HeyLo: Visualizing User Interests from Twitter Using Emoji in Mixed Reality

Hunter Harris, Makayla Thompson, Isaac Griffith, Paul Bodily
Department of Computer Science
Idaho State University
Pocatello, ID
harrhunt@isu.edu, thommaka@isu.edu, grifisaa@isu.edu, bodipaul@isu.edu

***Abstract*—We tackle the problem of analyzing a user's interests from social media content and subsequently visualizing these interests in an extended reality environment. We compare five models for extracting interests from Twitter users and how we can measure the effectiveness of these models. We also look at how these interest extraction models fit in the context of HeyLo, an extended reality computational creativity (XRCC) framework for visualizing potential conversational topics. The chosen interests for a particular person are visualized using emoji. We accomplish this by using an emoji2vector model to find the closest related emoji to a given interest. We perform a comparative analysis between the five interest extraction models on real-world users and their tweets, evaluating specificity, variance, and relevance.**

***Index Terms*—extend reality, computational creativity, sentiment analysis, twitter, emoji**

Source code is available at https://github.com/harrhunt/HeyLo

I. Introduction

The growth of AI for improving interactions on social media, online retail, and media-service recommendation systems signals the importance placed on the ability of a system to understand and leverage a user's interests for effective communication. Despite the value added from technology in these domains, relatively unexplored is how to improve real-life, day-to-day interactions using similar technologies in mixed reality—a medium with the potential to directly perceive and enhance a user's experience with the world. Already technology has emerged to translate written language in real time via augmented reality [1]. It is not difficult to imagine how similar systems may soon allow targeted advertising on billboards or storefronts. These technologies carry the potential to improve the effectiveness of all forms of day-to-day interaction.

Interaction, and conversation in particular, represents an important form of creativity. Effective dialogue hinges as much on generating novelty, value, surprise, and intention as effective artefacts in any artistic domain. Much of what is essential in an effective conversation focuses around finding areas of common interest.

We have designed an extended reality computational creativity (XRCC) framework called *HeyLo*. This system leverages the capabilities of mixed reality with the principles of computational creativity in order to analyze user interests and visually suggest (using emoji) potential conversation topics based on the areas of interest as determined from analyzing social media. Our main interest in developing HeyLo is to research how XRCC can be used to enable a user to meaningfully engage and interact with another randomly encountered user. In this paper, we focus on the design of the HeyLo system itself and a comparative analysis of interest-extraction algorithms for use in HeyLo, leaving the analysis of human interaction using HeyLo for future work.

Much of the design of HeyLo and the interest extraction and visualization algorithms we compare rely on previous work in these fields. Topic modeling, which is in some sense analogous to interest extraction, has been attempted using a variety of methods including Latent Dirichlet Allocation [2], [3], chi-square analysis [4], and online dynamic topic modeling [5].

Whereas both topic modeling and interest extraction represent forms of semantic analysis, the latter is also more specifically a form of *sentiment* analysis. Although several sentiment analysis methods aim to classify text as merely positive or negative (e.g., for categorizing product reviews), some models exist which seek to extract greater variety of sentiment. The Empath [6] algorithm analyzes text across customizable lexical categories. The system requires the user to define a mapping from sets of words and phrases to each category. Given some input text, Empath returns a set of key-value pairs where keys are lexical categories and values are scores for that category. This allows Empath to extract and rank topical and emotional content from a body of text.

Also critical are tools for identifying relationships between words on the basis of sentiment. ConceptNet [7] is a semantic network that helps computers understand the meaning humans put on words. It connects nodes represented as words and phrases with labeled edges that represent the relationship between the words and phrases. These edges can be traversed using the representational state transfer (REST) API.

Interest visualization is significantly aided by tools based around ideograms (e.g., emoji). Eisner et al. [8] trained an emoji2vector model (similar to a word2vec model but with emoji) using the Unicode emoji name and descriptions. For each word in the description, its vector representation retrieved from the Google News word2vec model is summed together to build a vector representation for the emoji.

