Twitter recommendations across dialects: Seeing the big board through Faiss

John Sigurd Hammer
The University of the Faroe Islands
History- and Social Sciences Faculty
0000-0008-8771-9654

Konstantinos Theodoropoulos
University of West Attica
ICE Department
0009-0001-1233-3019

Georgios Drakopoulos
University of West Attica
ICE Department
0000-0002-0975-1877

Phivos Mylonas
University of West Attica
ICE Department
0000-0002-6916-3129

***Abstract*—Twitter account recommendations rely on a broad spectrum of factors including language, which influences all spheres of social life. Most notably, virtually every action of informal human communication online is almost exclusively actualized by language. As it largely remains the main stage upon which such communicative acts unfold freely, language is subject to a myriad of minor changes eventually accumulating to discernible patterns indicating lexical, syntactic, or semantic changes giving rise to dialects. Since often they are readable by speakers of the original language, it is statutory to ask whether recommendations can be made based across dialects and, if so, whether it can be deployed on a scalable back end infrastructure. Two use cases using Faiss for storage and Go as the implementation language yield encouraging results.**

***Index Terms*—multilingual social networks, linguistic diffusion, cross cultural language change, personalization, recommendation, social data mining, Go, vector databases, Faiss, Twitter**

1. Introduction

Twitter recommendation is fundamental across a series of functions such as content personalization, digital cultural awareness, and digital campgains. In light of this, effective recommendations rely on a broad spectrum of attributes primarily pertaining to the use of language in tweets. Language is diachronically the fundamental property of human communication and by definition a complex social phenomenon. As the actualization of communication undergoes an ever-growing alteration by external factors such as the emergence of new digitized communication channels such as social media, language remains necessary yet fragile in that it is in constant flux through numerous pressures and constraints in usage. The ongoing rise of digital services and platforms of communication such as Twitter not only allows for new areas of expression to emerge along with pathways of new linguistic behaviors and new forms; emoticons, abbreviations, phonetic spellings, and other neologisms, but it rather imposes these alterations to happen on a cross lingual basis as social media is becoming increasingly multilingual.

In fact research indicates that multiple languages are used throughout Twitter for social interaction across cultures [1]. In particular the majority of tweets account for English [2] [3], while the rest account mostly for Portuguese [4], Spanish [5], Japanese [6], and Indonesian [7]. Consequently, for acts on languages on a cross-lingual and cross-cultural basis in online social networks using interdisciplinary, innovative qualitative and quantitative approaches can shed light in the multidimensionality of the event of language diffusion during a period of rapid contact-induced linguistic change.

Table 1. NOTATION SYNOPSIS.

Symbol	Meaning	First in
$\triangleq$	Definition or equality by definition	Eq. (1)
$\{s_1, \ldots, s_n\}$	Set with elements $s_1, \ldots, s_n$	Eq. (1)
$\lvert\cdot\rvert$	Set cardinality functional	Eq. (5)
$\tau(\cdot,\cdot)$	Tanimoto set similarity coefficient	Eq. (5)
$\nu(\cdot,\cdot)$	Tversky set similarity index	Eq. (6)
$S_1 \setminus S_2$	Asymmetric set difference	Eq. (6)
$\mathrm{prob}\{\Omega\}$	Probability of event Ω occurring	Eq. (8)
$\kappa_2(\cdot; f)$	Condition number of function $f(\cdot)$	Eq. (12)
$\lVert\cdot\rVert_2$	Euclidean vector norm	Eq. (15)

The primary research objective of this conference paper is a probabilistic graphical model for creating personalized Twitter account recommendations based on a combination of attributes of various types including linguistic, affective, and geospatial ones. Especially important are the dialects of the languages an account is using since in many cases the dialects can be understood by speakers of the original language. Account and language similarity was determined by a large number of attributes as appropriate pertaining to Twitter including linguistic, syntactic, functional, and affective ones. As a concrete case two groups of languages were used, namely British and American English as well as Spanish, Portuguese, and Brazilian Portuguese. As a secondary objective, it is described in detail how this model was implemented in Go using Faiss for embeddings storage.

The remainder of this conference paper is structured as follows. In section 2 the recent scientific literature regarding social network analysis (SNA), graph mining, and vector databases is briefly overviewed. In section 3 the proposed graphical model is explained. The dataset collected and the results obtained are described in section 4. Future research directions are given in 5. Acronyms are explained the first time they are encountered in text. In function and functional definitions parameters come after the formal arguments sep-

978-1-xxxx-xxxx-x/25/$31.00 ©2025 IEEE

arated by a semicolon from the latter. Small boldface letters denote vectors and small ordinary one scalars. The terms *feature* and *attribute* are used interchangeably. Finally, table 1 summarizes the notation of this work.

