Enhancing crowd wisdom using explainable diversity inferred from social media

Shreyansh Bhatt*, Manas Gaur*, Beth Bullemer*†, Valerie L. Shalin*†, Amit P. Sheth*, Brandon Minnery‡
*Kno.e.sis Center, Wright State University, USA.
†Department of Psychology, Wright State University, USA.
‡Wright State Research Institute, USA.

***Abstract*—A crowd sampled from a set of individuals can provide a more accurate prediction in aggregate than most individuals. This effect, referred to as *wisdom of crowd*, exists when crowd members bring diverse perspectives to decision making. Such diversity leads to uncorrelated prediction errors that cancel out in aggregate. As crowd members' judgments are often the result of solution strategies, diversity in solution strategies can enhance crowd wisdom. One of the most challenging tasks in sampling such a crowd is to determine the individual's solution strategy for a prediction problem. As participating individuals often share their perspectives through social media, we can use such data to identify an individual's solution strategy. In this paper, we propose a crowd selection approach using social media posts (tweets) indicating diverse solution strategies. We use tweet classification to identify participants' prediction strategies and categorize participants based on the binomial test to identify sets of participants that apply a similar strategy. We then form a *diverse* crowd by sampling participants from different sets. Using the domain of Fantasy Sports, we show that such a diverse crowd can outperform crowd selected at random and 90% of individual participants, and participant categorization schemes using word2vec. Further, we use a knowledge graph to investigate the factors forming such a diverse crowd and how these factors can lead to a better decision. Relative to bottom-up (data-driven) processes the approach presented here provides an explanation of diverse crowd behavior.**

***Index Terms*—Diversity, Wisdom of Crowds, Twitter, DBpedia, Collective Intelligence, Semantic analysis, Fantasy Sports**

I. Introduction

The aggregated judgment of a crowd sampled from a large set of individuals is more accurate than most individuals [1] [2]. One of the most critical conditions for such an effect is diversity among the individuals in the crowd [3], [4]. A diverse group of individuals brings different perspectives in decision making. However, the type of diversity that forms an intelligent crowd is still an open research area [1]. Diversity in *judgments* may help form an intelligent crowd [5] [6]. However, such judgment data may not always be available for several specific prediction problems. For example an upcoming election lacks *a priori* participant judgment (vote) data.

Social media data can serve as an effective proxy for judgment diversity as participants may to share their perspectives about a prediction problem. Recent research studies confirm that diversity in participants' posts support intelligent crowd formation [7] [8]. In particular, social media diversity quantified using word2vec allows *intelligent* crowd formation [7]. However, such diversity quantification is *bottom-up*, as its inferred from data. Although such diversity allows intelligent crowd formation, it does not provide an explanation for the effect of diversity on the *Wisdom of Crowd(WoC)*. Top-down diversity quantification and crowd selection may achieve better crowd wisdom than *bottom-up* approaches. A top-down approach also has the benefit of explaining the diversity that affects crowd wisdom.

In this study, we explore whether top-down diversity quantification can help assemble an *intelligent* crowd. We define diversity in terms of the solution strategies employed by a participant for generating a prediction. Specifically, we hypothesize that diverse solution strategies lead to a more robust aggregated crowd prediction, where solution strategies are inferred using participants' social media posts. We provide real-world evidence that such diversity can help achieve an accurate prediction. We first characterize each participant according to whether his/her tweets refer to a particular strategy by classifying individual tweets. Using a binomial test-based for participant categorization, we then identify a set of participants employing *similar* solution strategies. Finally, we form a diverse *virtual* crowd by selecting participants from each category.

We evaluated our approach in the domain of the Fantasy Premier League (FPL) for 3385 participants predicting top-performing captains. An FPL participant selects a team of 11 soccer players and assigns one soccer player as the *captain* of his/her team. A participant receives *FPL points* based on the performance of the players of his/her team in real soccer games. The captain receives twice the number of points compared to a non-captain player. Hence, a participant is motivated to select a captain providing the best score. Participants employ a solution strategy to choose a captain. In FPL, two strategies have been widely used to select a captain: Popular Choice and Differential Choice. According to the popular choice strategy, participants imitate other participants' captain selection, specifically pursuing a crowd favorite. According to the differential choice strategy, participants attempt to predict the crowd favorite, but leverage this information to actively avoid such player(s)[1]. We investigate these two strategies in the formation of a diverse crowd.

We found that a diverse crowd determined by strategy is

[1]http://biebek.blogspot.com/2015/08/premier-league-boasts-more-than-3.html

likely to perform better in the FPL captain prediction task than 90% of the individual participants. We also compared a diverse crowd with a randomly selected crowd of comparable size and found that a diverse crowd is 63% likely to outperform a randomly selected crowd. Crowds based on diverse strategies also perform favorably relative to standard word2vec methods for clustering users.

To explain the diversity in captain selection strategies and its effect on captain selection, we used a domain specific knowledge graph extracted from DBpedia [9]. The extracted knowledge graph is a concept hierarchy where a parent concept subsumes child concepts. To explain diversity, we mapped the keyword features used in classification to the knowledge graph and investigated the parent concepts that maximally subsume these keywords. We found that features identifying Popular choice and Differential choice tweets mapped to two parent soccer players who happened to be the top performers in terms of scoring FPL points. This supports the claim that diversity of strategy revealed at the feature levels converge to select a captain who is effective from both perspectives.