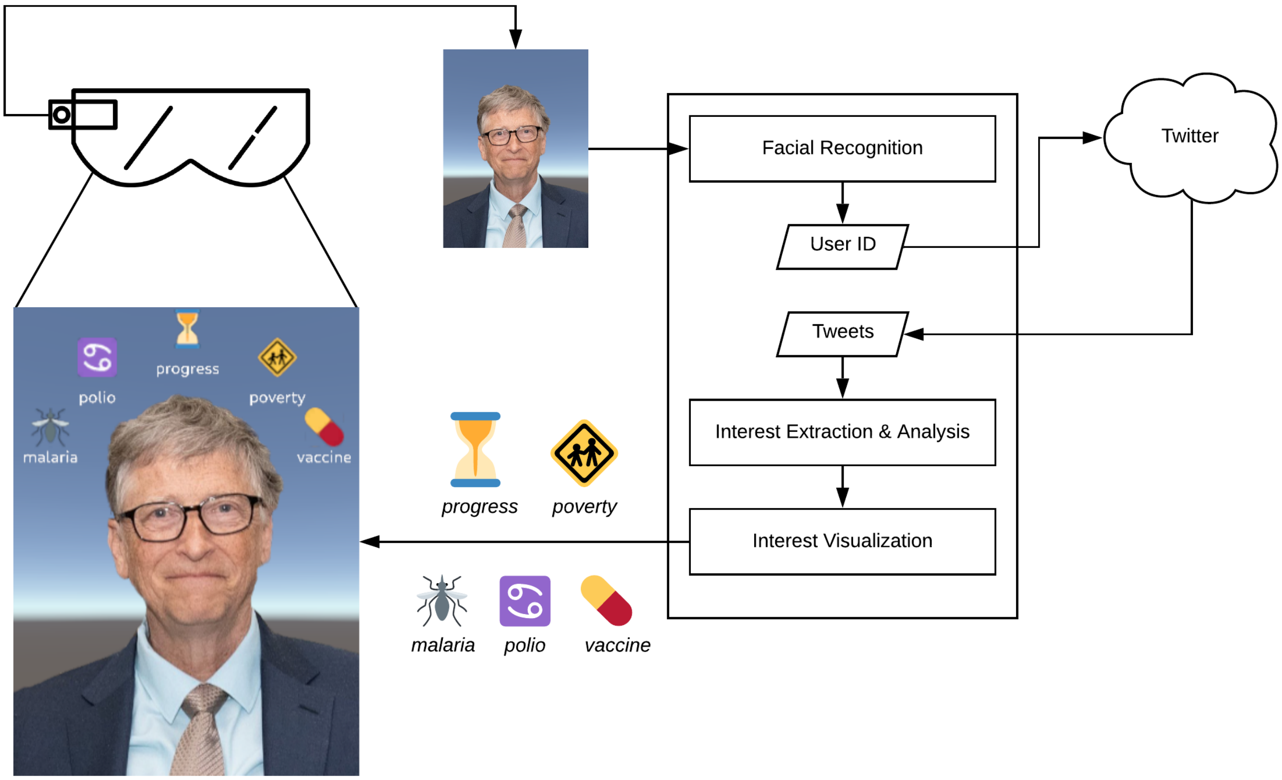

Fig. 1. An overview of the HeyLo system. Given an image of a person (e.g., from an XR headset), the system recognizes the user, gathers tweets, extracts interests, and outputs visualizations for display on the image. Picture of Bill Gates is in the public domain.

II. Methods

In this section we describe the design and operation of the HeyLo XRCC system. We first provide a high-level overview of the system. At the heart of this system is a model that extracts user interests from a set of tweets. For purposes of comparison, we define five possible implementations for this model. Finally we define four metrics by which we comparatively evaluate the performance of each of these five implementations.