2. Related work

SNA has evolved over time to a broad field which includes problems such as influence mining [8] and higher order dynamics [9] and interacts with fields like long supply chains [10], linguistics [11] [12], and even biology [13]. Twitter remains a popular microblogging platform even after part of its community moved to Bluesky[1] [2] or Threads[3] [4] with conversations about topics ranging from urban flood management [14] to migration [15] can be found there. Fake news can be discovered with an array of algorithmic tools examined in [16]. Moreover, cross-attention based transformers have been employed as classifiers for identifying offensive tweets [17]. Estimating the sentiment of a tweet using natural language processing (NLP) models is explored in [18]. Recommendation of cultural content can be done as in [19]. The connection between large language models (LLMs) and social networks is explored in [20]. Applications include exploring how Twitter sentiment influences the value of cryptocurrencies [21]. Finally identity bias in generative language models are explored in [22].

Graph mining is one of the algorithmic mainstays of SNA as well as a major research field on its own right [23] [24]. Current research directions include among others graph neural networks (GNNs) [25], graph convolutional networks (GCNs) [26], and graph kernels [27]. Other areas include resilience metrics [28], community discovery structure [29] [30], and approximating directed graphs with undirected ones [31]. Applications include mining for threat on the Web scale [32], multispectral and hyperspectral images [33], brain networks [34], and Industry 4.0 process graphs [35].

The Go language[5] has been developed by Google in order to write systems code [36] with an emphasis on concurrent programming [37]. Since then it has been extended for Web development [38] and code vulnerability analysis [39] among others. Vector databases [40] such as Pinecone have already found many applications including storing human personality embeddings generated from the Myers-Brigs taxonomy indicator (MBTI) [41].

1. https://www.techradar.com/computing/social-media/bluesky-is-the-new-home-for-millions-of-disillusioned-twitter-users-heres-how-to-make-the-switch

2. https://www.popsci.com/diy/how-to-leave-twitter-for-bluesky/

3. https://tbdconference.medium.com/how-to-move-your-following-from-x-to-threads-or-anywhere-50622ceadaa1

4. https://www.colorado.edu/today/2023/07/10/threads-surging-mass-migration-twitter-likely-remain-uphill-battle

5. https://go.dev

3. Methodlogy

3.1. Model overview

Communication is intrinsically bounded to language and thus the phenomenon of communication is a fundamentally a social phenomenon as well. For a new linguistic form such as a dialect to succeed, at least two things must happen: First, users must come into contact with the new form; second, they must decide to adopt it. The first condition implies that language change is related to the structure of social networks. But while modeling of the social factors of linguistic variation and change in terms of networks are being explored, questions like what changes are observed between the dialects of a spoken language in the context of social networks have not yet been fully addressed. The proposed model has been developed in order to account exactly for that using Twitter as an explanatory framework.

The following factors were taken into consideration in the proposed recommendation methodology. The former works in the same way as in most recommendation methodologies, whereas the latter reinforces or weakens account connections based on language similarity. This takes into considerations different language dialects, which may have geographical or even generational causes. Regarding the latter, as language evolves over time each successive generational cohort tends to have its own preferred spelling and syntax as well as its own preferred words.

- The similarity between accounts.
- The similarity between languages.

Regarding implementation the language of choice was Go, a lightweight language designed for systems and Web back end programming. As such, it has numerous APIs for databases including Faiss, a popular vector database. Although Go is an unconventional choice for this task, it has adequate functionality. Moreover, should the need arise if the existing implementation scales up, Python can be supplement more specialized data mining operations. In order to perform recommendation across language dialects, two language groups were considered based on their overall Twitter popularity as evaluated by their estimated total number of tweets, namely American and British English as well as Spanish, Portuguese, and Brazilian Portuguese. Said recommendation was evaluated using the actual tweet follow relationships as the ground truth.

3.2. Account and language similarity

Initially let U be the set of accounts under consideration. Since each such account may well tweet in more than one language, let L_u be the language set of an account u.

$$L_u \triangleq \{l_{u,1}, \ldots, l_{u,n}\} \tag{1}$$

In this case the set of languages under consideration L is given as the union of every L_a shown in equation (1).

$$L = \bigcup_{u \in U} L_u \tag{2}$$

The attributes collected from the dataset for the two language groups under consideration as well as their types are given in table 2. Said attributes cover a broad spectrum of written communication and pertain both to the language as well as to Twitter itself. Although the latter is not part of any language, it is the medium through which communication takes place and as such it shapes the message to an extent.

Table 2. ATTRIBUTES FOR ACCOUNT SIMILARITY.

