrams* to enrich word vectors with subword(n-grams) information if specified as 1; and *iter* or iterations which was the number of iterations (epochs) over the corpus.
  - *Glove* model only provisions for the *epochs* and *learning rate(lr)* to be defined.

  The model outputs were vectors for each word in the corpus. Since vector representation in *FastText* is linear, additive compositionality was possible in *FastText* based models. In addition, *FastText-CBOW* uses a distinct vector representation for each word whereas the skip-gram model ignores the internal structure of words. Therefore, each word w is represented as a bag-of-characters *n-grams*. The representation in FastText is such that the word itself is also included in the set of n-grams. *Word2Vec* vector representation works the same way albeit at word-level. In our implementation, $3 \geq n \leq 6$ window of characters was implemented in the FastText based frameworks. This way, most of the n-grams were considered based on average word lengths in tweets. Given a dictionary of n-grams i.e. size D and a word w, let $D_w \subset \{1, \ldots, D\}$. A vector representation z_d is associated with each n-gram d. Therefore, a word is represented as the sum of n-gram vectors. The scoring function as in [29] is formulated as below in this summation process :-

  $$s(w, c) = \sum_{d \subseteq D_w} z_{d_{v_c}}^{\top} \tag{1}$$

  where c is the context position of a word, and v the corresponding word vector.
  A preprocessed tweet ready for vectorization is made up of word tokens i.e. a list of all words split up per tweet. Therefore, its representation is the sum of the tweet's word vectors as they are linear. In our case, let W_x be the set of words in tweet x. The words are then vectorized. Therefore, w_x is the vector representation for a given word in the set. The tweet model w'_x is then represented as below:-

  $$w'_x = \sum_{w_x \subseteq W_x} w_x \tag{2}$$

  Specific values for training the model are elicited in the experimental framework section. At the end of this process, the model is ready for generation of vector representations for each word in the corpus.

3.2. Extraction of centroids and tweets clustering

A correlation of test tweets to the clusters of interest is computed in this step. To delimit the cluster of interest vector space, words in the cluster have to be grouped based on a semantic distance measure. Generally, clusters are defined via approximation and manual inspection of the underlying keywords. Each cluster is represented as a semantic topic where its keywords define it.

Tweet representation w'_x is used to define cluster numbers. In case clusters are to be labeled, then a manual inspection of defining keywords in each cluster was used to infer cluster names. To cluster tweets, *K-Means++* [33] was applied on the training corpus. This clustering algorithm optimizes the choice of cluster centers for k-means by spreading out the initial set of cluster centroids so that they are not close to each other. With the initialization of k-means++, an $O(\log k)$ solution was guaranteed. On the other hand, FastText uses a hierarchical softmax to reduce the computational complexity in the profiling process to $O(hlog(k))$. k is the number of classes and h the dimension of text representation. This made it possible to find the best set of centroids. The clustering insight is that objects within the cluster are as similar as possible, whereas those from different clusters are as different as possible. An optimal clustering process in most cases is dependent on the final purpose of the clusters i.e. the level of detail required of the clusters.

To determine the optimal cluster numbers, we used a heterogeneity convergence metric on the models [34]. We basically ran tests considering different k values (cluster numbers) on the dataset. The number of clusters that best identified the dataset at the elbow point were chosen. The cosine distance measure was then used to compute the intra-cluster distance between y points in a given cluster Z_k and centroid Z_z in that cluster.

The main aim of this phase was to identify words normally used in the cluster of interest and eventually compute its centroid map. Computation of centroids albeit by word as in [35] provided a reference point for computation of the distance between clusters and any other tweet.

3.3. Computation of a User's Degree of Interest (DoI)

To understand how the degree of interest in a cluster was computed, we first computed the similarity of test tweets with respect to cluster centroids.

- **Tweet Similarity to Cluster Centroids:** Computation of cluster centroids was important in classifying tweets with respect to the clusters. Let Y be the set of vectors representing centroids for clusters of interest i.e. $t \in Y$. Y was important in measuring the similarity of any given tweet to the generated clusters. In our case, the interest was to find the semantic distance between a tweet and clusters that semantically represented *sports betting, Swahili chatter* and *daily news chatter* clusters t_{SSC}. Analogy tests and manual

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

inspection of keywords in each cluster as in Section 4.1 informed the topical names of these clusters. Choice of cluster numbers was informed by the results in Table 2. Given a tweet x, let W_x be the set of words in the tweet. To find the vector of x, we computed the sum of vectors for words in W_x as represented in Eq. (2).

Cosine similarity was applied to measure the semantic similarity between a tweet x represented as vector w'_x to cluster centroids $t \in Y$ as earlier defined [36]. Cosine similarity was relevant in our case as similarity between tweets and the cluster centroid maps was possible irrespective of their vector sizes. This is because it measured the cosine of the angle between the two vectors (w'_x and $t \in Y$) projected in a multi-dimensional space. The advantage with cosine similarity measure is that despite the two vectors sets being far apart by say Euclidean distance due to their size, chances are that they may still be semantically close. The smaller the angle between the two sets, the higher the cosine similarity. Therefore two objects are presumed very similar if the cosine distance is close or equal to 1 and dissimilar if close or equal to 0.

$$s_{xt} = CosineDistance(w'_x, t), \forall t \in Y \tag{3}$$

Eq. (3) presents the cosine similarity computation. Therefore, it was possible to compute the semantic relevance of a tweet to clusters ($s_x t_{SCC}$).

- **DoI Computation:** Computing the DoI of a given tweeter u involved a few steps. *(i)* We first extracted tweets that user u posted on his/her timeline T via Twitter's Search API.[4] The API has the capability of returning a maximum of 3200 tweets for any given Twitter account.*(ii)* The second step involved preprocessing of tweets $x_u \in T_u$ as in Section 3.1. Similarity of the extracted tweets x_u to the clusters of interest t_{SCC} as in our case study is then computed. The Degree of Interest in the Sports betting, Swahili and Daily news related content clusters (DoiSCC) is then calculated as the average of vector similarities $s_{x_u t_{SCC}}$. Therefore,

$$DoiSCC_u = average(s_{x_u t_{SCC}}), \forall x_u \in T_u, t_{SCC} \in Y \tag{4}$$

Since the cosine distance was used, therefore, a tweeter's DoiSCC value close to 1 in relation to any of the clusters meant that the user's tweets were mostly inclined towards that specific topic. Values close or equal to 0 on the other hand meant minimal user interest to the domain. Algorithm 1 details the above processes.

3.4. Tweeter's DoiSCC vs their friendship network

A tweeter's profile is defined by the content he/she shares as well as with the intra-user relationships that exist from the 'mentions', 're-tweets' and 'follower' relationships. In essence, there theory of homophily is actualized in social networks, more so in Twitter with the above mentioned relationship entities following its follower–followee relationship on the platform [37]. Ideally, this is the same principle of "birds of the same feather" where users with either unidirectional or bidirectional relationship on the platform have shared common interests. Therefore, the assumption is that users with interest in *Swahili relate content* are likely to have friends with interest in related content.

For evaluation purposes in our approach, we computed the degree of interest for the friendship network of the initial users. This followed same process as in Algorithm 1 after downloading tweets from the timelines of the friends. Friendship network in this case is any tweeter who was "mentioned","retweeted" or "followed" the original tweeters who disseminated the 298835 tweets as in Section 4.1. To obtain tweets from the friendship network, the below processes were followed:-

1. Searched for tweeters fulfilling the friendship network conditions as above.
2. Downloaded a maximum of 3200 tweets and related metadata per user using Twitter's search API.[5]

The friendship network's DoI computation process follows the same modeling pattern as in Section 3.3. The end result is a computation of the friendship network DoIs and a comparative evaluation with the original user's DoI as in Section 5.1. This informs the DoI recommendations thresholds for users and their friendship networks. The results are fundamental in the design of short text based recommender systems for users, third-party content or groups and lists.

From Algorithm 1 inputs are tweet tokens W_x, that are vectorized and summed as w'_x (Line 4). Similarity of a tweet to a centroid t is computed by measuring the cosine angle between tweet vectors and the centroid of interest in centroid map Y (Line 8). The cosine distance is computed for all tweets under each user (Line 12). To compute the DoI, a summation of individual tweet DoIs for each user is made (Line 14). The average DoI (DoiSCC) across the tweets i.e. $DoiSCC = \{\sum DoI\}/len(x)$ (Line 14) is the end result in the computation i.e. a representation of the user's as well as friend's interest in the topic of interest.

