

Follow-back Recommendations for Sports Bettors: A Twitter-based Approach

Herman Wandabwa, M. Asif Naeem, Farhaan Mirza, Russel Pears
 School of Engineering, Computer and Mathematical Sciences
 Auckland University of Technology

[herman.wandabwa, mnaeem, farhaan.mirza, russel.pears]@aut.ac.nz

Abstract

Social network based recommender systems are powered by a complex web of social discussions and user connections. Short text microblogs e.g. Twitter present powerful frameworks for information consumption, due to their real-time nature in content throughput as well as user connections. Therefore, users on such platforms consume the disseminated content to a greater or lesser extent based on their interests. Quantifying this degree of interest is a difficult task based on the amount of information that such platforms generate at any given time. Thus, the generation of personalized profiles based on the Degree of Interest (DoI) that users have towards certain topics in such short texts presents a research problem. We address this challenge by following a two-step process in generation of personalized sports betting related user profiles in tweets as a case study. We (i) compute the Degree of Interest in Sports Betting (DoiSB) of tweeters and (ii) affirm this DoiSB by correlating it with their friendship network. This is an integral process in the design of a short text based recommender systems for users to follow i.e follow-back recommendations as well as content-based recommendations relying on the interests of users on such platforms. In this paper, we described the DoiSB computation and follow-back recommendation process by building a vector representation model for tweets. We then use this model to profile users interested in sports betting. Experiments using real Twitter dataset geolocated to Kenya shows the effectiveness of our approach in the identification of tweeter's DoiSBs as well as their correlation with their friendship network.

1. Introduction

Citizen journalism aided by the emergence of social networking platforms like Twitter and Facebook has led to the generation of massive and diverse online content e.g. text, videos, images. For example, 6000 tweets are

published every second, corresponding to over 350,000 tweets per minute and 500 million a day¹. Tweeters² in essence share photos, videos, hyperlinks and locations to members in their networks.

Tweeters extrinsically or/and intrinsically formulate online profiles. This mostly depends on the content they consume and disseminate over time in addition to their user follower-followee network. In general, tweeters show diversity in expressing their interests in certain topics. This can be based on hashtags that they follow at the time as well as time based event related information. On the hindsight, the dynamism in their friendship networks as well as the streaming nature of the platform makes it difficult to quantify their DoI in certain online topics. This is further compounded by the fact that such topics are always dynamic. For example, users who love outdoors activities, are likely to tweet about a mountaineering experience, may also often tweet in support of their favourite political candidate or sports team.

Therefore, interest identification for the purpose of better user profiling on such platforms is an important research problem. Precise profiling of a tweeter based on his/her interests and to what extent, largely alleviates personalization related problems. Twitter's recommender system normally discovers relevant followers to be suggested to tweeters or Twitter lists of interest based on the friendship network. However, this does not mean that explicit interests among the users are shared. The questions below elicit the need for this research:-

- Is it possible to group users based on their topical interests in short text microblogs?
- Do friendship connections in short text microblogs influence topical interests for such users?

¹<http://www.internetlivestats.com/twitter-statistics/>

²a person who posts on the social media application Twitter

Sports betting, just like lotteries is a huge industry in the world³. Tweeters with interest in online sports betting are assumed to propagate sports betting related content on Twitter. Despite their interest in sports betting, we are correct not to assume that all their followers are relevant to be followed back. Sports betting companies as tweeters, may have lots of followers but not all could be relevant. Besides, such users need to be aware of what other tweeters disseminate which partly helps them to rank important tweets or create a network of influencers in their domain. This is instrumental in enabling tweeters make decisions on who to follow as they are presented with the most relevant users to follow on such platforms. In addition, this model is vital in suggesting users to follow back in cold-start scenarios or in the case of lurkers⁴ as they normally do not have enough initial friendship connections.

Therefore, we present a method to compute the Degree of interest in Sports Betting (DoiSB) among tweeters. We consider the Kenyan Twitterspace as a case study where diversity of topics, and interest in sports betting is high in addition to the author's knowledge of Kenyan's tweeting patterns. To the best of our knowledge, this work presents the first attempt at quantifying tweeter's affinity towards sports betting by analyzing their disseminated content over time and corroborating with their friendship network.

We make the following scientific contributions in the paper:-

- We develop a novel framework for computing user profiles based on their content dissemination patterns.
- We proof the social theory of *homophily* by correlating tweeter's interests to those of their friendship network. Theoretically, this is the tendency for people with shared interests to be connected.
- We test our framework in deduction of user's interests in online sports betting. We also carried out an experimental study in formulation of user representative profiles.

The rest of the paper is organized as follows. Related literature of our study is in Section 2. Our methodology is described in Section 3. The experimental framework is presented in Section 4 while results are shown in Section 5. Conclusion and future work is summarized in Section 6.