Attribute	Type
Average number of words per tweet	Syntactic
Maximum number of words per tweet	Syntactic
Minimum number of words per tweet	Syntactic
Average number of words per sentence	Syntactic
Average number of punctuation marks per tweet	Syntactic
Average number of exclamation marks per tweet	Syntactic
Zipf exponent of word length distribution	Syntactic
Zipf exponent of word frequency distribution	Syntactic
Average percentage of words in all capitals	Spelling
Average percentage of words in the same alphabet	Spelling
Average percentage of emotionally charged words	Affective
Percentage of positive charged tweets	Affective
Percentage of negative charged tweets	Affective
Percentage of affectively neutral tweets	Affective
Percentage of tweets in the dataset	Functional
Average number of hashtags	Functional
Zipf exponent of hashtag length distribution	Functional
Number of trending hashtags	Functional
Percentage of tweets between midnight and noon	Functional

These attributes have been collected and computed for each account and constitute a mix of specific and general features. Every attribute has been normalized to a maximum value of one such that all features would be in the same range. The similarty score between every account pair is computed as the Euclidean distance in equation (3). Therein n is the total number of features. This normalization keeps metric distance h_n between zero and one.

$$h_n(\mathbf{u}_i, \mathbf{u}_j) \triangleq \frac{1}{\sqrt{n}} \left\| \mathbf{u}_i - \mathbf{u}_j \right\|_2 \tag{3}$$

The Gaussian kernel of equation (4) gives sharper bounds because of its decay rate. The variance σ_0^2 has been selected such that the numerator takes values between zero and four. The latter has been selected as it is very close to zero and it is a typical threshold in the exponential decay.

$$h_e\big(\mathbf{u}_i, \mathbf{u}_j; \sigma_0^2\big) \triangleq e^{-h_n^2(\mathbf{u}_i, \mathbf{u}_j)/\sigma_0^2} \tag{4}$$

The attributes used to quantifty language similarity are given in table 3. Notice that they can be also used to quantify similarity between languages, but in this case lower scores should be accounted for, especially between languages from different language families and in particular between any pair of proto-languages defining a family language. For example, Russian has a high number of cases and a low number of prepositions in contrast to German[6]. Additionally, the sets of most frequent words for a given language can be extended beyond those available in the dataset either on a larger Twitter dataset or across different social media. However, in the latter case care should be taken in order to account for the purpose of the social media under consideration as for instance LinkedIn in intended exclusively for public professional communication and as such the vocabulary is accordingly tailored and standardized. Another point which should be highlighted is that pronounciation is one of the determining factor between a language and its dialects. For instance the preposition "auf" is pronounced as /aʊf/ in German but as /ʊf/ in Swiss German.

6. The old Russian word for Germans немец means "mute" stemming from the fact that Germans did not speak Russian. This signifies the social role of language.

Table 3. ATTRIBUTES FOR LANGUAGE SIMILARITY.

Attribute	Type
Alphabet	Spelling
Set of available phonemes	Phonetic
Set of phoneme pronounciation	Phonetic
Distinct orders for the same phonemes	Phonetic
Set of cases	Grammatic
Set of prepositions	Grammatic
Set of numbers	Grammatic
Set of definite articles	Grammatic
Set of hundred most frequent words in dataset	Social
Set of accounts using the given language in dataset	Social
Set of countries where tweets in this language are used	Geospatial

In order for a tangible similarity score between any two accounts to be derived as in equation (4) each account has been linked to a numerical vector with the features shown in table 2. This is straightforward since tweets and account activity in general can be reduced to a collection of scores. On the other hand, a language being systemic and dynamic is hard to be mapped to one. Therefore it is more flexible to assign languages to a set of feature sets and rely on set similarity metrics. Recall that for two not necessarily distinct sets S_1 and S_2 the Tanimoto similarity coefficient as defined in equation (5) is one way to measure their similarity.

$$\tau\left(S_1, S_2\right) \triangleq \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} = \frac{|S_1 \cap S_2|}{|S_1| + |S_2| - |S_1 \cap S_2|} \tag{5}$$

The right hand side of equation (5) follows directly from Venn diagrams and it is frequently more efficient computationally as the set intersection is computed only once and when one of the sets is comparably smaller then the other one it is efficient. On the contrary, set union is not always efficient and depends on its arguments.

The Tanimoto similarity is symmetric with respect to is arguments, namely there is no way to discern between a template set and a variant thereof. In cases where this is desired, the asymmetric Tversky index of equation (6) for a reference or blueprint set S_1 and a modification or derivative set S_2 can be used. Using a language as reference adds an element of causality and allows setting a similarity chain where a dialect has dialects of its own.

$$\nu\left(S_1, S_2\right) \triangleq \frac{|S_1 \cap S_2|}{|S_1 \cap S_2| + w_1|S_1 \setminus S_2| + w_2|S_2 \setminus S_1|} \tag{6}$$

The positive factors w_1 and w_2 in the denominator of equation (6) determine the relative penalties of the elements

not present in S_1 and S_2 respectively. There is no general rules for setting their values, but as a rule the conditions of equation (7) often yield acceptable results and they are used here. The interpretation of these rules is that w_1 is β_0 times more important that w_2 in a constrained setting.

$$\begin{cases} w_1 + w_2 \;=\; 1 \\ \\ w_1 \;=\; \beta_0 w_2 \end{cases} \Rightarrow \begin{cases} w_1 \;=\; \dfrac{\beta_0}{1+\beta_0} \\ \\ w_2 \;=\; \dfrac{1}{1+\beta_0} \end{cases} \tag{7}$$

Given the above description connections between account pairs and language pairs as well as between accounts and languages can be defined and the appropriate connections under this model be computed. The above can be represented as a weighted graph with two kinds of vertices, namely languages and accounts as shown in figure 1. Therein neither the edge weights nor most of the connections between accounts are shown to avoid cluttering. The edges between accounts are bidirectional as the way account similarity is computed focuses primarily on linguistic features and precludes preferences like following influential accounts, which however can be an attribute in a suitable extension of the proposed model. The edges between accounts and languages indicate language preferences, in this case that u_1 tweets equally in l_1 and l_2. Finally, vertices denoting dialects point to the vertex of the proto-language, namely the language they are derived from. In the particular case shown in the figure below l_1 is a dialect of l_2.