Our contributions include:

- A social media-based diverse and intelligent crowd selection approach.
- A top-down approach to diversity quantification based on solution strategy, exploring its role in *intelligent* crowd formation.
- A domain-specific knowledge graph based approach to explain diversity and its effect on the accuracy of judgment.
- Evaluation in the Fantasy Premier League domain to support the role of top-down, strategic diversity in *Intelligent* crowd formation.

The rest of the paper is organized as follows. Section II provides background on FPL and captain selection strategies. Section III details the approach for generating a diverse crowd using participants' social media posts. Section IV provides details on the dataset, evaluation measures, and results, Section V describes related work, and Section VI concludes the paper.

II. Background

We used the commonly employed Fantasy Premier League domain to study diversity and *intelligent* crowd formation, primarily because the domain provides a scored outcome measure that makes it possible to evaluate the success of our algorithms [10] [7]. In FPL, participants construct their *fantasy* team consisting of real-world soccer players. The outcomes are determined by actual player performance in England's Professional Football league, also known as the English Premier League. One decision that occurs weekly over the course of the season is captain selection. Participants select a captain within their 11 soccer player fantasy team. As the captain receives twice as many points as a non-captain player, the decision of whom to select as captain can significantly influence a participant's reward.

Given the importance of captain selection in FPL, several blog posts and FPL portals suggest captain selection strategies. We focus on two strategies, both of which concern how FPL participants integrated information concerning other FPL participants' captain selections when selecting their own captain: 1. Popular choice and 2. Differential choice.

A. Popular choice (PC)

One approach is to imitate others' captain selections directly, specifically pursuing the crowd favorite. In other words, participants select the most popular player. Such a strategy works in general as the popular captains often yield higher points for the teams[2]. Table I shows a few popular choice tweets from our data set.

TABLE I
Sample tweets labeled as Popular Choice (PC) or Differential Choice (DC)

Tweets	Strategy
a look at the best captain contenders for gw6 #fpl	PC
some quality information on the main contenders for #fpl #gw21 #captains thanks!	PC
here are the leading #fpl #captains contenders for #gw35 thanks	PC
i have the urge to change my captain from a differential to a consensus pick.	DC
i think i am. unless i try to pull off a differential captain like ighalo.	DC
aguero right now; might be ighalo by tomorrow morning. #fpl #differentialcaptain	DC

B. Differential Choice (DC)

A differential choice strategy assumes that selecting a captain that is perceived to be less popular (not necessarily a rare selection), but still likely to score points, will set the participant apart from the crowd of participants opting for a popular choice candidate. Thus, such a strategy is considered to be more of a risk or a gamble. Table I also shows tweets that refer to differential choice.

Tweets that refer to Popular choice and Differential choice may have both explicit and implicit strategy indicators. Additionally, a tweet may have indicators of both popular choice as well as differential choice. As an example, *"kane or giroud for captain i can't decide!! kane probs the safe option but most peeps doing that. Giroud worth the gamble? #fpl #gw30"*. Here, the participant is not sure about the particular strategy to use in captain selection yet he/she indicates one of these two strategies shall be used in his/her captain selection.

III. Approach

Figure 1 describes our approach for generating diverse crowds based on FPL captain selection strategies in social media posts. Each participant is categorized based on the number of tweets indicating the two captain selection strategies. Hence, classifying tweets indicating a captain selection strategy and participant categorization are the two key components of our diverse crowd generation. Crowd selection from these categories completes the crowd formation process. Subsequent knowledge graph analysis provides an explanation for calculated diversity.

[2]https://fantasy-sports-info.com/fpl-guide-beginners/

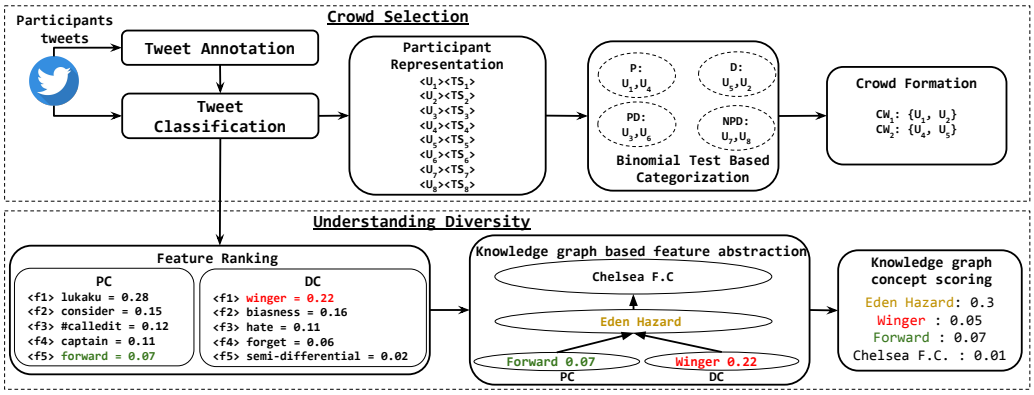

Fig. 1. Proposed crowd selection and diversity explanation approach. Each participant is represented by his/her tweets and processed for identifying tweets referring to two solutions strategies. A knowledge graph-based feature abstraction identifies relevant concepts subsuming PC and DC keywords. $< f >$:feature, $< U_i >$: Participant, $< TS_i >$:Tweet Set per user, CW: Crowd

A. Tweet Classification

Consistent with common practice [11], we used machine learning based text classification for identifying player selection strategy from tweets. We manually annotated 165 of tweets as Popular choice and 258 tweets as Differential choice tweets. We considered an equal number of non-popular and non-differential choice tweets in training the classifiers. The two classification categories (popular choice and differential choice) are not orthogonal; the same tweet may provide evidence for both popular as well as differential choice. Hence, we trained two tweet classification models. Model 1 identifies whether a tweet belongs to popular choice and Model 2 determines whether a tweet belongs to differential choice.