³<https://www.statista.com/topics/1740/sports-betting/>

⁴a member of an online community who observes, but does not participate

2. Background and Related Work

2.1. User Interests Preferences in Short Text Microblogs

The choice of content to be consumed in short text microblogs is largely influenced by the interests of the consumer(tweeter). Therefore, such interests are integral in the design of short text based recommender systems as they are developed to match users to resources of interest. Chen et al.,[1] proposed collaborative ranking in the capture of user interests through integration of useful contextual information such as tweet topic level factors in tweets. A User Interest Profile design methodology was also proposed by Goel et al.,[2]. In the design proposal, user generated tags were enriched with friendship information through vector representations. In the context of profiling users of malicious intent in short text microblogs, Sahoo et al.,[3] proposed a hybrid approach in leveraging classifications and Petri net structure. In addition, authors in [4] proposed a URL recommender system for Twitter users based on social voting and content sources. Results suggested that the generated topics and social interactions were more significant in presentation of recommendations. Recommendation of users to follow-back in short text microblogs was also addressed in [5],[6] and [7].

2.2. Twitter in Sports and Betting Recommendations

Twitter related activities have been instrumental in the domain of sports. Robert et al.,[8] predicted outcomes of NFL games using tweets. Technical stock market techniques were applied to sentiment gathered from social media for the predictions. On the other hand, Brown et al.,[9] evaluated the accuracy of social media forecasting in the English premier league soccer matches where the aim to assess whether tweet semantics could be used in predicting match outcomes. The authors further investigated whether the predictions were only restricted to large events i.e. when goals were scored. Findings indicated that if the combined tone of tweets was positive at any time of the match, then the likelihood of a team winning was higher than betting market prices implied. Vaughan et al., [10] work mirrors what Brown et al.,[9] did. The authors measured Twitter activity around unique, identifiable and newsworthy events and correlated the activities with betting prices fluctuations on Betfair. Their findings corroborated the initial assumptions that response of market prices appeared sluggish with little event related data compared to post-news drift times.

The goal of our research is to improve on methodologies that can be used to infer the level of interest denoted as *Degree of Interest (DoI)* that users may have towards certain topics in streaming microblog texts. DoI measure is instrumental in the design of short text based recommender systems in diverse domains. Lack of studies in short text influence-based recommender systems makes our contribution unique. In this work, we made use of neural-network based vector representations of short text word tokens to comprehend better, the underlying semantic structure of tweets. Vector representations via a neural-network based algorithm, *FastText*⁵ worked well with our type of textual data i.e. one with misspelled/shortened words reminiscent of tweets [11]. The algorithm typically makes intelligent guesses on even out of vocabulary words as long as some character level consistency is observed.

3. Our Approach

Inferring the extent of interests by a group of short text microblog users in certain topics involves a number of processes. The processes are listed below: -

- **Text Modelling** - In this step, we train a *FastText* based model using a corpus of tweets geolocated to Kenya. The output of the model is a vector space representation of tweet word tokens. In the vector space, a tweet is represented as a vector in an n-dimensional space, where each dimension represents a term. Similarity between terms or documents (tweets in our case) is measured as the cosine angle between the vectors being compared. We evaluated this modeling approach against *Word2Vec* and *Glove* baselines in choosing the best model for the task
- **Clustering and Extraction of Centroids** - Tweets are grouped based on their semantic similarity via a clustering algorithm. Cluster centroids represented as vectors for each cluster are extracted via the algorithm.
- **User's Degree of Interest (DoI)** - To compute the DoI, a tweeter's level of interest in a topic is measured. The tweeter's tweets are transformed to a vector format via the trained model and distance to the centroid of interest measured.
- **Correlation with tweeter's friendship network** - To proof a tweeter's interest in a certain topic, his/her friendship network DoI was

computed. This follows the *homophily theory* in social networks where similar nodes (friendship connections) may be more likely to share interests than than dissimilar ones.

3.1. Text Modeling

We based our text modeling methodology on a neural network model *FastText*⁶. *FastText* was the algorithm of choice based on its mode of extracting syntactic information in short, sparse and often misspelled words in a corpus. Unlike other word embedding algorithms like [12], *FastText* makes use of word morphologies where, word vectors are associated with each character n-gram and words are modelled and represented as the sum of character word vectors. Therefore, this algorithm proved to be an ideal model for learning misspelled or words out of the dictionary.

To model tweets via this neural network algorithm, the below procedure was followed: -

- **Text pre-processing** - Pre-processing text is necessary for a better corpus as model input. This process entails removal of unnecessary words and punctuation as they do not provide any contextual meaning for the model to learn. We followed the below steps in pre-processing the input text: -
 - Lower-cased all words in the corpus.
 - Removed all accented characters and numbers. Some of the accented characters were encoded to Unicode Transformation Format 8-bit(UTF-8) format.
 - Removed all hyperlinks. They were not of interest in this instance.
 - Removed all user mentions. They are words prefixed by the @ symbol in a tweet. Often, they refer to tweeters' usernames.
 - Removed all words with less than three characters. They were found not to be semantically relevant in most tweets.
 - Cleaned out all hashtags. These are words in a tweet that are prefixed by the hash (#) symbol.
 - Removed stopwords. These are the most common English words. Normally, they are not semantically significant. We used a custom list of stopwords in addition to the NLTK stopword list⁷.
 - Tokenized the remaining words in each tweet and stored them as a list ready for model training.