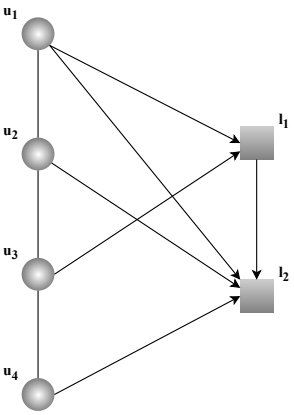

Figure 1. Conceptual representation of the proposed method.

Under the proposed model the probability $p_{i,k}$ of account u_i to tweet in language l_k is the normalized frequency of their tweets in that language as shown in equation (8).

$$p_{i,k} \triangleq \text{prob}\left\{u_i \text{ tweets in } l_k\right\} \triangleq \frac{\text{tweets from } u_i \text{ in } l_k}{\text{tweets from } u_i} \tag{8}$$

The weight $w_{i,j,k}$ between accounts u_i and u_j with respect to language l_k are given by equation (4) and between languages either by equation (5) or as in the case of our experiments by the latter because of its sharper bounds. The inclusion of language l_k in this weight allows assessing in a straightforward the similarity of these accounts over different ones in case the of multilingual accounts.

The weight $w_{k,r}$ between languages l_k and l_r in the set L of equation (1) are given either by the Tanimoto coefficient of (5) or by the Tversky index of (6) depending on the context. Specifically, if an account has a specific language preference or if an account tweets only in a single language ignoring its dialects, then (6) may be more appropriate.

Finally, the recommendation score $s_{i,j}$ of account u_j for u_i is given by equation (9). The highest the score u_j receives, then the most relevant their tweets are deemed to be to those of u_i and thus the higher they are in the account recommendation list presented to u_i.

$$s_{i,j} \triangleq \sum_{l_k \in L_{u_i}} p_{i,k} \sum_{l_r \in L_{u_j}} p_{j,r} \frac{1+\eta_0}{\dfrac{1}{w_{i,j,r}} + \dfrac{\eta_0}{w_{k,r}}} \tag{9}$$

Certain remarks about $s_{i,j}$ are in order. First, this score is pairwise and depends only on the profiles of the accounts directly involved without taking into consideration any communities they may belong to. Second, from a numerical perspective $s_{i,j}$ keeps computations in a well defined scale as it ranges over weighted sums whose positive weights sum up to one. Moreover, the harmonic mean in the second sum not only is quite resistant to zero or near zero values of $w_{i,j,r}$ or $w_{k,r}$ compared to the other two Pythagorean means, namely the arithmetic and the geometric mean, but also tends to be closer to the true mean compared to the other two means as it is much less prone to outliers. Third, the tunable parameter η_0 offers flexibility as it can favor either account or language similarity. In our experiments the former took a higher relative weight, being more dynamic.

3.3. Linguistic attributes

The Zipf exponent for the word length distribution ζ_k for a given language is defined the exponent γ_0 of the power law of equation (10). The Zipf exponent is intrinsically related to the principle of least action regarding human activity and motivation. In general, power laws tend to appear in human activity such as the distribution of road lengths in transportation networks and the distribution of number of sea connections between ports. On the contrary, machine-generated activity is frequently described by exponential laws because of feedback or transient phenomena.

$$\zeta_k \triangleq \alpha_0 k^{-\gamma_0} \tag{10}$$

Each natural language has its own Zipf exponent with minor variations accounting for local dialects. Typically such computations rely on some preprocessing phase such as using the Porter stemmer, but also exponents based on the original words have been reported. The actual computation can take place in a number of ways, linearization as shown in equation (11) being the most straightforward of them.

$$\ln \zeta_k \;=\; -\gamma_0 \ln k + \ln \alpha_0 \tag{11}$$

Despite the easy form of equation (11) which leads to linear least squares (LS) fitting with a linear system with rich structure and high interpretability, the numerical properties of the logarithm may lead to instability for small values of the argument k. One way to evaluate the numerical difficulty of evaluating a function taking and returning a single value is the condition number of equation (12). Observe that by definition the condition number need not be the same everywhere. Said difficulty is not in the sense of computational complexity, but instead it can be attributed to phenomena like floating point precision, catastrophic cancellation, and the implementation of floating point operations [42] [43].

$$\kappa_2(x; f) \triangleq \left| \frac{x\, f'(x)}{f(x)} \right| \tag{12}$$

Applying the above definition to the natural logarithm function results in the condition number of equation (13). From it it follows that when x is close to one, then numerical instability may be an issue. Still, for larger values of x, as is the case here, computations become more reliable.