A. *HeyLo System Overview*

Taking as input the image f of a person, the system outputs a user u and a set R of l weighted keywords representing interests of user u ($l = 5$ for all of our examples). Each interest is paired with a representative emoji. The flow of information through HeyLo is as follows. On input f:

1) Identify from the set of all users U (i.e., all available social media users) the user u represented in f using facial recognition[1].
2) Retrieve 1000 of the most recent social media posts created by u and represent the content of these posts as a single multiset of words S.
3) Filter S for stopwords, URLs, and non-alphabetic characters.
4) For each word $s \in S$, replace s with the NLTK [9] WordNetLemmatizer lemmatized form of s.
5) Apply interest extraction model M on S to obtain a set of pairs $I_{M,u} = \{(k_1, w_1), \ldots, (k_i, w_i), \ldots, (k_n, w_n)\}$ where k_i is a keyword or interest and $w_i \in \mathbb{R}_{\geq 0}$ represents a weight or level of interest for k_i.
6) Reduce $I_{M,u}$ to the pairs (k_i, w_i) with the l highest weights w_i.
7) For $(k_i, w_i) \in I_{M,u}$ select an emoji $e_i \in E$ (where E is the set of all emoji) as follows:
   a) Use word2vec to compute a vector representation v_i for k_i.
   b) Identify an emoji e_i from v_i by computing
   $$e_i = \underset{e \in E}{\operatorname{argmin}}\, d(v_i, v_e)$$
   where v_e is the vector obtained by applying word2vec to the text label for an emoji e (where the label is multiple words, v_e is the average of the vectors for each word), and d is the function computing the cosine distance of two vectors (i.e., a measure of dissimilarity).
   c) Add the triple (k_i, w_i, e_i) to the return set R.

[1]This step is envisioned as part of future work and is not currently employed as part of HeyLo.

8) Output (u, R)

A system overview is depicted in Fig. 1.

B. Interest Extraction Models

Here we describe five possible implementations for model M as described above in step 5. For each implementation we define how the method generates a set of pairs $I_{M,u} = \{(k_1, w_1), \ldots, (k_n, w_n)\}$ from a multiset of word lemmas S.

1) Empath: The Empath algorithm [6] uses a mapping function $f : \Sigma* \rightarrow C$ from the set $\Sigma*$ of word strings to a set C of categories. Given a multiset of words S, Empath computes a score $\omega(c_j)$ for each category $c_j \in C$ equal to the number of words $s_i \in S$ such that $f(s_i) = c_j$. Thus for the Empath model,

$$I_{M,u} = \{(k_i, w_i) | k_i \in C, w_i = \omega(k_i)\}.$$

In this implementation of M we use the pre-defined default set C (200 categories) and the default, human-validated mapping function f.

2) Retrained Empath: In a second implementation of M, we retrained Empath on an expanded set of categories C' that unions C with 1000 categories taken from Facebook's list of page categories[2]. We automatically generated a mapping $f' : S \rightarrow C'$ (an augmentation of the function f described in the previous section) by adding a mapping to each $c_j \in C'$ from words connected to c_j via any (English) relationship in ConceptNet [7]. Using C' and f', we retrained Empath to generate a set of weighted interests as

$$I_{M,u} = \{(k_i, w_i) | k_i \in C', w_i = \omega'(k_i)\},$$

where $\omega'(c_j)$ for a category $c_j \in C'$ equals the number of words $s_i \in S$ such that $f'(s_i) = c_j$.

3) Raw word counts: Given the multiset S, the raw word count model derives a keyword k_i for each unique $s_i \in S$ and derives a weight w_i equal to the number of times k_i appears in S:

$$I_{M,u} = \{(k_i, w_i) | k_i \in S, w_i = |[s | s \in S, s = k_i]|\}.$$

4) Bayesian: Given the multiset S for user u, the Bayesian model derives a keyword k_i for each unique $s_i \in S$ and derives a weight w_i equal to

$$w_i = Pr(u|k_i) = \frac{Pr(k_i|u) \times Pr(u)}{Pr(k_i)}.$$

The distributions $Pr(s|u)$, $Pr(u)$, and $Pr(s)$ were computed from the word counts for words s in the multisets S' derived from the last 1000 tweets from each of a set of 500 users $\tilde{U}$ (ensuring that $u \in \tilde{U}$). These users were selected from a list of the top 500 most followed Twitter handles[3].