⁵<https://fasttext.cc/>

⁶<https://fasttext.cc/>

⁷<http://www.nltk.org/>

- **Model Training** - The tokenized list of words in the corpus forms the input pipeline for model training. In model training, a machine learning algorithm (*FastText* in our case) is provided with training data to learn from. The model learns semantic knowledge in the dataset by mapping each word to a continuous vector space from its distributional properties observed in the the corpus.

Several parameters have to be specified in order to train *Word2Vec* and *FastText* models:-

- *size* or the number of dimensions in the vector space.
- *min_count* or minimum count of a term in the corpus for it to be included in the training. Terms with word counts lower than this value were excluded from training;
- *sg=1* for training a SkipGram model, otherwise Continuous Bag of Words (CBOW). In the SkipGram modeling, the algorithm loops over the list of words and uses current word to predict its neighbors (its context). However, in CBOW, the context is used to predict the current word.
- *window* parameter is the maximum distance between the current and predicted word in the list of word tokens;
- *word_ngrams* are specified in order to enrich word vectors with subword(n-grams) information. This enrichment is possible if the value is specified as 1 ;
- *iter* or iterations is the number of iterations (epochs) over the corpus. In essence, this parameter defines the number times that the learning algorithm goes through the entire training.
- *Glove* model only had the *epochs* and learning rate(lr) defined.

FastText is unique in its vector space representation as it ignores word structures. Each word w is represented in the vector space as a bag-of-character *n-grams* n where the word itself is included in the n-grams set. We used $3 \geq n \leq 6$ in our implementation as specified in [11]. This way, most of the n-grams were factored in the modeling.

For an n-grams dictionary of size B and word w , $B_w \subset \{1, \dots, B\}$. x_b is the vector representation for each n-gram b . The scoring function is formulated as in [11] :-

$$s(w, c) = \sum_{d \subseteq B_w} x_b^\top v_c \quad (1)$$

where c is the context position of a word, and v the corresponding word vector.

In our case, each tweet is made of word tokens. Therefore, its vector representation is the sum of its word vectors after pre-processing. Using the parameters elicited earlier, the model was ready to be used in the generation of vectors for each word in the corpus. The process is the same for *Word2Vec* and *Glove* baselines except that their vector space applies only at word level. They were trained for validation purposes.

- **Clustering and Extraction of Centroids** - A clustering algorithm was deployed to group most similar tweets as close as possible (clusters). Semantically dissimilar tweets were pushed as far away as possible from each other. The insight here is that objects in respective clusters are to be as similar as possible. Thereafter, manual inspection of the underlying keywords and analogy tests for terms in each of the clusters were carried out to identify the topic or closely related topic that the each of the clusters inclined towards.

To cluster tweets, *K-Means++* was applied on the training corpus. This 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 guaranteeing an $O(\log k)$ solution [13]. Therefore, finding the optimal set of centroids was guaranteed. We used a heterogeneity convergence metric to determine the optimal cluster numbers across the models [14]. In determining the numbers, we ran tests considering different k values as cluster numbers on a known test set. The cluster numbers that best represented the test set were chosen to be optimal. Intra-cluster distance between y points in a given cluster X_k and the cluster's centroid X_x was then computed as cosine distance. Our interest in the case study is with regard to finding a cluster that best represented sports betting content. To do so, we first have to identify the sports betting cluster. This as described earlier is done via analogy tests as well as manual inspection of terms in each cluster. Once the terms are identified, a *centroid map* that contains terms and their respective cluster numbers is computed as in [15]. Thereafter, centroids for each model are

generated via the trained models. For example, a *FastText* model with 100 dimensions and 3 clusters generated a JSON file with 3, 10×10 matrices.

- **User’s DoI in Sports Betting** - To understand the computation of user DoIs to the sports betting cluster, similarity between tweets and the cluster centroids had to be derived first.

– **Similarity to Cluster Centroids** - Similarity of a tweet to a cluster of interest, involved calculating the semantic distance of the specific tweet tokens to the centroid of the cluster of interest. To represent this process, let Q be the set of vectors for clusters $q \in Q$ in the model. Q is significant in getting the distance between the tweet and clusters. In our case study, the interest was in getting the distance between a given test tweet and the sports betting cluster q_{SB} . To represent this similarity computation process for a tweet s , let W_s be the set of word tokens in the tweet. The average of the vectors W_s was the vector space representation of tweet s as illustrated in Equation 2.

$$w'_s = average(w_s), \forall w_s \in W_s \quad (2)$$

In our case, we computed the cosine distance by measuring the similarity between the tweet vector w'_s and cluster centroids like q_{SB} . Cosine distance was the optimal similarity measurement metric between the tweet vectors and cluster centroids. The advantage with cosine similarity measure is that it works well despite the size of the two vectors being measured. 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 to or equal to 1 and dissimilar if close to or equal to 0.

$$s_{xs} = CosineDistance(w'_x, s), \forall s \in q_{SB} \quad (3)$$

Computation of the cosine distance between the tweet vector w'_x, s and cluster centroids q_{SB} is shown in Equation 3. This way, the

similarity between a tweet and the cluster of interest is computed.