$$\kappa_2(x; \ln x) \triangleq \left| \frac{x \dfrac{1}{x}}{\ln x} \right| \; = \; \left| \frac{1}{\ln x} \right| \tag{13}$$

It should be noted here that power laws have close connections to fractals. Generalizations of equation (10) include the multifractal distribution of equation (14) where multiple exponents are allowed. In this case the number of exponents n plays an important role in the distribution besides the actual numbers of the exponents γ_i themselves. The positive scaling factors α_i indicate the relative contribution of each component to the total distribution and sum up to one.

$$\zeta_k \triangleq \sum_{i=1}^{n} \alpha_i k^{-\gamma_i}, \quad \sum_{i=1}^{n} \alpha_i \; = \; 1, \alpha_i > 0 \tag{14}$$

As a general side note, power laws are also present in information retrieval (IR) in precision-recall diagrams, the distribution of the number of documents in a collection, as well as in the word distribution length inside documents. The latter is important in creating efficient graph embeddings of documents. In turn, this can lead to efficient algorithmic approximation and implementation of cosine similarity queries in vector databases such as Pinecone[7] and Faiss[8].

3.4. Affective attributes

Emotions are not only a major motivation of human action, but in the context of this work are also indicative of the overall Twitter public sentiment. The latter does not consist of a single affective polarity, but rather instead of the emotional distribution over a given time frame of reference along with major events which may potentially drive changes in it. Affective analysis is done by established models like Plutchik's wheel [44] and the universal emotion theory by Ekman [45]. The primary emotions according to the former model are given in table 4.

Table 4. PRIMARY EMOTIONS IN PLUTCHIK'S WHEEL.

Emotion	Sign
Neutral	Neutral (not counted)
Surprise	Positive or negative depending on the context
Anticipation	Positive or negative depending on the context
Anger	Negative
Fear	Negative
Disgust	Negative
Sadness	Negative
Happiness	Positive
Trust	Positive

The affective sign of the tweets of an account not only can reveal important information about how they perceive and react to events, but it also may be an indication of the generational cohort this account belongs to. A concrete example was the tragedy of Germanwings flight 9525[9] in 2015. Analysis of the Twitter sentiment in the wake of the incident indicted a differentiation in the reaction in terms of affective state and in the expression thereof [46] [47]. This can be used in the general direction of discerning generational causes in the uses of subdialects and the adaptation of a language to the needs of a specific cohort.

3.5. Geolocation attributes

Location plays an important role in SNA as it readily provides a clustering of accounts and tweets on multiple geographic levels such as continent, country, region, and even in some cases township in large urban metroplexes with their own domain names. Tweets collected via the Twitter API have been geo-tagged with the geographical "box" of origin with the extents circumscribing the borders of the regions under consideration, ensuring thereofre that only tweets originating from them were included in the dataset. The latitude and longitude coordinates of each tweet was checked with the coordinates of the national and regional borders as encoded by GIS files publicly available through the Global Administrative Areas database (GADM)[10].

4. Results

4.1. Dataset

The dataset has been created by obtaining tweets from the Twitter API and dowloading them in JSON[11] files for further processing, namely primarily the extrtaction of the vectors to be stored in a Faiss installation. At the time of obtaining the dataset (May 2020) the Twitter API was free[12]

7. https://pinecone.io
8. https://faiss.ai
9. https://en.wikipedia.org/wiki/Germanwings_Flight_9525
10. https://gadm.org
11. https://json.org
12. https://medium.com/newtargetinc/twitter-api-is-no-longer-free-now-what-2a57e157696f

and only the generation of the OAuth tokens was required. A summary of the dataset is given in table 5. Additionally, the values of the parameters β_0 and η_0 of equations (6) and (9) were $2/3$ and $1/3$ respectively.

Table 5. DATASET SYNOPSIS.

Property	Value
Number of British English tweets	6618
Number of American English tweets	6116
Number of Spanish tweets	1132
Number of Portuguese tweets	894
Number of Brazilian Portuguese tweets	1094
Average number of followers	719.6667
Average followers to followees ratio	37.25

As stated earlier Go is a high level language designed purposefully with a simple syntax aiming mainly at high performance as well as high maintainability even in extensive codebases spanning numerous projects. It is compiled and statically typed with inherent support for concurrency, especially on current multicore hardware, networking, and garbage collection. Although not the go to language (pun intended) for data science, its library has more than sufficient functionality for the purposes of this work and, more importantly, it can interface seamlessly with Faiss.