[2]https://www.facebook.com/pages/category/

[3]https://socialblade.com/twitter/top/500/followers

5) Chi-square: The chi-square approach derives a keyword k_i for each unique $s_i \in S$ and computes a weight w_i equal to the chi-square contribution of the occurrence of s_i for user u:

$$w_i = \frac{(\hat{x_i} - x_i)^2}{x_i}$$

where $\hat{x_i}$ is the expected occurrence of s_i for user u (as computed from $\tilde{U}$), and x_i is the actual number of times u said s_i. As this method proved particularly effective, we show an example of the chi-square analysis in Table I.

C. Evaluation Metrics

We evaluated the output from each model M according to four criteria which serve to define measures of quality and novelty for the set of extracted interests: *specificity*, *intra-user variance*, *inter-user variance*, and *relevance*. In defining these terms we use $I_{M,u}$ to represent the set of interests extracted by M for a particular user u. For the purposes of evaluation, the weights of items in $I_{M,u}$ (used for the pruning in step 6 from the system overview section) are ignored.

1) Specificity: In order to uniquely characterize the user's interests, a good model will return interests that are specific rather than general. For two words v and w, $v\ IsA\ w$ indicates that v is more specific than w (e.g., "field lacrosse" IsA "sport"). We define $in\text{–}degree(k)$ for a keyword k as the number of words v such that $v\ IsA\ k$ is a valid relationship catalogued in ConceptNet. From this we define the specificity of an interest k as

$$specificity(k) = 1/(in\text{–}degree(k) + 1)$$

Note that if k has an in-degree of 0 (e.g., as with k = "field lacrosse"), k cannot be further categorized or specified. In this scenario, k would receive the maximum specificity score of 1.0.

Using the definition of specificity for a keyword, we define the specificity of a model M. Let I_M represent the set of unique interests extracted by M across all users. Then the *specificity* of M is the average of the specificity values for each unique interest:

$$specificity(M) = \frac{\sum_{k \in I_M} specificity(k)}{|I_M|}.$$

2) Intra-user variance: In returning a set of interests representative of a user, a good model will extract a set of *diverse* interests. For a set of interests $I_{M,u}$ extracted by model M for user u, we define the *variance* of $I_{M,u}$ as

$$variance_{intra}(I_{M,u}) = \sum_{(k_i,w_i),(k_j,w_j) \in I_{M,u}} D_C(v_i, v_j)$$

where $D_C(v_i, v_j)$ is the cosine distance between two vectors v_i and v_j representing interests k_i and k_j. Using the definition of intra-user variance for a set of interests, we define the intra-user variance of a model M as the average of the intra-user variance values across all users:

$$variance_{intra}(M) = \frac{\sum_{u \in U} variance_{intra}(I_{M,u})}{|U|}.$$

	@DaveRamsey	@realDonaldTrump	@tonyhawk	@Nick_Offerman	@BillGates	…	Total
thank	12 (96.62)	123 (1.63)	90 (2.51)	110 (9.07)	17 (83.15)	…	33871
get	379 (471.78)	158 (0.32)	154 (12.16)	106 (2.68)	115 (0.87)	…	37070
impeachment	0 (N/A)	**140 (4166.12)**	0 (N/A)	1 (1.02)	0 (N/A)	…	1083
money	**236 (5560.14)**	20 (8.11)	11 (0.93)	4 (0.90)	18 (9.38)	…	2623
skate	0 (N/A)	0 (N/A)	**160 (30569.06)**	0 (N/A)	0 (N/A)	…	264
malaria	0 (N/A)	0 (N/A)	0 (N/A)	0 (N/A)	**69 (12311.65)**	…	113
sleep	3 (0.46)	0 (N/A)	1 (2.20)	0 (N/A)	1 (2.49)	…	1256
wood	0 (N/A)	0 (N/A)	1 (0.66)	**5 (61.53)**	1 (0.54)	…	144
time	120 (7.49)	95 (1.59)	82 (0.02)	52 (2.50)	94 (0.20)	…	26528
birdhouse	0 (N/A)	0 (N/A)	53 (16162.62)	0 (N/A)	0 (N/A)	…	55
⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
Total	11340	13106	10098	7846	10886	…	3216565