- **Computation of the Degree of Interest in Sports Betting (DoiSB)** - Computation of a tweeter u ’s DoiSB entailed following the below steps : -

1. We first extracted tweets from the user’s timeline T via Twitter’s Search API⁸. A maximum of 3200 tweets can be extracted from the timeline.
2. The extracted tweets $x_u \in T_u$ are then preprocessed and modeled as described in Section 3.1.
3. Similarity of the processed tweets q_u to the sports betting cluster q_{SB} is then computed as in Equation 3.

The DoiSB computation for user u is illustrated as in Equation 4.

$$DoiSB_u = average(s_{x_u q_{SB}}), \forall x_u \in T_u, q_{SB} \in Q \quad (4)$$

Interpretation of DoiSB values followed the same process as the cosine distance. Tweeter’s with DoiSBs close to 1 meant that they disseminated content that was largely related to sports betting. On the contrary, users with DoiSBs close to 0 meant that their disseminated content had very little sports betting related content.

- **Homophily Social Theory in DoiSBs** - Homophily is defined as the tendency for people to have positive ties with people who are similar to themselves in friendship networks. In our case, homophily is measured with regard to the extent to which users share interests [16]. In essence, users with high DoiSB values shows that they share interests thus can be recommended to each other as follow-backs.

4. Experimentation

In this section, we elicit all the practical steps that were followed in validation of the proposed approach in Section 3. *Word2Vec* and *Glove* baselines were also trained for validation purposes.

4.1. Datasets, Settings and Analogy Tests

We collected 298835 tweets geolocalized to Kenya for six months starting 1/9/2018 via Twitter’s streaming

⁸<https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html>

API. The collection was made up of tweets as single row records with their associated metadata. These among others included the mentions, hashtags, list of usernames and geo coordinates among others. Most of the tweets in the dataset were written in English with a few in Swahili and the rest using a mixture of the two. The choice of the Kenyan Twitterspace was firstly informed by the authors knowledge of the dynamic nature of Twitter topical content in the country. On the other hand, online sports betting related activities are also on the rise in the country thus necessitating the need to investigate interests of tweeters in this domain. In order to make sure that the sports betting content related cluster existed for evaluation purposes, betting related tweets had to be added to the generic pool of tweets. We collected 50639 sports betting related tweets to add to the dataset. The tweets were collected from timelines of online sports betting companies with presence in Kenya. They included tweets associated with *sportpesa*⁹, *betin*¹⁰, *eazibet*¹¹, *betika*¹² and *betwayke*¹³ Twitter handles. In addition to the generic set, the total number of tweets in the training corpus totalled 349474. Analogy tests validated the generalization and quality of the trained models. Therefore, we conducted several qualitative tests on the model to make sure that the model was relevant to the test scenario.

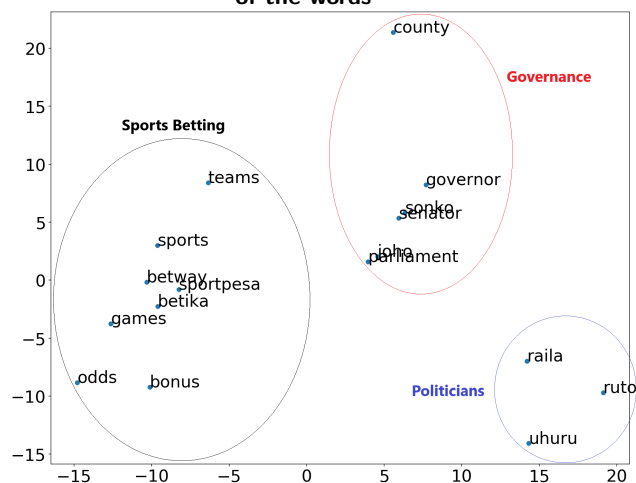
Table 1 summarizes one validation example in the context of *sports betting*. In the given example, *sportpesa*¹⁴ a Kenyan-based betting company, depicts high similarity with words like *tzsportpesa*, the Tanzanian wing of the same company.

We further used the FastText model to plot the semantic distance between words as an analogy test. Figure 1 shows the distance between words in the corpus. Words like *ruto*, *raila*, *uhuru* being grouped close to each other is semantically relevant since they are all politicians in Kenya. On the other hand, words like *county*, *governor*, *joho*, *senator*, *parliament*, *sonko* are all governance related. In fact *Sonko* and *Joho* are current governors in Kenya. *Sportpesa*, *betway*, *betika* are sports betting companies in Kenya, thus grouped together in addition to words like *odds*, *bonus*, *games*, *sports* and *teams* being semantically close.