We used a *Bag of Words* approach combined with term frequency and inverse document frequency (TF-IDF) for generating a feature vector for each tweet [12]. We considered uni-grams and bi-grams as features and found whether each feature is present in a tweet. These vectors are then processed for TF-IDF computation, and each feature is represented with its TF-IDF value instead of "1" or "0". To avoid over-fitting in training based on these sparse tweet vectors, we used k-best feature selection. This feature selection technique uses Singular Value Decomposition, widely used in feature selection for text classification [13]. We trained two models using a Random Forest classifier with ten-fold cross-validation with a 70%, 30% train-validation split. We also reserved 10% of the labeled data as test data to determine classifier performance accuracy. We report the final accuracies for each model, i.e., popular choice, and differential choice classification, accuracies in Section IV.

We processed the tweets for each participant using Model 1 to identify the number of tweets in popular choice and Model 2 to identify number of tweets in Differential choice. Formally, for each participant U_i, we have $\mathbf{U_i} = \{n_P, \overline{n}_P, n_D, \overline{n}_D\}$ (Same as TS_i in Figure 1). Here each n is a number where n_P is popular choice tweets, $\overline{n}_P$ non popular choice tweets, n_D Differential choice tweets, and $\overline{n}_P$ is non-differential choice tweets.

B. Binomial Test Based Categorization

Each n in U_i may result from different distributions. Hence, we normalize each n by computing the Z-score for each n with respect to all participants. These Z-scores represent each participant relative to others in terms of n_P, $\overline{n}_P$, n_D, and $\overline{n}_D$. As each participant provides tweets suggesting both strategies or neither, we require a decision rule to categorize each participant according to the different amounts of data they provide. We do not consider these strategies as mutually exclusive and investigate whether a popular (or differential) choice best characterizes the participant or non-popular (or non-differential). We assign a participant into either one of the following four *participant types*: 1. Popular Choice, 2. Differential Choice, 2. Popular and Differential, 3. Neither popular Nor differential. As the names suggest, Type 1 refers to participants most likely to exhibit popular choice tweets, Type 2 are participants most likely to exhibit Differential choice, Type 3 participants are ambiguous, providing substantial evidence for both, and Type 4 participants do not provide any evidence for either one of these strategies. We used a binomial test with a null hypothesis that the two strategies are equally likely to occur. Formally, a binomial test (BiTest) $\mathbf{B_q}$ for an event with q as the probability for an event to succeed and $\overline{q}$ as the probability of failure can be described as,

$$B_{q,\overline{q}} = \binom{N}{q} \cdot p_0^q (1-p_0)^{\overline{q}} \quad (1)$$

Here, N is defined as the sum of q and $\overline{q}$. p_0 is the probability of occurrence of a success in each one of the N trials. In our study, we set $p_0 = 0.5$ and the binomial test was performed at 5% alpha level[3]. This determines whether q is likely to occur more than $\overline{q}$, 95% of the time.

The participant exemplifies Popular choice (Type 1) or Differential choice (Type 2) when the Binomial tests find significant evidence for only the corresponding type. Specifically, $\mathbf{B(n_P, \overline{n}_P)}$ should indicate that likelihood of n_P over $\overline{n}_P$ is more than 95%, and the possibility of other events is less than 95% to consider a participant as Type 1. Algorithm 1 formalizes this participant type assignment process. It starts

[3]http://www.statisticshowto.com/what-is-an-alpha-level/

Input : $U_i = \{n_P, \overline{n}_P, n_D, \overline{n}_D\}$
Output : $T_1 \vee T_2 \vee T_3 \vee T_4$
$B_{n_P,\overline{n}_P} = BiTest(n_P, \overline{n}_P)$
$B_{\overline{n}_P,n_P} = BiTest(\overline{n}_P, n_P)$
$B_{n_D,\overline{n}_D} = BiTest(n_D, \overline{n}_D)$
$B_{\overline{n}_D,n_D} = BiTest(\overline{n}_D, n_D)$
$B_{\overline{n}_P,\overline{n}_D} = BiTest(\overline{n}_P, \overline{n}_D)$
$B_{\overline{n}_D,\overline{n}_P} = BiTest(\overline{n}_D, \overline{n}_P)$
if $B_{n_P,\overline{n}_P} > 0.05$ and $B_{n_D,\overline{n}_D} \leq 0.05$ and $B_{\overline{n}_P,\overline{n}_D} \leq 0.05$ **then**
  return T_2
else if $B_{n_P,\overline{n}_P} \leq 0.05$ and $B_{n_D,\overline{n}_D} > 0.05$ and $B_{\overline{n}_P,\overline{n}_D} \leq 0.05$ **then**
  return T_1
else if $B_{n_P,\overline{n}_P} \leq 0.05$ and $B_{n_D,\overline{n}_D} \leq 0.05$ **then**
  return T_3
else
  return T_4
end if

Algorithm 1: Binomial test based categorization of participants at 5% significance. Given the Z-scores of a participant U_i, the algorithm categorizes participant in one of following types: Type 1(T_1), Type 2(T_2), Type 3(T_3), Type 4(T_4)

with six binomial tests (lines 3-8). Based on the condition described above, it decides the participant type (line 9-17).