Faiss belongs to the emerging family of vector dabases where the geometry between embeddings representing documents and queries on them plays an important role. In particular the similarity between a query $\mathbf{q}$ and a document $\mathbf{s}$ is determined by the cosine similarity metric shown in equation (15). Therein ϑ is the angle between these two vectors. In fact, it is possible to compute the angle of a query to an entire document space efficiently provided that a basis has been computed, especially as Faiss can derive a result for dense spaces. Faiss has been used in this work to store both the numerical vectors resulting for each account as well as the profile set for each language and its dialects.

$$\frac{\mathbf{q}^T\mathbf{s}}{\|\mathbf{q}\|_2 \|\mathbf{s}\|_2} = \cos\vartheta(\mathbf{q},\mathbf{s}) \tag{15}$$

Through the Go API for Faiss used[13] bindings for the Faiss library were obtained. Having built and installed the Faiss C API which was a prerequisite the dynamic library libfaiss_c.so was created and used to utilize said libray. Other packages considered[14] [15] offer comparable functionality.

4.2. Recommendation evaluation

In this subsection the proposed recommendation system is assessed. As ground truth the The following two language groups have been considered. In each such group the protolanguage, namely the basis from which the others derived, is mentioned first.

- **G1:** British English and American English.
- **G2:** Spanish, Portuguese, and Brazilian Portuguese.

The ground truth for this dataset is the set of actual follow relationships. The accuracy, precision, recall, and F1 scores of the two groups are shown in table 6.

Table 6. EVALUATION METRICS.

Group	Accuracy	Precision	Recall	F1
G1	0.8725	0.8544	0.8746	0.8644
G2	0.8945	0.8717	0.8933	0.8824

Another question is what is the effect of adopting cross dialect recommendations. To this end only accounts tweeting in the same language of the account getting the recommendations are considered. Although this is a limiting case for the proposed model, it nevertheless remains a valid one. Similarly to the previous case, the results are in table 7.

Table 7. EVALUATION METRICS IGNORING DIALECTS.

Group	Accuracy	Precision	Recall	F1
G1	0.7945	0.7634	0.8011	0.7817
G2	0.8216	0.8133	0.8545	0.8333

From the tables above the following can be said. First, the proposed recommendation system achieves a considerable performance in all metrics considered. Also, when dialects are ignored, then recommendation quality is degraded, indicating their importance especially for the second group.

5. Conclusions and future work

This conference paper focuses on a probabilistic graph model for recommending accounts in social media based on account and language similarity based on a large number of attributes including linguistic, affective, and geospatial ones. An account is given recommendations based on the languages they use as well as in dialects thereof which is the novelty of this conference paper. The premise is that such recommendations can potentially be useful as dialects in many cases can be read by speakers of the original language. As a concrete case study two language groups, namely American and British English as well as Spanish, Portuguese, and Brazilian Portuguese, were examined on Twitter. These languages have been selected because of their high populariy on Twitter based on the total number of tweets. The ground truth was considered the set of actual follow relationships already existing on Twitter.

This work can be extended in a number of ways. Regarding the computational part, bigger datasets and more language groups can be used. Moreover, the proposed recommendation scheme can be extended to a framework with more similarity metrics between accounts and languages which can furthermore be aggregated to improve recommendations. Moreover, the proposed model can be extended to yield cross-language recommendations between language families. For instance, English has some similarity to Spanish, mainly through French. Such linguistic dependencies can be captured by the proposed model.

13. github.com/DataIntelligenceCrew/go-faiss package
14. github.com/DataIntelligenceCrew/go-faiss
15. github.com/blevesearch/go-faiss

Regarding the algorithmic part, the rise of social media has led to the increase of linguistic diversity in all levels from spelling, grammar, and semantics across the lexicon. Language change results primarily from the differential propagation of linguistic variants and determining the factors shaping constitute a primary research objective. Factors which may provide explanation are the following:

- **Twitter connections:** Ties between accounts may facilitate or inhibit linguistic change contributing respectively to cohesion and uniformity. These factors may well coexist giving to each language its own speed. Thus languages may be clustered according to their social media variation rate.
- **Diffusion model:** In contested change the diffusion spread depends on both the adamancy of those resisting it as well as the countervailing influence of those in favor. Thus the probability of adopting change is inversely proportional to the neighborhood size. Conversely, in uncontested models contagion spread is proportional to the number of connections.
- **Language resistance:** Some natural languages may be more resistant to change than others. Since online language contact and other social factors redounds in the speed of variation more types of change occurring more frequently in languages that are more affected by contact may be seen.

The ever broadening empirical investigation of language change as a social phenomenon allows sociolinguistics and other fields to establish more robust typologies of change and its diffusion. A final question would be whether the kind of change is uniform across languages and cultures and across time. The nature of the occurrence of the language change and its diversification not only on individual level but rather in the community seem to be starting points.

Acknowledgment

This conference paper is part of Project 451, a long term research initiative with a primary objective of developing novel, scalable, numerically stable, and interpretable higher order analytics.

References

[1] T. Louf, J. J. Ramasco, D. Sánchez, and M. Karsai, "When dialects collide: How socioeconomic mixing affects language use," *EPJ Data Science*, vol. 14, no. 1, p. 47, 2025.

[2] W. D. W. Gonzales, "Sociolinguistic analysis with missing metadata? Leveraging linguistic and semiotic resources through deep learning to investigate English variation and change on Twitter," *Applied Linguistics*, vol. 46, no. 3, pp. 411–434, 2025.