Algorithm 1 User DoI Computation Process

```
Require: Tweet Tokens $W_x$, Word vectors $w_x$, Tweet Model $w'_x$,
    Cluster of Interest $t$, Centroid Map $Y$
Ensure: Tweeters DoI $U_{DoI}$
 1: Initialization via K-Means++                     ▷ Word Vectors $w_x$
 2: Computation of Tweet Model $w'_x$
 3: for $w_x \in W_x$ do
 4:     $w'_x = \sum_{w_x \subseteq W_x} w_x$        ▷ Tweet Model as the sum of word
    vectors $w_x$ in tweet $x$
 5: end for
 6: Similarity of a tweet to a cluster of interest in Centroid Map
 7: for $t \in Y$ do
 8:     $s_{xt} = Cosine(w'_x, t), \forall t \in Y$  ▷ Computation of the
    cosine distance between the tweet's model $w'_x$ and cluster of
    interest $t$ in the Centroid Map $Y$
 9: end for
10: Degree of Interest (DoI) Computation
11: for $x_u \in T_u$ do
12:     $DoI+ = 1 - Cosine(x_{uv}, t_{SCC})$         ▷ Similarity via cosine
    measure between tweets $x_u$ with the clusters of interest $t_{SCC}$.
13: end for
14: $U_{DoI} = \{\sum DoI\}/len(x_u)$                ▷ Compute the Degree
    of Interest (DoI) for the tweeter by computing sum of a user's
    tweet level DOIs $DoI$ divided by the count of user's vectorized
    tweets $x_u$.
```

Computation of the friendship network *DoiSCCs* followed the same process as in Algorithm 1 except that the input was tweeters usernames in their respective clusters as illustrated in Section 3.4.

This process is illustrated in the computational workflow in Fig. 1. Friendship network as the input consists of all users who

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

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

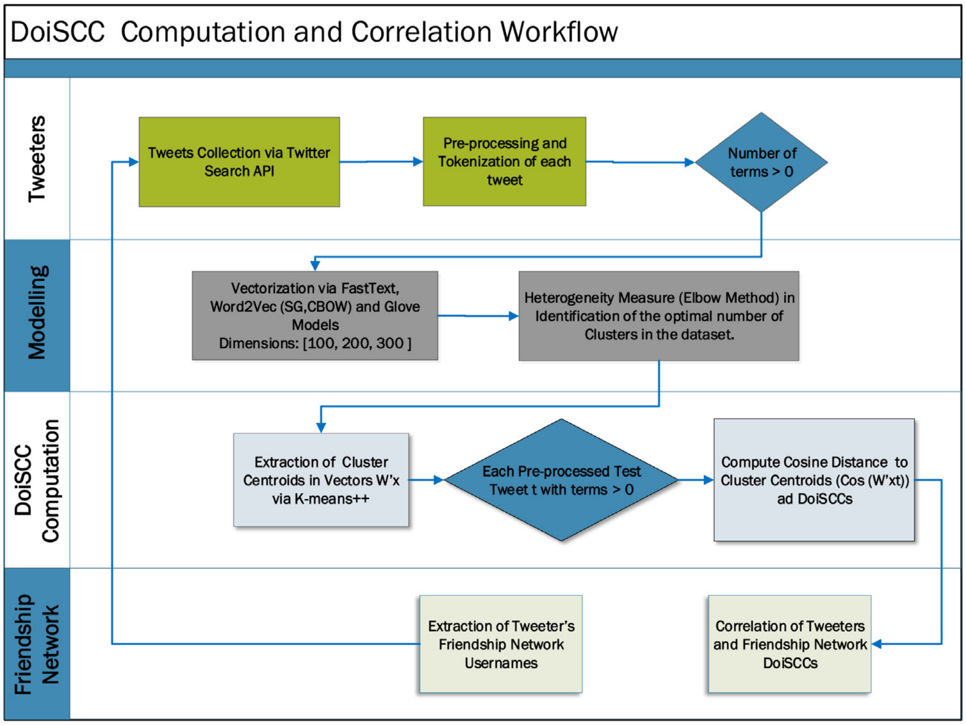

DoiSCC Computation and Correlation Workflow
Tweeters
Modelling
DoiSCC Computation
Friendship Network
Tweets Collection via Twitter Search API
Pre-processing and Tokenization of each tweet
Number of terms > 0
Vectorization via FastText, Word2Vec (SG,CBOW) and Glove Models Dimensions: [100, 200, 300]
Heterogeneity Measure (Elbow Method) in Identification of the optimal number of Clusters in the dataset.
Extraction of Cluster Centroids in Vectors W'x via K-means++
Each Pre-processed Test Tweet t with terms > 0
Compute Cosine Distance to Cluster Centroids (Cos (W'xt)) ad DoiSCCs
Extraction of Tweeter's Friendship Network Usernames
Correlation of Tweeters and Friendship Network DoiSCCs

Fig. 1. Computation workflow to generate tweeters and friendship network DoiSCCs.

were *mentioned by tweeters*, *replied to their tweets*, *their bidirectional followers* and those who *retweeted their content*. We limited the collection to 50 tweets per user due to Twitter search API limitations. The below process followed thereafter:-

1. Pre-processing of the extracted set of tweets as described in Section 3.1 and vectorization of tweet tokens from the pre-processing step.
2. Computation of the test tweet's similarity to the extracted model clusters where the distance between the pre-processed tweet vectors and cluster centroids is calculated. This process as described in Algorithm 1 is computed via cosine distance with the cosine value being the DoiSCC.

4. Experimentation

This section elicits the processes followed to validate the proposed approach presented in Section 3. The collected datasets for both tweeters and their friendship network mirrored real Twitter intra-user interactions. The data collection process and experimentation are described in the sections below : -

4.1. Datasets, settings and analogy tests

We collected 298835 unique and generic tweets in JSON format for a period of six months starting 1/9/2018. Each entry in the dataset was a single tweet with its associated metadata i.e. geo, mentions, hashtags etc. Retweets were filtered out from the collection. 90% of the collected tweets were written in English and 6% in Swahili. The remaining set of tweets were comprised of a mixture of English and Swahili vocabulary.

4.1.1. Ground truth - Sports betting data

To accurately measure the topical interest in the generic dataset in Section 4.1, ground truth was required. Therefore, we curate our data further by adding *sports betting* tweets to the generic dataset that we had collected. The purpose of this process was to make sure that a cluster that ground truth can be based on is present. This is important in the overall evaluation of the modeling frameworks. Therefore, a pool of 50639 unique sports betting related tweets were added to the generic dataset. These unique betting related tweets were collected from timelines of a few sports betting companies. They included tweets associated with *sportpesa*,[6] *betin*,[7] *eazibet*,[8] *betika*[9] and *betwayke*[10] Twitter handles. The total number of tweets in the training corpus was 349474.

4.1.2. Analogy tests

Fact checking through *analogy* tests validated the generalization and quality of the corpus and trained models. Therefore, we conducted several qualitative tests on the optimized models to ascertain their relevance to the test scenario. Table 1, summarizes validation examples in the context of *sports betting* and *politics* respectively. In these diverse examples, *odds*.[11] is a term used in the betting industry. On the other hand, *uhuru*[12] is a politician in Kenya. We used the terms to compute inter-term similarity using the models. Top five most similar terms per cosine distance were then generated per model and with different dimensions as shown in Table 1 Its worth noting that the values in this table are simply guidance to the generalization process for users and do not depict comparative inter-model accuracy. For example, some words may have higher values as distance to the test terms though may be irrelevant contextually to the term. Therefore, the analogy tests in Table 1 just guided model choices for further evaluation by the authors.

[6] https://www.sportpesa.org/.
[7] https://www.betin.co.ke/.
[8] https://www.eazibet.co.ke.
[9] https://www.betika.com/.
[10] https://www.betway.co.ke.
[11] https://mybettingsites.co.uk/learn/betting-odds-explained/.
[12] https://en.wikipedia.org/wiki/Uhuru_Kenyatta.

Table 1
Sample analogy test with two terms; *odds* - a betting Markets related term and *uhuru*, the president and politician in Kenya in models of *100,200* and *300* dimensions.