C. Diverse Crowd Selection

Type 1 (Popular) and Type 2 (Differential) participants provide strong evidence of using the corresponding strategy in player selection. About half of the participants are Type 3 (Both) or Type 4 (Neither), who provide ambiguous or unclear strategy indicators that will likely muddy diversity. Hence, we avoided participants belonging to these two types in our diverse crowd formation and created our diverse crowd using clear Type 1 and Type 2 participants. We selected n participants from Type 1 and Type 2 to build our diverse crowd.

D. Understanding Diversity

The diverse crowds that we generated in the previous step depend on tweet classification. To understand the kind of diversity such a tweet classification captures in Fantasy Premier League domain, we extracted the top most informative features (keywords) from our Random Forest Classifier [14]. We mapped these keywords to an English Premier League domain-specific knowledge graph extracted from DBpedia using a domain-specific knowledge graph extraction tool [15]. Such a knowledge graph provides a good representation of a corresponding domain [15] [16]. The resulting hierarchy for the English Premier League in Figure 1 indicates that the top concept *Chelsea F.C.* subsumes *Eden Hazard* who has two attributes (subsumes) *Forward* and *Winger*. Hence, a *parent* concept subsuming multiple child concepts explains child concepts. We use the parents in this hierarchical structure to encompass the keywords identifying multiple strategies.

We seek the concept in a knowledge graph *subsuming* most of these keywords. As we had two classification models corresponding to two of the captain selection strategies, we obtain two lists of keywords from each model. Each keyword has an importance value between 0 and 1 from the Random Forest Classifier. We use these values to assign each concept in the domain-specific knowledge graph a weight, as shown in Figure 1. Specifically, we compute 3-hop parents of each concept and assign a score for each of the parent concept as follows,

$$S = \frac{C_w}{P_l} \tag{2}$$

Here, C_w refers to the concept weight indicated by the keyword weight and P_l indicates a parent level of the current concept. For each concept C associated with the keyword, we get the parents of C and compute its corresponding S. If we find that the parent concept being processed as part of C is already identified as a parent for another concept, then we add this score to the existing score. Hence, a score associated with each parent concept indicates the number of keywords the particular parent concept subsumes. We repeat the same procedure for two hops, and three hop parents. As 1-hop parents are more relevant than 2-hop and 3-hop parents, the concept score is multiplied by the inverse of parent level P_l. As shown in Figure 1, we found *Forward* and *Winger* as important keywords to distinguish popular choice tweets and differential choice tweets respectively. In the English Premier League knowledge graph, we start with these two concepts and ascend the hierarchy (extracting *Eden Hazard*, and *Chelsea F.C.*) and compute S for each concept. As concepts with high scores can best explain multiple keywords and hence the player selection strategies, we score each concept in the knowledge graph and consider the top-N concepts for understanding diversity. This allows us to identify concepts that unify the categories with implicit contents that are not explicit in the tweets themselves.

IV. Dataset and Results

In this section, we describe the dataset, evaluation measure, and results.

A. Dataset

We collected FPL related tweets using the Twitter streaming API with two FPL related keywords, *FPL*, and *@OfficialFPL*. We determined captain pick data from the FPL portal[4] by matching Twitter usernames associated with these tweets to their FPL usernames. We used FPL captain pick data for the 2016-17 season.

We manually verified 3385 participant matches based on Twitter username and FPL username and collected their additional tweets by crawling their Twitter timelines. For each participant, we collected tweets ranging from 2014 to August 2016 (before the start of FPL 2016 season). We filtered these tweets using FPL related keywords to consider only relevant tweets in the participant representation. We obtained ∼1M participant tweets for the 3385 participants with 1282 median

[4]fantasy.premierleague.com

tweets and 1385 average tweets per participant. Hence, for each participant, we have a set of his/her FPL related tweets, and captain picks for 25 game weeks. [5]

B. Evaluation Measure

Consistent with Goldstein et al. [10], we computed a crowd's wisdom score (**WS**) as the FPL score of a captain receiving the greatest number of votes from a crowd. For a crowd of participants, $G = \{U_1, U_2, \ldots, U_n\}$, we extracted their captain picks for a week w_{index} as $C_{index} = \{c_1, c_2, \ldots, c_n\}$ where c_i is a captain picked by participant U_i in week w_{index}. The wisdom score for a crowd is computed as,

$$WS = \frac{\sum_1^{25} Mod(C_{index})}{25} \tag{3}$$

Here, $Mod(C_{index})$ represents the corresponding real-world points of the individual captain receiving the most votes from the crowd in the $index$ game week. In case of a non-unique mode - i.e., for a tie, we selected one of these modes randomly. A crowd's wisdom score was the average of its weekly scores over the 25 game weeks considered in our analysis.

C. Results and Analysis

We first show the results for tweet classification. We used ten-fold cross-validation for training and an unseen labeled test data for validating the resulting classifier.

TABLE II
TWEET CLASSIFICATION CROSS-VALIDATION RESULTS. LABELS '0' AND '1' INDICATE RESULTS FOR CLASSIFYING A TWEET TO $\overline{P}(\overline{D})$ AND $P(D)$, RESPECTIVELY. THE CLASSIFIER ACHIEVED 0.85 AVERAGE F-SCORE.