Table 1. Sample analogy in the sports betting domain. Sportpesa is a sports betting company operating in several countries in the world

Most similar to "sportpesa")	Similarity Score
tzsportpesa	0.9649
sportpesajp	0.9615
sportpesasa	0.9536
sportpesacup	0.9284
sportpesarewind	0.8904
hullcitysportpesa	0.8891
sportpesahullcity	0.8607
sportpesanews	0.8535
sportpesashield	0.8456
sportpesapic	0.8442

Figure 1. Sample plot showing the semantic relevance of words in the training set. Semantic distance between words is depicted by the closeness of the words



4.2. Samples of Sports Betting Related Tweets

The analogy tests in Section 4.1 provided a general semantic view of the dataset. However, before selection of the best performing model to compute the DoiSBs, the model had to be subjected to a known dataset. Two baselines i.e. *Word2Vec* and *Glove* models were introduced for validation purposes. All models were trained on the same dataset with the same parameters.

We sampled 100 sports betting tweets to test the three model's accuracy with different parameters as well as get the optimal number of clusters. The hypothesis was that the selected tweets had to be as close as possible to sports betting related content. This way, it was easier to distinguish classification performances across the models. Manual inspection of the test tweets by

⁹<https://www.sportpesa.org/>

¹⁰<https://www.betin.co.ke/>

¹¹<https://www.eazibet.co.ke>

¹²<https://www.betika.com/>

¹³<https://www.betway.co.ke>

¹⁴<https://www.sportpesa.org/>

three human judges indicated that they were all centered around sports betting. This process was significant in identification of optimal model dimension sizes in model training as well as cluster numbers that best represented the corpus. The hypothesis in this step was such that the higher the number of correctly classified tweets in the sports betting cluster, the better the modeling algorithm and related parameters. Therefore, it was a matter of iterative trialing of varied model parameters in picking the best performing model for use in computing follow-back recommendations. The 100 test tweets were subjected to *FastText-CBOW*, *FastText-SkipGram(SG)*, *Word2Vec-CBOW*, *Word2Vec-SkipGram(SG)* and *Glove* models trained with 100, 200 and 300 dimensions consistent with [12]. We tested the model dimensions with the number of clusters set to 3, 4, 5 and 6 based on the elbow method for identification of the optimal number of clusters [14].

In computing the classification accuracies, we followed the below processes: -

1. A comparative evaluation was performed for each test tweet to cluster labels using each model as identified by K-means++. For example, FastText's (100 dimensions, 3 clusters):- *cluster 0* represented the *sports betting domain*, *cluster 1*, *Swahili Related Chatter*, *cluster 2*, *General/News* based on the analogy tests in Section 4.1. Therefore, *cluster 0* denoted by 0 in our experiments was the ground truth (*true labels*).
2. Each tweet vector was computed and the distance to the three clusters derived to generate *predicted labels*. The Fowlkes-Mallows Index (FMI-Score) was used to derive correlations between the labels. The FMI-Score is interpreted as the geometric mean of pairwise precision and recall between the true and predicted labels. The score just like cosine distance ranges from 0 to 1. A higher value indicates better similarity between two points [17].

We report the values in Figure 3. *FastText-SkipGram* with 100 dimensions and 3 clusters, reported the highest FMI-Score in relation to the sports betting cluster. From the table, we can infer that models with more than 3 clusters reported lower FMI-Scores. Therefore, we selected *FastText-SkipGram* with 100 dimensions and 3 clusters to compute *DoiSBs* further.

4.3. Samples of Tweeters in the Kenyan Twitter-sphere

We simulated a real Twitter environment by collecting sample tweets geolocated to Kenya. The

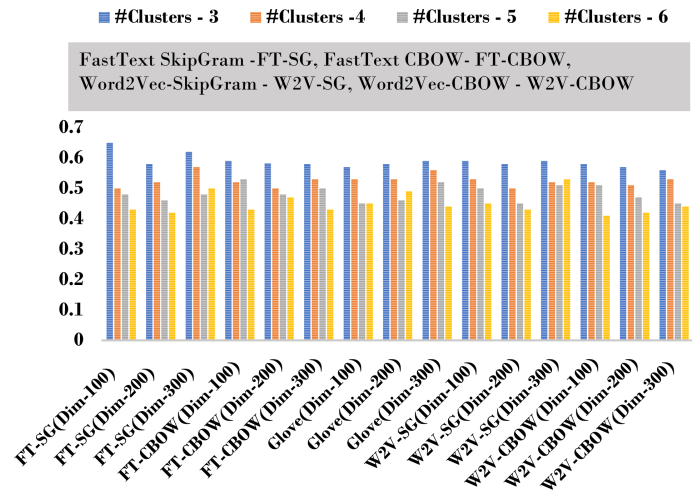


Figure 2. Model's classification scores with respect to model dimensions (100,200,300) consistent to [12] and cluster numbers (3,4,5 and 6).

aim of this process was to help us derive tweeters in the Kenyan Twitter space who would fit this study i.e. have interest in online sports betting. Computation of *DoiSBs* for sample tweeters as in Section 4.1 involved the collection, pre-processing and modeling of tweets disseminated by the tweeters. We collected a maximum of 3200 tweets from 137 users who tweeted from/near Kenya from 1/1/2019 to 1/04/2019, via Twitter's search API. Our assumption in the tweets collection process was that a three month period was sufficient to collect enough data with diverse topics as most tweets were disseminated as a reaction to certain events within that timeframe. Another assumption was that there was a likelihood for sports related content to be tweeted in addition to other topics within that timeframe.