Model dimensions	Most similar to “odds” - (Betting term)						Most similar to “uhuru” - (Politician in Kenya)					
	100		200		300		100		200		300	
	Word	Relevance	Word	Relevance	Word	Relevance	Word	Relevance	Word	Relevance	Word	Relevance
Glove	Finns	0.84	Finns	0.79	Finns	0.76	Wheelpicinns	0.95	Wheelpicinns	0.94	Wheelpicinns	0.92
	Available	0.66	Acesse	0.61	Acesse	0.582	Graftheadlines	0.83	Opanga	0.77	Opanga	0.74
	Acesse	0.65	Defy	0.60	Noen	0.581	Opanga	0.80	Graftheadlines	0.76	Graftheadlines	0.73
	Defy	0.64	Noen	0.59	Defy	0.576	Kenyatta	0.79	Deactivated	0.75	Exits	0.72
	Competition	0.62	Competition	0.58	Competition	0.55	Deactivated	0.76	Exits	0.73	Scolds	0.64
Word2Vec-SG	2odds	0.66	Price	0.50	Sportpesatips	0.49	Uhuru	0.65	Ruto	0.60	Deactivated	0.44
	Guaranteed	0.65	Markets	0.443	Betnba	0.48	Kenyatta	0.59	Uhuru	0.55	Kenyatta	0.439
	Bets	0.634	Bets	0.44	Chekiodds	0.46	Kamama	0.56	dp	0.50	Inaleta	0.423
	Price	0.633	Stake	0.43	In-play	0.459	Corrupt	0.55	Aache	0.46	Ocsragira	0.411
	2020qualifiers	0.63	Evens	0.40	Evens	0.45	President	0.545	Murkomen	0.44	Puga	0.407
Word2Vec-CBOW	stake	0.58	Price	0.50	Price	0.46	Ruto	0.64	Ruto	0.52	Ruto	0.55
	Price	0.57	Markets	0.44	Stake	0.42	Uhuru	0.63	Dp	0.47	Dp	0.45
	Games	0.53	Bets	0.42	Markets	0.40	Dp	0.58	Murkomen	0.44	Alisema	0.409
	bets	0.52	Evens	0.41	Bets	0.37	Ruto	0.54	Jubilee	0.436	Hamjui	0.408
	Markets	0.50	Prices	0.39	Evens	0.36	Raila	0.50	Raila	0.428	Murkomen	0.40
FastText-CBOW	2odds	0.72	2odds	0.72	2odds	0.70	UhuRuto	0.82	Uhurus	0.793	Uhurus	0.808
	Bestodds	0.69	3odds	0.67	3odds	0.68	Uhurus	0.81	UhuRuto	0.77	UhuRuto	0.75
	80/1	0.68	Bestodds	0.64	Oddset	0.611	UhuruKenyatta	0.76	UhuruKenyatta	0.72	UhuruKenyatta	0.652
	3odds	0.67	Odds-on	0.63	Odds-on	0.60	Kenyatta	0.756	Huru	0.68	Huru	0.648
	Chekiodds	0.66	Chekiodds	0.62	Chekiodds	0.583	uKenyatta	0.698	Kenyatta	0.65	Uhuruhighway	0.61
FastText-SkipGram	2odds	0.97	2odds	0.968	2odds	0.966	Uuru	0.88	Huru	0.89	Huru	0.89
	3odds	0.965	3odds	0.967	3odds	0.965	Odm-Uhuru	0.87	Odm-Uhuru	0.86	Odm-Uhuru	0.86
	Chekiodds	0.88	Chekiodds	0.87	Odds-on	0.889	Ushuru	0.869	Ushuru	0.85	Ushuru	0.847
	Oddset	0.855	Oddset	0.86	Chekiodds	0.868	Uhurus	0.855	Uhurus	0.84	Uhurus	0.836
	Bestodds	0.76	Bestodds	0.74	Oddset	0.859	UhuRuto	0.85	UhuRuto	0.82	UhuRuto	0.82

4.2. Ground truth tweet samples

Analogy tests in Section 4.1.2 provided a general semantic overview of the dataset. However, before selecting the best modeling framework for our research problem, the models had to be subjected to a known dataset. Therefore, we sampled 1000 tweets from the five Twitter handles of betting companies mentioned in Section 4.1.1 extracted via Twitter's search API. Manual inspection of the 1000 tweets was done by three judges to make sure that all tweets were semantically close to the sports betting domain at least by human understanding. This sample dataset was set up purposefully to test the accuracy of the approaches in terms of best parameter sets in relation to cluster numbers. Cosine distances between the sample tweets and the centroids of the sports betting clusters in models were then computed. This part was significant in helping achieve two objectives:-

1. Identification of the most representative model and dimension sizes in training the models.
2. Identification of the number of clusters that were the best representation of the corpus.

The assumption in this case was that the more the number of correctly classified tweets in relation to the sports betting cluster, the better the parameter adjustments. Therefore, it was a matter of iterative trialing of varied parameters in the models. Cluster assignment followed the process of centroids extraction and clustering as described in Section 3. Regarding the affinity of tweeters to the sports betting cluster, we had to verify that the interest per sports bettor was greater than that of the other clusters. This was important in the verification of the trained model accuracy as well as optimization of model parameters. The sample tweets we collected were modeled with *FastText-CBOW*, *FastText-SkipGram(SG)*, *Word2Vec-CBOW*, *Word2Vec- SkipGram(SG)* and *Glove* state-of-the-art baselines trained with 100, 200 and 300 dimensions consistent with [38]. We tested the models with the number of clusters set to 3, 4, 5 and 6 based on the elbow method heterogeneity measure for identification of cluster numbers befitting the dataset [34]. The accuracy of classifications followed a two-step process: -

- Comparative evaluation with known sports betting cluster labels in all the models, with different cluster numbers. For example, FastText's *cluster 0* represented the *Sports Betting Content*, *cluster 1* represented the *Swahili Related Chatter*, and *cluster 2* represented the *Daily News Chatter*. This was based on comparative analogy tests and manual vocabulary inspection in each of the hard clusters. The assumption was that the test set had tweets that closely inclined towards *cluster 0*, thus 0 was assumed to be the true cluster labels(ground truth). These labels were then compared with the predicted labels with different dimensions and cluster numbers. We used the Fowlkes–Mallows Index (FMI Score) [39] to compute this comparison. The FMI Score is the geometric mean of the pairwise precision and recall between the true and predicted labels. The score ranges from 0 to 1. A higher value indicates better similarity between two points.
- We also computed the Silhouette Coefficient (S-Score) for each model with vectors of the test tweets, assuming the ground truth unknown [40]. The computed Silhouette Coefficient consisted of two scores: *(a) The mean distance between a sample and all other points in the same topical cluster. (b) The mean distance between a sample and all other points in the next nearest cluster.* The best value is 1 indicating the best cluster quality and the worst value is −1.

We report the values in Table 2. *FastText-SkipGram* with 100 dimensions and 3 clusters, reported the highest FMI and S-Scores for the sports betting cluster. The general consensus was that, lower FMI and S-Scores were recorded in models with more than 3 clusters. Therefore, we selected *FastText-SkipGram* with 100 dimensions and 3 clusters to further compute *DoiSCCs*. For comparative validation against related state-of-the-art methodologies, *FastText-CBOW (100,3)*, *Word2Vec-CBOW (200,3)* and *Word2Vec-SG (100,3)* were also selected for validation based on the consistency in their **FMI** and **S-Scores**.

From the values in Table 2, Glove models were not selected for further evaluation because of the inconsistencies in the FMI and S-Score values. For example, Glove model with 300 dimensions and 3 clusters had the highest FMI-Score but also the lowest S-Score for the same setting thus skipped for selection. The highest and consistent values in each modeling framework are highlighted in bold text in the table. In addition, the above results meant that the dataset's best representation was three topics especially with the consistency in the reduction of S-Score across the models and cluster numbers. For the baseline, we made use of a Bag-of-Words (BOW) Kmeans based clustering algorithm [41]. The maximum iterations were 100 across the dataset. As much as the FMI score was quite high compared to the other models, the S-Scores were quite on the contrary across the clusters compared to the other models. This was an indication of a subpar performance in terms of the clusters quality. Therefore, the models with the highest and consistent FMI and S-Scores across the clusters were selected for further evaluation.

4.3. Test set collection and computation

Computation of *DoiSCCs* for sample tweeters in Section 4.1 entailed the collection of tweets disseminated by the very tweeters. The intuition in this experiment is to have generic tweeters who might fit this profiling. A maximum of 3200 tweets from each tweeter were collected via Twitter's search API. The numbers varied depending on how many tweets individual users had disseminated over the time period of interest. We sampled and collected tweets from 200 tweeters between 1/1/2019 to 1/04/2019. There assumption was that there was a possibility of identifying generic tweeters who could have been disseminating content related to the topics of interest in this study.

4.3.1. Homophily in tweeter's friendship network

Analysis of a tweeter's friendship network (*mentions, retweets, replies, lists* and *hashtags*) helps identify whether the presented tweeter's profile was relevant in terms of positive correlation with the *DoiSCCs* of his/her friends. Homophily [42] is evident in social networks based on the fact that users tend to follow those whom they share interests with. We extracted 62,275 tweets from the timelines of tweeters who bidirectionally retweeted, listed, followed and were mentioned by the 200 tweeters as detailed in Section 4.1. This was to compute their *DoiSCCs* and correlation with the *DoiSCCs* of their friendship network. The assumption was that a positive correlation to a large extent indicated better performance in the chosen modeling framework in terms of profiling tweeters with respect to the generated topics of interest, thus their topical affinity.

4.4. Parameter settings and experiments

The optimized *FastText-SkipGram* model had the following parameters setup: *size = 100, minimum count = 2, learning rate (lr) = 0.1* and *iteration(iter) =30*. Minimum count of characters was two. Words with lesser character counts were excluded in the modeling as they were likely to be stop words. Descriptions of the above

Table 2
Models classification scores with respect to model dimensions (*100,200,300*) consistent to [38] and cluster numbers (*3, 4, 5* and *6*).