Strategies	Label	Precision	Recall	F-score
Popular	0	0.87	0.90	0.93
	1	0.97	0.66	0.76
Differential	0	0.97	0.67	0.79
	1	0.87	0.99	0.93

TABLE III
TWEET CLASSIFICATION RESULTS ON THE TEST DATASET. LABELS HAVE THE SAME MEANING AS IN TABLE II. THE CLASSIFIER ACHIEVED 0.79 AVERAGED F-SCORE FOR LABELS '0' AND '1'.

Strategies	Label	Precision	Recall	F-score
Popular	0	0.86	0.92	0.89
	1	0.86	0.76	0.81
Differential	0	0.77	0.50	0.61
	1	0.78	0.92	0.84

Table II shows the cross-validation results for tweet classification and Table III shows results for tweet classification on the unseen test data. We report the classifier's performance for both identifying popular (or differential) and non-popular (or non-differential) judgments. The label '0' indicates Popular choice (or Differential choice) tweets and label '1' indicates non-popular choice (or non-differential choice) tweets. Rows with '1' indicate the classifier's performance for identifying Popular choice (or Differential choice) and Rows with '0' the indicate classifier's performance for identifying non-popular choice (or non-differential choice) tweets. For each participant, we computed the four U_i values with the count of tweets identified in '1' and '0' using Model 1 and Model 2 as follows: 1) n_P is the number of tweets identified by Model 1 in class '1'. 2) $\overline{n}_P$ is the number of tweets identified by Model 1 in class '0'. 3) n_D is the number of tweets identified by Model 2 in class '1'. 4) $\overline{n}_D$ is the number of tweets identified by Model 2 in class '0'. In other words, we ignored tweets for which the classifier was not able to decide which class (either '1' or '0') it belongs.

Model 1 achieved a 84.5% F-score for the Popular Choice tweet classification model and Model 2 achieved 86% F-score for the Differential Choice tweet classification in cross-validation. A classification model with high training accuracy may also indicate over-fitting. To guard against over-fitting, we measured these results on test data that the classifier did not encounter while training. On test data, the models achieved adequate 85% and 72.5% F-scores for Popular choice and Differential Choice, respectively which rules out over-fitting.

We used these classifier models to generate participant representations and select diverse crowds (see III). Out of 3385 total participants, we had 895 participants identified as Type 1 (Popular choice) and 789 participants identified as Type 2 (Differential choice). Next, we evaluate diverse crowd formation using the *wisdom score* achieved by crowds selected from these types. As described in Section III-C we generated *diverse* crowds by randomly picking n participants from the two participant types. We generated l such crowds, where $l = 5000$ referred to as D (Diverse crowds). We compared these crowds with R (Random crowds), i.e., crowds generated by randomly selecting the same number of participants from the complete set of 3385 participants. Figure 2 shows box-plots for different crowd sizes. Here, crowd size is a multiple of 2 as we had two types of participant categories to generate diverse crowds. For each crowd, we computed its wisdom score based on the *Wisdom Score* formula resulting in two score lists (the diverse crowd score list and random crowd score list). On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the upper and lower quartile, respectively for these lists. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the '+' symbol.

Diverse crowd lists always achieved a better median wisdom score than Random crowds for crowd sizes ranging from 8 to 200. Such an effect was also indicated by $p < 0.05$ for the T-test between Diverse crowds and Random crowds for each crowd size. We had a modest yet statistically significant effect of the diverse crowd out-performing Random crowds. As larger crowds tend to be more accurate than smaller crowds [1], we get better wisdom scores for large crowds than small crowds.

Figure 2 indicates that diverse crowds achieved a better median wisdom score than Random crowds. However, it is also of interest to know how likely a diverse crowd is to produce a better wisdom score than a Random crowd. We used Monte Carlo simulation for this purpose. Specifically, we randomly selected a single diverse crowd from the Diverse crowd set

[5]We have not provided unrestricted access to the dataset, as it contains actual tweets and usernames. However, the dataset is available from the corresponding author upon request.

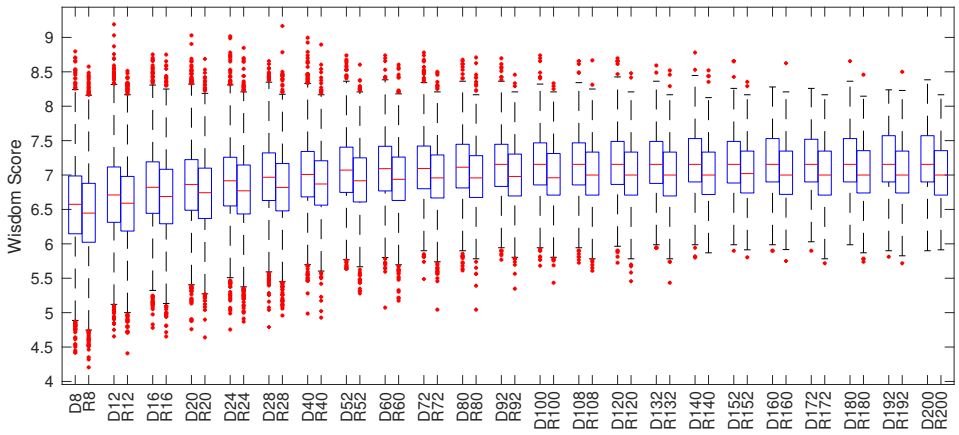

Fig. 2. Box plots comparing wisdom scores of diverse crowds (D) and random crowds (R). Diverse crowds achieved a better wisdom score with a smaller standard deviation compared to random crowds.{D8,R8}: Diverse and Random crowd of size 8.