4.4. Proof of the Homophily

Follower-followee relationships define connections in social networks. Therefore, homophily is evident in social networks based on the fact that tweeters tend to follow other users whom they share interests with [18]. In Twitter, friendship connections are in form of *mentions*, *retweets*, *replies*, *hashtags etc.* Proof of homophily in connections between tweeters and their friendship networks is in the form of shared interests. Therefore, a positive correlation in *DoiSB*, was proof that the identified interests in tweeters were realistic in addition to the good model performance in the Degree of Interest identification with respect to sports betting.

4.5. Parameter Settings and Experiments

The selected *FastText* (100,3) model had the following parameters setup : $size = 100$, $minimum\ count = 2$, $learning\ rate\ (lr) = 0.1$ and $iter = 30$. In depth descriptions of the above parameters are in Section 3. *FastText* default parameters were assumed in cases where the above parameters were not explicitly defined. The output of our modeling process was a vector representation of 22816 unique words in the training corpus. The number of clusters as well as the initialization mechanism i.e. k-means++ was specified in the clustering process to generate cluster centroid maps. The process of choosing cluster numbers K was as described in Section 4.2. Centroid maps consisted of words in the corpus and their respective cluster assignments. With the centroid maps in place, words in specific clusters are able to be placed as close as possible each each other.

In modeling tweeters, the training corpus in Section 3 was used. The optimal number of clusters in our case was 3 where each of the clusters had a unique identifier. The sports betting related one was cluster 0 and consisted of 3123 unique words. Clusters 1 and 2 had 12518 and 7175 unique words respectively. This made it easier to compute cluster centroids.

Resultant tweet vectors were then used to compute the tweet clusters similarity. The similarity as pointed out earlier is the distance between the average tweet vector and the cluster centroid of interest. This process 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/>
- **Preprocessed Tweet** - *away multibet football tips odds kenya january*
- **Cluster Similarity values** s_{xy} - [0.496, 0.196, 0.434] where the value in the array position 0 is the tweet similarity measure to the sports betting related cluster ($s_{xy\ DoiSB}$).

The value shows that the tweet is semantically close to the sports betting cluster as compared to other clusters.

5. Results

5.1. DoiSBs for Follow-Back Recommendations

Short text microblog users tend to have positive ties as evidenced by follower-followee relationships.

Normally, such users tend to have common interests. Group interests based on user *DoiSBs* were preferred compared to individual analyses as depicted in Figure 4.

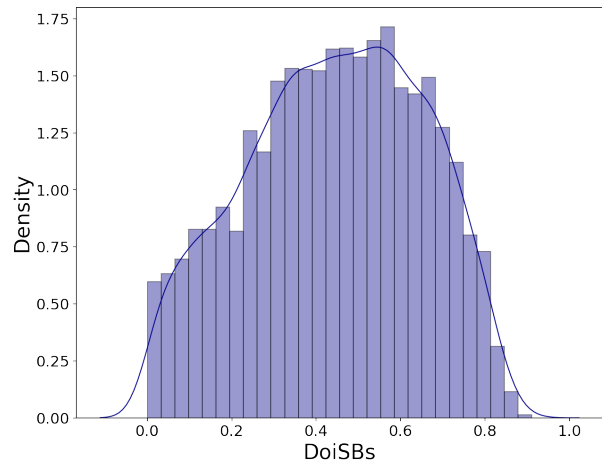


Figure 3. Overall Distribution of DoiSBs

The graph in Figure 3 shows an almost symmetrical distribution necessitating the grouping of *DoiSBs* in the below groups: - a) users with *DoiSB* equal to 0 (*Group i*), b) users with *DoiSB* greater than 0 but less than or equal to 0.3 (*Group ii*), c) users with *DoiSB* greater than 0.3 but less than 0.5 (*Group iii*), d) users with *DoiSB* greater than or equal to 0.5 (*Group iv*).

Results in Figure 4 show the correlation distribution between the *DoiSBs* of tweeters and their friendship network. From the box plot, tweeters with $DoiSBs = 0$ correlated with friends whose median $DoiSB = 0.37$. The same can be said of tweeters with $0 < DoiSB \leq 0.3$ who shared sports betting interests with friends whose median $DoiSB = 0.36$. The third and fourth groups showed stronger ties between tweeters and their friendship networks. Tweeters with $0.3 < DoiSB < 0.5$ correlated with friendship networks whose approximate median $DoiSB = 0.47$. Tweeters depicting high interest in sports betting coincidentally had friendship connections who showed the same level of interest. This is shown in the *Group iv*, where tweeters $DoiSB \geq 0.5$. They shared interests with friends having a median $DoiSB$ of 0.62. The output corroborated with the expectations in the homophily social theory where users with shared interests follow are more likely to connect.