		Fowlkes–Mallows scores(FMI), Silhouette Coefficient (S-Score)							
	#Clusters	*3*		*4*		*5*		*6*	
	Dim	FMI Score	S-Score	FMI Score	S-Score	FMI Score	S-Score	FMI Score	S-Score
FastText-Skip Gram	100	**0.65**	**0.21**	0.50	0.16	0.48	0.13	0.43	0.19
	200	0.58	0.15	0.52	0.14	0.46	0.13	0.42	0.13
	300	0.62	0.18	0.57	0.11	0.48	0.14	0.50	0.15
FastText-CBOW	100	**0.59**	**0.15**	0.52	0.16	0.53	0.15	0.43	0.18
	200	0.59	0.13	0.50	0.16	0.48	0.14	0.47	0.17
	300	0.58	0.13	0.53	0.16	0.50	0.14	0.43	0.17
Glove	100	0.57	0.12	0.53	0.12	0.45	0.11	0.45	0.12
	200	0.58	0.13	0.53	0.08	0.46	0.13	0.49	0.11
	300	0.59	0.10	0.56	0.11	0.52	0.10	0.44	0.13
Word2Vec-SkipGram	100	**0.59**	**0.15**	0.53	0.17	0.50	0.13	0.45	0.19
	200	0.58	0.13	0.50	0.14	0.45	0.11	0.43	0.12
	300	0.59	0.12	0.52	0.13	0.51	0.14	0.53	0.13
Word2Vec-CBOW	100	0.57	0.14	0.52	0.15	0.51	0.12	0.41	0.14
	200	**0.57**	**0.15**	0.51	0.14	0.47	0.12	0.42	0.13
	300	0.57	0.13	0.53	0.13	0.45	0.12	0.44	0.13
BOW (KMeans) baseline	100 Iterations	**0.72**	**0.019**	0.59	0.015	0.48	0.016	0.53	0.016

parameters are given in Section 3. Other parameters that are not explicitly specified, assumed *FastText's* and *Word2Vec's* default parameters. The model outputs were vector representations of vocabulary in the 62,275 tweets. The number of clusters as well as the initialization mechanism via k-means++ was specified in the clustering process to generate cluster centroid maps. Basically, the centroid maps are a list of terms in the dictionary embedded with their respective cluster indexes.

The training corpus in Section 3 was used to model tweets and tweeters. The most representative number of clusters in the chosen modeling framework i.e. *FastText-SG* with 100 was 3 as per the results in Table 2. Each of the clusters had a unique identifier. *Cluster 0* represented the *Sports Betting Content*, *Cluster 1*, *Swahili Related Chatter*, and *Cluster 2*, *Daily News Chatter* (**DoiSCC**). This was after manual inspection of the vocabulary under each cluster as well as results from test terms analogy tests on the model. One assumption relating to *Cluster 0* i.e. *Sports Betting Content* was that this particular cluster was expected as we added content relevant to this domain in the generic dataset as ground truth for validation. However, *Cluster 1*, *Swahili Related Chatter*, and *Cluster 2*, *Daily News Chatter* clusters were generalized from the dataset by the modeling framework.

The resultant vectors from the modeling frameworks were then used to compute the tweet-cluster similarity. The similarity as pointed out earlier is the distance between the average tweet vector and the centroid map of interest. This process on a sample tweet is illustrated below:-

- **Original tweet** - *Away Win 3 Multibet Football Tips Odds Kenya January 11 2019* http://www.zuribet.com/away-win-3-multibet-football-tips-odds-kenya-january-11-2019/pic.twitter.com/1at2nLy8je
- **Preprocessed Tweet** - [*away, multibet, football, tips, odds, kenya, january*]
- **Cluster Similarity values** - [**0.496**,0.196,0.434] where the value in the array index 0 represents the similarity of the tweet to the sports betting related cluster. The second value is in relation to *Swahili related chatter* and the third, *daily news chatter*. From the above example, the modeled tweet had seven terms. Kenya was a dominant term in *Cluster 1* (*Swahili related chatter*) justifying the score of 0.196. On the other hand, words like *football* and *away* were present in clusters with daily news updates and betting. This justifies the closeness in their similarities scores.

From the output, we can infer that the above tweet was more closely related to the sports betting domain than the other two clusters.

5. Results

5.1. Group recommendations

Social media users tend to have ties with their friends based on follower–followee relationships. Therefore, the assumption is that users with such relationships tend to share interests and by extension *DoiSCC* values for both groups. The assumption is that user *DoiSCCs* for tweeters and their friendship network correlate.

In this evaluation, group topical recommendations were of choice as opposed to individual analyses. Grouping and correlating *DoiSCCs* for both tweeters and their friendship network was the best way to represent such recommendations. Ideally, its easier to make recommendations over a large spectrum of users as compared to individuals. The logic is that users with 10 to 15 percent interest in certain content, would better be grouped together as their interests similarities match. Each modeling framework's vector representation differed in the computation of topical affinity of users to the three topics of interest. Therefore, *DoiSCC* values also differed. The overall distribution per model for the *DOiSCC* values was computed to inform group-level recommendations. The values varied per topic simply because interest in some topical clusters e.g. *daily news chatter* was greater than the *Swahili chatter* across the modeling frameworks. Density distributions for each of the model's *DoiSCCs* in each topical cluster were computed to inform the grouping process as shown in Table 3. For example, with FastText-SkipGram(100,3), interest in *sports betting* content ranged from 0 to 0.3 while in *daily news chatter*, it ranged between 0.4 to 0.8. In as much as the dataset was enriched with *sports betting* related tweets, a vast majority of tweets depicted a higher similarity to the *daily news chatter* topical cluster. This can be quantified by observing tweeting patterns where tweets mostly related to daily news were shared. One other reason is that the vocabulary in the two cluster centroids were also contextually shared. The latter values meant that the model was able to accurately identify users with interest in daily news. However, interest in *Swahili related chatter* was quite low i.e. between 0 to 0.2 across the modeling frameworks. The dashes (-) in Table 3 meant that the model did not find *DoiSCC* for that

Table 3
DoiSCCs with respect to the modeling frameworks and topical clusters in the dataset.

Modeling framework	FastText- CBOW (100,3)			FastText- SG (100,3)			Word2Vec- SG (100,3)			Word2Vec- CBOW (200,3)			BOW - KMeans		
Topical cluster	Sports betting (DoiSB)	Swahili related chatter (DoiSRC)	Daily news chatter (DoiDNC)	Sports betting (DoiSB)	Swahili related chatter (DoiSRC)	Daily news chatter (DoiDNC)	Sports betting (DoiSB)	Swahili related chatter (DoiSRC)	Daily news chatter (DoiDNC)	Sports betting (DoiSB)	Swahili related chatter (DoiSRC)	Daily news chatter (DoiDNC)	Sports betting (DoiSB)	Swahili related chatter (DoiSRC)	Daily news chatter (DoiDNC)
Group 1	0.0	–	0.0–0.1	0.0	0.0–0.1	0.4–0.5	0.0	–	0.4–0.5	0.0	0.0–0.1	0.0–0.1	0.0–0.5	0.8–0.9	0.0–0.1
Group 2	0.0–0.1	–	0.1–0.2	0.0–0.1	0.1–0.2	0.5–0.6	0.0–0.1	–	0.5–0.6	0.0–0.1	0.1–0.2	0.1–0.2	0.5–0.7	–	0.1–0.3
Group 3	0.1–0.2	–	0.2–0.3	0.1–0.2	–	0.6–0.7	0.1–0.2	–	–	0.1–0.2	–	0.2–0.3	0.7–0.9	–	0.3–0.5
Group 4	0.2–0.3	–	0.3–0.4	0.2–0.3	–	0.7–0.8	–	–	–	0.2–0.3	–	0.3–0.4	–	–	0.5–0.7
Group 5	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.7–0.9

specific topical cluster. For example, *FastText-CBOW(100,3)* and *Word2Vec-SG (100,3)* could not positively identify *Swahili related* topical content resulting in null values in the interest groups.

5.2. Follow-back recommendations for tweeters vs friendship network

Following the theory of homophily, tweeters and their friendship network *DoiSCCs* should ideally be close. The *DoiSCCs* extraction process follows the same procedure as in Section 3.3. In our setup in Table 3, interest computation between tweeters and their friendship network was in relation to the three topical clusters. The three topics of interest were the most optimal in the dataset as per the FMI and S-Scores in Table 2 thus were not manually selected.

Therefore, follow-back recommendations were pegged on the group relations between tweeters and their friendship network in terms of *DoiSCCs*. This in essence details the correlation of group *DoiSCCs* in both tweeters and their network to the three clusters of interest like in our case. The DoI generation process is synonymous to the one in Section 3.4.

To analyze these correlations better, we present intra-user *DoiSCCs* correlations in relation to the topical clusters in Sections 5.2.1–5.2.3.