TABLE I
WORD COUNTS (AND CHI-SQUARE CONTRIBUTIONS) FROM USER TWEETS

3) Inter-user variance: In addition to extracting a set of diverse interests, a good model will also extract diverse sets across users or, in other words, avoid repeatedly extracting the same interests for multiple users. To measure this inter-user variance, we find the sum of unique words for each user interest set divided by the total unique words across all user sets:

$$variance_{inter}(M) = \frac{|\bigcup_{u \in U} I_{M,u}|}{\sum_{u \in U} |I_{M,u}|}.$$

4) Relevance: A model's success depends on extracting interests that are not only specific and varied, but which also reflect the user's actual interests. This final metric may be the most important of all, but is also one of the most challenging aspects to measure. A model's predictions for a user's interests can only be accurately assessed by the user him/herself. We are planning to conduct such a study as future work.

III. RESULTS

To compare the performance of each of these methods, we preselected five Twitter handles for five widely-recognized users. We chose to perform the analysis on these users on the basis that their interests are generally well-known and therefore the results could be more easily compared to known results. A more thorough quantitative and qualitative analysis of these models using real users is envisioned as future work.

The top five interests for each model, together with interest weights, are shown in Table II. Weights are only useful for comparing the level of interest for each keyword for a particular user and model (i.e., the way in which weights are derived does not allow for comparison across usernames or across models). Comparing the keywords and orderings for each model against general knowledge about users' interests forms the basis for making arguments about a model's relevance.

To objectively evaluate the specificity, intra-, and inter-user variance, we extracted interests using each model for over 500+ publicly available Twitter users. Results of these calculations are shown in Table III.

By the specificity and variance metrics the Bayesian model looks to have performed the best. Looking at Dave Ramsey's results, for example, the Bayesian model extracts words that are specific and varied (e.g., *belay*, *godly*, *backache*, *variable*, and *toolbox*). These results, however, have very low relevance. Knowing that Dave Ramsey is a businessman, author, and a renowned financial advisor, the interests extracted by the chi-square model are of significantly higher relevance (e.g., *money*, *advice*, *financial*, *debt*, and *millionaire*).

Both the default and retrained Empath models struggle significantly with inter-use variance, reflecting that these models extract many of the same words across several users. The low specificity score betrays that these model also extracts words which are relatively non-descript (e.g., *play* and *party*). The raw word count model also suffers from low specificity (e.g., *thank*, *thanks*, and *people*). The Bayesian model has more specific words for each user, however the nature of this model leads it to prefer words that given the data are uniquely used