5.2. Practical Application Areas

Results in this setup and experiments are applicable in several areas related to short text microblogs based recommender systems.

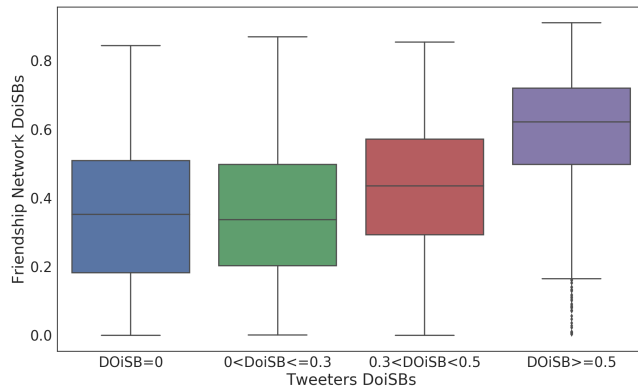


Figure 4. Correlation between users' DoiSBs and their friendship network

- **Follow back recommendations** - From the experimental results, users with $DoiSBs \geq 0.5$ can be recommended to other users with $DoiSB \geq 0.62$ or vice-versa. The two sets of users correlated with each other thus, is plausible to have suggestions for similar interests.
- **Cold-start Scenarios** - New users on short text microblogs are always in need of accurate recommendations regarding users to follow, hashtags and even lists. Correlating $DoiSBs$ with other factors such as geolocation is an ideal process in suggesting pages of interest for such users.

5.3. Qualitative Evaluation of Homophily

For affirmation of quantitative results in Section 5, we presented a sample list of 40 randomised and anonymised clean tweets to five judges/evaluators with good English command for evaluation. This was based on topics that we felt were representative of the dataset. Out of the 40 tweets, 20 were from tweeters while the remaining 20 were extracted from the specific tweeter's friendship network. Overall, each evaluator received a unique set of randomised tweets. The intuition behind this process was for presentation of a dataset that mirrored a real twitter stream in terms of content diversity in both tweeters and their friendship network. Upon manual inspection of the tweets, we identified three classes in the tweets. *Swahili related*, *Sports Betting* and *General News* classes. Evaluators were expected to classify the tweets based on the three clusters, where a tweet could only fall in one class for consistency purposes and in line with the hard clustering approach in the model.

In Table 2, $X1$ to $X5$ represents the

evaluators/judges. x_{1u} to x_{5u} represents individual tweeter classifications in the topics of interest as per the evaluators. On the other hand, $M_{1u}..M_{5u}$ represents respective *FastText* model classifications for the same tweets subjected to evaluators as described in Section 4.5. For example, according to evaluator $X1$, tweeter x_{1u} had 4 tweets classified under the sports betting topic. Their friendship network i.e. x_{1f} had two tweets under the same topic. k_{1u} to k_{5u} are the Cohen Kappa scores which in this instance is the inter-topic agreement between judge's and model's classifications [19]. This evaluation score was an indicator of the extent to which tweeter's and their friendship network tweets contextually correlated. $k_{1f}..k_{5f}$ represents correct topical classifications of the friendship network tweets by the evaluators. $M_{1f}..M_{5f}$ just like in the tweeter's instance represents the model classifications for the same tweets in the friendship network that were subjected to evaluators. Kappa score k was derived as follows; $k = \frac{p_o - p_e}{1 - p_e}$ where p_e was the hypothetical probability of chance agreement. p_o was the relative observed agreement between tweeters and their friendship network ratings.

From the results in Table 2, 56.67 percent of the inter-topic ratings depicted a weak to perfect agreement as per the Kappa statistic scale [20]. This was quite impressive based on the small sample of tweets in both groups. The results corroborate the *homophily theory* in social networks. A positive correlation to a certain level between the two sets of data is proof that friends share interests thus follow back recommendations can be made among such users. This correlation was also evident in the model in addition to the output from evaluators.

6. Conclusion and Future Work

Twitter as a short text micro-blogging platform is instrumental in disseminating event related information or news. Tweeters in essence show preference towards certain topics to a lesser or greater extent based on their level of interest in them. In addition, tweeters with shared interests are deemed to correlate when they follow-back each other.

We developed a model framework that can be used in identification of interests that microblog users have based on their disseminated content. A *FastText* model was deployed to learn tweet semantics as well as compute the level of interest that a tweeter has in sports betting. Experimental results were inline with the homophily social theory whereby users with shared interests also shared connections, a fundamental principle in user recommendations.