5.2.1. Sports Betting Topical Cluster Interest (DoiSB)

Results in Fig. 2 show the correlation distribution between the *DoiSB* of tweeters and their friendship network in the models. A comparative evaluation was made with four other modeling frameworks for validation. From the box plots in Fig. 2, correlations between the different groups' *DoiSBs* showed marginal variance to those of their friendship network. In the case of *FastText-SkipGram (100,3)*, tweeters with *DoiSBs* equal to 0 in Fig. 2(a) correlated with friends whose *DoiSBs* median is approximately 0.08 . The same can be observed for the second group i.e. with tweeters' *DoiSBs* greater than 0 and less than or equal to 0.1. Tweeters with *DoiSBs* between 0.1 and 0.2 correlated better with friends whose *DoiSBs* were about 0.09. Considering the last group i.e. tweeters with *DoiSBs* between 0.2 and 0.3 were expected to show higher friendship correlation values. However, their best set of friendships were the same as those in the first and second group i.e. with $DoiSB = 0.08$. This was attributed to the number of tweeters in this specific group. Generally, they were fewer compared to the ones in Group 3 in the dataset. This further influenced their friendship connections thus explains Group 4 results in Fig. 2(a). Results in Fig. 2(e) are a bit inconsistent with vector representations as the algorithm is purely based on the bag-of-words approach. This explains why the *DoiSBs* are quite high across the groups. This may seem great results but the fact that no semantics are factored in the model, overfitting in the word re-usage across the classifications becomes a problem. This explains the inconsistencies in Table 3 when humans evaluate the same model and in the FMI-Scores for BOW in Table 2. Therefore, in as much as the model exhibited positivity in correlations, the same could not be justified when humans validated the same model. In addition the rest of the other models except *FastText-SkipGram (100,3)* depicted negative correlations between tweeters' and their friendship networks' *DoiSBs* as much as they were contextually similar to it. This shows that as much as some tweeters had interest in online sports betting, their friends did not, which contradicted the *homophily theory*. This can also be partly attributed to the model quality which is evident in Table 2 for the models in Figs. 2(b), (c) and (d). Fig. 2(a) shows otherwise.

5.2.2. Swahili Related Chatter Topical Cluster Interest ((DoiSRCs)

Results in Fig. 3 show the same intra-user correlations with regard to the Swahili Related Chatter (*DoiSRC*) topical cluster. From the results, only modeling frameworks based on *FastText-SkipGram (100,3)* and *Word2Vec-CBOW(200,3)* depicted a positive correlation in their *DoiSRCs*. *Word2Vec-SkipGram(100,3)*, *FastText-CBOW(200,3)* and the Bag of Words (BOW) baseline frameworks could not entirely model any positive correlations in relation to the *Swahili Related Chatter* topical cluster as evidenced in Table 3. In Fig. 2, *FastText-SkipGram (100,3)* depicted the most positive correlation in the values thus was the best performing framework in the identification of the tweeters and their friendship network propagating Swahili related content.

5.2.3. Daily News Chatter Topical Cluster Interest (DoiDNC)

We computed the Degree of Interest in *Daily News Chatter* (*DoiDNC*) in the relation to the third and final topical cluster. This cluster entailed users who disseminated content related to daily news over the collection period. Due to the volatile nature of content dissemination patterns on short text microblogs, citizen journalism has been on the rise. Tweeters share daily news items with their networks forming a large chunk of these news related topical cluster. Since the *Daily News Chatter* was one of the extracted generic clusters from the vector representation of the dataset, there was a need to compute the interest in the same. The aim just like in Sections Section 5.2.1, 5.2.2 was to correlate the interest of users and their friendship network in news related content.

From the results in Fig. 4, *FastText-SkipGram (100,3) (4a)* modeling framework extracted and correlated the interests better than the other three models shown in the figure. *DoiDNC* values were the highest of all the modeling frameworks as well as the depiction of better correlations in both tweeters and friendship network. *Word2Vec-SkipGram(100,3)* only identified *DoiDNCs* falling in two categories as evidenced in Table 3. On the other hand, *FastText-CBOW(100,3) (4b)* and *Word2Vec-CBOW(200,3) (4c)* had the lowest *DoiDNCs* for both tweeters and their friendship network.

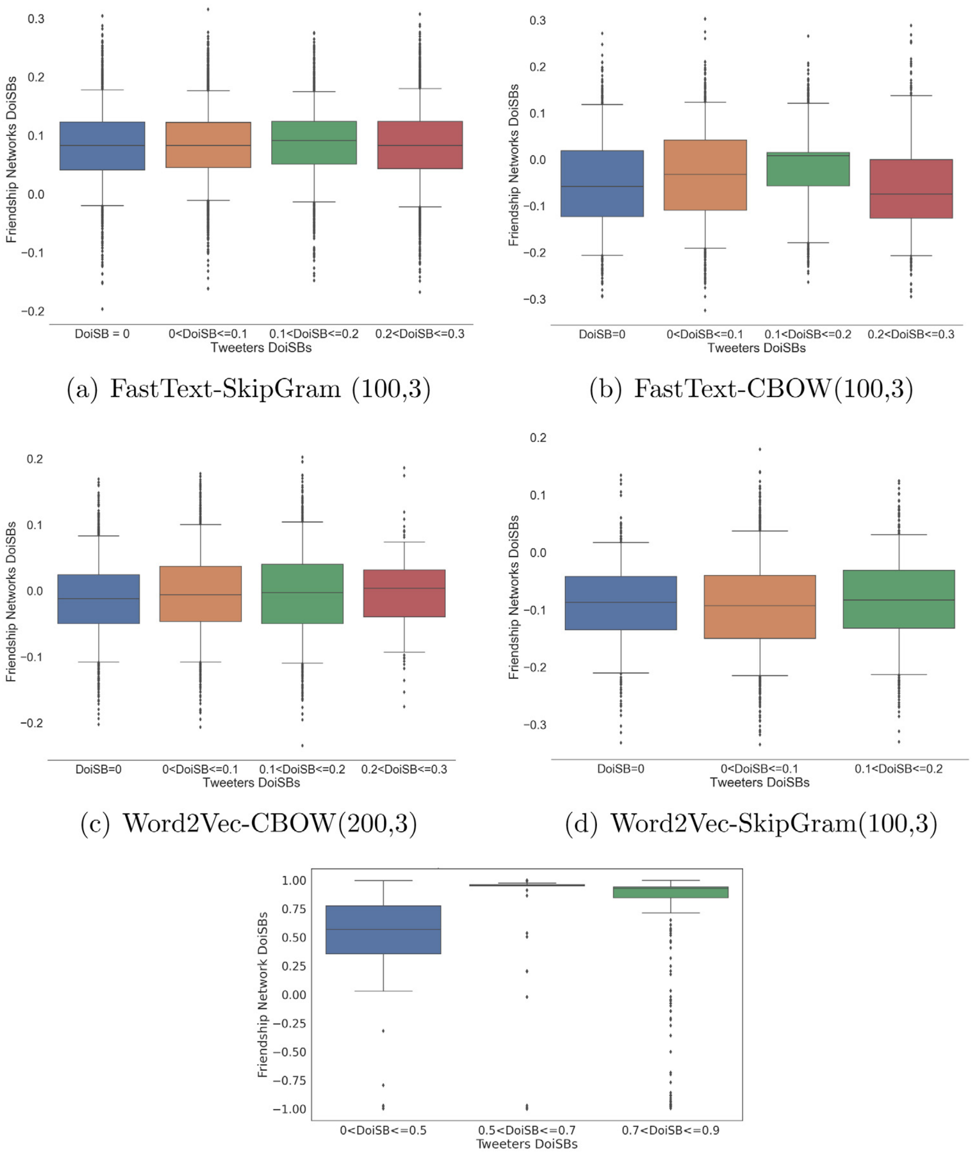
Evaluations in Sections 5.2.1, 5.2.2, 5.2.3 justify our choice of the modeling framework. *FastText-SkipGram (100,3)* framework depicted the best performance among vector representation techniques as well as with the *Bag-of-Words* baseline.

5.3. Qualitative evaluation in follow-back recommendations

Homophily is the tendency of users to have positive ties with other similar users in socially significant ways. Currently, the term is used to refer to an observable behavioral tendency whose causes can include preference or opportunity [43].