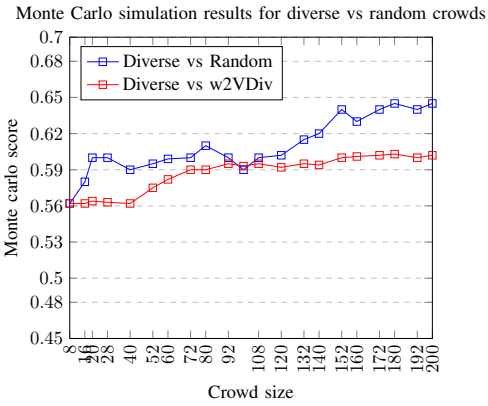

Fig. 3. A diverse crowd consistently out performed Random crowds for group sizes 10 to 100. Our diverse crowds also outperformed crowds sampled using word2vec based diversity, w2VDiv.

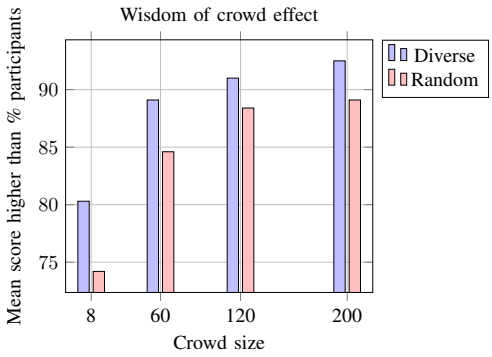

Fig. 4. Wisdom of crowd effect. Diverse crowd of size 60 outperforms 89% of the individual participants.

and a single random crowd from the Random crowd set. We then computed a *win* if the diverse crowd had a higher wisdom score than the random crowd. We repeated this for 1000 times and calculated a Monte Carlo simulation score as a ratio of the number of wins to 1000. A Monte Carlo simulation score of $\sim$ 0.5 indicates that the two sets of crowds are equally likely to beat each other. A Monte Carlo score of $\sim$ 1.0 suggests that a crowd from set one almost always beats a crowd from set two. Figure 3 shows the results for these Monte Carlo simulations.

We observed a $\sim$ 0.63 Monte Carlo simulation score indicating that a Diverse crowd is 63% likely to outperform a Random crowd. We also compared our Diverse crowds generated from solution strategy to a diverse crowd formed using our previous word2vec based diverse crowd selection approach. For this analysis, we created word2vec representations of participants based on their tweets and generated $l \times 10 = 50,000$ random crowds. For each crowd, we computed an average pairwise cosine distance between participants of the crowd using their word2vec representations [7]. We selected the top 10% ($l = 5000$) of the crowds having the highest average pairwise distance, referred to as w2VDiv, and compared them with the Diverse crowds. A Monte Carlo simulation score of $\sim$ 0.59 indicated that the proposed strategy-based method for assembling diverse crowds can assemble a better crowd than word2vec based diverse crowd selection.

Next, we evaluated our Diverse crowds regarding the *wisdom of crowd* effect. In other words, we measured the number of participants that a crowd, on an average, can outperform. Specifically, we computed the season score achieved by each participant (using the same formula as **WS**) and found the number of participants that had a lower season score than an average wisdom score of a comparison crowd. Figure 4 plots the number of participants a crowd outperforms on an average.

On an average, a random crowd of size 8 outperforms 74.2% of the individual participants while a diverse crowd of the same size outperforms 80.3% of the individual participants. On an average, a diverse crowd of size 60 is better than almost 90% of the individual participants and approximates the performance of a random crowd more than three times as large.

We also examined whether diverse crowds produce diverse judgments. The intuition is that crowds producing diverse judgments likely imply a less biased sample of participants, which in turn likely yields a better-aggregated opinion. For this purpose, we used a *judgment diversity* measure proposed by Merayo et al. [17]. Formally the measure is defined as follow,

$$M = \frac{\sum_{i,j} d(u_i, u_j)}{n(n-1)} \tag{4}$$

Here, $d(u_i, u_j)$ is the difference between wisdom scores of participants u_i and u_j and n is the total number of participants

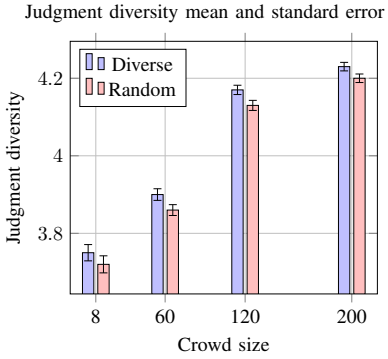

Fig. 5. Judgment diversity comparison Diverse crowds and Random crowds. Diverse crowds had more judgment diversity than Random crowds.

TABLE IV
HIGH RANKED KNOWLEDGE GRAPH CONCEPTS SUBSUMING BOTH POPULAR CHOICE AND DIFFERENTIAL CHOICE KEYWORDS.