[3] B. Tahir and M. A. Mehmood, "TepiSense: A social computing based real-time epidemic surveillance system using artificial intelligence," *IEEE Access*, 2025.

[4] F. Carneiro, D. Vianna, J. Carvalho, A. Plastino, and A. Paes, "BERTweet.BR: A pre-trained language model for tweets in Portuguese," *Neural Computing and Applications*, vol. 37, no. 6, pp. 4363–4385, 2025.

[5] C. Maíz-Arévalo, "Striking the balance between friendliness and professionality: Pragmatic uses and functions of emoji by spanish MPs on Twitter/X bios," *Internet Pragmatics*, 2025.

[6] T. Murayama, K. Miyazaki, Y. Matsubara, and Y. Sakurai, "Linguistic landscape of generative AI perception: A global Twitter analysis across 14 languages," in *AAAI*, vol. 19, 2025, pp. 1262–1294.

[7] E. Utami, I. Oyong, S. Raharjo, A. Dwi Hartanto, and S. Adi, "Supervised learning and resampling techniques on DISC personality classification using Twitter information in Bahasa Indonesia," *Applied Computing and Informatics*, vol. 21, no. 1/2, pp. 141–151, 2025.

[8] A. Zareie and R. Sakellariou, "Fuzzy influence maximization in social networks," *ACM Transactions on the Web*, vol. 18, no. 3, pp. 1–28, 2024.

[9] I. Iacopini, M. Karsai, and A. Barrat, "The temporal dynamics of group interactions in higher-order social networks," *Nature Communications*, vol. 15, no. 1, 2024.

[10] G. Ramya, K. S. H. Nathan, A. Srinithi, and P. Jaykrishna, "Social network analysis (1980–2023)," *Advances in Communication and Applications: Proceedings of ERCICA 2024, Volume 2*, vol. 1398, p. 115, 2025.

[11] C. Li, Z. Li, and X. Gao, "The application of social network analysis in applied linguistics research: A systematic review," *Applied Linguistics Review*, vol. 16, no. 4, pp. 1449–1479, 2025.

[12] K. K. Terry and R. Bayley, *Social network analysis in second language research: Theory and methods*. Routledge, 2024.

[13] F. Beghini, J. Pullman, M. Alexander, S. V. Shridhar, D. Prinster, A. Singh, R. Matute Juárez, E. M. Airoldi, I. L. Brito, and N. A. Christakis, "Gut microbiome strain-sharing within isolated village social networks," *Nature*, vol. 637, no. 8044, pp. 167–175, 2025.

[14] M. W. Boota, H. M. Zwain, X. Shi, J. Guo, Y. Li, M. Tayyab, M. H. A. A. Soomro, C. Hu, C. Liu, Y. Wang *et al.*, "How effective is twitter (X) social media data for urban flood management?" *Journal of Hydrology*, vol. 634, 2024.

[15] J. E. Blumenstock, G. Chi, and X. Tan, "Migration and the value of social networks," *Review of Economic Studies*, vol. 92, no. 1, pp. 97–128, 2025.

[16] S. Baribi-Bartov, B. Swire-Thompson, and N. Grinberg, "Supersharers of fake news on Twitter," *Science*, vol. 384, no. 6699, pp. 979–982, 2024.

[17] J. Paul, S. Mallick, A. Mitra, A. Roy, and J. Sil, "Multi-modal Twitter data analysis for identifying offensive posts using a deep cross-attention–based transformer framework," *ACM Transactions on Knowledge Discovery from Data*, vol. 19, no. 3, pp. 1–30, 2025.

[18] A. Albladi, M. Islam, and C. Seals, "Sentiment analysis of Twitter data using NLP models: A comprehensive review," *IEEE Access*, 2025.

[19] G. Drakopoulos, I. Giannoukou, S. Sioutas, and P. Mylonas, "Self organizing maps for cultural content delivery," *NCAA*, vol. 31, no. 7, 2022.

[20] Z. Shu, X. Sun, and H. Cheng, "When LLM meets hypergraph: A sociological analysis on personality via online social networks," in *CIKM*. ACM, 2024, pp. 2087–2096.

[21] B. Amirshahi and S. Lahmiri, "Investigating the effectiveness of Twitter sentiment in cryptocurrency close price prediction by using deep learning," *Expert Systems*, vol. 42, no. 1, 2025.

[22] T. Hu, Y. Kyrychenko, S. Rathje, N. Collier, S. van der Linden, and J. Roozenbeek, "Generative language models exhibit social identity biases," *Nature Computational Science*, vol. 5, no. 1, pp. 65–75, 2025.

[23] G. Lee, F. Bu, T. Eliassi-Rad, and K. Shin, "A survey on hypergraph mining: Patterns, tools, and generators," *ACM Computing Surveys*, vol. 57, no. 8, pp. 1–36, 2025.

[24] C. Niu, G. Pang, and L. Chen, "Affinity uncertainty-based hard negative mining in graph contrastive learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 9, pp. 11 681–11 691, 2024.