	X1			X2			X3			X4			X5		
Topics(Tweeters)	X_{1u}	M_{1u}	k_{1u}	X_{2u}	M_{2u}	k_{2u}	X_{3u}	M_{3u}	k_{3u}	X_{4u}	M_{4u}	k_{4u}	X_{5u}	M_{5u}	k_{5u}
<i>Swahili Related</i>	5/20	4/20	0.286	6/20	4/20	0.474	7/20	6/20	0.659	2/20	3/20	1.00	9/20	7/20	0.588
<i>Sports Betting</i>	4/20	6/20	0.474	4/20	8/20	0.545	7/20	5/20	0.765	5/20	4/20	0.857	4/20	6/20	0.474
<i>General/News</i>	11/20	10/20	0.3	10/20	8/20	0.6	6/20	9/20	0.479	13/20	13/20	0.468	7/20	7/20	0.341
Topics (Friendship Network)	X_{1f}	M_{1f}	k_{1f}	X_{2f}	M_{2f}	k_{2f}	X_{3f}	M_{3f}	k_{3f}	X_{4f}	M_{4f}	k_{4f}	X_{5f}	M_{5f}	k_{5f}
<i>Swahili Related</i>	4/20	4/20	0.688	8/20	10/20	0.6	4/20	3/20	0.828	5/20	3/20	0.692	2/20	4/20	0.615
<i>Sports Betting</i>	2/20	3/20	0.773	5/20	6/20	0.625	5/20	4/20	0.571	6/20	9/20	0.479	4/20	3/20	0.828
<i>General/News</i>	14/20	13/20	0.205	7/20	4/20	0.634	11/20	13/20	0.271	9/20	8/20	0.490	14/20	13/20	0.432

Table 2. Shows the correlation between curated topics and their share of sample tweets among users and their friends.

As part of our future work, we plan on automatically modeling multi-topic user profiles based on varied interests that short text microblog users may have over time. Twitter specific features such as bi-directional network metadata could be used in this computation process.

References

- [1] K. Chen, T. Chen, G. Zheng, O. Jin, E. Yao, and Y. Yu, "Collaborative personalized tweet recommendation," in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pp. 661–670, ACM, 2012.
- [2] S. Goel and R. Kumar, "Folksonomy-based user profile enrichment using clustering and community recommended tags in multiple levels," *Neurocomputing*, vol. 315, pp. 425–438, 2018.
- [3] S. R. Sahoo and B. Gupta, "Hybrid approach for detection of malicious profiles in twitter," *Computers & Electrical Engineering*, vol. 76, pp. 65–81, 2019.
- [4] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi, "Short and tweet: experiments on recommending content from information streams," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1185–1194, ACM, 2010.
- [5] Y. Liu, X. Chen, S. Li, and L. Wang, "A user adaptive model for followee recommendation on twitter," in *Natural Language Understanding and Intelligent Applications*, pp. 425–436, Springer, 2016.
- [6] S. Takimura, R. Harakawa, T. Ogawa, and M. Haseyama, "Twitter followee recommendation based on multimodal ffm considering social relations," in *2018 IEEE 7th Global Conference on Consumer Electronics (GCCE)*, pp. 204–205, IEEE, 2018.
- [7] D. P. Karidi, Y. Stavarakas, and Y. Vassiliou, "Tweet and followee personalized recommendations based on knowledge graphs," *Journal of Ambient Intelligence and Humanized Computing*, vol. 9, no. 6, pp. 2035–2049, 2018.
- [8] R. P. Schumaker, C. S. Labeledz Jr, A. T. Jarmoszko, and L. L. Brown, "Prediction from regional angst—a study of nfl sentiment in twitter using technical stock market charting," *Decision Support Systems*, vol. 98, pp. 80–88, 2017.
- [9] A. Brown, D. Rambaccussing, J. J. Reade, and G. Rossi, "Forecasting with social media: evidence from tweets on soccer matches," *Economic Inquiry*, vol. 56, no. 3, pp. 1748–1763, 2018.
- [10] L. V. Williams and J. J. Reade, "Prediction markets, social media and information efficiency," *Kyklos*, vol. 69, no. 3, pp. 518–556, 2016.
- [11] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.
- [12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
- [13] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1035, Society for Industrial and Applied Mathematics, 2007.
- [14] P. Bholowalia and A. Kumar, "Ebk-means: A clustering technique based on elbow method and k-means in wsn," *International Journal of Computer Applications*, vol. 105, no. 9, 2014.
- [15] L. Recalde and 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*, pp. 94–99, ACM, 2017.
- [16] Y. Halberstam and B. Knight, "Homophily, group size, and the diffusion of political information in social networks: Evidence from twitter," *Journal of Public Economics*, vol. 143, pp. 73–88, 2016.
- [17] M. Z. Rodriguez, C. H. Comin, D. Casanova, O. M. Bruno, D. R. Amancio, L. d. F. Costa, and F. A. Rodrigues, "Clustering algorithms: A comparative approach," *PloS one*, vol. 14, no. 1, p. e0210236, 2019.
- [18] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, vol. 27, no. 1, pp. 415–444, 2001.
- [19] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica: Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.
- [20] A. J. Viera, J. M. Garrett, et al., "Understanding interobserver agreement: the kappa statistic," *Fam med*, vol. 37, no. 5, pp. 360–363, 2005.