DBpedia Concept	**Score**
Eden_Hazard	0.34
Category:Chelsea_F.C._players	0.17
Romelu_Lukaku	0.17
Category:West_Bromwich_Albion_F.C._players	0.17
Category:Manchester_United_F.C._players	0.08
Midfielder(Winger)	0.05
Daniel_Sturridge	0.005

in the crowd. Figure 5 shows the results for this metric for Diverse and Random crowds. We found that diverse crowds were more generally more diverse regarding the judgment diversity metric $\mathbf{M}$. This suggests that a crowd of participants with a diverse player selection strategy ends up producing more diverse judgments than Random crowds. Our method assembles a crowd who can produce diverse and accurate judgment *using only their social media data*.

To explain this diversity, we used an English Premier League domain specific knowledge graph. We mapped the keywords identifying Popular choice and Differential choice strategies to the knowledge graph concepts using DBPedia lookup[6]. We then found concept scores using Equation 2. The concept that subsumes most of these keywords and are not far up in the concept hierarchy are ranked higher according Equation 2. We ranked each concept in the knowledge graph and sorted them in reverse order based on these scores.

Table IV shows the resulting concept scores for a few concepts receiving high scores. We found *Eden Hazard* and *Romelu Lukaku*, two soccer players, subsuming both popular choice and differential choice features (keywords). These two players happened to be in the top 10 players scoring the most FPL points for the 2016-17 season. As our diverse crowd reflects both popular choice and differential choice strategies, they select a better player than Random crowds, albeit for different reasons. Hence, diversity in solution strategies leads to a better captain selection. Moreover, the concepts with high scores also help us interpret these two strategies. For FPL, we can determine whether diversity in solution strategy is related to the specific English Premier League teams or locations. We can also learn about teams whose players are chosen by both popular and differential choice. This information is helpful especially in deciding the kind of factors one should focus on in decision making so that the decision is not biased.

[6]https://wiki.dbpedia.org/lookup

V. RELATED WORK

Our work can be understood within a larger body of work that explores collective intelligence and *intelligent* crowd formation. The traditional wisdom of crowd research has explored the correlation between diversity and collective judgment accuracy [18]. These experiments explore the role of explicitly indicated participant diversity instead of inferring diversity. In one of the recent studies, Teng et al. reported that diverse teams were more creative than less diverse teams [19]. However, they *asked* the users (within a group) to define their similarity to other group members. Such an analysis does not apply to the problem of inferring diversity from social media data and its correlation with the accuracy of collective judgment.

Hong et al. recently reported that user-generated content diversity is positively correlated with crowd performance [8]. However, they did not explore solution strategy as a facet of content diversity. Moreover, they used a traditional word vector-based user representation and computed diversity as the cosine distance between these users [20]. It is also difficult to explain the effect of diversity on collective intelligence. Finally, such a representation does not capture the context of short social media text. Robert Jr. et al. also reported that diverse users could effectively generate a quality Wikipedia article [21] [22]. They compute user diversity based on author specified topics of interest instead of inferring strategy-specific user diversity from their social media content.

We have also previously found that social media content can help form diverse user groups with a word2vec based user representation [7]. However, in this paper, rather than relying on such a *bottom-up* diversity, we define a *top-down* diversity measure based on solution strategies. We found that a proposed top-down solution strategy based diversity measure indeed forms a more *intelligent* crowd than a word2vec based bottom-up measure [7]. Moreover, such a top-down diversity is easy to interpret in the domain of soccer and fantasy sports and provides additional evidence for the superiority of diversity based crowd formation.

Several research studies explore crowd formation based on the crowd members' current judgment and past judgment data [6]. Using the same FPL domain, Goldstein et al. reported that a crowd formed using experts, based on their historical performance data, can accurately predict top-performing FPL captain [10]. Warnaar et al. also explored expertise and crowd wisdom [23]. Davis-Stober et al. found that diversity in crowd judgments can be used to assemble *intelligent* crowds [5]. Galesic et al. also showed that a small crowd could be assembled using their current judgments for coming up with an accurate prediction [24]. Similar to this, Nguyen et al. reported that diversity in crowd judgments is correlated with crowd performance [17]. These works exploit current or historical judgment data that may not be available for several prediction tasks. However, they do report a correlation between judgment

diversity and crowd performance. We also found that our diverse crowd was diverse in their judgments as well. The binomial is a commonly applied model. For example, [25] models the chance game of die as a binomial experiment. Mahjong can also be considered a game of chance. [26] applied the binomial to both define the game and then develop a mechanism to select players. Some prior work has employed a binomial-test for assessing hypotheses with respect to social media content. In [27], chances of Post-traumatic stress disorder is high for military personnel compared to civilians was evaluated using binomial-test on the Twitter text. Furthermore, [28] utilized a binomial test to estimate the probability of a feature and determines its chances of being spam or non-spam based on p-value. However, these studies have not utilized a binomial test for categorizing users based on their content diversity for effective decision making.

VI. Conclusions and Future work

Social media data can be used to infer diversity in top-down defined solution strategies. Using such diversity, we formed crowds that actually produce diverse judgment. More importantly, such a method can be used to assemble an *intelligent* crowd that can collectively make an accurate judgment *before* they render a judgment. Our strategy-based diversity framework can be used to interpret diversity in several domains, explaining the correlation between various domain features and collective intelligence. We also demonstrated that machine learning based tweet classification methods work for classifying tweets for solution strategy. As the proposed approach only requires strategy characterization and training data, it applies in domains other than Fantasy Soccer.