5.3.1. Homophily in short text microblog users

To evaluate the results qualitatively, we selected a *DoiSCC* topical classification group that was most consistent across the modeling frameworks. Group 3 ($0.1 < DoiSBs <= 0.2$) in Table 3 under the *sports betting* topical cluster fit this evaluation. We also had ground truth tweets in the sports betting domain added to the dataset as in Section 4.1.1. This means that users with 0.1 to 0.2 interest in the sports betting domain could be identified in the dataset despite the modeling algorithm. This is partly attributed to the introduction of the ground truth sample data as described in Section 4.2. We followed the below process in the selection of tweeter's and their friendship network tweets for *DoiSBs* computation and correlation : -



(a) FastText-SkipGram (100,3)

(b) FastText-CBOW(100,3)

(c) Word2Vec-CBOW(200,3)

(d) Word2Vec-SkipGram(100,3)

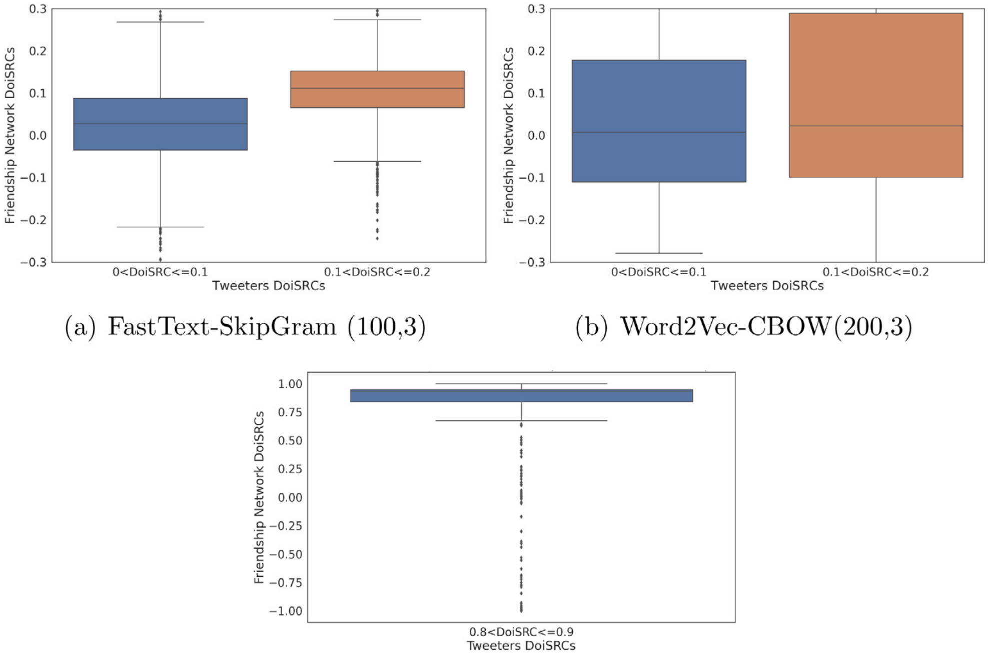
(e) BOW-KMeans Baseline

Fig. 2. Correlation of tweeters and their friendship network Degree of Interest in Sports Betting (DoiSB) in four models.

1. Three random users with $0.1 < DoiSBs <= 0.2$ and disseminated at least 30 tweets in the initial collection were selected for further evaluation. To the best of our knowledge, 30 tweets could somewhat be modeled thus the user's interest level in a certain topical cluster could be computed. Tweets collected were recorded under each of their usernames and had to be in English. Non-English ones were removed from the set.
2. A collection of English tweets from the user's friendship network's list i.e. those who *"mentioned"*,*"retweeted"*, *"replied"* to tweets disseminated by the tweeters as in point 1. One other condition was that the users in the friendship network should have disseminated at least 30 English tweets too. A minimum of 30 tweets in friends and their network provided sufficient data for modeling as it depicted better online user activity.
3. The qualitative evaluation process centered around one topical cluster i.e. the *sports betting* one as mentioned earlier. Therefore, tweets fell in either the *betting* or *others* classes for consistency. This meant that any tweet that did not depict any betting related content as per human understanding was placed in the *others* class.

On the other hand, evaluators were selected to specifically look at the semantic correlation between tweeters and their friendship network's tweets in relation to the *sports betting* topical cluster. The *sports betting* cluster was chosen as it was the ground truth as detailed in Section 4.2 thus labeling the tweets was faster and more accurate. The semantic correlation in the friendship network was averaged across the evaluators. To ascertain this, the level of agreement in terms of the semantic relation between user's tweets and their friendship network's computed. The three evaluators were selected based on their knowledge of the English language and familiarity with betting related terms. They followed a two-step process in assigning tweets to respective topical clusters:-

1. Firstly, the evaluators were presented with a list of 100 tweets from the *sports betting* set as described in Section 4.1.1. They were required to go through the list at least



(a) FastText-SkipGram (100,3)

(b) Word2Vec-CBOW(200,3)

(c) BOW-KMeans Baseline

Fig. 3. Correlation of tweeters and their friendship network Degree of Interest in Swahili Related Chatter (DoiSRC).

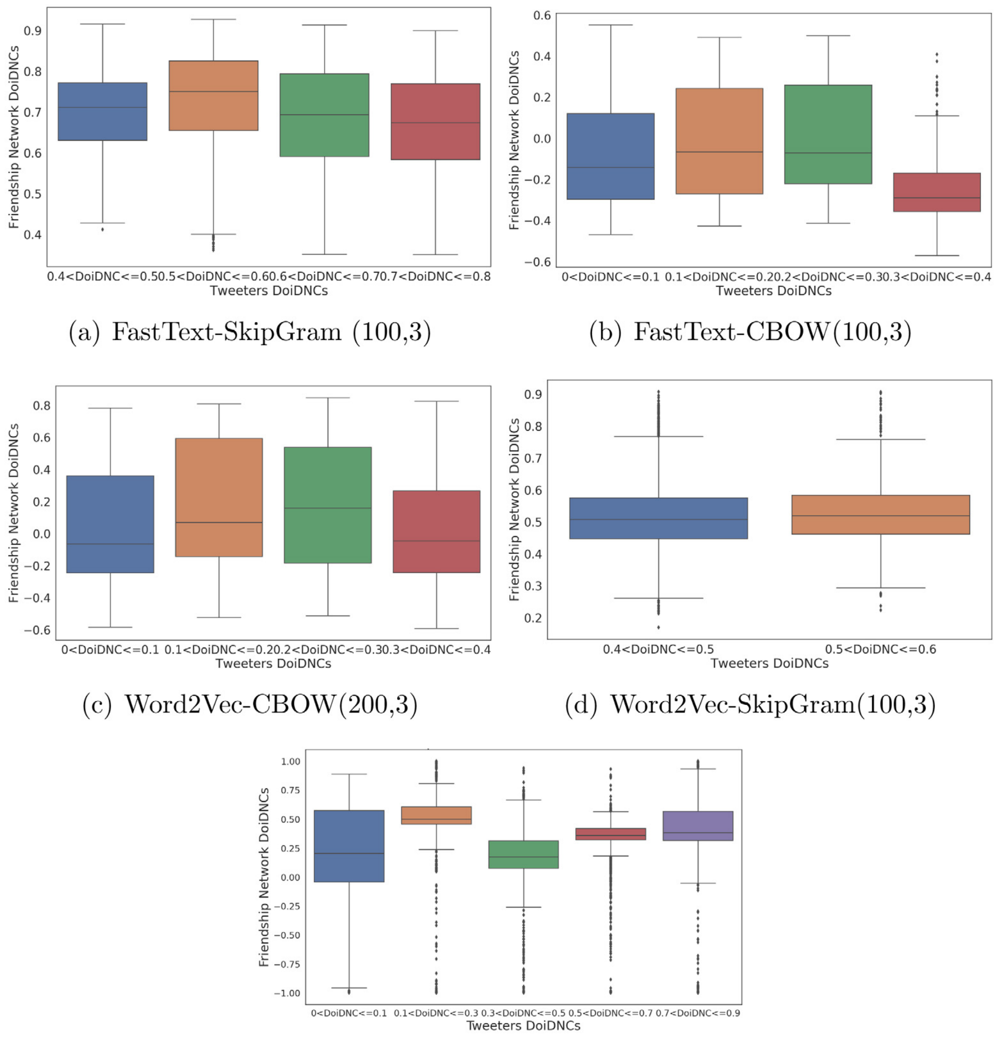
three times for a deeper contextual interpretation of tweet content in this domain.

2. Secondly, the evaluators were required to label 30 random tweets assigned to them from the tweeters and 30 from respective friendship network. The assignment was in either the *sports betting* or *others* topical clusters. The assumption was that, once an evaluator picked a tweeter and his/her tweets, he/she was also obliged to classify tweets of that specific tweeter's friendship network in the presented list. This further provided a deeper contextual understanding of the disseminated content between tweeter's and their friendship network's.
3. Finally, we computed the inter-evaluator agreement across the *sports betting* topical cluster in form of Cohen Kappa scores [44]. A higher inter-evaluator agreement value means more relevant friendship networks. Primarily, this process was the final affirmation of *homophily* among users on short text microblogs validating the follower–followee relationships. It proved that topical affinity of users can not only be measured by the disseminated content, but by analysis of the friendship network content.

Fig. 5 shows the classification accuracies with respect to *online sports betting* for three randomized users. The values on the *y*-axis are the percentages on a scale of 0 to 1. Consistency and percentage of accuracies was the measure of qualitative performance for tweeters as well as their friendship network. Across the selected tweeters and their friendship network, ***FastText-SkipGram(100,3)*** consistently performed well in identification of sports betting related content as depicted in Fig. 5(a). For example, for one random user (labeled User1 in the graph), his/her classification results for tweeters were 65% while those of their friendship network were approximately 75%, a strong indicator of users with shared interests. The same was replicated for another user (labeled User3 in the graph) disseminating approximately 70% of sports betting related content with around 60% of his/her friendship network disseminating related content. ***Word2Vec-CBOW (200,3)*** had the lowest accuracies in identification of sports betting related content in tweeters and their friendship network as shown in Fig. 5(c). *User 1* in this framework recorded very minimal interest in sports betting i.e. about 4% while his/her friendship network recorded a 5% interest in such content. On the other hand, ***FastText-CBOW (100,3)*** in Fig. 5(b) consistently depicted better classification accuracies compared to ***Word2Vec-SkipGram (100,3)***(Fig. 5(d)). The *Bag-of-Words* baseline results were also inconsistent with the *homophily theory*. For example, evaluation by *User3* placed sports betting related tweets at 50% sports betting related content. The comparative friendship network disseminated just 21% sports betting related content. These results corroborate quantitative findings in Section 3.4 where ***FastText-SkipGram (100,3)*** depicted the best **FMI** and **S-Scores**. Generally, *FastText* makes use of a sliding window of characters making it relevant in datasets with shortened or misspelled words like in tweets. This is a plausible conclusion as to why *FastText* based models performed better.