[25] X. Hu, H. Chen, J. Zhang, H. Chen, S. Liu, X. Li, Y. Wang, and X. Xue, "GAT-COBO: Cost-sensitive graph neural network for telecom fraud detection," *IEEE Transactions on Big Data*, vol. 10, no. 4, pp. 528–542, 2024.

[26] R. Bhattacharya, N. K. Nagwani, D. S. Asudani, G. S. Chhabra, S. Bhattacharya, and S. Kadam, "A comprehensive overview of graph convolutional network." Springer, 2025, pp. 1–19.

[27] L. Wang, W. Fan, J. Li, Y. Ma, and Q. Li, "Fast graph condensation with structure-based neural tangent kernel," in *Web Conference*. ACM, 2024, pp. 4439–4448.

[28] G. Drakopoulos, X. Liapakis, G. Tzimas, and P. Mylonas, "A graph resilience metric based on paths: Higher order analytics with GPU," in *ICTAI*. IEEE, 2018, pp. 884–891.

[29] Y. Liu, J. Li, Y. Chen, R. Wu, E. Wang, J. Zhou, S. Tian, S. Shen, X. Fu, C. Meng *et al.*, "Revisiting modularity maximization for graph clustering: A contrastive learning perspective," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 2024, pp. 1968–1979.

[30] G. Drakopoulos and P. Mylonas, "A genetic algorithm for Boolean semiring matrix factorization with applications to graph mining," in *Big Data*. IEEE, 2022.

[31] G. Drakopoulos, E. Kafeza, P. Mylonas, and S. Sioutas, "Approximate high dimensional graph mining with matrix polar factorization: A Twitter application," in *IEEE Big Data*. IEEE, 2021, pp. 4441–4449.

[32] S. Freitas and A. Gharib, "Web scale graph mining for cyber threat intelligence," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[33] R. Guan, W. Tu, Z. Li, H. Yu, D. Hu, Y. Chen, C. Tang, Q. Yuan, and X. Liu, "Spatial-spectral graph contrastive clustering with hard sample mining for hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[34] C. Li, T. Tang, Y. Pan, L. Yang, S. Zhang, Z. Chen, P. Li, D. Gao, H. Chen, F. Li *et al.*, "An efficient graph learning system for emotion recognition inspired by the cognitive prior graph of EEG brain network," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 4, pp. 7130–7144, 2024.

[35] G. Drakopoulos, E. Kafeza, P. Mylonas, and S. Sioutas, "Process mining analytics for Industry 4.0 with graph signal processing," in *WEBIST*. SCITEPRESS, 2021, pp. 553–560.

[36] J. Bodner, *Learning Go*. O'Reilly Media, Inc., 2024.

[37] J. Cutajar, *Learn Concurrent Programming with Go*. Simon and Schuster, 2024.

[38] X. Gu, M. Chen, Y. Lin, Y. Hu, H. Zhang, C. Wan, Z. Wei, Y. Xu, and J. Wang, "On the effectiveness of large language models in domain-specific code generation," *ACM Transactions on Software Engineering and Methodology*, vol. 34, no. 3, pp. 1–22, 2025.

[39] J. Hu, L. Zhang, C. Liu, S. Yang, S. Huang, and Y. Liu, "Empirical analysis of vulnerabilities life cycle in golang ecosystem," in *International Conference on Software Engineering*. IEEE/ACM, 2024, pp. 1–13.

[40] Z. Jing, Y. Su, and Y. Han, "When large language models meet vector databases: A survey," in *Conference on Artificial Intelligence x Multimedia (AIxMM)*. IEEE, 2025, pp. 7–13.

[41] G. Drakopoulos and P. Mylonas, "Clustering MBTI personalities with graph filters and self organizing maps over Pinecone," in *Big Data*. IEEE, 2024.

[42] D. J. Higham, "Condition numbers and their condition numbers," *Linear Algebra and its Applications*, vol. 214, pp. 193–213, 1995.

[43] W. Cao and W. Zhang, "An analysis and solution of ill-conditioning in physics-informed neural networks," *Journal of Computational Physics*, vol. 520, 2025.

[44] H. Park, "A study on facial expression design guidelines for digital human modeling-focusing on Plutchik's eight primary emotions," *International Journal of Advanced Culture Technology*, vol. 13, no. 1, pp. 201–217, 2025.

[45] P. Ekman, "Facial expressions of emotion: An old controversy and new findings," *Philosophical transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 335, no. 1273, pp. 63–69, 1992.

[46] P. Masip, C. Ruiz, and J. Suau, "Contesting professional procedures of journalists: Public conversation on Twitter after Germanwings accident," *Digital journalism*, vol. 7, no. 6, pp. 762–782, 2019.

[47] V. Hoste, C. Van Hee, and K. Poels, "Towards a framework for the automatic detection of crisis emotions on social media: A corpus analysis of the tweets posted after the crash of Germanwings flight 9525," in *HUSO*, 2016, pp. 29–32.