	Interest Extraction Model									
Username	Empath		Retrained Empath		Raw Word Count		Bayesian		Chi-square	
@DaveRamsey	*money*	683	*business*	972	*get*	287	*belay*	1	*money*	5569.32
	economics	638	*party*	728	*money*	236	*godly*	1	*advice*	4630.43
	business	517	*wedding*	679	*today*	166	*backache*	1	*financial*	4504.55
	banking	515	*school*	556	*thanks*	129	*variable*	0.66	*debt*	3642.45
	valuable	459	*work*	551	*life*	120	*toolbox*	0.6	*millionaire*	2991.81
@realDonalTrump	*negative_emotion*	459	*healthcare*	655	*president*	321	*legitimize*	1	*transcript*	5749.73
	communication	386	*leader*	566	*trump*	186	*unfairness*	1	*witch*	5618.87
	speaking	365	*stand*	538	*impeachment*	140	*cudgel*	1	*impeachment*	4170.22
	achievement	321	*work*	497	*people*	130	*admissibility*	1	*democrat*	3759.48
	positive_emotion	298	*business*	491	*thank*	123	*frack*	1	*shifty*	3486.31
@tonyhawk	*party*	294	*party*	609	*thanks*	223	*staab*	1	*skate*	30569.06
	play	284	*stand*	419	*video*	99	*primus*	1	*birdhouse*	16162.62
	celebration	270	*play*	418	*see*	88	*clubman*	1	*hawk*	13786.73
	messaging	251	*night*	406	*today*	86	*fameless*	1	*demolition*	9016.61
	sports	229	*wedding*	401	*thank*	86	*lasher*	1	*vert*	5619.85
@Nick_Offerman	*positive_emotion*	242	*party*	458	*nick*	158	*woodworker*	1	*nick*	21095.84
	communication	198	*producer*	364	*thank*	108	*pawnee*	1	*mirth*	9908.16
	party	197	*stand*	345	*get*	77	*unsettling*	1	*berry*	4759.14
	friends	196	*healthcare*	336	*please*	76	*gumption*	1	*woodworker*	2039.82
	giving	185	*joint*	334	*see*	67	*approbation*	1	*pawnee*	1223.89
@BillGates	*optimism*	572	*school*	1082	*world*	245	*diagnostic*	1	*malaria*	12311.65
	positive_emotion	473	*wedding*	871	*people*	196	*shapeshifter*	1	*polio*	7288.12
	reading	443	*show*	698	*help*	132	*thacker*	1	*progress*	3269.68
	school	442	*work*	682	*progress*	112	*teletype*	1	*poverty*	2588.42
	science	401	*business*	668	*energy*	108	*pathogen*	1	*vaccine*	1970.48

TABLE II

TOP FIVE INTERESTS (AND INTEREST SCORES) PER USER BY INTEREST EXTRACTION MODEL

Metric	Empath	Retrained Empath	Raw Word Count	Bayesian	Chi-square
Average specificity	0.013 ± 0.020	0.020 ± 0.076	0.093 ± 0.238	$\mathbf{0.677} \pm 0.376$	$\mathbf{0.245} \pm 0.350$
Average intra-user variance	$\mathbf{0.174} \pm 0.069$	0.141 ± 0.052	$\mathbf{0.171} \pm 0.072$	0.113 ± 0.068	0.140 ± 0.081
Average inter-user variance	0.054	0.090	0.23	**1.00**	**0.75**

TABLE III

COMPARATIVE ANALYSIS OF INTEREST EXTRACTION MODELS

by a particular user, regardless of their relevance. It stands to reason that as the amount of data in the model increases that the Bayesian model will improve, but in its current state, this model suffers from over-specificity. Like the results of the Bayesian model, those of the chi-square model exhibit keywords that are varied and specific but without being too specific.

The NULL emoji for Nick Offerman's word *pawnee* (a fictional city in a TV series featuring Offerman) was a result of the word *pawnee* not being in the Google News word2vec model. The model is unable to find the closest associated emoji because a vector for the word could not be determined. Future work will seek to address this issue.

We concluded that as a combination of relevance, specificity, intra-, and inter-user variance, the chi-square model was the only model that performed reasonably well across all metrics. The results of the full HeyLo interest visualization system using the chi-square model are depicted in Table IV.

IV. DISCUSSION AND CONCLUSION

The relevance of the keywords identified by the chi-square model is noticeably higher than that of other models, in some cases to the point of being surprising. When the chi-square model was applied to Bernie Sanders, one of the interests it returned was *insulin*. Upon further investigation, we learned that Bernie Sanders is diabetic. This shows that the chi-square model is able to find relevant interests to a given user that are not obvious or trivial.

Because the models are defined to operate on text obtained from social media, their ability to extract relevant interests is limited to what users are open and willing to discuss on social media. This is a known limitation of our approach. The approach is not inherently limited to text from this medium, however, and as other suitable means for obtaining textual representations of users interests (e.g., journals, private blogs, etc.) become available, our method is easily adapted to take these sources as input.