5.3.2. *Semantic correlations*

In addition to the inter-rater agreements in evaluation of homophily in Section 5.3.1, we evaluated the semantic correlation between the selected users and their friendship network tweets. The assumption was that for the two sets to be semantically similar, then their representation should ideally be in the same semantic space. Therefore, the Pearson Correlation Coefficient (PCC) between the sets of user tweets and their friendship network was computed [45]. There is an assumption of linearity of users and their friendship network with the same in the same semantic space. For example, users with *DoiSB* or *DoiDNC* values between 0 and 0.3 should ideally have values in the same range in their friendship networks. Therefore, a correlation between *DoiDNCs* of the two groups assumes linearity at least to a certain level. The *PCC* is then computed as the covariance of the two variables per classification groups of *DOIs* as indicated in Section 5.1 divided by the product of their standard deviations. This is represented as $PCC = \text{Cov}(a, b)/\sigma(a) * \sigma(b)$. σ is the standard deviation that is applied to variables a and b. The correlation is a value between -1 and 1, where 1 is an indicator



(a) FastText-SkipGram (100,3)

(b) FastText-CBOW(100,3)

(c) Word2Vec-CBOW(200,3)

(d) Word2Vec-SkipGram(100,3)

(e) BOW - KMeans Baseline

Fig. 4. Correlation of tweeters and their friendship network Degree of Interest in Daily News Chatter (DoiDNC).

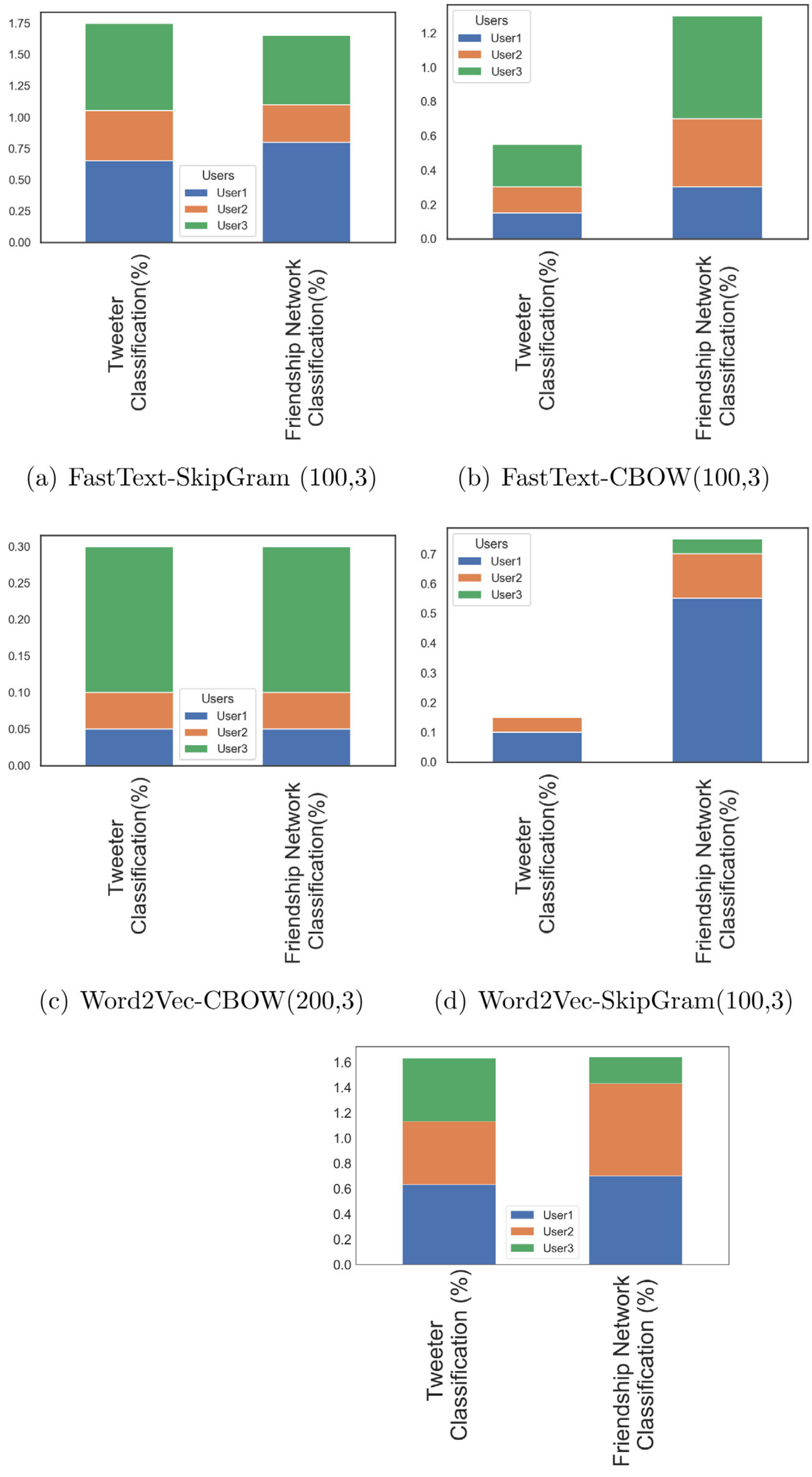
of positive correlation, −1, negative correlation and 0 indicating no correlation. In the computation of *PCCs* to validate follow-back recommendations, specific groups were selected from both sets of classifications. For example, if users in classification $0.7 <$ *DoiDNC* $<= 0.8$ are selected, then the comparison is only made with their friendship network *DoiDNCs* in the same classification group. A perfect scenario would be almost equal values in the comparative groups.

In Table 4, it is evident that the *Bag-of-Words* baseline results are dismal in correlating what users and their friendship networks disseminated. This output corroborates the results in the classification scores in Table 2. *FastText-SkipGram (100,3)* performed the best in identification of semantic correlations between users and their friendship networks with a *PCC* of 0.7 in the classification group. The choices of the classification group of interest and related interest range was based on the results in Table 3. *DoiDNC* were uniformly identified across the modeling frameworks this was the most optimal for evaluation. The interest range was one with the highest values in the group i.e. the most representative of the interest and had at least 30 users and related friendship network. The only exception was with *FastText-CBOW (100,3)* where the lowest interest value was selected as no other interest range fulfilled the 30 user count in both users and friendship network except this range. The comparisons were based on the most representative users in the lowest range and in the same interest group. Such a measure is statistically significant in this evaluation.

5.4. Application areas

Quantitative and qualitative results obtained in Sections 5 affirm the possibility of extracting relevant interests in terms of topical clusters in short text microblog content. Experimental processes in our setup affirmed our choice for a FastText-based modeling framework as the ideal modeling framework in such scenarios. However, the same modeling process can be applicable in a number of other ways on short text microblogs as below :-

- **Follower–Followee recommendations** — User interactions on short text microblogs help in content propagation on such platforms. Content in terms of *hashtags*, *"trend"* based on the rate at which users on the platforms re-share the content. Users with interest in certain content should therefore be known to each other. Basing our argument on the ***FastText-SkipGram(100,3)*** (Fig. 2(a)), tweeters with $0.1 <$ *DoiSBs* $<= 0.2$ can be recommended for follow-back to users whose *DoiSBs* $>= 0.09$. This will help in building more solid semantic connections which ultimately propagates content to a wider and desired audience.



(a) FastText-SkipGram (100,3)

(b) FastText-CBOW(100,3)

(c) Word2Vec-CBOW(200,3)

(d) Word2Vec-SkipGram(100,3)

(e) BoW (K-Means) Baseline

Fig. 5. Qualitative evaluation of sports betting content classification accuracy for tweeters and their friendship network.

- **Cold-start recommendations** - New users on microblog platforms struggle to find users with aligned interests due to lack of content and sometimes irrelevant connection suggestions. Based on the same FastText model, tweeters neutral to sports betting i.e *DoiSBs* $= 0$ tend to correlate well with friendship networks with *DoiSBs* $>=$ 0.0.08. Factoring in other measures such as location metadata, such tweeters can be recommended to the cold-start users for follow-back.
- **Third-party content propagation** - Content-based recommender systems are reinforced by relevant shared content

Table 4
Follow-back correlations as Pearson Correlation Coefficients (PCCs) between users and their friendship networks in with interest in Daily news chatter classification group.