The proposed diverse crowd selection achieved a statistically significant effect. Though of potential practical significance in some domains, the effect size was modest compared with simple random crowd selection strategy. One of the possible reasons is the limited number of strategies and the likely presence of additional strategic differences. Another explanation lies in the crowd selection strategies. Our ongoing work includes optimally diverse crowd selection algorithms to achieve even more significant effect sizes as well as more sophisticated and complete models of strategic expertise and participant categorization. Moreover, while fantasy sports provide an ideal test bed for examining diversity-based approaches due to the availability of outcome measures, follow-on research should extend and validate these findings in other domains having more practical relevance, such as marketing, election prediction, and geopolitical forecasting.

VII. Acknowledgments

This work was supported by Army Research Office Grant No. W911NF-16-1-0300.

References

[1] J. Surowiecki, *The wisdom of crowds*. Anchor, 2005.
[2] A. E. Mannes, R. P. Larrick, and J. B. Soll, "The social psychology of the wisdom of crowds." 2012.
[3] L. Hong and S. E. Page, "Groups of diverse problem solvers can outperform groups of high-ability problem solvers," *NAS*, 2004.
[4] C. P. Davis-Stober, D. V. Budescu, J. Dana, and S. B. Broomell, "When is a crowd wise?" *Decision*, 2014.
[5] C. P. Davis-Stober, D. V. Budescu, S. B. Broomell, and J. Dana, "The composition of optimally wise crowds," *Decision Analysis*, 2015.
[6] H. Olsson and J. Loveday, "A comparison of small crowd selection methods." in *CogSci*, 2015.
[7] S. Bhatt, B. Minnery, S. Nadella, B. Bullemer, V. Shalin, and A. Sheth, "Enhancing crowd wisdom using measures of diversity computed from social media data," in *IEEE/WIC/ACM*, 2017.
[8] H. Hong, Q. Du, G. Wang, W. Fan, and D. Xu, "Crowd wisdom: The impact of opinion diversity and participant independence on crowd performance," 2016.
[9] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann *et al.*, "Dbpedia: A nucleus for a web of open data," in *The semantic web*, 2007.
[10] D. G. Goldstein, R. P. McAfee, and S. Suri, "The wisdom of smaller, smarter crowds," in *ACM CEC*, 2014.
[11] S. A. Salloum, M. Al-Emran, A. A. Monem, and K. Shaalan, "A survey of text mining in social media: facebook and twitter perspectives," *ASTESJ*, 2017.
[12] A. Bordes, Y.-L. Boureau, and J. Weston, "Learning end-to-end goal-oriented dialog," *arXiv preprint arXiv:1605.07683*, 2016.
[13] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *JMLR*, 2003.
[14] D. Petkovic, R. Altman, M. Wong, and A. Vigil, "Improving the explainability of random forest classifier–user centered approach," in *Pacific Symposium on Biocomputing*, 2018.
[15] S. Lalithsena, P. Kapanipathi, and A. Sheth, "Harnessing relationships for domain-specific subgraph extraction: A recommendation use case," in *IEEE Big Data*, 2016.
[16] P. Kapanipathi, P. Jain, C. Venkataramani, and A. Sheth, "User interests identification on twitter using a hierarchical knowledge base," in *ESWC*, 2014.
[17] M. G. Merayo, N. T. Nguyen *et al.*, "Intelligent collective: The role of diversity and collective cardinality," in *ICCCI*, 2017.
[18] I. Lorge, D. Fox, J. Davitz, and M. Brenner, "A survey of studies contrasting the quality of group performance and individual performance, 1920-1957." *Psychological bulletin*, 1958.
[19] T. Ye and L. P. Robert Jr, "Does collectivism inhibit individual creativity?: The effects of collectivism and perceived diversity on individual creativity and satisfaction in virtual ideation teams," in *CSCW*, 2017.
[20] C. De Boom, S. Van Canneyt, T. Demeester, and B. Dhoedt, "Representation learning for very short texts using weighted word embedding aggregation," *Pattern Recognition Letters*, 2016.
[21] L. Robert and D. M. Romero, "Crowd size, diversity and performance," in *ACM CHI*, 2015.
[22] L. P. Robert and D. M. Romero, "The influence of diversity and experience on the effects of crowd size," *JAIST*, 2017.
[23] D. B. Warnaar, E. C. Merkle, M. Steyvers, T. S. Wallsten, and E. R. Stone, "The aggregative contingent estimation system: Selecting, rewarding, and training experts in a wisdom of crowds approach to forecasting." in *AAAI Spring Symposium: Wisdom of the Crowd*, 2012.
[24] M. Galesic, D. Barkoczi, and K. Katsikopoulos, "Smaller crowds outperform larger crowds and individuals in realistic task conditions." *Decision*, 2018.
[25] I. Etukudo and J. Usen, "The generalized die binomial experiment," *ASRJETS*, 2016.
[26] H. Sato, T. Shirakawa, and M. Kubo, "Play style classification of the strong mahjong players," in *AIP Conference Proceedings*, 2015.
[27] G. Harman and M. H. Dredze, "Measuring post traumatic stress disorder in twitter," *ICWSM*, 2014.
[28] J. Yang, Y. Liu, Z. Liu, X. Zhu, and X. Zhang, "A new feature selection algorithm based on binomial hypothesis testing for spam filtering," *Knowledge-Based Systems*, 2011.