Inter-user variance, while a helpful metric, can become problematic if given too much emphasis. Where the goal is ultimately to identify interests that users have in common. This problem should ideally be mitigated by finding ways to identify similar but non-identical interests or to propose ways to bridge dissimilar interests. In this case, inter-user variance maintains its value as a metric for extracting variety and specificity in users' interests.

In this paper, we explored several different models for extracting a user's interests from their tweets and visualizing

User	*Interests* & **Emoji Visualizations**				
@DaveRamsey	*money* **money-bag.png**	*advice* **warning.png**	*financial* **bank.png**	*debt* **credit-card.png**	*millionaire* **man-farmer.png**
@realDonaldTrump	*transcript* **memo.png**	*witch* **woman-fairy.png**	*impeachment* **file-cabinet.png**	*democrat* **man-office-worker.png**	*shifty* **wavy-dash.png**
@tonyhawk	*skate* **ice-skate.png**	*birdhouse* **snowman.png**	*hawk* **owl.png**	*demolition* **building-construction.png**	*vert* **skateboard.png**
@Nick_Offerman	*nick* **man-bouncing-ball.png**	*mirth* **cat-with-tears-of-joy.png**	*berry* **cherries.png**	*woodworker* **man-artist.png**	*pawnee* **NULL**
@BillGates	*malaria* **mosquito.png**	*polio* **cancer.png**	*progress* **hourglass-not-done.png**	*poverty* **children-crossing.png**	*vaccine* **pill.png**

TABLE IV
TOP FIVE INTERESTS AND EMOJI VISUALIZATIONS FROM CHI-SQUARE INTEREST EXTRACTION MODEL

those interests as emoji. We discovered that the chi-square model had the best results across the metrics we looked at despite the Bayes model seeming to be a better fit. We also found the emoji2vector model described in [8] to be an excellent way to visualize the interests as emoji.

Our system is limited to only being able to analyze text to discover a user's interests. Our system also becomes less effective if the user does not write about their core interests or has limited text in their social media. As seen in Table IV, our system can't always provide a visualization for a given interest a user has.

Despite these limitations, our newfound tools will aid us in developing HeyLo as a system for used in social interactions. Our next steps are to find a way to compare interests between users and bridge their seemingly dissimilar interests in a novel and surprising way. Our goal is to create an XRCC framework capable of enhancing social interaction between its users by suggesting similarities that would otherwise go unnoticed.

REFERENCES

[1] P. KR, R. Mukhopadhyay, J. Philip, A. Jha, V. Namboodiri, and C. Jawahar, "Towards automatic face-to-face translation," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1428–1436.

[2] Q. Mao, B. Feng, and S. Pan, "Modeling user interests using topic model," *Journal of Theoretical and Applied Information Technology*, vol. 48, no. 1, pp. 600–606, 2013.

[3] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[4] L. Leydesdorff and K. Welbers, "The semantic mapping of words and co-words in contexts," *Journal of Informetrics*, vol. 5, no. 3, pp. 469–475, 2011.

[5] K. Sasaki, T. Yoshikawa, and T. Furuhashi, "Twitter-TTM: An efficient online topic modeling for Twitter considering dynamics of user interests and topic trends," *2014 Joint 7th International Conference on Soft Computing and Intelligent Systems, SCIS 2014 and 15th International Symposium on Advanced Intelligent Systems, ISIS 2014*, pp. 440–445, 2014.

[6] E. Fast, B. Chen, and M. S. Bernstein, "Empath: Understanding topic signals in large-scale text," *Conference on Human Factors in Computing Systems - Proceedings*, pp. 4647–4657, 2016.

[7] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," 2017, pp. 4444–4451. [Online]. Available: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972

[8] B. Eisner, T. Rocktäschel, I. Augenstein, M. Bosnjak, and S. Riedel, "emoji2vec: Learning Emoji Representations from their Description," pp. 48–54, 2016.

[9] E. Loper and S. Bird, "Nltk: The natural language toolkit," in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, 2002, pp. 63–70.