Modeling framework	Degree of Interest in Daily news chatter (DoiDNC) Classification Group	Pearson Correlation Coefficient
FastText CBOW (100,3)	0.0–0.1	0.657
FastText SkipGram (100,3)	0.7–0.8	**0.700**
Word2Vec SkipGram (100,3)	0.5–0.6	0.106
Word2Vec CBOW (200,3)	0.3–0.4	0.500
Bag of Words - KMeans Baseline	0.7–0.9	−0.110

over time. Therefore, with better computation of interest levels in certain topics, more semantically relevant third-party is bound to be recommended to users with varied *DoiSCCs*.

6. Conclusion and future work

Twitter just like other social media platforms has proven to be instrumental in online data dissemination. In addition, users on the same platforms show preference towards several topics to a lesser or greater extent. This is influenced by their online interactions i.e. based on the interests of other tweeters or largely by personal preferences. The interactions form relationships among tweeters. This relationship may either be unidirectional or bi-directional. Therefore, such tweeters need support of recommender systems to identify the right tweeter to follow back or content to propagate. In this paper, we introduced an approach that can be used to learn the semantic relevance of tweets. The new approach identifies the degree of interest a tweeter has in a particular topic considering the semantic relevance of the user's tweets. We considered a set of neural network based models in accomplishing this task. In our experiments, a *FastText* model with 100 dimensions and a little bit of hyper-parameter tuning worked best in the extraction of words with high semantic relevance to the generated topical clusters. To do so, we modeled the input corpus for model training and vector representation. We then performed clustering on the corpus to extract the centroids maps representative of topical clusters. These centroid maps were used in determining the overall closeness of tweets to the generated clusters.

Results in the chosen model (*FastText*) corroborated the *homophily theory* whereby users who were highly involved in disseminating for example sports betting related tweets, had friends who did the same. This supports our argument stated in this paper which is also aligned with a fundamental principal in follower–followee based social networks. In our experiments we also determined thresholds (denoted by the interest groups in the paper) to define the *DoI*. These *DoI* values can be used for the design of recommender systems in this domain.

In future, we plan to automatically generate multi-topic profiles for users. The profiles could be based on tweeter's interests as well as with their friendship networks. Aggregation of Twitter specific features such as retweets, mentions, lists as well as bi-directional network structure could be used to improve on the computation of multi-topic degrees of interest.

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.

References

[1] P. Symeonidis, A. Nanopoulos, Y. Manolopoulos, Feature-weighted user model for recommender systems, in: International Conference on User Modeling, Springer, 2007, pp. 97–106.

[2] G. Jawaheer, P. Weller, P. Kostkova, Modeling user preferences in recommender systems: A classification framework for explicit and implicit user feedback, ACM Trans. Interact. Intell. Syst. (TiiS) 4 (2) (2014) 8.

[3] H. Yin, B. Cui, L. Chen, Z. Hu, X. Zhou, Dynamic user modeling in social media systems, ACM Trans. Inf. Syst. (TOIS) 33 (3) (2015) 10.

[4] P. Brusilovsky, E. Millán, User models for adaptive hypermedia and adaptive educational systems, in: The Adaptive Web, Springer, 2007, pp. 3–53.

[5] H. Liu, Z. Hu, A. Mian, H. Tian, X. Zhu, A new user similarity model to improve the accuracy of collaborative filtering, Knowl.-Based Syst. 56 (2014) 156–166.

[6] Y. Cai, H.-f. Leung, Q. Li, H. Min, J. Tang, J. Li, Typicality-based collaborative filtering recommendation, IEEE Trans. Knowl. Data Eng. 26 (3) (2013) 766–779.

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

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

[9] K. Chen, T. Chen, G. Zheng, O. Jin, E. Yao, Y. Yu, Collaborative personalized tweet recommendation, in: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2012, pp. 661–670.

[10] S. Goel, R. Kumar, Folksonomy-based user profile enrichment using clustering and community recommended tags in multiple levels, Neurocomputing 315 (2018) 425–438.

[11] S.R. Sahoo, B. Gupta, Hybrid approach for detection of malicious profiles in twitter, Comput. Electr. Eng. 76 (2019) 65–81.

[12] J. Chen, R. Nairn, L. Nelson, M. Bernstein, E. Chi, Short and tweet: experiments on recommending content from information streams, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2010, pp. 1185–1194.

[13] 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.

[14] F. Abel, Q. Gao, G.-J. Houben, K. Tao, Analyzing user modeling on twitter for personalized news recommendations, in: International Conference on User Modeling, Adaptation, and Personalization, Springer, 2011, pp. 1–12.

[15] Y. Liu, X. Chen, S. Li, L. Wang, A user adaptive model for followee recommendation on Twitter, in: Natural Language Understanding and Intelligent Applications, Springer, 2016, pp. 425–436.

[16] S. Takimura, R. Harakawa, T. Ogawa, M. Haseyama, Twitter Followee recommendation based on multimodal FFM considering social relations, in: 2018 IEEE 7th Global Conference on Consumer Electronics (GCCE), IEEE, 2018, pp. 204–205.

[17] D.P. Karidi, Y. Stavrakas, Y. Vassiliou, Tweet and followee personalized recommendations based on knowledge graphs, J. Amb. Intell. Humaniz. Comput. 9 (6) (2018) 2035–2049.

[18] F. Figueiredo, A. Jorge, Identifying topic relevant hashtags in Twitter streams, Inform. Sci. 505 (2019) 65–83.

[19] K. Sasaki, T. Yoshikawa, T. Furuhashi, Online topic model for twitter considering dynamics of user interests and topic trends, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1977–1985.

[20] T. Iwata, S. Watanabe, T. Yamada, N. Ueda, Topic tracking model for analyzing consumer purchase behavior, in: Twenty-First International Joint Conference on Artificial Intelligence, 2009.

[21] N. Ben-Lhachemi, et al., Using tweets embeddings for hashtag recommendation in Twitter, Procedia Comput. Sci. 127 (2018) 7–15.

[22] P. Liu, L. Zhang, J.A. Gulla, Real-time social recommendation based on graph embedding and temporal context, Int. J. Hum.-Comput. Stud. 121 (2019) 58–72.

[23] S.-H. Yang, A. Kolcz, A. Schlaikjer, P. Gupta, Large-scale high-precision topic modeling on Twitter, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, ACM, New York, NY, USA, 2014, pp. 1907–1916, http://dx.doi.org/10.1145/2623330.2623336,.

[24] 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.

[25] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (Jan) (2003) 993–1022.

[26] J. Li, H. Xu, X. He, J. Deng, X. Sun, Tweet modeling with LSTM recurrent neural networks for hashtag recommendation, in: 2016 International Joint Conference on Neural Networks (IJCNN), IEEE, 2016, pp. 1570–1577.

[27] S. Mishra, M.-A. Rizoiu, L. Xie, Modeling popularity in asynchronous social media streams with recurrent neural networks, in: Twelfth International AAAI Conference on Web and Social Media, 2018.

[28] Z. Zhang, D. Robinson, J. Tepper, Detecting hate speech on twitter using a convolution-gru based deep neural network, in: European Semantic Web Conference, Springer, 2018, pp. 745–760.

[29] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Trans. Assoc. Comput. Linguist. 5 (2017) 135–146.

[30] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, 2016, CoRR arXiv:1607.04606.

[31] 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.

[32] 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.

[33] D. Arthur, S. Vassilvitskii, k-means++: The advantages of careful seeding, in: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[34] 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).

[35] L. Recalde, A. Kaskina, Who is suitable to be followed back when you are a twitter interested in politics?, in: Proceedings of the 18th Annual International Conference on Digital Government Research, ACM, 2017, pp. 94–99.

[36] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O.P. Patel, A. Tiwari, M.J. Er, W. Ding, C.-T. Lin, A review of clustering techniques and developments, Neurocomputing 267 (2017) 664–681.

[37] 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.

[38] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, 2013, arXiv preprint arXiv:1301.3781.

[39] M.Z. Rodriguez, C.H. Comin, D. Casanova, O.M. Bruno, D.R. Amancio, L.d.F. Costa, F.A. Rodrigues, Clustering algorithms: A comparative approach, PLoS One 14 (1) (2019) e0210236.

[40] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.

[41] D. Arthur, S. Vassilvitskii, k-means++: The Advantages of Careful Seeding, Tech. Rep., 2006.

[42] M. McPherson, L. Smith-Lovin, J.M. Cook, Birds of a feather: Homophily in social networks, Annu. Rev. Sociol. 27 (1) (2001) 415–444.

[43] S. van den Beukel, S.H. Goos, J. Treur, An adaptive temporal-causal network model for social networks based on the homophily and more-becomes-more principle, Neurocomputing 338 (2019) 361–371.

[44] M.L. McHugh, Interrater reliability: the kappa statistic, Biochemia Med.: Biochemia Med. 22 (3) (2012) 276–282.

[45] J. Benesty, J. Chen, Y. Huang, I. Cohen, Pearson correlation coefficient, in: Noise Reduction in Speech Processing, Springer, 2009, pp. 1–4.